Kernel Based Nonparametric Coefficient Estimation in Diffusion Models

Dissertation

zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

vorgelegt
der Fakultät für Mathematik
der Technischen Universität Dortmund

von

Dipl.-Math. Benedikt Funke

Dortmund, Juli 2015

Datum der mündlichen Prüfung

18.11.2015

Veröffentlichte Fassung vom

25.11.2015

Acknowledgments

For the successful completion of this thesis, I firstly wish to thank my advisor Prof. Dr. Jeannette H.C. Woerner. She suggested this interesting topic to me and gave me constant support and inspiration by interesting and constructive discussions throughout the development of this thesis. Beyond that, she always encouraged me to participate at several different workshops and, thereby, offered me the possibility to meet colleagues from diversified fields of stochastics and statistics.

Moreover, I am deeply grateful for the financial support and the fruitful research environment by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823) of the German Research Foundation (DFG), through which I also came in contact with Prof. Herold Dehling.

I also wish to thank my colleagues from the sixth floor Daniel Kobe, Sven Glaser, and Christian Palmes for their assistance, friendship and for sharing many happy and enjoyable times.

Finally, I wish to thank my family for supporting me all the time. Last but not least, I wish to thank Fabienne for her love and her tireless support.

Contents

1 Introduction 3

1.1 Motivation 3

1.2 Preliminaries and basic notations 4

2 Nonparametric drift estimation in a Lévy-driven diffusion model 10

2.1 First intuitions in scalar diffusion models 10

2.2 Scalar jump diffusion models driven by a finite activity jump process 13

2.3 Drift estimation in a Lévy-driven diffusion model 15

2.4 Consistency of $\hat{b}(x)$ 21

2.5 Derivation of the asymptotic distribution 26

2.6 Examples of possible Lévy processes 36

2.7 Bandwidth selection 39

2.8 Comparison to an alternative nonparametric estimation approach 41

2.9 Estimation of the asymptotic variance 43

2.10 Asymptotic distribution of the variance estimator 50

2.11 The case of noisy data 53

2.12 Formulation of the pre-averaged drift estimator 54

2.13 Drift estimation of an integrated jump diffusion process 64

3 Using varying bandwidths in nonparametric regression and density estimation 71

3.1 Sample smoothing estimators in nonparametric density estimation 71

3.2 Sample smoothing estimators in nonparametric regression models 74

3.3 Asymptotic properties of $\hat{m}_{ANW}(x)$ 75

3.4 Asymptotic mean squared error 91

3.5 Adaptive Nadaraya-Watson like estimators for stochastic differential equations 93

4 Boundary bias correction methods 95

4.1 Boundary bias of nonparametric kernel estimators 95

4.2 Multivariate density estimation via asymmetric kernels 100
4.3 MBC techniques for multivariate density estimation 102
4.4 Bandwidth selection for MBC estimators 120
4.5 Finite sample performance 121
4.6 Nonparametric inference for multivariate diffusions 124

5 Nonparametric estimation of copula densities 131

5.1 Copula functions and multivariate distributions 131
5.2 Nonparametric estimation of copulas 135
5.3 Copula density estimation and the boundary bias effect 136
5.4 Beta kernel based copula density estimation 138
5.5 Bernstein copula density estimator 146
5.6 Nonparametric estimation approaches via copula based representations . . 148

6 Conclusion 154

7 References 156

1 Introduction

1.1 Motivation

The present thesis mainly focuses on nonparametric estimation methods for certain classes of stochastic processes. It is divided into three main subject areas. We start with the formulation of a kernel based nonparametric estimation procedure for jump diffusions. Afterwards we will focus on bias reduction techniques for this class of estimators. Finally, we will work with multivariate models and introduce the concept of copula functions.

As already mentioned, the first part deals with nonparametric kernel estimators for solutions of stochastic differential equations. We will start with a model based on a Brownian motion driven diffusion process as a motivation and will afterwards mainly deal with jump diffusions. Nonparametric kernel estimation for the coefficients of jump diffusions has not attracted much attention in the literature, yet. Most of the existing articles are concerned with the case of an additive independent finite activity jump process, which means -roughly speaking- that they focus on compound Poisson processes as a source of additive jumps. We will extend the existing results to the case of Lévy-driven jump diffusion models. The class of Lévy processes provides many possibilities of modeling certain jump behavior of, for instance, economic processes like stock prices, volatilities or interest rates.

We propose Nadaraya-Watson like estimators based on kernel functions as well as a bandwidth and will explore their asymptotic properties, i.e., consistency and asymptotic normality. The latter allows us to construct pointwise asymptotic confidence intervals, which are very useful for practical issues.

Based on these results, we will subsequently extend them to the case of noisy data. This kind of data plays a significant role, especially in high frequency settings; see for example Jones (2003) and Zhou (1996). In particular, we will not observe the diffusion itself but rather a sample containing an additive white noise process. Using a pre-averaging approach, we are able to get rid of the noise and make use of the asymptotic results in the non-noisy case.

The last section of the first part contains another example of a very interesting and, for practical issues, relevant jump diffusion model. Particularly, we will focus on integrated jump diffusion processes. These processes appear naturally in problems of engineering and physics. One can, for example, think of a velocity of a particle as the original jump diffusion process and the coordinate of the particle as the integral up to a fixed time point $t \geq 0$. Moreover, we assume that we are only able to observe the coordinate whereas the velocity is hidden. Making again use of the pre-averaging approach, we see that under appropriate assumptions, the results of the first section can also be used in this context.

The second part is mainly concerned with bias reduction techniques for nonparametric kernel based estimators. We will start with an adaptive version of the well-known Nadaraya-Watson estimator for nonparametric regression estimation. We will see that an appropriate choice of the newly introduced bandwidth function will asymptotically lead to a significant reduction of the order of the bias term. We show that this appealing property withstands even under weak dependency of the available sample. For our proofs we will make use of techniques borrowed from classical discrete time series analysis and will afterwards construct an adaptive drift estimator for a continuous diffusion process in view of the discrete findings.

Subsequently, we will leave the univariate case and focus on multivariate stochastic processes. We will especially deal with the so-called "boundary bias" effect, which occurs when we want to estimate densities possessing bounded or compact support by the use of symmetric kernel based estimators. The effect describes, for instance, the fact that symmetric kernel based estimators smear over probability mass to the negative real line or outside the unit square, although the corresponding densities are only supported on $\mathbb{R}^+$ or respectively on $[0,1]^2$. This problem has been attained a lot of attention in the literature and many methods for avoiding this effect have been published. We will focus on an approach by Chen (1999, 2000), who uses asymmetric probability densities like Beta and Gamma kernels for the estimation of unknown univariate densities. Furthermore, we will develop an extension of this approach due to the multivariate case and will afterwards introduce two non-negative multiplicative bias correction methods improving the rate of the bias term significantly. We explore bias and variance approximations and focus on several choices for the kernel functions. Regression estimators based on this method are then suggested for the estimation of the drift vector of a multivariate diffusion process.

The last section mainly focuses on the estimation of compact supported densities. Particularly, we will look at densities whose support is the unit hyper cube $[0,1]^d$ and construct nonparametric estimators via the use of Bernstein polynomials. In the literature, this approach has been used for the estimation of copula densities. We will briefly introduce the class of copula functions and will afterwards introduce various estimation approaches. Finally, due to Sklar´s theorem, we are able to represent conditional densities as well as conditional expectations in terms of the corresponding copula densities. This representation will lead us to the use of a Bernstein polynomial based estimator for conditional expectations which are in contrast -as we have already seen- approximations for the unknown drift and diffusion coefficient in diffusion models. Hence, this gives us another possibility to face the problem of estimating unknown conditional expectations or regression functions.

1.2 Preliminaries and basic notations

In this section we will shortly introduce our used notations and will define the basic tools for our subsequent analysis. There exists a variety of useful books introducing the

concept of stochastic processes. We only refer to those definitions being useful and in any way needed for our following analysis. For this purpose we restrict ourselves to the most important definitions and tools given in the books by Karatzas and Shreve (1996) as well as Cont and Tankov (2004). The following section is based on their introduction chapters.
We will subsequently work on a filtered probability space $(\Omega, \mathcal{A}, (\mathcal{A}_t)_{t\geq 0}, P)$ where $\mathcal{A}$ denotes a σ-algebra, $(\mathcal{A}_t)_{t\geq 0}$ a filtration of sub-σ-algebras, and P a probability measure.
A stochastic process $X = (X_t(\omega))_{t\geq 0} := (X_t)_{t\geq 0}$ is a family of random variables, which means that X_t is a random variable for every $t \geq 0$. We omit the dependency of the randomness ω, but always keep in mind that we work on a probability space. Moreover, a process X is called adapted to $(\mathcal{A}_t)_{t\geq 0}$, if X_t is $\mathcal{A}_t$ measurable for all $t \geq 0$.
For a fixed $\omega \in \Omega$, the map $t \to X_t(\omega)$ is called a path of the process X. During our following analysis, we will always observe a discrete sample of a path of an underlying stochastic process and will construct estimators for characteristics of this process based on this sample.
Another mentionable definition are the $L^p(\Omega, \mathcal{A}, P) := L^p(P)$ spaces. We say that a random variable $X \colon \Omega \to \mathbb{R}^d$ belongs to $L^p(P)$, if

$$E[|X|^p] = \int_\Omega |X(w)|^p dP(w) = \int_{\mathbb{R}^d} |x|^p P^X(dx) < \infty,$$

where

$$P^X(A) = P(X^{-1}(A)),\ A \in \mathcal{B}(\mathbb{R}^d)$$

denotes the image measure of X.
A very important class of stochastic processes, which additionally plays a major role in the following chapters, are martingales. We will omit the concept in terms of discrete time and will restrict ourselves to the continuous-time case. We will at first introduce the Brownian motion, which acts as a fundamental stochastic process and is also very important in the context of diffusion processes, which will play a central role within this work later on. For the sake of simplicity, we will restrict ourselves to the case $d = 1$.

Definition 1.1. *A one-dimensional Brownian motion $W = (W_t)_{t\geq 0}$ on a filtered probability space $(\Omega, \mathcal{A}, (\mathcal{A}_t)_{t\geq 0}, P)$ is an adapted stochastic process fulfilling*

a) W has independent increments, which means that $W_t - W_s$ is independent of $\mathcal{A}_s$, for $s \leq t$,

b) W has stationary increments, which are normally distributed such that

$$W_t - W_s \overset{\mathcal{D}}{=} \mathcal{N}(0, t-s),$$

c) W has almost surely continuous paths and

d) $W_0 = 0$ almost surely (a.s.).

A Brownian motion is one basic example of a continuous-time martingale. Other examples of martingales will become very important later on. For this purpose let us now define another fundamental class of stochastic processes, namely the class of Lévy processes. These processes play a central role in many scientific fields. Applications can be found for instance in physics (analysis of turbulence data), in engineering (construction of dams protecting the environment against flood catastrophes), in economics as a toy-example of a discontinuous asset price, and, of course, in mathematical finance (in particular in risk theory). Several advisable books introduce the interested reader into this field of stochastic processes. We refer to Sato (1999), Cont and Tankov (2003), and Barndorff-Nielsen et al. (2001). A very useful survey about Lévy processes and their applications in finance has been published by Papapantoleon (2008), from which the ideas how to introduce Lévy processes in the following way are extracted.
We start with a formal definition of a Lévy process and concentrate only on the one-dimensional case due to the sake of simplicity.

Definition 1.2. *A real-valued, adapted, and càdlàg (French "continue à droite, limite à gauche"; English RCLL: "right continuous with left limits") stochastic process $L = (L_t)_{t\geq 0}$ on a filtered probability space $(\Omega, \mathcal{A}, (\mathcal{A}_t)_{t\geq 0}, P)$ such that $L_0 = 0$ a.s. is called Lévy process, if*

a) L has independent increments, which means that $L_t - L_s$ is independent of $\mathcal{A}_s$, $0 \leq s < t$,

b) L has stationary increments, i.e., the distribution of $L_{t+h} - L_t$ is independent of t for all $h > 0$ and

c) L is stochastically continuous, i.e.

$$\forall \varepsilon > 0: \lim_{t \to s} P(|L_t - L_s| > \varepsilon) = 0, \quad \textit{for every } t \geq 0.$$

The literature features different definitions for Lévy processes. The definition above seems to be most adequate for our purposes, because the handling of small increments of stochastic processes will play a central role in our subsequent analysis; see c) in Definition 1.2.

Example 1.3. *The easiest examples of Lévy processes are given by a deterministic linear drift $L_t = at$, $a \in \mathbb{R}$, by the Brownian motion $L_t = W_t$, and by a (compound) Poisson process.*

A fundamental characterization of Lévy processes provides the Lévy-Khintchine representation of the corresponding characteristic function of L. Due to its importance, we will state this result here. Further, it acts as an elegant tool for the derivation of higher moments of L and will be important subsequently.

Theorem 1.4 (Cont and Tankov (2003), Theorem 3.1)**.** *Let $L = (L_t)_{t\geq 0}$ be a Lévy process, then there exists a triplet (b, σ, ν) where $b \in \mathbb{R}, \sigma \in \mathbb{R}^+$ and additionally ν is a measure on $\mathbb{R}$ fulfilling*

$$\nu(\{0\}) = 0 \text{ and } \int_{\mathbb{R}} (1 \wedge x^2)\nu(dx) < \infty$$

such that the characteristic function of L_t can be represented as

$$\varphi_{L_t}(u) := E[e^{iuL_t}] = \exp\left[t\left(ibu - \frac{u^2\sigma}{2} + \int_{\mathbb{R}} \left(e^{iux} - 1 - iux1_{\{|x|\leq 1\}}\right)\nu(dx)\right)\right].$$

Remark 1.5. *This representation does not only hold for Lévy processes but rather for the characteristic function of a random variable whose distribution is infinitely divisible. We do not go into detail here, but remark that the law of L_t is infinitely divisible and (in turn) for every infinitely divisible random variable X one can construct a Lévy process $\tilde{L}$ such that $\tilde{L}_1 \overset{\mathcal{D}}{=} X$.*

We will now introduce the concept of Poisson random measures of Lévy processes, which provide a very useful tool to work with these processes by using martingale techniques under appropriate assumptions.

Definition 1.6. *Let $L = (L_t)_{t\geq 0}$ be a Lévy process and $\Delta L := (\Delta L_t)_{t\geq 0} := (L_t - L_{t_-})_{t\geq 0}$ the corresponding pure jump process, where $L_{t_-} := \lim_{s\uparrow t} L_s$. Let $B \in \mathcal{B}(\mathbb{R}\backslash\{0\})$ a Borel set such that $0 \notin \overline{B}$, where $\overline{B}$ denotes the closure of the Borel set B. We set*

$$\mu^L(\omega, t, B) := \#\{0 \leq s \leq t : \ \Delta L_s \in B\} = \sum_{0\leq s\leq t} 1_B(\Delta L_s)$$

and call μ^L the Poisson random measure of L.

Roughly speaking, μ^L is the non-negative integer-valued random measure, which counts the jumps of L up to a certain time t such that $\Delta L_s \in B$.
The name *Poisson* random measure is justified by the following observations (cf. Papapantoleon (2008)):

a) μ^L has independent increments, which means that $\mu^L(\omega, t, B) - \mu^L(\omega, s, B)$ is independent of $\mathcal{A}_s$. This can be seen in the fact that

$$\mu^L(\omega, t, B) - \mu^L(\omega, s, B) \in \sigma(\{L_u - L_v; \ s \leq v < u \leq t\})$$

and that L has independent increments.

b) μ^L has stationary increments because

$$\mu^L(\omega, t, B) - \mu^L(\omega, s, B) = \#\{0 \leq u \leq s - t : \ \Delta(L_{s+u} - L_s) \in B\}$$

and that L has stationary increments.

Hence, $(\mu^L(\omega, t, B))_{t\geq 0}$ is a non-negative integer-valued Lévy process whose jumps are all of height 1. Consequently, μ^L is, for a fixed Borel set B and indexed in $t \geq 0$, a Poisson process with intensity

$$\nu(B) := E[\#\{0 \leq s \leq 1 : \ \Delta L_s \in B\}],$$

which means that

$$P(\mu^L(\omega, \cdot, B) = k) = e^{-\nu(B)} \frac{\nu(B)^k}{k!}, \ \forall k \in \mathbb{N}_0.$$

For further reading on this topic, we refer to Cont and Tankov (2004), Section 2.6.
ν is called the Lévy measure or the Lévy density of L and has already been used in the Lévy-Khintchine representation as well as in the characteristic triplet.
The Lévy measure provides information about the expected number of jumps of a certain height within a time interval of length 1. Since $\int_{\mathbb{R}}(1 \wedge x^2)\nu(dx) < \infty$, the Lévy process L can have infinitely many small jumps, but does only have finitely many jumps of size $J \geq 1$. A Lévy process is called infinitely active when $\nu(\mathbb{R}) = \infty$. Finite activity processes $(\nu(\mathbb{R}) < \infty)$ are, in general, compound Poisson processes and will not play a significant role in our following analysis.
We now state another fundamental representation of a Lévy process, namely the Lévy-Itô decomposition where Poisson random measures as well as the Lévy measure appear, too.

Proposition 1.7 (Cont and Tankov (2004), Proposition 3.7)**.** *Let $L = (L_t)_{t\geq 0}$ be a Lévy process with Poisson random measure μ^L and characteristic triplet (b, σ, ν). Then, L can be decomposed into the sum of four independent Lévy processes $L^{(i)}, i = 1, ..., 4$, in the following way*

$$\begin{aligned} L_t &= L_t^{(1)} + L_t^{(2)} + L_t^{(3)} + L_t^{(4,\varepsilon)} \\ &= bt + \sqrt{\sigma} W_t + \int_0^t \int_{\{|x|\geq 1\}} x\mu^L(ds, dx) + \lim_{\varepsilon \downarrow 0} \int_0^t \int_{\{\varepsilon<|x|<1\}} x(\mu^L - \nu^L)(ds, dx) \\ &=: bt + \sqrt{\sigma} W_t + \sum_{0<s\leq t} \Delta L_s 1_{\{|\Delta L_s|\geq 1\}} + \int_0^t \int_{\{|x|<1\}} x(\mu^L(ds, dx) - \nu(dx)ds), \end{aligned}$$

where $L^{(1)}$ is a linear drift term, $L^{(2)}$ is a scaled Brownian motion, $L^{(3)}$ is a compound Poisson process with arrival rate $\lambda = \int_{\{|x|\geq 1\}} \nu(dx)$, and $L^{(4)} := \lim_{\varepsilon\downarrow 0} L^{(4,\varepsilon)}$ is a pure jump martingale.

The fact that the last part $L^{(4)}$ is a martingale can be seen for instance through Proposition 2.16 in Cont and Tankov (2004). Moreover, the reason why we stated this decomposition here is that we will work with diffusions driven by Lévy processes, which are time-continuous martingales. Additionally, we assume that the driving Lévy process possesses finite moments of certain orders. The relation between the existence of moments of L and its corresponding Lévy measure ν can be found in the following proposition.

Proposition 1.8 (Cont and Tankov (2004), Proposition 3.13). *Let $L = (L_t)_{t\geq 0}$ be a Lévy process with triplet (b, σ, ν), then*

$$E[|L_t|^n] < \infty \Leftrightarrow \int_{\{|x|\geq 1\}} |x|^n \nu(dx) < \infty.$$

In this case, the moments can be deduced in a rather simple way by differentiation of the characteristic function. In particular, we can state that

$$E[L_t] = t\left(b + \int_{\{|x|\geq 1\}} |x|\nu(dx)\right)$$

as well as

$$Var(L_t) = t\left(\sigma + \int_{\mathbb{R}} x^2\nu(dx)\right).$$

Finally, we want to conclude the mentioned properties of Lévy processes and use them for our purposes. In the following section, we will work with a Lévy process L of the form

$$\begin{aligned} L_t &= \int_{(0,t]}\int_{\mathbb{R}} x(\mu^L(ds,dx) - \nu(dx)ds) \\ &:= \int_0^t \int_{\mathbb{R}} x(\mu^L(ds,dx) - \nu(dx)ds) := \int_0^t \int_{\mathbb{R}} x\bar{\mu}^L(ds,dx) \end{aligned}$$

possessing the property that $\int_{\mathbb{R}} x^4\nu(dx) < \infty$. $\bar{\mu}^L$ denotes the compensated Poisson random measure of L. We are able to rewrite L in view of the Lévy-Itô decomposition as follows:

$$\begin{aligned} L_t &= \int_0^t \int_{\mathbb{R}} x(\mu^L(ds,dx) - \nu(dx)ds) \\ &= \int_0^t \int_{\{|x|<1\}} x(\mu^L(ds,dx) - \nu(dx)ds) + \int_0^t \int_{\{|x|\geq 1\}} x\mu^L(ds,dx) \\ &\quad - \int_0^t \int_{\{|x|\geq 1\}} x\nu(dx)ds \\ &:= L^{(3)} + L^{(4)} + \tilde{b}t, \end{aligned}$$

where $\tilde{b} := -\int_{\{|x|\geq 1\}} x\nu(dx) \in \mathbb{R}$.
Hence, L has the triplet $(\tilde{b}, 0, \nu)$ and is a martingale possessing a finite fourth moment

$$E[L_t^4] = t\left(3\left(\int_{\mathbb{R}} y^2\nu(dy)\right)^2 + \int_{\mathbb{R}} y^4\nu(dy)\right).$$

In addition, the characteristic function $\varphi_{L_t}(u)$ of L is given by

$$\varphi_{L_t}(u) = E[e^{iuL_t}] = \exp\left(t\int_{\mathbb{R}}(e^{iux} - 1 - iux)\nu(dx)\right).$$

2 Nonparametric drift estimation in a Lévy-driven diffusion model

2.1 First intuitions in scalar diffusion models

In this chapter we will focus on the nonparametric estimation of the coefficients in a certain jump diffusion model. To motivate our procedure, we will at first have a look at an ordinary diffusion model driven by a Brownian motion. To be precise, consider a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t^W)_{t\geq 0}, P)$ equipped with a Brownian Motion $W = (W_t)_{t\geq 0}$ and the canonical filtration $\mathcal{F}_t^W := \sigma(W_s, s \leq t)$. Let $X = (X_t)_{t\geq 0}$ be a diffusion process given by the time-homogeneous stochastic differential equation

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t,\ X_0 \overset{\mathcal{D}}{=} \eta,$$

where b and $\sigma > 0$ are unknown functions, which are globally Lipschitz-continuous such that the equation possesses a pathwise unique strong solution (see Karatzas and Shreve (1996), Proposition 2.13). Moreover, let the initial condition $\eta \in L^2(P)$ be independent of W. Now define

$$\mathcal{F}_t := \sigma(\eta, W_s; 0 \leq s \leq t),\ 0 \leq t \leq \infty$$

and additionally the "collection of null sets"

$$\mathcal{N} := \mathcal{N}_\infty := \{N \subset \Omega; \exists F \subset \mathcal{F}_\infty \text{ with } N \subset F \text{ and } P(F) = 0\}.$$

We call

$$\mathcal{F}_t^X := \sigma(\mathcal{F}_t \cup \mathcal{N})$$

the augmented filtration fulfilling the usual assumptions. This means that it is complete and right-continuous (see Karatzas and Shreve (1996), Proposition 7.7). The solution

$$X_t = X_0 + \int_0^t b(X_s)ds + \int_0^t \sigma(X_s)dW_s$$

is $\mathcal{F}_t^X$ adapted and also a semimartingale. Furthermore, the process is determined by the functions b and σ and, hence, it seems natural to be interested in their behavior and shape. By additional assumptions within this model, which cause that X is stationary and equipped with a stationary density, one can see that this density can explicitly be represented only in dependence of b and σ^2; see for example Karatzas and Shreve (1996), pp. 352.

Due to the importance of the drift b and the volatility σ, several different approaches for the (non-)parametric estimation of diffusion models have been published. In this work we are only interested in nonparametric approaches. For parametric estimation procedures, we refer to a recently published book by Kessler et al. (2012).

In the nonparametric setting, any list of existing methods would be incomplete, so we will only mention those approaches playing a significant role for our subsequent analysis. Very fundamental and crucial work has been done by Comte and Genon-Catalot; see for instance Comte et al. (2009), (2010) and Comte and Genon-Catalot (2007). Their approach is based on model selection and provides an adaptive estimator for both, the drift and the volatility function, for which the L^2-risk can be bounded dependent on the smoothness of these functions as well as the sampling frequency. Due to the adaptivity of their estimator, an asymptotic distribution is not derivable and, therefore, confidence intervals cannot be determined. Nevertheless, simulation issues reveal that their estimators provide good results. Later on, we will compare our findings with those of Schmisser (2014), who developed the same estimation procedure as Comte and Genon-Catalot in a jump diffusion setting.

In contrast to the upper approach, other authors focused on kernel based estimation of diffusion models. Their findings will act as a basis for our approach in the Lévy-driven model. In this context, very important work has been done by Bandi. His articles provide a complete asymptotic analysis of kernel based estimators for ordinary diffusion models; see Bandi and Phillips (2001), (2003). He also extends his results to jump diffusion models driven by compound Poisson processes and to multivariate diffusion models; see Bandi and Nguyen (2003) as well as Bandi and Moloche (2008). Due to the importance of his results, we will briefly describe the motivation behind a kernel based approach below.

To develop nonparametric estimators for b and σ, Stanton (1997) uses approximations of the infinitesimal generator $\mathcal{L}$ of X defined by

$$\begin{aligned}\mathcal{L}(f)(x,t) &:= \lim_{\tau\downarrow t} \frac{E[f(X_\tau,\tau)|X_t = x] - f(x,t)}{\tau - t}\\ &= \frac{\partial f(x,t)}{\partial t} + \frac{\partial f(x,t)}{\partial x} b(x) + \frac{1}{2}\frac{\partial^2 f(x,t)}{\partial x^2}\sigma^2(x).\end{aligned}$$

The last equation is a classical result for Itô-diffusions and can, for instance, be found in Øksendal (2000, Theorem 7.3.3). For our purposes, it is sufficient to assume that $f \in C_0^2(\mathbb{R}\times\mathbb{R}^+)$. Under this assumption, the above limit exists and is therefore contained in the domain of $\mathcal{L}$. Using a Taylor expansion of f, one can finally deduce a first order approximation for the drift b (by setting $\frac{\partial f(x,t)}{\partial x} = x$) and for the volatility function σ (by setting $\frac{\partial f(x,t)}{\partial t} = (x - X_t)^2$) via

$$\begin{aligned}b(x) &= \frac{1}{\Delta}E[X_{t+\Delta} - X_t|X_t = x] + O(\Delta),\ \text{ as } \Delta \to 0\\ \sigma^2(x) &= \frac{1}{\Delta}E[(X_{t+\Delta} - X_t)^2|X_t = x] + O(\Delta),\ \text{ as } \Delta \to 0.\end{aligned}$$

We refer to Stanton (1997) for higher order approximations of b and σ. For our purposes, these approximations are sufficient and are the basis for both Bandi´s and our estimation procedure in more involved models.

Using the above approximations, it is quite intuitive to use nonparametric regression techniques, which are well developed for discrete time series analysis. There is a plethora of different regression estimators like local polynomial estimators, splines, jackknife estimators or neural networks. Due to its popularity and the well-known asymptotic properties, we focus on a local constant estimator, which is also known as the Nadaraya-Watson estimator (cf. Nadaraya (1965), Watson (1964) or Härdle (1990)) for the estimation of conditional moments in regression frameworks. The intuition of this estimator is to use weighted averages of infinitesimal increments of X, which lie in the vicinity of the spatial point x at which we want to estimate both functions b and σ. The quantification of the rather imprecise term "vicinity" will be handled later on. The weighting will be realized by a kernel function K which is in general a symmetric probability density function possessing a finite second moment. For all following derivations, one can think of a Gaussian density as a toy-example for which all assumptions will hold true.

To further illustrate the idea of the previously described procedure, we suppose that we observe the process X at equidistant time points $0, \Delta, 2\Delta, ..., n\Delta$. An initial estimator for the drift b at point x according to the above description would be given by

$$\hat{b}_1(x) := \frac{\frac{1}{nh}\sum_{i=0}^{n-1} 1_{(|X_{i\Delta}-x|\leq h)} \frac{(X_{(i+1)\Delta}-X_{i\Delta})}{\Delta}}{\frac{1}{nh}\sum_{i=0}^{n-1} 1_{(|X_{i\Delta}-x|\leq h)}} = \frac{\frac{1}{nh}\sum_{i=0}^{n-1} 1_{\left(\frac{|X_{i\Delta}-x|}{h}\leq 1\right)} \frac{(X_{(i+1)\Delta}-X_{i\Delta})}{\Delta}}{\frac{1}{nh}\sum_{i=0}^{n-1} 1_{\left(\frac{|X_{i\Delta}-x|}{h}\leq 1\right)}},$$

where $h = h_n$ is a bandwidth and $\Delta = \Delta_n$ is the sampling frequency. The bandwidth will regulate what is meant by the term "vicinity".

To explore asymptotic properties like consistency and the derivation of the asymptotic distribution, it seems intuitive to ensure the following points:

i) It is a well-known fact that the drift function cannot consistently be (nonparametrically) estimated on a compact interval; see for instance Bandi and Phillips (2003). Therefore, it seems natural to impose that $n\Delta := T \to \infty$ for the examination of $\hat{b}_1(x)$.

ii) To reproduce our initial approximation of b, we have to impose that $\Delta \to 0$ as $n \to \infty$.

iii) The bandwidth h has to decrease to zero as $n \to \infty$ to ensure that only observations "near" x are included for estimating $b(x)$.

iv) The rescaled denominator

$$\sum_{i=0}^{n-1} 1_{\left(\frac{|X_{i\Delta}-x|}{h}\leq 1\right)}$$

has to diverge as $n \to \infty$ in order to guarantee that the process X infinitely often hits a neighborhood of x.

Later on, we will specify all these intuitive conditions to derive the desired asymptotic properties of our proposed drift estimator.
From a heuristic point of view, it seems to be reasonable that observations near x should be equipped with higher weights than others. Therefore, it would be a canonical approach to substitute the indicator kernel $1_{\left(\frac{|X_{i\Delta}-x|}{h}\leq 1\right)}$ by a smooth kernel function K whose basic properties have already been mentioned. Bandi and Phillips (2003) and Bandi and Nguyen (2003) finally suggested the estimators

$$\hat{b}(x) := \frac{\frac{1}{nh}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)\frac{(X_{(i+1)\Delta}-X_{i\Delta})}{\Delta}}{\frac{1}{nh}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)}$$

as well as

$$\hat{\sigma}^2(x) := \frac{\frac{1}{nh}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)\frac{(X_{(i+1)\Delta}-X_{i\Delta})^2}{\Delta}}{\frac{1}{nh}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)}.$$

They explored the strong consistency and the asymptotic normality under usual regularity assumptions on b and σ like Lipschitz-continuity and ellipticity of σ (which means that $\sigma(x) > \sigma_0 > 0$). Following our first intuitions, their derivations are based on the double asymptotics scheme

$$\Delta \to 0 \text{ and } T = n\Delta \to \infty.$$

Remark 2.1. *For the consistent nonparametric estimation of σ, the assumption $T \to \infty$ is not necessary; see for example Florens-Zmirou (1993) or Bandi and Phillips (2003) for an exact mathematical justification of this fact. Due to this reason, the second estimator possesses a faster rate of convergence. Later on, we will see that, in contrast to the ordinary diffusion model, the double asymptotics scheme is necessary for both estimators in our considered Lévy-driven model.*

2.2 Scalar jump diffusion models driven by a finite activity jump process

To the best of our knowledge, the first paper which investigated nonparametric kernel based estimators for the coefficients in a jump diffusion model is the one by Bandi and Nguyen (2003). We will shortly present their approach and the used techniques. After that, we will use analogous arguments for the derivation in a more general model.
Adding jumps in a diffusion model seems to be very important from a practical point of view. As we want to model interest rates or stock prices, macroeconomic news, endogenous as well as exogenous shocks can cause abrupt changes during the evolution of the process X. Hence, it is quite intuitive to include an additive (independent) jump component. Bandi and Nguyen (2003) focused on the following class of processes

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t + \int c(X_{s-}, y)\bar{\nu}(ds, dy),\ X_0 \overset{\mathcal{D}}{=} \eta,$$

where $\bar{\nu}(ds, dy) = \mu(ds, dy) - \Gamma(dy)ds$ is a compensated Poisson random measure with intensity $\Gamma(dy)ds$, where $\Gamma(dy)$ is a probability distribution. To coincide with their notation, we set here the intensity of the underlying Poisson process equal to 1. Because of this assumption on the jump size distribution $\Gamma(dy)$, X is a process with finite activity. As a consequence, they allow for jumps occurring due to an independent additive compound Poisson process.

In the following section, we will present our main results, where we explicitly not assume that the additive jump component is of finite activity but is rather a Lévy process possessing certain finite moments.

Bandi and Nguyen (2003) focused on the same double asymptotics scheme as in the Brownian case before and used nonparametric regression techniques for the estimation of the first and second conditional moment of infinitesimal changes of the process X. The approximations of the infinitesimal generator for the above discontinuous process X are now given by (Bandi and Nguyen (2003), p.297):

$$\begin{aligned} b(x) &= \lim_{\Delta \to 0} \frac{1}{\Delta} E[X_{t+\Delta} - X_t | X_t = x] \\ \sigma^2(x) + E[c^2(x, Y)] &= \lim_{\Delta \to 0} \frac{1}{\Delta} E[(X_{t+\Delta} - X_t)^2 | X_t = x] \\ E[c^k(x, Y)] &= \lim_{\Delta \to 0} \frac{1}{\Delta} E[(X_{t+\Delta} - X_t)^k | X_t = x], \;\; k \geq 3, \end{aligned} \tag{2.1}$$

where we naturally assume that the above conditional moments exist and $Y \overset{\mathcal{D}}{=} \Gamma(dx)$ denotes the jump size distribution. This should only be understood as a motivation how we can construct Nadaraya-Watson like estimators even in this jump diffusion setting. The dependency of these approximations of the jump component is not surprising. In the jump diffusion case, the approximation of the infinitesimal generator of X can be decomposed into the approximation of the continuous part $\mathcal{L}$ and the one for the discontinuous part $\mathcal{M}$, where

$$\mathcal{M}(\phi)(x) := \int_{\mathbb{R}} (\phi(x + c(x, y)) - \phi(x) - \phi'(x)c(x, y))\Gamma(dy).$$

Bandi and Nguyen (2003) provide an asymptotic analysis of the appropriate Nadaraya-Watson like estimators based on symmetric kernels. It should also be mentioned that a comparable semiparametric analysis has been conducted by Johannes (2004), who proposed estimators for the coefficients of the above mentioned jump diffusion model, too. For this purpose, he parametrized the jump size distribution and used the proposed approximations to construct estimators based on the method of moments.

Remark 2.2. *The results in Bandi and Phillips (2003) as well as in Bandi and Nguyen (2003) hold true even for a relatively broad class of stochastic processes. Their central identifiability assumption is that the process X is Harris-recurrent. This ensures that X infinitely often hits a neighborhood of an already visited point x; see Bandi and Phillips*

(2003) for more details. Subsequently, we will assume that X is stationary and σ is elliptical. Moreover, we weaken their assumption that b has to be bounded.

2.3 Drift estimation in a Lévy-driven diffusion model

We now turn to our main section, namely the nonparametric estimation of the coefficients in a Lévy-driven jump diffusion model.

Let, therefore, $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, P)$ be a filtered probability space equipped with a Brownian Motion $W = (W_t)_{t\geq 0}$ and a Lévy process $L = (L(t))_{t\geq 0} = (L_t)_{t\geq 0}$ of the already introduced form

$$dL_t = \int_{\mathbb{R}} y(\mu(dt, dy) - \nu(dy)dt) := \int_{\mathbb{R}} y\bar{\mu}(dt, dy),$$

where μ is a Poisson random measure compensated by its intensity measure $\nu(dy)dt$. We assume that the Lévy measure $\nu(dy)$ satisfies

$$E[L^2(1)] = Var(L(1)) = \int_{\mathbb{R}} y^2\nu(dy) < \infty.$$

The representation of $Var(L(1))$ by its corresponding Lévy measure can easily be derived by differentiation of the characteristic function as it was presented in the introduction. We remark that L is a martingale with respect to its canonical filtration and possessing finite variance.

Now consider a stochastic process $X = (X_t)_{t\geq 0}$ such that

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t + \xi(X_{t-})dL_t, \quad X_0 \overset{\mathcal{D}}{=} \eta, \tag{2.2}$$

where b, σ and ξ are unknown functions and X possesses the initial distribution $\eta \overset{\mathcal{D}}{=} \Gamma(dx)$. We also assume that W and L are independent processes and that $\eta \in L^2(\Gamma(dx))$ is independent of W as well as L. In addition, we choose as filtration $\mathcal{F}_t := \sigma(\eta, (L_s, W_s),\ s \leq t)$ such that X_t is $\mathcal{F}_t$ adapted.

We should remark that the integral

$$Z_t := \int_0^t \int_{\mathbb{R}} \xi(X_{s-})y\bar{\mu}(ds, dy)$$

denotes a martingale with respect to $\mathcal{F}_t$ and satisfies the isometry formula

$$E[Z_t^2] = E\left[\int_0^t \int_{\mathbb{R}} \xi^2(X_s)y^2\nu(dy)ds\right].$$

Furthermore, by imposing that ξ is bounded (cf. Assumption A1,ii)), we easily conclude that

$$E\left[\int_0^t \int_{\mathbb{R}} \xi^2(X_s)y^2\nu(dy)ds\right] \leq t||\xi^2||_\infty \int_{\mathbb{R}} y^2\nu(dy).$$

Moreover, X_{t-} denotes the càglàd version of the process X which ensures that the integrand is a predictable process such that the stochastic integral can be defined in the usual manner; see Cont and Tankov (2004), Section 8.1.4. for further details on stochastic integrals with respect to Poisson random measures. Suppose that we observe X in a high frequency setting on the interval $[0,T]$ on an equidistant grid at time points $0,\Delta,2\Delta,...,n\Delta = T$. Our first aim will be the construction of a meaningful pointwise estimator of the drift function b at a design point x based on the available sample $X_0, X_\Delta, X_{2\Delta}, ..., X_{n\Delta} = X_T$.
This problem originates from Schmisser (see Schmisser (2013), (2014)), who proposed adaptive nonparametric estimators for b as well as $\sigma^2 + Var(L(1))\xi^2$ on a compact subset $A \subset \mathbb{R}$ by the use of a model selection approach. Our procedure differs substantially from this technique, as we are interested in pointwise estimators. Moreover, we are concerned with kernel based estimators for which we incorporate an additional regularization parameter, namely the bandwidth h. To the best of our knowledge, there is no work done yet covering the nonparametric estimation of $b(x)$ in a Lévy-driven diffusion model. It should be mentioned that for the nonparametric estimation of $\sigma^2(x)$ in an infinite activity model, there is a reasonable and alternative approach by Mancini and Renó (2011). They used a threshold estimator, which disentangles the discontinuous from the continuous part of the process X. The new thinned out sample is used for the construction of an estimator of $\sigma^2(x)$. In contrast to our approach, they are only interested in the fixed T case, namely when one observes the process X only on a compact interval of the form $[0,T]$. Their estimator converges with rate $\sqrt{nh}$. Furthermore, and to the best of our knowledge, minimax rates for b, σ and ξ are not available in the literature, yet. We are now ready to list our assumptions on the considered model, which will lead us to the first very useful proposition.

Assumption A1

i) The functions b, σ and ξ are globally Lipschitz-continuous.

ii) The function σ is bounded away from zero (ellipticity condition) as well as uniformly bounded for all x:

$$\exists\ \sigma_1, \sigma_0 \in \mathbb{R} : \forall x \in \mathbb{R} : \quad 0 < \sigma_1 \leq \sigma(x) \leq \sigma_0.$$

iii) The function ξ is non-negative and also bounded:

$$\exists\ \xi_0 \in \mathbb{R} : \forall x \in \mathbb{R} : \quad 0 \leq \xi(x) \leq \xi_0.$$

iv) The function b is elastic (cf. Masuda (2007)). This means that

$$\exists\ M > 0 : \forall x \in \mathbb{R}, |x| > M : \quad xb(x) \lesssim -x^2,$$

where the relation $\lesssim$ is defined as follows:
let S be a set and $f, g : S \to [0, \infty)$ two functions. Then,

$$f \lesssim g \quad :\Leftrightarrow \ \exists\, C \in \mathbb{R}^+ : \ f(x) \leq C \cdot g(x) \ \forall x \in S.$$

Especially, b cannot be bounded as required in Bandi and Nguyen (2003).

v) The Lévy measure ν possesses the properties that

$$Var(L(1)) = \int_{\mathbb{R}} y^2 \nu(dy) < \infty, \qquad \nu(\{0\}) = 0.$$

Under Assumption A1,i) a unique strong solution X of (2.2) it exists (cf. Masuda (2007)). Moreover, under A1,i)-v), this solution is equipped with a unique invariant probability distribution $\Gamma(dx)$. Moreover, X fulfills a β-mixing condition. In general, a homogeneous Markov process X with transition semigroup $(P_t)_{t \in \mathbb{R}_+}$ and initial distribution η is said to be β-mixing with coefficient $\beta_X(t)$, if

$$\beta_X(t) := \sup_{s \in \mathbb{R}_+} \int ||P_t(x, \cdot) - \eta P_{s+t}(\cdot)|| \eta P_s(dx) \longrightarrow 0, \text{ as } t \to \infty,$$

where ηP_t denotes the distribution of X_t and $||\lambda||$ defines the total variation norm of a signed measure λ; see Masuda (2007). This mixing property describes the temporal dependence of the process. Due to Masuda (2007), X is, in addition, exponentially β-mixing, which means that there exists a constant $\gamma > 0$ such that

$$\beta_X(t) = O(e^{-\gamma t}), \text{ as } t \to \infty.$$

Using Theorem 2.1 in Masuda (2007) we can deduce the ergodicity of X, which means that for all measurable functions $g \in L^1(\Gamma(dx))$:

$$\frac{1}{T} \int_0^T g(X_s) ds \longrightarrow \int_{\mathbb{R}} g(x) \Gamma(dx) \qquad \text{a.s., as } T \to \infty.$$

For further information and especially equivalent reformulations of A1,iv), we recommend Masuda (2007).
Due to our assumptions on the Lévy measure ν and the Lipschitz-continuity of the coefficients b, σ and ξ, we have that $E[X_t^2] < \infty$. This can easily be proven by applying the Cauchy-Schwarz inequality successively. We will focus on this property later on.
Moreover, we impose that

vi) Γ is absolutely continuous with respect to the Lebesgue measure and, thus, possesses a Lebesgue density π such that $\Gamma(dx) = \pi(x)dx$.

For sufficient conditions on A1, vi) see Ishikawa and Kunita (2006), which ensure that under our assumptions on ξ (especially the boundedness from below) a smooth transition density exists.
We remark that the process X is also stationary, because we assumed that $X_0 \sim \Gamma(dx)$. These assumptions are largely congruent to those in Schmisser (2014), which make comparisons of the derived results much easier.
We are now ready to state our first proposition. It turns out that this result is one of the key elements for our asymptotic analysis. The following result originates from Schmisser (2014) but is not proven there.

Proposition 2.3. *Let $X = (X_t)_{t\geq 0}$ be the solution of* (2.2). *Under assumptions A1,i)-vi) a constant $C > 0$ exists, such that*

$$E\left[\sup_{|s-t|\leq\Delta} (X_s - X_t)^2\right] \leq C\Delta,$$

provided that $\Delta \leq 1$.

For the proof of this statement, we need the following two inequalities of the Burkholder-Davis-Gundy type (see Schmisser (2014), Result 11):

Lemma 2.4. *Let C_1, C_2 and Δ be positive constants and recall that*

$$\mathcal{F}_t = \sigma(X_0, (W_s, L_s); s \leq t)$$

denotes the underlying filtration of sub-σ-algebras. For $\Delta \leq 1$ it holds that

$$1.)\quad E\left[\sup_{|s-t|\leq\Delta}\left(\int_t^s \sigma(X_u)dW_u\right)^2 \middle| \mathcal{F}_t\right] \leq C_1 E\left[\int_t^{t+\Delta} \sigma^2(X_u)du \middle| \mathcal{F}_t\right]$$

$$2.)\quad E\left[\sup_{|s-t|\leq\Delta}\left(\int_t^s \xi(X_{u_-})dL_s\right)^2 \middle| \mathcal{F}_t\right] \leq C_2 E\left[\int_t^{t+\Delta} \xi^2(X_u)du \middle| \mathcal{F}_t\right]\left(1 + \int_{\mathbb{R}} y^2\nu(dy)\right).$$

Proof of Proposition 2.3. Recall that σ and ξ are bounded and derive the desired result in the following way:

$$
\begin{aligned}
&E\left[\sup_{|s-t|\le\Delta}(X_s-X_t)^2\right]\\
&= E\left[\sup_{|s-t|\le\Delta}\left(\int_t^s b(X_u)du+\int_t^s\sigma(X_u)dW_u+\int_t^s\xi(X_{u-})dL_u\right)^2\right]\\
&\lesssim E\left[\sup_{|s-t|\le\Delta}\left(\int_t^s b(X_u)du\right)^2\right]+E\left[\sup_{|s-t|\le\Delta}\left(\int_t^s\sigma(X_u)dW_u\right)^2\right]\\
&\quad+E\left[\sup_{|s-t|\le\Delta}\left(\int_t^s\xi(X_{u-})dL_u\right)^2\right]\\
&= E\left[\sup_{|s-t|\le\Delta}\left(\int_t^s b(X_u)1_{[t,s]}(u)du\right)^2\right]+E\left[E\left[\sup_{|s-t|\le\Delta}\left(\int_t^s\sigma(X_u)dW_u\right)^2\middle|\mathcal{F}_t\right]\right]\\
&\quad+E\left[E\left[\sup_{|s-t|\le\Delta}\left(\int_t^s\xi(X_{u-})dL_u\right)^2\middle|\mathcal{F}_t\right]\right]\\
&\lesssim \Delta E\left[\sup_{|s-t|\le\Delta}\int_t^s b^2(X_u)du\right]+E\left[E\left[\int_t^s\sigma^2(X_u)du\middle|\mathcal{F}_t\right]\right]+E\left[E\left[\int_t^s\xi^2(X_u)du\middle|\mathcal{F}_t\right]\right]\\
&\le \Delta\int_t^{t+\Delta}E[b^2(X_u)]du+\int_t^{t+\Delta}E[\sigma^2(X_u)]du+\int_t^{t+\Delta}E[\xi^2(X_u)]du\\
&\le \Delta\left(\Delta E[b^2(X_0)]+\sigma_0+\xi_0\right)\lesssim\Delta.
\end{aligned}
$$

□

We want to remark some facts we have used. Δ is specified as the sampling frequency of X, so one can think of a small number. The functions σ and ξ are quite well manageable. The only term which can cause trouble is the first one involving the drift function. For our last deduction, we implicitly used the Lipschitz-continuity of b, the stationarity, and the fact that $E[X_0^2]<\infty$:

$$
\begin{aligned}
E[b^2(X_0)] &= E[(b(X_0)-b(0)+b(0))^2]\lesssim E[(b(X_0)-b(0))^2]+b^2(0)\\
&\le L_bE[X_0^2]+b^2(0)\le C,
\end{aligned}
$$

where L_b denotes the Lipschitz constant of b and C is a generic constant.

Remark 2.5. *We can also bound the increments of X of order $2p$, $p\ge 1$, by imposing that $\int_{\mathbb{R}}y^{2p}\nu(dy)<\infty$ and the successive use of the Hölder inequality. Moreover, the Burkholder-Davis-Gundy inequalities hold true even for higher moments of the mentioned integrals; see Schmisser (2014). For our purposes, it suffices to bound the squared increments.*

Now we are ready to define the new drift estimator $\hat{b}(x)$ based on the high frequency sample $X_0, X_\Delta, ..., X_{n\Delta}$ such that $\Delta \to 0$ and $n\Delta \to \infty$ as $n \to \infty$. Using the available sample we define

$$\hat{b}(x) := \frac{\frac{1}{nh}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)\frac{(X_{(i+1)\Delta}-X_{i\Delta})}{\Delta}}{\frac{1}{nh}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)}.$$

We now state assumptions on the kernel function K as well as on the speed of convergence of Δ and h.

Assumption A2

i) Let K be a bounded probability density function, which is symmetric around zero, differentiable and Lipschitz-continuous. Hence, K possesses a bounded derivative.

ii) Let K fulfill

$$\int_{\mathbb{R}} z^2 K(z)dz < \infty, \quad \int_{\mathbb{R}} K^2(z)dz < \infty.$$

iii) Let Δ and h fulfill

$$n\Delta h^2 = Th^2 \to \infty \text{ and } \Delta^{1/2}h^{-2} \to 0 \text{ as } n \to \infty.$$

Let us shortly remark that A2,i) and A2,ii) are standard in kernel based estimation procedures. One could also allow higher order kernels. A kernel function K is of order $l \in \mathbb{N}$, if

$$\int_{\mathbb{R}} K(z)dz = 1, \int_{\mathbb{R}} z^j K(z)dz = 0, \quad j = 1, ..., l-1 \text{ and } \int_{\mathbb{R}} z^l K(z)dz < \infty.$$

This would cause an asymptotic bias reduction, but the proofs are rather long and invoke tedious calculations. Further, such kernels lack the property of being a probability density anymore. These are reasons why we will only focus on kernels of order two. Certainly, all proofs can be carried over to more general kernels.
For our purposes, one can always think of the Gaussian kernel

$$K_G(z) := \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

as a toy-example for which all assumptions hold true. Of course, there are many other choices for kernel functions possible, but in the literature it is quite reputable that the choice of the kernel function is not as important as the choice of the bandwidth parameter. Later on, we will present an example for which assumption A2, iii) is fulfilled. In general, it turns out that Δ has to converge relatively fast compared to h, which means that

$$h >> \Delta.$$

2.4 Consistency of $\hat{b}(x)$

We are now ready to state our first theorem, namely the weak consistency of $\hat{b}(x)$.

Theorem 2.6. *Let assumptions A1 and A2 hold true. Then, provided that $\pi(x) > 0$, we can conclude that*

$$\hat{b}(x) \xrightarrow{P} b(x), \quad \textit{as } n \to \infty.$$

We remark an essential difference to the findings in Bandi and Nguyen (2003) and to Mancini and Renó (2011): due to the possible occurrence of infinitely many jumps on a finite interval, we are only able to prove convergence in probability of our proposed estimator. The reason for this weaker result is relatively intuitive. It turns out that one of the key points of their proofs of the strong consistency is the uniform boundedness of increments of X on small intervals. They defined for this purpose the value

$$\delta_{n,T} := \max_{i \leq n-1} \sup_{i\Delta \leq s \leq (i+1)\Delta} |X_{s-} - X_{i\Delta}|.$$

Using Lévy´s modulus of continuity of the Brownian motion and due to the fact that only finitely many jumps occur on every small interval, they deduced that

$$\limsup_{n\to\infty} \frac{\delta_{n,T}}{(\Delta \log(\Delta^{-1}))^{1/2}} = C, \quad a.s.$$

for some constant C; see Bandi and Nguyen (2003), equation (95).
We are not able to specify a comparable almost surely bound in our setting and, hence, will only deduce the convergence in probability of our proposed estimator.

Proof of Theorem 2.6. We will now derive the weak consistency of $\hat{b}(x)$. For this purpose we will divide the proof into different steps. At first, we decompose $\hat{b}(x)$ due to its definition:

$$\begin{aligned}
\hat{b}(x) &= \frac{\frac{1}{nh}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)\left(X_{(i+1)\Delta} - X_{i\Delta}\right)}{\frac{\Delta}{nh}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)} \\
&= \frac{\frac{1}{h}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)\left(\int_{i\Delta}^{(i+1)\Delta} b(X_s)ds + \int_{i\Delta}^{(i+1)\Delta} \sigma(X_s)dW_s + \int_{i\Delta}^{(i+1)\Delta} \xi(X_{s-})dL_s\right)}{\frac{\Delta}{h}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)} \\
&:= \frac{\text{I+II+III}}{\text{IV}}.
\end{aligned}$$

Each term will be handled separately and we will start with the derivation of the denominator

$$\begin{aligned}
\mathrm{IV} &= \frac{\Delta}{h}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right) \\
&= \int_0^T \frac{1}{h}K\left(\frac{x-X_s}{h}\right)ds + \frac{\Delta}{h}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right) - \int_0^T \frac{1}{h}K\left(\frac{x-X_s}{h}\right)ds \\
&= \int_0^T \frac{1}{h}K\left(\frac{x-X_s}{h}\right)ds + \frac{1}{h}\sum_{i=0}^{n-1}\int_{i\Delta}^{(i+1)\Delta}\left(K\left(\frac{X_{i\Delta}-x}{h}\right) - K\left(\frac{x-X_s}{h}\right)\right)ds \\
&:= \int_0^T \frac{1}{h}K\left(\frac{x-X_s}{h}\right)ds + F_1^n. \qquad (2.3)
\end{aligned}$$

We are interested in the rate of convergence of the approximation error F_1^n and consider therefore its L^1-distance. Under the assumption that K is Lipschitz-continuous, by Proposition 2.3, the Cauchy-Schwarz as well as the Jensen inequality, we conclude that

$$\begin{aligned}
E[|F_1^n|] &\leq \frac{1}{h}\sum_{i=0}^{n-1} E\left[\int_{i\Delta}^{(i+1)\Delta}\left|K\left(\frac{X_{i\Delta}-x}{h}\right) - K\left(\frac{x-X_s}{h}\right)\right|ds\right] \\
&\leq \frac{1}{h}\sum_{i=0}^{n-1} E\left[\int_{i\Delta}^{(i+1)\Delta}||K'||_\infty\left|\frac{X_{i\Delta}-X_s}{h}\right|1_{[i\Delta,(i+1)\Delta]}(s)ds\right] \\
&\leq \frac{||K'||_\infty}{h^2}\sum_{i=0}^{n-1} E\left[\left(\int_{i\Delta}^{(i+1)\Delta}(X_{i\Delta}-X_s)^2 ds\right)^{1/2}\right]\Delta^{1/2} \\
&\leq \frac{\Delta^{1/2}||K'||_\infty}{h^2}\sum_{i=0}^{n-1}\int_{i\Delta}^{(i+1)\Delta}\left(E[X_{i\Delta}-X_s)^2]\right)^{1/2}ds \\
&\lesssim \frac{\Delta^{1/2}}{h^2}\sum_{i=0}^{n-1}\int_{i\Delta}^{(i+1)\Delta}\Delta^{1/2}ds = \frac{n\Delta^2}{h^2} = \frac{T\Delta}{h^2}.
\end{aligned}$$

Using the Markov inequality, we can deduce that

$$\begin{aligned}
&P(|F_1^n| > \varepsilon) \leq \frac{E[|F_1^n|]}{\varepsilon} = O(T\Delta h^{-2}), \quad \varepsilon > 0 \\
&\Rightarrow F_1^n = O_P\left(\frac{T\Delta}{h^2}\right), \text{ as } n \to \infty.
\end{aligned}$$

As we want to use the ergodic theorem for the first term of (2.3), we will multiply both terms by $\frac{1}{T}$. Definitely, we will do the same for the derivation of the numerator. As $T \to \infty$

and $h \to 0$ we conclude that

$$\frac{1}{T} \cdot \mathrm{IV} = \frac{1}{T} \int_0^T \frac{1}{h} K\left(\frac{x - X_s}{h}\right) ds + O_P\left(\frac{\Delta}{h^2}\right)$$
$$\longrightarrow \pi(x), \text{ as } n, T \to \infty,$$

which means that

$$\frac{1}{T} \cdot \mathrm{IV} \xrightarrow{P} \pi(x), \text{ as } n, T \to \infty.$$

Now we will derive the three parts in the numerator of $\hat{b}(x)$. We start with the drift term I:

$$\begin{aligned}
\mathrm{I} &= \frac{1}{h} \sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta} - x}{h}\right) \int_{i\Delta}^{(i+1)\Delta} b(X_s) ds \\
&= \int_0^T \frac{1}{h} K\left(\frac{x - X_s}{h}\right) b(X_s) ds + \frac{1}{h} \sum_{i=0}^{n-1} \int_{i\Delta}^{(i+1)\Delta} \left(K\left(\frac{X_{i\Delta} - x}{h}\right) - K\left(\frac{x - X_s}{h}\right)\right) b(X_s) ds \\
&:= \int_0^T \frac{1}{h} K\left(\frac{x - X_s}{h}\right) b(X_s) ds + F_b^n. \qquad (2.4)
\end{aligned}$$

We again derive the L^1-distance of the approximation error F_b^n. We will use comparable techniques as before, namely Proposition 2.3 and the Cauchy-Schwarz inequality. We will also rely on the fact that $b(X_0) \in L^2(\Gamma(dx))$ and conclude as follows:

$$\begin{aligned}
E[|F_b^n|] &\leq \frac{1}{h} \sum_{i=0}^{n-1} \int_{i\Delta}^{(i+1)\Delta} E\left[\left|K\left(\frac{X_{i\Delta} - x}{h}\right) - K\left(\frac{x - X_s}{h}\right)\right| \cdot |b(X_s)|\right] ds \\
&\leq \frac{1}{h} \sum_{i=0}^{n-1} \int_{i\Delta}^{(i+1)\Delta} \left(E\left[\left(K\left(\frac{X_{i\Delta} - x}{h}\right) - K\left(\frac{x - X_s}{h}\right)\right)^2\right]\right)^{1/2} \left(E\left[b^2(X_s)\right]\right)^{1/2} ds \\
&\leq \frac{1}{h} \sum_{i=0}^{n-1} \int_{i\Delta}^{(i+1)\Delta} \left(E\left[||K'||_\infty^2 \left(\frac{X_{i\Delta} - X_s}{h}\right)^2\right]\right)^{1/2} \left(E\left[b^2(X_s)\right]\right)^{1/2} ds \\
&= \frac{\left(E\left[b^2(X_0)\right]\right)^{1/2} ||K'||_\infty}{h^2} \sum_{i=0}^{n-1} \int_{i\Delta}^{(i+1)\Delta} \left(E\left[(X_{i\Delta} - X_s)^2\right]\right)^{1/2} ds \\
&\lesssim \frac{\Delta^{1/2} T}{h^2}.
\end{aligned}$$

We multiply again by T^{-1} and conclude that

$$\frac{1}{T} \cdot F_b^n = O_P\left(\frac{\Delta^{1/2}}{h^2}\right).$$

Again, this term converges to zero due to A2, iii).
Finally, we receive for (2.4), as n and T diverge simultaneously, that

$$\frac{1}{T} \cdot \mathrm{I} \xrightarrow{P} b(x)\pi(x), \text{ as } n, T \to \infty,$$

where we used the standard substitution from above and recalled that b as well as π are continuous functions and that $K(z)dz$ is a probability distribution.
Now we will handle term II:

$$\mathrm{II} = \frac{1}{h}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta} - x}{h}\right)\int_{i\Delta}^{(i+1)\Delta} \sigma(X_s)dW_s.$$

We remark that this term is a martingale-difference sequence. Moreover, due to the stationary and independent increments of the Brownian motion W, we are allowed to derive the L^2-distance of this term in an easy manner. At first, remember that the probability space was endowed with a filtration of sub-σ-algebras $\mathcal{F}_t = \sigma(X_0, (W_s, L_s), s \leq t)$. With respect to this filtration, we can make use of conditional expectations to conclude that

$$E[\mathrm{II}] = E[E[\mathrm{II}|\mathcal{F}_{i\Delta}]] = E\left[\frac{1}{h}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta} - x}{h}\right)\underbrace{E\left[\int_{i\Delta}^{(i+1)\Delta} \sigma(X_s)dW_s \middle| \mathcal{F}_{i\Delta}\right]}_{=0}\right] = 0.$$

Now, the variance can be decomposed into

$$\begin{aligned}
&E[\mathrm{II}^2] = E\left[\left(\frac{1}{h}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta} - x}{h}\right)\int_{i\Delta}^{(i+1)\Delta} \sigma(X_s)dW_s\right)^2\right] \\
&= \frac{1}{h^2}\sum_{i=0}^{n-1} E\left[K^2\left(\frac{X_{i\Delta} - x}{h}\right) E\left[\int_{i\Delta}^{(i+1)\Delta} \sigma^2(X_s)ds \middle| \mathcal{F}_{i\Delta}\right]\right] \\
&\quad + \frac{1}{h^2}E\left[\sum_{i,j=0; i\neq j}^{n-1} K\left(\frac{X_{i\Delta} - x}{h}\right)\int_{i\Delta}^{(i+1)\Delta} \sigma(X_s)dW_s \; K\left(\frac{X_{j\Delta} - x}{h}\right)\int_{j\Delta}^{(j+1)\Delta} \sigma(X_u)dW_u\right] \\
&\leq \sigma_1 \frac{\Delta}{h^2}\sum_{i=0}^{n-1} E\left[K^2\left(\frac{X_{i\Delta} - x}{h}\right)\right] \\
&\quad + \frac{2}{h^2}E\left[\sum_{i<j} K\left(\frac{X_{i\Delta} - x}{h}\right)\int_{i\Delta}^{(i+1)\Delta} \sigma(X_s)dW_s \; K\left(\frac{X_{j\Delta} - x}{h}\right)\int_{j\Delta}^{(j+1)\Delta} \sigma(X_u)dW_u\right] \\
&\leq \frac{\sigma_1 n\Delta ||K^2||_\infty}{h^2}.
\end{aligned}$$

For the last step, we used the tower property of conditional expectations as follows:

$$E\left[\sum_{i<j} K\left(\frac{X_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta}\sigma(X_s)dW_s\ K\left(\frac{X_{j\Delta}-x}{h}\right)\int_{j\Delta}^{(j+1)\Delta}\sigma(X_u)dW_u\right]$$

$$= E\Bigg[\sum_{i<j} K\left(\frac{X_{i\Delta}-x}{h}\right)$$

$$E\Bigg[\int_{i\Delta}^{(i+1)\Delta}\sigma(X_s)dW_s\ K\left(\frac{X_{j\Delta}-x}{h}\right)\underbrace{E\left[\int_{j\Delta}^{(j+1)\Delta}\sigma(X_u)dW_u\Big|\mathcal{F}_{j\Delta}\right]}_{=0}\Big|\mathcal{F}_{i\Delta}\Bigg]\Bigg] = 0.$$

Using Chebyshev´s inequality and assumption A2, ii) we are finally able to conclude that

$$P\left(T^{-1}|\mathrm{II}|>\varepsilon\right)\leq\frac{E[\mathrm{II}^2]}{T^2\varepsilon^2}=O\left(\frac{T}{T^2h^2}\right)=O\left(\frac{1}{Th^2}\right)=o(1),\ \text{ as } n\to\infty$$

and therefore

$$\frac{1}{T}\cdot\mathrm{II}=O_P((Th^2)^{-1/2})=o_P(1),\ \text{ as } n\to\infty.$$

Finally, the term

$$\mathrm{III}=\frac{1}{h}\sum_{i=0}^{n-1}K\left(\frac{X_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta}\xi(X_{s-})dL_s$$

can be handled in the exact same manner as the Brownian part. We assumed that L is an L^2-martingale, which can be represented as an integral with respect to a compensated random measure. This is why term III is a martingale-difference sequence, too. Using the Burkholder-Davis-Gundy type inequality in Lemma 2.4 for Lévy driven stochastic integrals, we can derive the same asymptotic rate of convergence as for term II:

$$\frac{1}{T}\cdot\mathrm{III}=O_P((Th^2)^{-1/2})=o_P(1),\ \text{ as } n\to\infty.$$

Now we can summarize our recent findings and are able to complete the proof by the quotient limit theorem for stationary Markov processes; see Bandi and Nguyen (2003), p.

317:

$$
\begin{aligned}
\hat{b}(x) &= \frac{\frac{1}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)(X_{(i+1)\Delta} - X_{i\Delta})}{\frac{\Delta}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)} \\
&= \frac{\frac{1}{T}\int_0^T \frac{1}{h}K\left(\frac{X_s-x}{h}\right) b(X_s)ds + \frac{1}{T}\cdot F_b^n + \frac{1}{T}\cdot \mathrm{II} + \frac{1}{T}\cdot \mathrm{III}}{\frac{1}{T}\int_0^T \frac{1}{h}K\left(\frac{X_s-x}{h}\right) ds + \frac{1}{T}\cdot F_1^n} \\
&= \frac{\frac{1}{T}\int_0^T \frac{1}{h}K\left(\frac{X_s-x}{h}\right) b(X_s)ds + O_P\left(\frac{\Delta^{1/2}}{h^2}\right) + O_P\left(\frac{1}{Th^2}\right)}{\frac{1}{T}\int_0^T \frac{1}{h}K\left(\frac{X_s-x}{h}\right) ds + O_P\left(\frac{\Delta}{h^2}\right)} \\
&= \frac{b(x)\pi(x) + o_P(1)}{\pi(x) + o_P(1)} = b(x) + o_P(1), \text{ as } n, T \to \infty.
\end{aligned}
$$

Recall that we assumed that $\pi(x) > 0$ holds true. □

2.5 Derivation of the asymptotic distribution

From a practical point of view, it is desirable to be able to construct confidence intervals for estimated values of $b(x)$. To derive our second important theorem, namely the asymptotic normality of $\hat{b}(x)$, we have to strengthen our assumptions in a slightly different manner.

Assumption A3

i) Let the drift function b as well as the stationary density π be twice continuously differentiable.

ii) Let Δ and h satisfy

$$n\Delta h^5 \to 0 \text{ and } n\Delta^2 h^{-3} \to 0 \text{ as } n \to \infty.$$

iii) Let the Lévy-measure ν fulfill

$$\int_{\mathbb{R}} y^4 \nu(dy) < \infty.$$

Now we are ready to state our next important theorem.

Theorem 2.7. *Under Assumptions A1-A3, provided that $\pi(x) > 0$, it holds that*

$$\sqrt{n\Delta h}(\hat{b}(x) - b(x)) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{||K||_2^2(Var(L(1))\xi^2(x) + \sigma^2(x))}{\pi(x)}\right), \text{ as } n \to \infty.$$

We shortly remark that Assumption A3, ii) ensures that the bias term is negligible.
It will turn out that the key point of the derivation of the asymptotic distribution will be a central limit theorem (CLT) for arrays of martingale-difference sequences. We will make use of the following version stated in Shiryayev, Probability (1995), page 511.

Theorem 2.8. *Let $(Y_{in}, \mathcal{F}_{in})_{n\in\mathbb{N}}$ be a square-integrable array of martingale-difference sequences satisfying the Lindeberg condition: if for $\varepsilon > 0$*

$$\sum_{i=0}^{\lfloor nt \rfloor} E[Y_{i,n}^2 \cdot 1(|Y_{i,n}| > \varepsilon)|\mathcal{F}_{i-1,n}] \xrightarrow{P} 0 \text{ as } n \to \infty \tag{2.5}$$

for a $0 < t \leq 1$, it holds that

$$\begin{aligned}
&1.) \quad \sum_{i=0}^{\lfloor (n-1)t \rfloor} E[Y_{i,n}^2|\mathcal{F}_{i-1,n}] \xrightarrow{P} \sigma_t^2 \Longrightarrow \sum_{i=0}^{\lfloor (n-1)t \rfloor} Y_{i,n} \overset{\mathcal{D}}{\to} \mathcal{N}(0, \sigma_t^2)\\
&2.) \quad \sum_{i=0}^{\lfloor (n-1)t \rfloor} Y_{i,n}^2 \xrightarrow{P} \sigma_t^2 \Longrightarrow \sum_{i=0}^{\lfloor (n-1)t \rfloor} Y_{i,n} \overset{\mathcal{D}}{\to} \mathcal{N}(0, \sigma_t^2).
\end{aligned}$$

Another interesting result, stated below, will also be used and is concerned with moment properties of Lévy-driven stochastic integrals.

Lemma 2.9. *Let $X = (X_t)_{t\geq 0}$ be the solution of (2.2) and let f be a bounded and continuous function. Moreover, let $\Xi = (\Xi_t)_{t\geq 0}$ be a centered pure jump Lévy process possessing the property that $\int_{\mathbb{R}} y^4 \nu(dy) < \infty$. Define the process $Y(t) = \int_0^t f(X_{s-})d\Xi_s$, then it holds that*

$$\begin{aligned}
&i) \quad E[Y^2(t)] = E\left[\left(\int_0^t f(X_{s-})d\Xi_s\right)^2\right] = \int_0^t E[f^2(X_s)]dsVar(\Xi(1))\\
&ii) \quad E[Y^4(t)] = 6\int_0^t E[Y^2(s)f^2(X_s)]dsVar(\Xi(1))\\
&\qquad + 4\int_0^t E[Y(s)f^3(X_s)]ds\int_{\mathbb{R}} y^3\nu(dy) + \int_0^t E[f^4(X_s)]ds\int_{\mathbb{R}} y^4\nu(dy).
\end{aligned}$$

For the derivation of this result we will need the Itô-formula for general semimartingales. We will state this fundamental result below, which can be found for instance in Protter (2005), Chapter 7, Theorem 32. We have already implicitly used it in the context of the approximation of the infinitesimal generator of the considered jump process X.

Theorem 2.10. *Let $X = (X_t)_{t\geq 0}$ be a semimartingale and let g be a real valued C^2-function. Then $g(X) = (g(X_t))_{t\geq 0}$ is again a semimartingale and the following formula* *holds true:*

$$g(X_t) - g(X_0) = \int_{(0,t]} g'(X_{s-})dX_s + \frac{1}{2}\int_{(0,t]} g''(X_{s-})d[X,X]_s^c$$
$$+ \sum_{0<s\leq t} (g(X_s) - g(X_{s-}) - g'(X_{s-})\Delta X_s).$$

Proof of Lemma 2.9. Recall that Ξ is a martingale with respect to its augmented canonical filtration.

Using the Itô-formula for semimartingales and especially choosing $g(x) = x^2$, we are able to derive $Y^2(t)$ as

$$\begin{aligned} E[Y^2(t)] = E\left[\int_0^t \int_{\mathbb{R}} \left((Y(s_-)+y)^2 - Y^2(s_-) - 2Y(s_-)y\right) \tilde{\nu}_s^Y(dy)ds\right] \\ + E\left[\int_0^t \int_{\mathbb{R}} \left((Y(s_-)+y)^2 - Y^2(s_-) - 2Y(s_-)y\right) (\tilde{\mu}_s^Y - \tilde{\nu}_s^Y)(dy,ds)\right] \\ := E\left[\int_0^t \int_{\mathbb{R}} \left(Y^2(s_-) + 2Y(s_-)y + y^2 - Y^2(s_-) - 2Y(s_-)y\right) \tilde{\nu}_s^Y)(dy)ds\right] \\ = E\left[\int_0^t \int_{\mathbb{R}} y^2 \tilde{\nu}_s^Y(dy)ds\right], \end{aligned}$$

where $\tilde{\mu}_s^Y(dy,ds)$ is the Poisson random measure corresponding to the process Y with its intensity measure $\tilde{\nu}_s^Y(dy)ds$. Furthermore, $\bar{\tilde{\mu}}_s^Y(dy,ds)$ denotes the compensated random measure. Observe that the intensity measure is now time-dependent and that the second term is a martingale by construction. To derive a closed form of the time-dependent random measure, we make use of the following representation of $\tilde{\nu}_s^Y(A)$

$$\tilde{\nu}_s^Y(A) = \int_{\mathbb{R}} 1_A(f(X_s)x)\nu(dx), \quad \forall\ A \in \mathcal{B}(\mathbb{R}),\ 0 \notin A,$$

which can, for example, be found in Kallsen (2006), Proposition 2.4.

Therefore, we can conclude that

$$\begin{aligned} E[Y^2(t)] &= E\left[\int_0^t \int_{\mathbb{R}} y^2 \tilde{\nu}_s^Y(dy)ds\right] \\ &= \int_0^t \int_{\mathbb{R}} E[f^2(X_s)]x^2\nu(dx)ds = \int_0^t E[f^2(X_s)]dsE[L^2(1)]. \end{aligned}$$

For the derivation of the second statement, we will again make use of the Itô-formula but now by the use of $g(x) = x^4$:

$$\begin{aligned}
Y^4(t) &= \int_0^t \int_{\mathbb{R}} \left((Y(s_-) + y)^4 - Y^4(s_-) - 4Y^3(s_-)y\right) \tilde{\nu}_s^Y(dy)ds \\
&\quad + \int_0^t \int_{\mathbb{R}} \left((Y(s_-) + y)^4 - Y^4(s_-) - 4Y^3(s_-)y\right) (\tilde{\mu}_s^Y - \tilde{\nu}_s^Y)(dy, ds) \\
&= \int_0^t \int_{\mathbb{R}} \left(6Y^2(s_-)y^2 + 4Y(s_-)y^3 + y^4\right) \tilde{\nu}_s^Y(dy)ds \\
&\quad + \int_{\mathbb{R}} \left(6Y^2(s_-)y^2 + 4Y(s_-)y^3 + y^4\right) \bar{\tilde{\mu}}_s^Y(dy, ds) \\
&= 6\int_0^t \int_{\mathbb{R}} Y^2(s_-)y^2\tilde{\nu}_s^Y(dy)ds + 4\int_0^t \int_{\mathbb{R}} Y(s_-)y^3\tilde{\nu}_s^Y(dy)ds \\
&\quad + \int_0^t \int_{\mathbb{R}} y^4\tilde{\nu}_s^Y(dy)ds + M_t,
\end{aligned}$$

where M_t denotes the martingale part. Using analogous arguments, we are now able to deduce that

$$\begin{aligned}
E[Y^4(t)] &= 6\int_0^t E[Y^2(s)f^2(X_s)]ds \int_{\mathbb{R}} y^2\nu(dy) \\
&\quad + 4\int_0^t E[Y(s)f^3(X_s)]ds \int_{\mathbb{R}} y^3\nu(dy) + \int_0^t \int_{\mathbb{R}} E[f^4(X_s)]ds \int_{\mathbb{R}} y^4\nu(dy) \\
&= 6\int_0^t E[Y^2(s)f^2(X_s)]dsE[L^2(1)] + 4\int_0^t E[Y(s)f^3(X_s)]ds \int_{\mathbb{R}} y^3\nu(dy) \\
&\quad + \int_0^t \int_{\mathbb{R}} E[f^4(X_s)]ds \int_{\mathbb{R}} y^4\nu(dy),
\end{aligned}$$

which finishes the proof. □

An elementary consequence is the following upper bound

Corollary 2.11. *Under the assumptions of Lemma 2.9, we can derive the following upper bounds*

$$E[Y^2(t)] \leq ||f||_\infty^2 Var(L(1))t$$

as well as

$$E[Y^4(t)] \leq 3Var^2(L(1))||f||_\infty^4 t^2 + \frac{8\sqrt{Var(L(1))}||f||_\infty^4}{3} \int_{\mathbb{R}} y^3\nu(dy)t^{3/2} + ||f||_\infty^4 \int_{\mathbb{R}} x^4\nu(dx)t.$$

Proof of Corollary 2.11. We only state the proof of the second statement:

$$\begin{aligned}
E[Y(t)^4] &= 6\int_0^t E[Y^2(s)f^2(X_s)]dsVar(L(1)) \\
&\quad + 4\int_0^t E[Y(s)f^3(X_s)]ds\int_{\mathbb{R}} y^3\nu(dy) + \int_0^t E[f^4(X_s)]ds\int_{\mathbb{R}} y^4\nu(dy) \\
&\leq 6||f||_\infty^2 Var(L(1))\int_0^t E[Y^2(s)]ds + 4\sqrt{Var(L(1))}||f||_\infty^4\int_{\mathbb{R}} y^3\nu(dy)\int_0^t \sqrt{s}ds \\
&\quad + \int_0^t E[f^4(X_s)]\int_{\mathbb{R}} x^4\nu(dx) \\
&\leq 6Var^2(L(1))||f||_\infty^4\int_0^t sds + \frac{8\sqrt{Var(L(1))}||f||_\infty^4}{3}\int_{\mathbb{R}} y^3\nu(dy)t^{3/2} \\
&\quad + \int_0^t E[f^4(X_s)]\int_{\mathbb{R}} x^4\nu(dx) \\
&= 3Var^2(L(1))||f||_\infty^4 t^2 + \frac{8\sqrt{Var(L(1))}||f||_\infty^4}{3}\int_{\mathbb{R}} y^3\nu(dy)t^{3/2} \\
&\quad + \int_0^t E[f^4(X_s)]\int_{\mathbb{R}} x^4\nu(dx).
\end{aligned}$$

□

Now we are able to state the proof of the asymptotic normality.

Proof of Theorem 2.7. We will start with a decomposition, comparable to derivation of the consistency of $\hat{b}(x)$:

$$\begin{aligned}
\sqrt{Th}(\hat{b}(x) - b(x)) &= \sqrt{Th}\left(\frac{\frac{1}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)(X_{(i+1)\Delta} - X_{i\Delta})}{\frac{\Delta}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)} - b(x)\right) \\
= \sqrt{Th}&\left(\frac{\frac{1}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta}(b(X_s) - b(x))ds}{\frac{\Delta}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)}\right) \\
+ \sqrt{Th}&\left(\frac{\frac{1}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)\left(\int_{i\Delta}^{(i+1)\Delta}\sigma(X_s)dW_s + \int_{i\Delta}^{(i+1)\Delta}\xi(X_{s-})dL_s\right)}{\frac{\Delta}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)}\right) \\
= \sqrt{Th}&\left(\frac{\frac{1}{T}\int_0^T \frac{1}{h}K\left(\frac{x-X_s}{h}\right)(b(X_s) - b(x))ds + F_b^n + \frac{1}{T}F_1^n}{\frac{\Delta}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)}\right) \qquad (2.6) \\
+ &\frac{\frac{1}{\sqrt{Th}}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)\left(\int_{i\Delta}^{(i+1)\Delta}\sigma(X_s)dW_s + \int_{i\Delta}^{(i+1)\Delta}\xi(X_{s-})dL_s\right)}{\frac{\Delta}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)}
\end{aligned}$$

The first term in (2.6) is a bias term and negligible as we will see in our subsequent analysis, because of $Th^5 = n\Delta h^5 \to 0$ as $n, T \to \infty$. By the use of a Taylor expansion of the functions b and π around x, we are able to handle the first term. Due to the fact that b and π are twice continuously differentiable, the remainder terms are negligible and we can conclude that

$$\begin{aligned}
&\int_{\mathbb{R}} \frac{1}{h} K\left(\frac{x-y}{h}\right)(b(y)-b(x))\,\pi(y)dy = \int_{\mathbb{R}} K(z)(b(x-zh)-b(x))\pi(x-zh)dz \\
&= \int_{\mathbb{R}} K(z)\left(-zhb'(x) + \frac{z^2h^2}{2}b''(x) + O(h^3)\right)\left(\pi(x) - zh\pi'(x) + O(h^2)\right)dz \\
&= \int K(z)\left(b'(x)\pi(x)(-zh) + z^2h^2b'(x)\pi'(x)\frac{z^2h^2}{2}b''(x)\pi(x)\right)dz + O(h^3) \\
&= h^2 \int z^2 K(z)dz \left(b'(x)\pi'(x) + \frac{b''(x)\pi(x)}{2}\right) + O(h^3) = O(h^2). \qquad (2.7)
\end{aligned}$$

Now, by letting $T \to \infty$, we can use the ergodic property of the process X and are able to derive the asymptotic rate of convergence of the numerator of (2.6) by the use of (2.7):

$$\begin{aligned}
&\sqrt{Th}\left(\frac{1}{T}\int_0^T \frac{1}{h}K\left(\frac{x-X_s}{h}\right)(b(X_s)-b(x))\,ds + F_b^n + \frac{1}{T}F_1^n\right) \\
&= O((n\Delta h)^{1/2}h^2) + O_P((n\Delta h)^{1/2}\Delta^{1/2}h^{-2}) = o_P(1), \text{ as } n \to \infty.
\end{aligned}$$

Where the last equality follows from A3, ii).
We already know that the denominator is a consistent estimate of $\pi(x)$. Hence, we will now treat the remaining two terms of the numerator, which we define as

$$\sum_{i=0}^{n-1} \eta_{i+1,n} := \frac{1}{\sqrt{Th}} \sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta} \sigma(X_s)dW_s = \sum_{i=0}^{n-1} \tilde{\eta}_{i+1,n} + F_W^n$$

as well as

$$\sum_{i=0}^{n-1} \zeta_{i+1,n} := \frac{1}{\sqrt{Th}} \sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta} \xi(X_{s-})dL_s := \sum_{i=0}^{n-1} \tilde{\zeta}_{i+1,n} + F_n^L,$$

where

$$F_n^L = \frac{1}{\sqrt{Th}} \sum_{i=0}^{n-1} \int_{i\Delta}^{(i+1)\Delta} \xi(X_{s-})\left(K\left(\frac{x-X_{s-}}{h}\right) - K\left(\frac{x-X_{i\Delta}}{h}\right)\right)dL_s$$

and

$$F_n^W = \frac{1}{\sqrt{Th}} \sum_{i=0}^{n-1} \int_{i\Delta}^{(i+1)\Delta} \sigma(X_s)\left(K\left(\frac{x-X_s}{h}\right) - K\left(\frac{x-X_{i\Delta}}{h}\right)\right)dW_s$$

are denoting the approximation errors. Both terms are negligible in probability, which can be seen through

$$
\begin{aligned}
E[(F_n^L)^2] &= \frac{1}{Th}\sum_{i=0}^{n-1}\int_{i\Delta}^{(i+1)\Delta} E\left[\xi^2(X_s)\left(K\left(\frac{x-X_s}{h}\right)-K\left(\frac{x-X_{i\Delta}}{h}\right)\right)^2\right] Var(L(1))ds \\
&\leq \frac{||K'||_\infty^2||\xi^2||_\infty Var(L(1))}{Th^3}\sum_{i=0}^{n-1}\int_{i\Delta}^{(i+1)\Delta} E\left[(X_{i\Delta}-X_s)^2\right]ds \\
&\lesssim \frac{n\Delta^2}{Th^3} = \frac{\Delta}{h^3} = o(1), \text{ as } n\to\infty,
\end{aligned}
$$

where the last equation follows from the fact that $\Delta^{1/2}h^{-2}\to 0$. Therefore, this term is negligible in probability by the fact that $F_n^L = O_P(\Delta^{1/2}h^{-3/2}) = o_P(1)$ as $n, T\to\infty$. This suffices, because we are interested in convergence in distribution. We only used the fact that L is a process with stationary and independent increments such that the Lévy measure $\nu(dy)$ integrates y^4 to a finite number. Therefore, the order of F_n^W can be found in a similar manner.

Now we will focus on the array of martingale difference sequences $(\tilde{\eta}_{i+1,n}, \mathcal{F}_i)$ and $(\tilde{\zeta}_{i+1,n}, \mathcal{F}_i)$. Both can be treated by the use of central limit theorem 2.8. Recall that we chose $\mathcal{F}_t = \sigma(X_0, (W_s, L_s); s\leq t)$ as filtration. We will start with the verification of the Lindeberg condition (2.5) for $t=1$ by the following useful observation:

$$
\begin{aligned}
\sum_{i=0}^{n-1} E[\tilde{\zeta}_{i+1,n}^2 1(|\tilde{\zeta}_{i+1,n}|>\varepsilon)|\mathcal{F}_{i\Delta}] &= \sum_{i=0}^{n-1}\int_{\mathbb{R}} y^2 1(|y|>\varepsilon)P^{\tilde{\zeta}_{i+1,n}|\mathcal{F}_{i\Delta}}(dy) \\
&\leq \sum_{i=0}^{n-1}\int_{\mathbb{R}}\frac{y^4}{\varepsilon^2}1(|y|>\varepsilon)P^{\tilde{\zeta}_{i+1,n}|\mathcal{F}_{i\Delta}}(dy) \leq \sum_{i=0}^{n-1}\int_{\mathbb{R}}\frac{y^4}{\varepsilon^2}P^{\tilde{\zeta}_{i+1,n}|\mathcal{F}_{i\Delta}}(dy) \\
&= \frac{1}{\varepsilon^2}\sum_{i=0}^{n-1}E[\tilde{\zeta}_{i+1,n}^4|\mathcal{F}_{i\Delta}].
\end{aligned}
$$

Now recall the statement of Lemma 2.4. This lemma enables us to derive upper bounds for moments of Lévy as well as Brownian driven stochastic integrals. Instead of the function ξ, we will now make use of the lemma by the utilization of the function $\tilde{\xi} := K\cdot\xi$, which is again bounded and also continuous. For the determination of the fourth conditional moment of $\tilde{\zeta}_{i+1,n}$ we need the statement of Lemma 2.4 in a slightly more general form. By Schmisser (2014), p.895, it holds, because $\int_{\mathbb{R}} y^4\nu(dy)<\infty$, that

$$
\begin{aligned}
&E\left[\left(\int_t^{t+\Delta} K\left(\frac{x-X_u}{h}\right)\sigma(X_u)dW_u\right)^4\middle|\mathcal{F}_{i\Delta}\right] \\
&\leq E\left[\sup_{|t-s|\leq\Delta}\left(\int_t^{s} K\left(\frac{x-X_u}{h}\right)\sigma(X_u)dW_u\right)^4\middle|\mathcal{F}_{i\Delta}\right]
\end{aligned}
$$

$$\leq \tilde{C}_1 E\left[\left(\int_t^{t+\Delta} K^2\left(\frac{x-X_u}{h}\right)\sigma^2(X_u)du\right)^2\middle|\mathcal{F}_{i\Delta}\right] \leq \tilde{C}_1||K^2\sigma^2||_\infty^2\Delta^2.$$

For the Lévy driven integral we find analogously

$$\begin{aligned}
&E\left[\left(\int_t^{t+\Delta} K\left(\frac{x-X_{u_-}}{h}\right)\xi(X_{u_-})dL_u\right)^4\middle|\mathcal{F}_{i\Delta}\right]\\
&\leq E\left[\sup_{|t-s|\leq\Delta}\left(\int_t^{s} K\left(\frac{x-X_{u_-}}{h}\right)\xi(X_{u_-})dL_u\right)^4\middle|\mathcal{F}_{i\Delta}\right]\\
&\leq \tilde{C}_2(Var(L(1)))^2E\left[\left(\int_t^{t+\Delta} K^2\left(\frac{x-X_u}{h}\right)\xi^2(X_u)du\right)^2\middle|\mathcal{F}_{i\Delta}\right]\\
&\quad+\tilde{C}_2\int_{\mathbb{R}}y^4\nu(dy)E\left[\int_t^{t+\Delta}K^4\left(\frac{x-X_u}{h}\right)\xi^4(X_u)du\middle|\mathcal{F}_{i\Delta}\right]\\
&\leq \tilde{C}_2\left((Var(L(1)))^2||K^2\sigma^2||_\infty^2\Delta^2+\int_{\mathbb{R}}y^4\nu(dy)||K^4\xi^4||_\infty\Delta\right),
\end{aligned}$$

where $\tilde{C}_2$ denotes a deterministic constant.
We shortly remark that both integrands are non-negative and bounded functions such that the result of Schmisser (2014), p.895, can be transferred. These inequalities can also be found in Dellacherie and Meyer (1980), Theorem 92, Chapter VIII and Applebaum (2004), Theorem 4.4.23, p. 265.
Now we are able to verify the Lindeberg condition:

$$\begin{aligned}
&\sum_{i=0}^{n-1}E[\tilde{\zeta}_{i+1,n}^2 1(|\tilde{\zeta}_{i+1,n}|>\varepsilon)|\mathcal{F}_{i\Delta}] \leq \frac{1}{\varepsilon^2}\sum_{i=0}^{n-1}E[\tilde{\zeta}_{i+1,n}^4|\mathcal{F}_{i\Delta}]\\
&=\frac{1}{\varepsilon^2}\sum_{i=0}^{n-1}E\left[\left(\frac{1}{\sqrt{Th}}\int_{i\Delta}^{(i+1)\Delta}K\left(\frac{x-X_{s_-}}{h}\right)\xi(X_{s_-})dL_s\right)^4\middle|\mathcal{F}_{i\Delta}\right]\\
&\leq\frac{1}{T^2h^2\varepsilon^2}\sum_{i=0}^{n-1}\left(\tilde{C}_2\left((Var(L(1)))^2||K^2\sigma^2||_\infty^2\Delta^2+\int_{\mathbb{R}}y^4\nu(dy)||K^4\xi^4||_\infty\Delta\right)\right)\\
&\lesssim\frac{n\Delta^2}{T^2h^2}+\frac{n\Delta}{T^2h^2}=O\left(\frac{1}{nh^2}\right)+O\left(\frac{1}{T^2h^2}\right)=o(1),\text{ as } n,T\to\infty.
\end{aligned}$$

For the derivation of the asymptotic variance of the asymptotic distribution, we make use of the second condition of central limit theorem 2.8:

$$
\begin{aligned}
&\sum_{i=0}^{n-1} E[\tilde{\zeta}^2_{i+1,n}|\mathcal{F}_{i\Delta}] \\
&= \sum_{i=0}^{n-1} E\left[\left(\frac{1}{\sqrt{Th}}\int_{i\Delta}^{(i+1)\Delta} K\left(\frac{x-X_{s_-}}{h}\right)\xi(X_{s_-})dL_s\right)^2 \middle| \mathcal{F}_{i\Delta}\right] \\
&:= \frac{1}{Th}\sum_{i=0}^{n-1} E\left[(\zeta'_{(i+1)\Delta}-\zeta'_{i\Delta})^2|\mathcal{F}_{i\Delta}\right] \\
&= \frac{1}{Th}\sum_{i=0}^{n-1} E\left[(\zeta'_{(i+1)\Delta})^2-(\zeta'_{i\Delta})^2|\mathcal{F}_{i\Delta}\right] - 2E[\zeta'_{i\Delta}(\zeta'_{(i+1)\Delta}-\zeta'_{i\Delta})|\mathcal{F}_{i\Delta}] \\
&= \frac{1}{Th}\sum_{i=0}^{n-1} E\left[(\zeta'_{(i+1)\Delta})^2-(\zeta'_{i\Delta})^2|\mathcal{F}_{i\Delta}\right] - 2\zeta'_{i\Delta}E[\zeta'_{(i+1)\Delta}-\zeta'_{i\Delta}|\mathcal{F}_{i\Delta}] \\
&= \frac{1}{Th}\sum_{i=0}^{n-1} E\left[(\zeta'_{(i+1)\Delta})^2-(\zeta'_{i\Delta})^2|\mathcal{F}_{i\Delta}\right] \\
&= \frac{1}{\Delta h}\frac{2}{n}\sum_{i=1}^{n}\int_{i\Delta}^{(i+1)\Delta} E\left[\zeta'_{s_-}d\zeta'_s \middle| \mathcal{F}_{i\Delta}\right] \\
&\quad + \frac{1}{\Delta h}\frac{1}{n}\sum_{i=1}^{n}\int_{i\Delta}^{(i+1)\Delta} E\left[K^2\left(\frac{x-X_{s_-}}{h}\right)\xi^2(X_{s_-})y^2\bar{\mu}(ds,dy)\middle|\mathcal{F}_{i\Delta}\right] \\
&\quad + \frac{1}{\Delta h}\frac{1}{n}\sum_{i=1}^{n}\int_{i\Delta}^{(i+1)\Delta} E\left[K^2\left(\frac{x-X_s}{h}\right)\xi^2(X_s)\middle|\mathcal{F}_{i\Delta}\right]ds\int_{\mathbb{R}} y^2\nu(dy).
\end{aligned}
$$

By using the ergodicity of X we see that the first two terms converge to zero due to the fact that they are martingales. The third summand converges by invoking the stationarity of X to

$$Var(L(1))\xi^2(x)\pi(x)\int_{\mathbb{R}} K^2(z)dz, \text{ as } n,T\to\infty.$$

According to central limit theorem 2.8, this value denotes the asymptotic variance of this part.

Now we have almost finished the proof of the asymptotic normality. For the Lévy driven part, we are now able to summarize that

$$
\frac{\sqrt{Th}\left(\frac{1}{Th}\sum_{i=0}^{n-1}K\left(\frac{x-X_{i\Delta}}{h}\right)\int_{i\Delta}^{(i+1)\Delta}\xi(X_{s_-})dL_s\right)}{\frac{\Delta}{Th}\sum_{i=0}^{n-1}K\left(\frac{x-X_{i\Delta}}{h}\right)}
$$

$$
\frac{\frac{1}{\sqrt{Th}}\sum_{i=0}^{n-1}K\left(\frac{x-X_{i\Delta}}{h}\right)\int_{i\Delta}^{(i+1)\Delta}\xi(X_{s_-})dL_s}{\pi(x)+O_P\left(\frac{\Delta}{h^2}\right)}
$$

$$\xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{||K||_2^2 Var(L(1))\xi^2(x)}{\pi(x)}\right) \text{ as } n \to \infty,$$

where we used Slutsky´s lemma and the quotient limit theorem; see Bandi and Nguyen (2003), p.317.
Since $(\tilde{\eta}_{i,n}, \mathcal{F}_i)$ is a martingale difference sequence possessing a finite fourth moment, too, the exact analogous argumentation fits. For the sake of brevity we leave the corresponding proof out and only state the crucial point in the following. The Brownian part converges also in distribution to a normal distributed random variable and the asymptotic variance can also be deduced quite easily. In detail, we have

$$\frac{\sqrt{Th}\left(\frac{1}{Th}\sum_{i=0}^{n-1} K\left(\frac{x-X_{i\Delta}}{h}\right)\int_{i\Delta}^{(i+1)\Delta}\sigma(X_s)dW_s\right)}{\frac{\Delta}{Th}\sum_{i=0}^{n-1} K\left(\frac{x-X_{i\Delta}}{h}\right)}$$
$$\xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{||K||_2^2\sigma^2(x)}{\pi(x)}\right) \text{ as } n \to \infty.$$

We recall that L and W are assumed to be independent. This has the advantage that we can sum up the asymptotic variances of both martingale difference sequences to get the final variance. In fact, we have finished the proof and it finally holds that

$$\sqrt{Th}(\hat{b}(x) - b(x)) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \bar{\sigma}^2(x)\right) \text{ as } n \to \infty,$$

where we set $\bar{\sigma}^2(x) := \frac{||K||_2^2(Var(L(1))\xi^2(x)+\sigma^2(x))}{\pi(x)} := \frac{||K||_2^2\tilde{\sigma}^2(x)}{\pi(x)}$. □

Example 2.12. *An interesting question is whether the restrictions on Δ and h can be satisfied and which rate finally occurs. By letting*

$$\Delta \sim n^{-\alpha}, \quad h \sim n^{-\beta}, \quad \alpha,\ \beta > 0,$$

assumptions A2, iii) and A3, ii) can be reformulated according to

$$\alpha + 5\beta > 1,\ 2\alpha - 3\beta < 1,\ \alpha + \beta < 1\ \text{,and } 4\beta < \alpha.$$

One possible answer is to solve the following linear optimization problem. Let

$$G(\alpha, \beta) := \frac{1 - \alpha - \beta}{2}$$

and maximize the function G with respect to the upper restrictions. One possibility to solve this problem is to make use of a simplex algorithm as in Figure 1 providing the approximately optimal result

$$\alpha^* \approx 0.615, \quad \beta^* \approx 0.077,$$

which leads to an optimal rate of $n^{0.154}$.

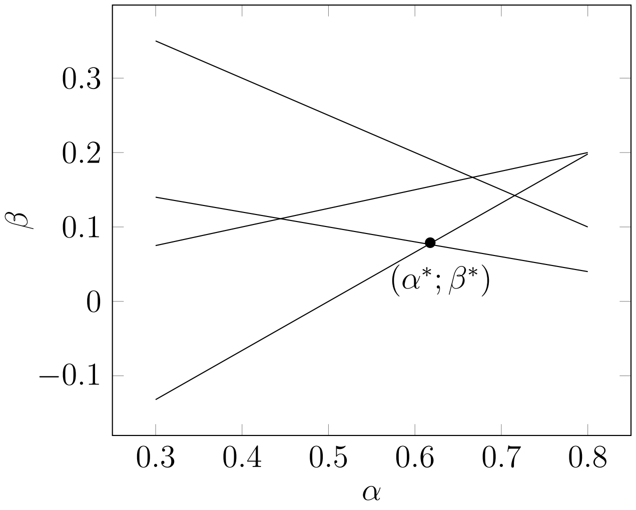

Figure 1: Plot of the constraints on $G(\alpha,\beta)$ and the corresponding optimal coordinates $(\alpha^*;\beta^*)$.

2.6 Examples of possible Lévy processes

In this section, we briefly want to state possible drivers for the stochastic differential equation (2.2) fulfilling our used assumptions. The crucial property is that L has to possess moments up to order four.

Examples

I) Let L be a compound Poisson process with intensity 1 and jump size distribution $\nu(dy)$, which means that

$$L_t = \sum_{i=1}^{N_t} \varsigma_i,$$

where ς_i are independent and identically distributed random variables with $\varsigma_i \sim \nu(dy)$ and N_t is a Poisson process with intensity 1 being independent of ς_i $\forall$ $i = 1, ..., n$. Now we specify the jump size distribution. Note that L is a process of finite activity, which means that $\nu(\mathbb{R}) < \infty$.

i) Let $\delta_a(dy)$ be the Dirac measure in $a \in \mathbb{R}$, then define

$$\nu(dy) := \frac{1}{2}\left(\delta_1(dy) + \delta_{-1}(dy)\right),$$

such that

$$Var(L(1)) = \int_{\mathbb{R}} y^2 \nu(dy) = \frac{1}{4} \text{ and } \int_{\mathbb{R}} y^4 \nu(dy) = \frac{1}{8}.$$

ii) Let

$$\nu(dy) = \frac{1}{2}\exp(-\lambda|y|)dy, \ \lambda > 0,$$

be the Laplace or double-exponential measure, then

$$Var(L(1)) = \int_{\mathbb{R}} y^2\nu(dy) = \frac{1}{2}\int_{\mathbb{R}} y^2 \exp(-\lambda|y|)dy = \frac{2}{\lambda^2} \text{ and } \int_{\mathbb{R}} y^4\nu(dy) = \frac{24}{\lambda^4}.$$

iii) Let

$$\nu(dy) = \varphi(y)dy,$$

where $\varphi(y)$ denotes the density of a standard normal distributed random variable. Then it holds that

$$Var(L(1)) = \int_{\mathbb{R}} y^2\varphi(y)dy = 1 \text{ and } \int_{\mathbb{R}} y^4\nu(dy) = \int_{\mathbb{R}} y^4\varphi(y)dy = 3.$$

II) Now we state examples of possible choices of Lévy processes not being compound Poisson processes.

i) Let L be an infinite activity process possessing a discrete Lévy measure

$$\nu(dy) := \sum_{k=0}^{\infty} 2^{k+2}\left(\delta_{\frac{1}{2^k}}(dy) + \delta_{-\frac{1}{2^k}}(dy)\right).$$

Though $\nu(\mathbb{R}) = \infty$, the required moments exist. Namely, we have that

$$Var(L(1)) = \int_{\mathbb{R}} y^2\nu(dy) = \sum_{k=0}^{\infty} 2^{k+2}\frac{2}{2^k} = 4$$

and

$$\int_{\mathbb{R}} y^4\nu(dy) = \sum_{k=0}^{\infty} 2^{k+2}\frac{2}{2^{2k}} = \sum_{k=0}^{\infty} 2^{3-3k} = \frac{64}{7}.$$

ii) Let L be a Gamma process, which is a purely discontinuous subordinator of relatively low activity and finite variation possessing the Lévy measure

$$\nu(dy) = \gamma y^{-1}\exp(-\lambda y)1_{(0,\infty)}(y)dy, \ \gamma, \lambda > 0.$$

Moreover, it holds that

$$Var(L(1)) = \int_{\mathbb{R}} y^2\nu(dy) = \gamma\int_0^{\infty} y\exp(-\lambda y)dy = \frac{\gamma}{\lambda^2}$$

and

$$\int_{\mathbb{R}} y^4 \nu(dy) = \gamma \int_0^\infty y^3 \exp(-\lambda y) dy = \frac{6\gamma}{\lambda^4}.$$

This process is a special case of the tempered stable subordinators possessing the Lévy measure

$$\nu(dy) = \gamma y^{-\alpha-1} \exp(-\lambda y) 1_{(0,\infty)}(y) dy, \ 0 \leq \alpha < 2.$$

For further details on this class of processes we refer to Cont and Tankov (2004), Section 4.4.

iii) In order to allow an asymmetric behavior of small jumps and also to model flexible decay rates for positive and negative big jumps, a widely referred model is the class of CGMY processes; see Carr et al. (2002). It is a purely discontinuous Lévy process equipped with Lévy measure

$$\nu(dy) = \left(C \frac{\exp(-G|y|)}{|y|^{1+Y}} 1_{(-\infty,0)}(y) + C \frac{\exp(-M|y|)}{|y|^{1+Y}} 1_{(0,\infty)}(y) \right) dy,$$

where $C > 0, G, M \geq 0$ and $Y < 2$. In dependence of Y, this process can have finite or infinite activity and, moreover, is of finite or infinite variation. Furthermore, it holds that

$$Var(L(1)) = \int_{\mathbb{R}} y^2 \nu(dy) = C\Gamma(2-Y) \left(\frac{1}{M^{2-Y}} + \frac{1}{G^{2-Y}} \right).$$

Note that all stated examples fulfill the assumption

$$\int_{|y| \geq 1} |y| \nu(dy) < \infty$$

such that the used representation

$$\begin{aligned} L_t &= \int_0^t \int_{\mathbb{R}} x(\mu^L(ds,dx) - \nu(dx)ds) \\ &= \int_0^t \int_{\{|x|<1\}} x(\mu^L(ds,dx) - \nu(dx)ds) + \int_0^t \int_{\{|x| \geq 1\}} x\mu^L(ds,dx) \\ &\quad - \int_0^t \int_{\{|x| \geq 1\}} x\nu(dx)ds \end{aligned}$$

of L is well defined.

2.7 Bandwidth selection

A very important question for practical issues is how to choose a proper bandwidth h in our model. There is an immense amount of papers exclusively dealing with this topic for nonparametric estimation procedures as density or regression estimation. We will restrict ourselves to three methods, which will be introduced in this section.
In general, the practitioner has n observations sampled at a given frequency Δ. Thus, both parameters are determined by the available data and a third parameter h has to be chosen by the use of certain procedures based on this sample. The question which procedure is optimal is hard to answer. Some selection methods are highly computable, whereas others are based on unknown quantities, which in turn have to be estimated. An overview concerning this problem in the context of nonparametric density estimation can be found in Jones et al. (1996). For kernel based regression estimation, we refer to Vieu (1993).
First of all, we recall Assumption A3, ii), where we assumed that

$$n\Delta h^5 = o(1), \text{ as } n \to \infty.$$

This assumption guarantees that the numerator of the bias term fulfills

$$\frac{1}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta}(b(X_s)-b(x))ds = o_P((n\Delta h)^{-1/2}), \text{ as } n \to \infty.$$

By choosing $h = T^{-1/5} = (n\Delta)^{-1/5}$, this term is not negligible and denotes the occurring bias. Hence, according to Theorem 2.7, the asymptotic mean squared error (AMSE) of $\hat{b}(x)$ is of the following form

$$\begin{aligned}\text{AMSE}(\hat{b}(x)) &= \left(\text{ABIAS}(\hat{b}(x))\right)^2 + \text{AVAR}\left(\hat{b}(x)\right)\\ &= h^4\mu_2^2(K)\Lambda(x) + \frac{\int_{\mathbb{R}} K^2(z)dz\tilde{\sigma}^2(x)}{n\Delta h\pi(x)},\end{aligned}$$

where

$$\Lambda(x) := \frac{b'(x)\pi'(x)}{\pi(x)} + \frac{b''(x)}{2}$$

denotes one part of the bias term and

$$\mu_2(K) = \int_{\mathbb{R}} z^2K(z)dz$$

the second moment of K.
We are able to recognize the well-known tradeoff between these two parts and, hence, in order to minimize the AMSE, we will differentiate the sum with respect to h. The

resulting bandwidth is called "oracle bandwidth" and in our case this optimal bandwidth $h_{opt,oracle}(x)$ has the form

$$h_{opt,oracle}(x) = (n\Delta)^{-1/5}\left(\frac{\tilde{\sigma}^2(x)\int_{\mathbb{R}}K^2(z)dz}{4\mu_2^2(K)\Lambda^2(x)\pi(x)}\right)^{-1/5}.$$

To ensure that this bandwidth fulfills A3, ii), Δ and T have to fulfill

$$T^{8/5}\Delta \to 0.$$

Using this bandwidth, the optimal AMSE is of order

$$AMSE\left(\hat{b}(x)\right) = O\left((n\Delta)^{-4/5}\right).$$

Assuming higher order smoothness properties of b and π, this rate can be fastened. Obviously, $h_{opt,oracle}(x)$ depends on the unknown quantities $\Lambda(x)$, $\tilde{\sigma}^2(x)$ and $\pi(x)$, which all have to be estimated. This task is quite challenging in practical issues, because it is often unclear how two build an appropriate pilot estimator which is a first-stage estimator. One possibility would be to use kernel based estimators again, where the occurring bandwidths are chosen by a rule of thumb, for example $h_{ROT} \equiv (n\Delta)^{-1/5}$. Moreover, recall that $h_{opt,oracle}(x)$ is a local plug-in choice for h and, hence, for every x at which b has to be estimated, a new bandwidth has to be computed. Thus, this method is highly computable, although it provides a natural choice of the bandwidth according to the minimization of the (asymptotic) mean squared error.

An alternative approach is provided by the selection of h invoking a global performance criterion, namely by selecting h such that the integrated mean squared error ("IMSE")

$$\text{IMSE}(\hat{b}) = \int MSE(\hat{b}(x))dx = E\left[\int\left(\hat{b}(x) - b(x)\right)^2 dx\right] = \text{MISE}(\hat{b})$$

is minimized, where the integration takes place over the support of the stationary density π of X. Moreover, we mention that the order of integration can be reversed due to the positivity of the integrand. In our case, the asymptotic IMSE of $\hat{b}$ (AIMSE($\hat{b}$)) has the form

$$\text{AIMSE}(\hat{b}) = h^4\mu_2^2(K)\int\Lambda(x)^2\pi(x)dx + \frac{\int_{\mathbb{R}}K^2(z)dz\int\tilde{\sigma}^2(x)dx}{n\Delta h}.$$

Analogously, we find the representation of the optimal bandwidth parameter $\tilde{h}_{opt,oracle}$ as follows:

$$\tilde{h}_{opt,oracle} = (n\Delta)^{-1/5}\left(\frac{\int_{\mathbb{R}}K^2(z)dz\int\tilde{\sigma}^2(x)dx}{4\int\Lambda^2(x)\pi(x)dx\mu_2^2(K)}\right)^{-1/5}.$$

This choice for the bandwidth is now x-independent, but the appearing integrals have to be discretized and the integrands have to be replaced by estimators afterwards. These estimators are then again dependent on a bandwidth, which, in turn, can also be chosen by an appropriate rule of thumb.

The third presented method is cross-validation. For independent and identically distributed data, this procedure is quite standard in the literature and has extensively been studied; see for example Härdle and Marron (1985) for a pioneering work in the context of nonparametric regression. In our case, the leave-one-out cross-validation method (see Härdle and Marron (1985)) is not appropriate because the available data set contains non-independent copies. Nevertheless, due to our assumptions, the jump-diffusion X is exponentially β-mixing (and, hence, also strong mixing or α-mixing). Thus, the dependency in terms of the correlation decreases as the lag between two observations increases. In the context of mixing data, Chu and Marron (1991) as well as Burman et al. (1994) introduced a generalization of the leave-one-out cross-validation method for dependent (strong mixing) data. Burman et al. (1994) initially called this method the H-block cross-validation, which provides a method for choosing an optimal global bandwidth parameter h_{H-CV}.

The intuition adapted to our model is as follows: Fix an $l \in 1, ..., n$ and estimate $b(X_{l\Delta})$ by a subsample of the available data set $\{X_{i\Delta}\}_{i=1,...,n}$ such that H observations on both sides are removed and $b(X_{l\Delta})$ is then estimated by the remaining $n-(2H+1)$ observations. To ensure asymptotic optimality, H has to be an increasing integer-valued positive sequence. Now define $\hat{b}_{-(l+H)\Delta:(l+H)\Delta}(X_{l\Delta})$ as the estimate of $b(X_{l\Delta})$ based on the sample $\{X_\Delta, X_{2\Delta}, ..., X_{(l-H-1)\Delta}, X_{(l+H+1)\Delta}, ..., X_{n\Delta}\}$. Then, the smoothing parameter h_{H-CV} is selected by

$$\text{H-CV}(h) = \text{argmin}_h \sum_{i=H+1}^{n-H} \left(\frac{X_{(i+1)\Delta} - X_{i\Delta}}{\Delta} - \hat{b}(X_{i\Delta}) \right)^2,$$

see Burmann et al. (1994), where also an ad-hoc choice of the sequence H is given as $H = \lfloor n^{1/4} \rfloor$.

For practical issues, H can, for instance, be selected by analyzing the empirical autocorrelation function of the data.

2.8 Comparison to an alternative nonparametric estimation approach

In this section, we briefly want to compare our approach to the one of Schmisser (2014). Schmisser focuses on the same class of stochastic processes given by the stochastic differential equation

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t + \xi(X_{t-})dL_t, \quad X_0 \overset{\mathcal{D}}{=} \eta.$$

Moreover, the same assumptions on the model have been imposed; see assumption A1. Hence, Schmisser also works with a stationary and ergodic process X as the unique solution of the considered stochastic differential equation. The aim of Schmisser is the construction of a nonparametric estimator of the drift function b, although her approach substantially differs from ours. We will now shortly introduce the ideas behind the alternative approach.

Schmisser (2014) is interested in constructing an estimator for b on a compact subset $A \subseteq \mathbb{R}$ and, for this purpose, introduces an increasing sequence of linear subspaces S_m of the vector space $L^2(A)$, whose finite dimension D_m is increasing in terms of m. As examples serve piecewise defined polynomials, compact supported wavelets, and splines. All of them are satisfying certain regularity assumptions; see Assumption A4, p.885 of Schmisser (2014).

Based on the high-frequency sample $X_\Delta, X_{2\Delta}, ..., X_{n\Delta}$, a contrast function $\gamma_n(t)$ is defined as

$$\gamma_n(t) := \frac{1}{n}\sum_{i=1}^{n-1}\left(\frac{X_{(i+1)\Delta}-X_{i\Delta}}{\Delta} - t(X_{i\Delta})\right)^2 := \frac{1}{n}\sum_{i=1}^{n-1}\left(Y_{i\Delta} - t(X_{i\Delta})\right)^2, \ t \in S_m.$$

$\gamma_n(t)$ has to be minimized on the subspace S_m in terms of t. Due to the fact that $\gamma_n(t)$ can always be minimized on S_m, but the minimizer may not be unique, Schmisser (2014) introduced the empirical risk

$$\mathcal{R}_n(t) := E\left[||t-b_A||_n^2\right], \text{ where } ||t||_n^2 := \frac{1}{n}\sum_{i=1}^{n-1} t^2(X_{i\Delta}) \text{ and } b_A(x) := b(x)1_A(x).$$

Now a least squares regression type estimator $\hat{b}_m$ is defined according to

$$\hat{b}_m := \operatorname{argmin}_{m\in\mathcal{M}_n} \gamma_n(t),$$

where $\mathcal{M}_n := \{m, \ D_m \leq \sqrt{n\Delta}/\log(n)\}$.

Under the double asymptotics scheme

$$\Delta \to 0 \text{ and } n\Delta \to \infty,$$

for fixed m, the risk of the estimator can be bounded as follows (see Schmisser (2014), Theorem 2, p.887):

$$\mathcal{R}_n(\hat{b}_m) \leq C_1||b_m - b_A||_{L^2}^2 + C_2\left(||\sigma^2||_\infty + ||\xi^2||_\infty Var(L(1))\right)\frac{D_m}{n\Delta} + C_3\Delta,$$

where $||\cdot||_{L^2}$ denotes the L^2-distance of the linear subspace S_m. Furthermore, b_m is the orthogonal L^2-projection of b_A over the vectorial subspace S_m and the constants C_i, $i = 1, 2, 3$, are independent of m, n, and Δ.

The last term is, under the assumption that $n\Delta^2 = o(1)$ as $n \to \infty$, negligible. In our case, by assuming $T^{8/5}\Delta = o(1)$, this assumption holds true.

Now consider this risk bound in more detail. The first term is a bias term, whereas the second term denotes the variance. Using the bias and variance decomposition of the AMSE, we are able to compare our rate and the above bound adequately.
The bias term decreases while the regularization parameter D_m increases. The variance behaves contrary and is, thus, proportional to D_m. By imposing certain smoothness assumptions on b, the rate of the bias can be quantified in terms of D_m. In particular, by assuming that b belongs to a ball of the Besov space $\mathcal{B}^2_{2,\infty}$, Schmisser (2014) concludes that

$$||b_m - b_A||^2_{L^2} \leq D_m^{-4}.$$

To balance the bias and the variance terms, the optimal choice of the dimension D_m is

$$D_{m,opt} \sim (n\Delta)^{1/5}$$

such that

$$\mathcal{R}_n(\hat{b}_{opt}) \lesssim (n\Delta)^{-4/5}.$$

In our case, under the assumption that π and b are twice continuously differentiable, the optimal pointwise AMSE is of order

$$AMSE\left(\hat{b}(x)\right) = O\left((n\Delta)^{-4/5}\right)$$

by choosing

$$h_{opt} \sim (n\Delta)^{-1/5}.$$

We see that the relation

$$D_{m,opt} \sim h_{opt}^{-1}$$

for the optimal regularization parameters holds true.
We will now compare the advantages and drawbacks of both methods. At first, Schmisser (2014) provides an adaptive selection method for choosing the dimension D_m by introducing a penalty function, too. This approach allows to choose m in an adaptive manner, whereas our approach relies on a rule-of-thumb or an oracle bandwidth, respectively.
But in contrast, we are able to derive the asymptotic distribution, which allows us to construct confidence intervals. From a practical point of view, this is an advantage compared to Schmisser´s approach.
Moreover, we are interested in pointwise estimators, whereas Schmisser provides some kind of uniform approximation on a compact interval of the drift function. Concerning this substantially different approach, practitioners should decide which approximation is needed for their purposes.

2.9 Estimation of the asymptotic variance

From a statistical point of view, it is desirable that the asymptotic variance of the asymptotic distribution of $\hat{b}(x)$ can be consistently estimated in order to construct feasible pointwise confidence intervals. In fact, the findings during the previous section lead us to the following approximation:

$$P\left(-\Phi^{-1}\left(1-\frac{\alpha}{2}\right) \leq \frac{\sqrt{n\Delta h}(\hat{b}(x)-b(x))}{\sqrt{\bar{\sigma}^2(x)}} \leq \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right) \approx 1-\alpha,$$

by choosing $n\Delta h^5 = o(1)$ as $n \to \infty$. Using this assumption on the bandwidth h, the bias term is negligible. In fact, we undersmooth $\hat{b}(x)$ to get rid of the bias.
In view of this approximation, an asymptotic confidence interval at point x is given by

$$I_x := \left[\hat{b}(x) - \frac{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\bar{\sigma}(x)}{\sqrt{n\Delta h}} \leq b(x) \leq \hat{b}(x) + \frac{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\bar{\sigma}(x)}{\sqrt{n\Delta h}}\right].$$

As we have seen, the asymptotic variance $\bar{\sigma}^2(x) = \frac{||K||_2^2\tilde{\sigma}^2(x)}{\pi(x)}$ is a quotient of unknown functions. Our aim is now to provide a consistent estimator for $\bar{\sigma}^2(x)$ based on consistent estimators for numerator and denominator. As we have seen, the denominator $\pi(x)$ can be consistently estimated by

$$\hat{\pi}(x) = \frac{\Delta}{Th}\sum_{i=0}^{n-1} K\left(\frac{x - X_{i\Delta}}{h}\right)$$

using Theorem 2.6. Thus, we focus on the estimation of the numerator of $\bar{\sigma}^2(x)$, in particular on $\tilde{\sigma}^2(x)$.
In view of approximation (2.1), we propose the estimator of $\tilde{\sigma}^2(x)$ as follows:

$$\hat{\tilde{\sigma}}^2(x) := \frac{\frac{1}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)(X_{(i+1)\Delta} - X_{i\Delta})^2}{\frac{\Delta}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)}.$$

We remark that $\hat{\tilde{\sigma}}^2(x)$ acts as an estimator for the second infinitesimal conditional moment $\tilde{\sigma}^2(x) = \sigma^2(x) + Var(L(1))\xi^2(x)$. In contrast to the threshold approach by Mancini and Renò (2011), where only the state-dependent volatility $\sigma^2(x)$ is of interest. Moreover, due this different problem, a faster rate of convergence, $\sqrt{nh}$, is attained in this case.
For the derivation of the asymptotic properties of this estimator, we have to strengthen the assumptions of Theorem 2.6 a little bit.

Theorem 2.13. *Under Assumptions A1, A2, and additionally under the assumption that the Lévy measure ν fulfills*

$$\int_{\mathbb{R}} y^4\nu(dy) < \infty,$$

$\hat{\tilde{\sigma}}^2(x)$ is a (weak) consistent estimator for $\tilde{\sigma}^2(x)$, provided that $\pi(x) > 0$. In particular, we find out that

$$\hat{\tilde{\sigma}}^2(x) \xrightarrow{P} \tilde{\sigma}^2(x), \ \text{as } n \to \infty.$$

Proof of Theorem 2.13. To establish the consistency, we start with the determination of an explicit form of a squared increment of the process X. Such increments appear in the denominator of $\hat{\tilde{\sigma}}^2(x)$. We have already presented the Itô-formula in Theorem 2.10. This will act as the first crucial tool for the derivation of the consistency.
Now we will decompose the squared increments in the following way:

$$(X_{(i+1)\Delta} - X_{i\Delta})^2 = X^2_{(i+1)\Delta} - X^2_{i\Delta} - 2X_{i\Delta}(X_{(i+1)\Delta} - X_{i\Delta}) \tag{2.8}$$

The first two terms can now be represented by the use of Itô´s formula as follows:

$$\begin{aligned} dX_t^2 &= 2X_t b(X_t)dt + \sigma^2(X_t)dt + 2X_t\sigma(X_t)dW_t \\ &+ \xi^2(X_t)\int_{\mathbb{R}} y^2\nu(dy)dt + \int_{\mathbb{R}}((X_{t_-} + \xi(X_{t_-})y)^2 - X^2_{t_-})\bar{\mu}(dy, dt). \end{aligned}$$

Now we are ready to find an explicit form of the considered squared increments according to (2.8):

$$\begin{aligned} &(X_{(i+1)\Delta} - X_{i\Delta})^2 = X^2_{(i+1)\Delta} - X^2_{i\Delta} - 2X_{i\Delta}(X_{(i+1)\Delta} - X_{i\Delta}) \\ &= 2\int_{i\Delta}^{(i+1)\Delta} X_s b(X_s)ds + 2\int_{i\Delta}^{(i+1)\Delta} X_s\sigma(X_s)dW_s + \int_{i\Delta}^{(i+1)\Delta}\sigma^2(X_s)ds \\ &\quad + \int_{i\Delta}^{(i+1)\Delta}\xi^2(X_s)ds\int_{\mathbb{R}} y^2\nu(dy) + \int_{i\Delta}^{(i+1)\Delta}\int_{\mathbb{R}}((X_{s_-} + \xi(X_{s_-})y)^2 - X^2_{s_-})\bar{\mu}(dy, ds) \\ &\quad - 2X_{i\Delta}\left(\int_{i\Delta}^{(i+1)\Delta} b(X_s)ds + \int_{i\Delta}^{(i+1)\Delta}\sigma^2(X_s)dW_s + \int_{i\Delta}^{(i+1)\Delta}\xi(X_{s_-})dL_s\right) \\ &= 2\int_{i\Delta}^{(i+1)\Delta}(X_s - X_{i\Delta})b(X_s)ds + 2\int_{i\Delta}^{(i+1)\Delta}(X_s - X_{i\Delta})\sigma(X_s)dW_s \\ &\quad + \int_{i\Delta}^{(i+1)\Delta}(\sigma^2(X_s) + \xi^2(X_s)Var(L(1)))ds \\ &\quad + \int_{i\Delta}^{(i+1)\Delta}\int_{\mathbb{R}}((X_{s_-} + \xi(X_{s_-})y)^2 - X^2_{s_-})\bar{\mu}(dy, ds) - 2\int_{i\Delta}^{(i+1)\Delta} X_{i\Delta}\int_{\mathbb{R}}\xi(X_{s_-})y\bar{\mu}(dy, ds). \end{aligned}$$

The last two terms can be summed up according to

$$\begin{aligned} &\int_{i\Delta}^{(i+1)\Delta}\int_{\mathbb{R}}((X_{s_-} + \xi(X_{s_-})y)^2 - X^2_{s_-})\bar{\mu}(dy, ds) - 2\int_{i\Delta}^{(i+1)\Delta} X_{i\Delta}\int_{\mathbb{R}}\xi(X_{s_-})y\bar{\mu}(dy, ds) \\ &= \int_{i\Delta}^{(i+1)\Delta}\int_{\mathbb{R}}(2X_{s_-}\xi(X_{s_-})y + \xi^2(X_{s_-})y^2)\bar{\mu}(dy, ds) - 2\int_{i\Delta}^{(i+1)\Delta} X_{i\Delta}\int_{\mathbb{R}}\xi(X_{s_-})y\bar{\mu}(dy, ds) \\ &= \int_{i\Delta}^{(i+1)\Delta}\xi^2(X_{s_-})\int_{\mathbb{R}} y^2\bar{\mu}(dy, ds) + 2\int_{i\Delta}^{(i+1)\Delta}\xi(X_{s_-})(X_{s_-} - X_{i\Delta})\int_{\mathbb{R}} y\bar{\mu}(dy, ds) \\ &= \int_{i\Delta}^{(i+1)\Delta}\xi^2(X_{s_-})\int_{\mathbb{R}} y^2\bar{\mu}(dy, ds) + 2\int_{i\Delta}^{(i+1)\Delta}\xi(X_{s_-})(X_{s_-} - X_{i\Delta})dL_s. \end{aligned}$$

Now we summarize the derived decomposition:

$$\begin{aligned}
(X_{(i+1)\Delta} - X_{i\Delta})^2 &= 2\int_{i\Delta}^{(i+1)\Delta} (X_{s_-} - X_{i\Delta}) dX_s \\
&\quad + \int_{i\Delta}^{(i+1)\Delta} (\sigma^2(X_s) + \xi^2(X_s) Var(L(1)))ds + \int_{i\Delta}^{(i+1)\Delta} \xi^2(X_{s_-}) \int_{\mathbb{R}} y^2 \bar{\mu}(dy, ds) \\
&= 2\int_{i\Delta}^{(i+1)\Delta} (X_s - X_{i\Delta}) b(X_s) ds + 2\int_{i\Delta}^{(i+1)\Delta} (X_s - X_{i\Delta}) \sigma(X_s) dW_s \\
&\quad + 2\int_{i\Delta}^{(i+1)\Delta} (X_{s_-} - X_{i\Delta}) \xi(X_{s_-}) dL_s + \int_{i\Delta}^{(i+1)\Delta} (\sigma^2(X_s) + \xi^2(X_s) Var(L(1)))ds \\
&\quad + \int_{i\Delta}^{(i+1)\Delta} \xi^2(X_{s_-}) \int_{\mathbb{R}} y^2 \bar{\mu}(dy, ds).
\end{aligned}$$

Due to this representation, we are able to decompose the estimator $\hat{\tilde{\sigma}}^2(x)$ into the following five parts

$$\begin{aligned}
\hat{\tilde{\sigma}}^2(x) &= \frac{\frac{1}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)(X_{(i+1)\Delta} - X_{i\Delta})^2}{\frac{\Delta}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)} \\
&= \frac{\frac{2}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta}(X_s - X_{i\Delta}) b(X_s) ds}{\frac{\Delta}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)} \\
&\quad + \frac{\frac{2}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta}(X_s - X_{i\Delta}) \sigma(X_s) dW_s}{\frac{\Delta}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)} \\
&\quad + \frac{\frac{2}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta}(X_{s_-} - X_{i\Delta}) \xi(X_{s_-}) dL_s}{\frac{\Delta}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)} \\
&\quad + \frac{\frac{1}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta}(\sigma^2(X_s) + \xi^2(X_s) Var(L(1)))ds}{\frac{\Delta}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)} \\
&\quad + \frac{\frac{1}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta} \xi^2(X_{s_-}) \int_{\mathbb{R}} y^2 \bar{\mu}(dy, ds)}{\frac{\Delta}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)} \\
&:= \frac{I' + II' + III' + IV' + V'}{VI'}.
\end{aligned}$$

The meaning of the six terms seems intuitive: the first three summands will converge to zero in probability according to the fact that they are additionally dependent on "small" increments of X. The rate of convergence of these increments has already been studied. The fourth term will tend to $\tilde{\sigma}^2(x)$ and the last term will also tend to zero according to the fact that this is an integral with respect to a martingale. Last but not least, the

denominator is a consistent estimator of $\pi(x) > 0$, as we have already seen.
These intuitions will now be proved and we will start with the derivation of the first three terms. In particular we have that

$$\begin{aligned}
E[|I'|] &= E\left[\left|\frac{2}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta}(X_s - X_{i\Delta})b(X_s)ds\right|\right] \\
&\leq \frac{2}{Th}\sum_{i=0}^{n-1}\int_{i\Delta}^{(i+1)\Delta} E\left[\left|K\left(\frac{X_{i\Delta}-x}{h}\right)\right|\cdot\left|(X_s - X_{i\Delta})b(X_s)\right|\right]ds \\
&\leq \frac{2||K||_\infty}{Th}\sum_{i=0}^{n-1}\int_{i\Delta}^{(i+1)\Delta} E\left[|X_s - X_{i\Delta}|\cdot|b(X_s)|\right]ds \\
&\leq \frac{2||K||_\infty}{Th}\sum_{i=0}^{n-1}\int_{i\Delta}^{(i+1)\Delta} \left(E\left[(X_s - X_{i\Delta})^2\right]\right)^{1/2}\left(E[b^2(X_s)]\right)^{1/2}ds \\
&\lesssim \frac{\Delta^{1/2}}{Th}\sum_{i=0}^{n-1}\int_{i\Delta}^{(i+1)\Delta}\left(E[b^2(X_s)]\right)^{1/2}ds = \frac{\Delta^{1/2}n\Delta}{Th}\left(E[b^2(X_0)]\right)^{1/2} \\
&= O\left(\frac{n\Delta^{3/2}}{Th}\right) = O\left(\frac{\Delta^{1/2}}{h}\right) = o(1), \text{ as } n\to\infty.
\end{aligned}$$

The Brownian term II' is handled in the same way as before. Particularly, we derive the order of its L^2-distance by using conditional expectations. Recall that we chose $\mathcal{F}_{i\Delta} = \sigma(X_0, (W_s, L_s); s \leq i\Delta)$ such that only the squared terms remain:

$$\begin{aligned}
E[(II')^2] &= E\left[\left(\frac{2}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta}(X_s - X_{i\Delta})\sigma(X_s)dW_s\right)^2\right] \\
&= \frac{4}{{}^2Th^2}\sum_{i=0}^{n-1} E\left[K^2\left(\frac{X_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta} E\left[(X_s - X_{i\Delta})^2\sigma^2(X_s)ds\middle|\mathcal{F}_{i\Delta}\right]\right] \\
&\leq \frac{4||\sigma^2||_\infty}{{}^2Th^2}\sum_{i=0}^{n-1}\int_{i\Delta}^{(i+1)\Delta} E\left[K^2\left(\frac{X_{i\Delta}-x}{h}\right)(X_s - X_{i\Delta})^2\right]ds \\
&\leq \frac{4||\sigma^2||_\infty||K^2||_\infty}{{}^2Th^2}\sum_{i=0}^{n-1}\int_{i\Delta}^{(i+1)\Delta} E\left[(X_s - X_{i\Delta})^2\right]ds \\
&\lesssim \frac{n\Delta^2}{T^2h^2} = O\left(\frac{\Delta}{Th^2}\right) = O\left(\frac{1}{nh^2}\right) = o(1), \text{ as } n\to\infty.
\end{aligned}$$

The third term can be examined in an analogous manner. We only make use of the fact that L is a martingale:

$$\begin{aligned}
E[(III')^2] &= E\left[\left(\frac{2}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta}(X_{s_-}-X_{i\Delta})\xi(X_{s_-})dL_s\right)^2\right] \\
&= \frac{4Var(L(1))}{T^2h^2}\sum_{i=0}^{n-1} E\left[K^2\left(\frac{X_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta} E\left[(X_{s_-}-X_{i\Delta})^2\xi^2(X_{s_-})ds\middle|\mathcal{F}_{i\Delta}\right]\right] \\
&\leq \frac{4Var(L(1))||\xi^2||_\infty}{T^2h^2}\sum_{i=0}^{n-1}\int_{i\Delta}^{(i+1)\Delta} E\left[K^2\left(\frac{X_{i\Delta}-x}{h}\right)(X_s-X_{i\Delta})^2\right]ds \\
&\leq \frac{4Var(L(1))||\xi^2||_\infty||K^2||_\infty}{T^2h^2}\sum_{i=0}^{n-1}\int_{i\Delta}^{(i+1)\Delta} E\left[(X_s-X_{i\Delta})^2\right]ds \\
&\lesssim \frac{n\Delta^2}{T^2h^2} = O\left(\frac{\Delta}{Th^2}\right) = O\left(\frac{1}{nh^2}\right) = o(1), \text{ as } n\to\infty.
\end{aligned}$$

Using the Markov inequality, we are now ready to conclude that

$$I' + II' + III' = O_P\left(\frac{\Delta^{1/2}}{h}\right) + O_P\left(\sqrt{\frac{1}{nh^2}}\right) = o_P(1), \text{ as } n\to\infty.$$

Now we deal with the fourth term. This term is responsible for the consistent estimation of $\tilde{\sigma}^2(x)$. At first, we will approximate the discrete observation $X_{i\Delta}$ by its continuous counterpart X_s, where s lies in the vicinity of $i\Delta$:

$$\begin{aligned}
IV' &= \frac{1}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta}(\sigma^2(X_s)+\xi^2(X_s)Var(L(1)))ds \\
&= \frac{1}{T}\int_0^T \frac{1}{h}K\left(\frac{X_s-x}{h}\right)(\sigma^2(X_s)+\xi^2(X_s)Var(L(1)))ds \\
&\quad + \frac{1}{Th}\sum_{i=0}^{n-1}\int_{i\Delta}^{(i+1)\Delta}\left(K\left(\frac{X_{i\Delta}-x}{h}\right)-K\left(\frac{X_s-x}{h}\right)\right)(\sigma^2(X_s)+\xi^2(X_s)Var(L(1)))ds \\
&:= \frac{1}{T}\int_0^T \frac{1}{h}K\left(\frac{X_s-x}{h}\right)\tilde{\sigma}^2(X_s)ds \\
&\quad + \frac{1}{Th}\sum_{i=0}^{n-1}\int_{i\Delta}^{(i+1)\Delta}\left(K\left(\frac{X_{i\Delta}-x}{h}\right)-K\left(\frac{X_s-x}{h}\right)\right)\tilde{\sigma}^2(X_s)ds \\
&:= \frac{1}{T}\int_0^T \frac{1}{h}K\left(\frac{X_s-x}{h}\right)\tilde{\sigma}^2(X_s)ds + F^n_{\tilde{\sigma}^2}.
\end{aligned}$$

The first term converges according to the ergodicity of X to

$$\begin{aligned}
&\frac{1}{T}\int_0^T \frac{1}{h}K\left(\frac{X_s - x}{h}\right)\tilde{\sigma}^2(X_s)ds \\
&\quad \longrightarrow \tilde{\sigma}^2(x)\pi(x) = \pi(x)(\sigma^2(x) + \xi^2(x)Var(L(1))), \text{ a.s. as } n, T \to \infty.
\end{aligned}$$

Recall that $\tilde{\sigma}^2$ is also bounded and continuous due to the assumed properties of σ and ξ. The approximation error is negligible in probability:

$$\begin{aligned}
&E[|F^n_{\tilde{\sigma}^2}|] \\
&= E\left[\left|\frac{1}{Th}\sum_{i=0}^{n-1}\int_{i\Delta}^{(i+1)\Delta}\left(K\left(\frac{X_{i\Delta} - x}{h}\right) - K\left(\frac{X_s - x}{h}\right)\right)\tilde{\sigma}^2(X_s)ds\right|\right] \\
&\leq \frac{1}{Th}\sum_{i=0}^{n-1}\int_{i\Delta}^{(i+1)\Delta} E\left[\left|K\left(\frac{X_{i\Delta} - x}{h}\right) - K\left(\frac{X_s - x}{h}\right)\right| \cdot \tilde{\sigma}^2(X_s)\right] ds \\
&\leq \frac{||K'||_\infty ||\tilde{\sigma}^2||_\infty}{Th^2}\sum_{i=0}^{n-1}\int_{i\Delta}^{(i+1)\Delta} E\left[|X_{i\Delta} - X_s| \cdot 1_{[i\Delta,(i+1)\Delta]}(s)\right] ds \\
&\leq \frac{||K'||_\infty ||\tilde{\sigma}^2||_\infty}{Th^2}\sum_{i=0}^{n-1}\int_{i\Delta}^{(i+1)\Delta} \left(E\left[|X_{i\Delta} - X_s|\right]\right)^{1/2} ds\Delta^{1/2} \\
&\lesssim \frac{n\Delta^2}{Th^2} = O\left(\frac{\Delta}{h^2}\right) = o(1), \text{ as } n \to \infty.
\end{aligned}$$

Let us summarize that

$$IV' = \tilde{\sigma}^2(x) + O_P\left(\sqrt{\frac{\Delta}{h^2}}\right) = \tilde{\sigma}^2(x) + o_P(1), \text{ as } n \to \infty.$$

Now focus on the fifth term which is a martingale with respect to the augmentation of the filtration $\mathcal{F}_t = \sigma((W_s, L_s), X_0; s \leq t)$ and, moreover, let

$$d\tilde{L}_s := \int_{\mathbb{R}} y^2\bar{\mu}(dy, ds) = \int_{\mathbb{R}} y^2(\mu(dy, ds) - \nu(dy)ds)$$

denote the squared compensated jumps of L. Using this abbreviation, we can conclude that

$$\begin{aligned}
V' &:= \frac{1}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta} - x}{h}\right)\int_{i\Delta}^{(i+1)\Delta} \xi^2(X_{s_-})\int_{\mathbb{R}} y^2\bar{\mu}(dy, ds) \\
&= \frac{1}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta} - x}{h}\right)\int_{i\Delta}^{(i+1)\Delta} \xi^2(X_{s_-})d\tilde{L}_s.
\end{aligned}$$

We derive again the L^2-distance of this term:

$$
\begin{aligned}
E\left[(V')^2\right] &= E\left[\left(\frac{1}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta} \xi^2(X_{s_-})d\tilde{L}_s\right)^2\right] \\
&= \frac{1}{T^2h^2}\sum_{i=0}^{n-1} E\left[K^2\left(\frac{X_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta} E[\xi^4(X_s)|\mathcal{F}_{i\Delta}]ds\right]\int_{\mathbb{R}} y^4\nu(dy) \\
&\leq \frac{||\xi^4||_\infty ||K^2||_\infty n\Delta}{T^2h^2}\int_{\mathbb{R}} y^4\nu(dy) = O\left(\frac{1}{Th^2}\right) = o(1), \text{ as } n\to\infty.
\end{aligned}
$$

We summarize our findings as follows:

$$
\begin{aligned}
\hat{\tilde{\sigma}}^2(x) &= \frac{\tilde{\sigma}^2(x)\pi(x) + O_P\left(\frac{\Delta^{1/2}}{h}\right) + O_P\left(\sqrt{\frac{1}{nh^2}}\right) + O_P\left(\frac{\Delta}{h^2}\right) + O_P\left(\sqrt{\frac{1}{Th^2}}\right)}{\pi(x) + O_P\left(\frac{\Delta}{h^2}\right)} \\
&= \tilde{\sigma}^2(x) + o_P(1), \text{ as } n\to\infty.
\end{aligned}
$$

□

2.10 Asymptotic distribution of the variance estimator

We are finally able to derive the asymptotic distribution of $\hat{\sigma}^2(x)$ by making use of the same technique as for the drift estimator $\hat{b}(x)$. To this end, we have to impose additional smoothness assumptions on σ and ξ as well as on the speed of convergence of Δ and h.

Assumption A4

i) Let the functions σ, ξ, and π be twice continuously differentiable.

ii) Let Δ and h satisfy

$$
n\Delta h^5 \to 0, \quad n\Delta^2 h^{-1} \to 0, \quad n\Delta^{3/2} \to 0, \quad n\Delta^3 h^{-3} \to 0.
$$

iii) Let the Lévy-measure ν fulfill

$$
\int_{\mathbb{R}} y^8\nu(dy) < \infty.
$$

Now we state the following Theorem.

Theorem 2.14. *Under Assumptions A1,A2, and A4, provided that $\pi(x) > 0$, we have that*

$$\sqrt{Th}\left(\tilde{\sigma}^2(x) - \tilde{\sigma}^2(x)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{||K||_2^2\xi^4(x)\int_{\mathbb{R}} y^4\nu(dy)}{\pi(x)}\right), \text{ as } n \to \infty.$$

Proof of Theorem 2.14. Using the smoothness assumptions on σ, ξ, and π as well using A4, ii), we find out that only the last part is responsible for the asymptotic distribution. In particular, we can derive that:

$$\begin{aligned}
&\sqrt{Th}\left(\hat{\tilde{\sigma}}^2(x) - \tilde{\sigma}^2(x)\right) \\
&= \sqrt{Th}\left(\frac{\frac{1}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)(X_{(i+1)\Delta} - X_{i\Delta})^2}{\frac{1}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)}\right) \\
&= \sqrt{Th}\left(\frac{\tilde{\sigma}^2(x)\pi(x) + O(h^2) + O_P\left(\frac{\Delta^{1/2}}{h}\right) + O_P\left(\sqrt{\frac{1}{nh^2}}\right) + O_P\left(\frac{\Delta}{h^2}\right)}{\pi(x) + O_P\left(\frac{\Delta}{h^2}\right)} - \tilde{\sigma}^2(x)\right) \\
&+ \sqrt{Th}\left(\frac{\frac{1}{Th}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta}\xi^2(X_{s_-})\int_{\mathbb{R}} y^2\bar{\mu}(dy, ds)}{\pi(x) + O_P\left(\frac{\Delta}{h^2}\right)}\right) \\
&= \frac{1}{\sqrt{Th}\pi(x)}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta}\xi^2(X_{s_-})\int_{\mathbb{R}} y^2\bar{\mu}(dy, ds) + \text{s.o.}.
\end{aligned}$$

The abbreviation "s.o." denotes the remaining terms, which are all negligible in probability compared to the first one. This follows directly by Assumption A4, ii) and we, therefore, omit this derivation.

For the examination of the term, which is responsible for the distribution, we will again make use of the central limit theorem 2.8 for martingale difference sequences. At first, define

$$\begin{aligned}
\sum_{i=0}^{n-1}\zeta'_{i+1,n} &= \frac{1}{\sqrt{Th}\pi(x)}\sum_{i=0}^{n-1} K\left(\frac{X_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta}\xi^2(X_{s_-})\int_{\mathbb{R}} y^2\bar{\mu}(dy, ds) \\
&:= \sum_{i=0}^{n-1}\tilde{\zeta}'_{i+1,n} + F_n^{\tilde{L}},
\end{aligned}$$

where $F_n^{\tilde{L}}$ again denotes the approximation error, which is negligible in probability by the same arguments as before. We always keep in mind that we assumed here that ν has moments up to order 8. We start with the derivation of the Lindeberg condition for the martingale difference sequence $(\tilde{\zeta}'_{i+1,n}, \tilde{\mathcal{F}}_{i+1})$ with respect to the filtration $\tilde{\mathcal{F}}_{i+1} := \sigma(X_0, (W_s, L_s); s \leq i+1)$. Now the Lindeberg condition can be derived by using Lemma

2.4:

$$\begin{aligned}
&\sum_{i=0}^{n-1} E\left[(\tilde{\zeta}'_{i+1,n})^4 | \mathcal{F}_{i\Delta}\right] \\
&= \sum_{i=0}^{n-1} E\left[\left(\frac{1}{\sqrt{Th}\pi(x)} \sum_{i=0}^{n-1} \int_{i\Delta}^{(i+1)\Delta} K\left(\frac{X_{s_-}-x}{h}\right) \xi^2(X_{s_-}) \int_{\mathbb{R}} y^2 \bar{\mu}(dy,ds) \tilde{\zeta}'_{i+1,n}\right)^4 \middle| \mathcal{F}_{i\Delta}\right] \\
&= \frac{1}{T^2h^2} \sum_{i=0}^{n-1} \left(3\Delta^2 ||K\xi^2||_\infty^6 \left(\int_{\mathbb{R}} y^4 \nu(dy)\right)^2 + \Delta ||K\xi^2||_\infty^8 \int_{\mathbb{R}} y^8 \nu(dy)\right) \\
&\lesssim \frac{n\Delta^2}{T^2h^2} + \frac{n\Delta}{T^2h^2} = \frac{1}{nh^2} + \frac{1}{n\Delta h^2} = o(1), \text{ as } n \to \infty.
\end{aligned}$$

The asymptotic variance can analogously be derived as in the drift case. Using the ergodicity of X, we are able to deduce that

$$\sum_{i=0}^{n-1} E\left[(\tilde{\zeta}'_{i+1,n})^2 | \mathcal{F}_{i\Delta}\right] \longrightarrow \frac{||K||_2^2 \xi^4(x) \int_{\mathbb{R}} y^4 \nu(dx)}{\pi(x)} \text{ a.s. as } n \to \infty.$$

Now we can apply the central limit theorem 2.8 and are ready to deduce that

$$\sqrt{Th}(\tilde{\sigma}^2(x) - \tilde{\sigma}^2(x)) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{||K||_2^2 \xi^4(x) \int_{\mathbb{R}} y^4 \nu(dy)}{\pi(x)}\right) \text{ as } n \to \infty.$$

□

Remark 2.15. *Wee see that the second conditional moment, which is the sum of the volatility function σ and a jump part, can only be consistently estimated when the double asymptotics scheme holds true. This is contrary to the ordinary diffusion case ($\xi \equiv 0$), where only $\Delta \to 0$ is required; see for example Florens-Zmirou (1993). The result that the double asymptotics scheme is needed is not surprising, due to the fact that even in the finite activity case, which is considered in Bandi and Nguyen (2003), it is necessary for the consistent estimation of the first two conditional moments.*

Remark 2.16. *It is often quite satisfactory when simpler models can be recovered in more difficult ones by leaving out occurring parameters. By setting $\xi \equiv 0$ our results are consistent with those in Bandi and Phillips (2003). By assuming that ν is a probability measure, our results are also consistent with those in Bandi and Nguyen (2003).*

Remark 2.17. *After discussing the feasible estimation of the first two conditional moments, one is able to hypothesize how higher moments can be estimated. One crucial point will be that for the estimation of the k-th conditional moment, one has to require that*

$$\int_{\mathbb{R}} y^{2k} \nu(dy) < \infty$$

for the Lévy measure ν of the driving Lévy process L. Moreover, the speed of convergence of the occurring regularization parameters has to be adjusted, too.

Example 2.18. *Again, it would be interesting to visualize the appearing constraints on α and β when writing*

$$\Delta \sim n^{-\alpha} \text{ and } h \sim n^{-\beta}.$$

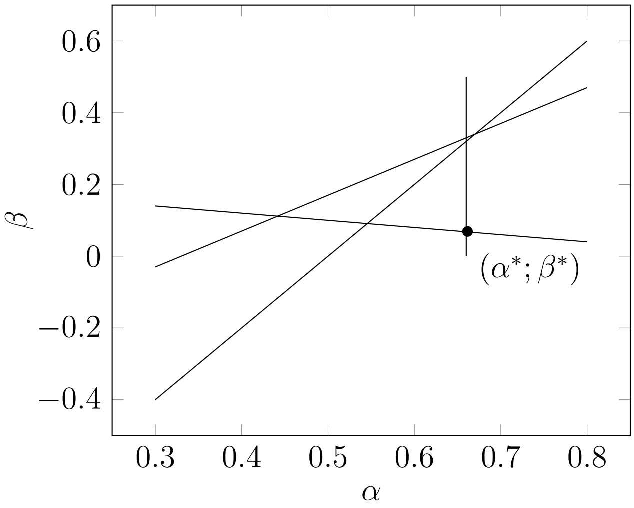

Figure 2: Plot of the constraints in assumption A4, ii) the corresponding optimal coordinates $(\alpha^*, \beta^*) \approx (0.67, 0.07)$.

2.11 The case of noisy data

In this section, we will extend our derived results to the case where we only observe noisy data. We will see that, by a slight modification of our proposed drift estimator, our estimation procedure is robust under measurement errors.

In particular, it is widely known that in the case of high-frequency observation schemes, measurement errors as well as the so-called microstructure noise play significant roles and can be found in several financial data sets; see Zhang et al. (2005), Jacod et al. (2009) or Jones (2003). In contrast to our considered high-frequency setting, this effect is not significant in low frequency models in such a way. There are quite a lot of articles dealing with noisy data in parametric as well as in nonparametric models. We are only interested in nonparametric models and shortly want to summarize three different approaches for the nonparametric estimation of integrated volatility ("IV") as well as integrated quarticity ("IQ") for Itô-diffusions of the form

$$X_t = X_0 + \int_0^t a_s ds + \int_0^t \sigma_s dW_s,$$

where $0 \leq t \leq 1$, a is a predictable drift function and σ is a càdlàg volatility process. Consider a process Y, defined on the same filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F})_{t\in[0,1]}, P)$, which is observed at time points $i/n,\ i = 0, ..., n$, and can be decomposed into

$$Y_{\frac{i}{n}} = X_{\frac{i}{n}} + U_{i,n}.$$

The additional white noise process $U = (U_{i,n})_{0\leq i\leq n}$ is centered and exhibits a finite variance and is, in addition, independent of X. In this setting, Podolskij and Vetter (2006) introduced the so-called pre-averaging approach for the nonparametric estimation of IV and IQ based on realized bipower variation in the setting of Itô-diffusions with as well as without an additional jump process. This widely used approach has for instance been studied in Jacod et al. (2009) in a more general setting. We will focus on their approach later on for the estimation of $b(x)$ in our considered Lévy-driven diffusion model.
Two additional approaches for the handling of noisy data have been proposed by Zhang et al. (2005) and Zhang (2006) where a subsampling based method has been considered. Another possibility to estimate the values of interest in noisy models was suggested by Barndorff-Nielsen et al. (2006). They derived an asymptotic theory for multipower variation based estimators for IV in the simultaneous presence of jumps and measurement errors by linear combinations of autocovariances.

2.12 Formulation of the pre-averaged drift estimator

Recall that our model is given by

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t + \xi(X_{t_-})dL_t,\ X_0 \overset{\mathcal{D}}{=} \eta$$

and that we observe a high-frequency sample at time points $0, \Delta, 2\Delta, ..., n\Delta := T$ such that $\Delta \to 0$ as well as $n\Delta \to \infty$. Now assume that the process X is contaminated by an additional noise process $\varepsilon = (\varepsilon_t)_{t\geq 0}$ such that we observe a process $Y = (Y_t)_{t\geq 0}$ instead of X, which can be decomposed as

$$Y_{i\Delta} = X_{i\Delta} + \varepsilon_{i\Delta},\ i = 1, ..., n,$$

where we assume that $\{\varepsilon_{i\Delta}\}_{i=0,...,n}$ are independent and identically distributed random variables such that

$$E[\varepsilon_{i\Delta}] = 0, \quad E[\varepsilon_{i\Delta}^2] := \sigma_\varepsilon^2$$

for all i and Δ. Moreover, the process ε is independent of X.
We will now focus in detail on the idea of the pre-averaging approach by Podolskij and Vetter (2006). Let us at first decompose the available sample $Y_{i\Delta},\ i = 1, ..., n$ into $m_n := m$ subgroups of length $r_n := r$, such that $mr = n$. An example would be that we observe a sample of a certain asset price in 36 months ($\widehat{=}\ m$) and 60 times ($\widehat{=}\ r$) in each month.

Without loss of generality, we assume that n can be decomposed into the product of m and r. Otherwise, one would introduce a first and a last block whose lengths are smaller than r. Moreover, the block length as well as the number of blocks fulfill

$$r \to \infty, \quad m = \left\lfloor \frac{n}{r} \right\rfloor \to \infty, \quad \Delta r \to 0, \text{ as } n \to \infty.$$

Instead of working with the contaminated rare data set $\{Y_{i\Delta}\}$, we build averages inside every block j, where $j = 1, ..., r$. Thus, we define

$$\bar{Y}_j := \bar{Y}_{j,\Delta} := \frac{1}{r} \sum_{i=1}^{r} Y_{((j-1)r+i)\Delta}$$

and analogously

$$\bar{X}_j := \bar{X}_{j,\Delta} := \frac{1}{r} \sum_{i=1}^{r} X_{((j-1)r+i)\Delta}, \quad \bar{\varepsilon}_j := \bar{\varepsilon}_{j,\Delta} := \frac{1}{r} \sum_{i=1}^{r} \varepsilon_{((j-1)r+i)\Delta}.$$

Using these notations, we obtain

$$\bar{Y}_j = \bar{X}_j + \bar{\varepsilon}_j.$$

The motivation for this approach is rather simple. Due to the fact that the noise process $\{\varepsilon_t\}_{t\geq 0}$ is a centered i.i.d. process with finite variance, the averages $\bar{\varepsilon}_j$ tend to zero in probability (we only need this kind of convergence) as $r \to \infty$ for every $j = 1, .., m$. Because of that, the averages $\bar{Y}_j$ act as approximations of the averages $\bar{X}_j$, which are, in turn, approximations of the original sample $\{X_{i\Delta}\}$. Recall that $\{\bar{X}_{j,\Delta}, j = 1, ..., m\}$ can be seen as a discrete sample of the original time continuous process (X_t) with sampling frequency $r\Delta \to 0$ on the time interval $[0, mr\Delta] = [0, n\Delta]$ with $mr\Delta \to \infty$.
Let us now define a new drift estimator $\hat{b}_Y(x)$ based on the sample $\bar{Y}_j$, $j = 1, ..., m$, as

$$\hat{b}_Y(x) := \frac{\frac{1}{mh} \sum_{j=1}^{m-1} K\left(\frac{\bar{Y}_j - x}{h}\right) \left(\bar{Y}_{j+1} - \bar{Y}_j\right)}{\frac{r\Delta}{mh} \sum_{j=1}^{m-1} K\left(\frac{\bar{Y}_j - x}{h}\right)}.$$

In the ordinary diffusion setting, kernel estimators occur in the approach by Barndorff-Nielsen et al. (2006) as covariances estimators. Moreover, Greenwood et al. (2015) and Lee (2014) focused on kernel estimators in a noisy diffusion setting without jumps.
In order to derive the asymptotic properties of this estimator, we will make use of the results in our previous section. In particular, our strategy will be to prove that

$$\hat{b}_Y(x) - \hat{b}(x) = o_P(1), \text{ as } n \to \infty$$

by a proper rate such that the results concerning the asymptotic normality of $\hat{b}(x)$ can be transferred.

To derive these results, we have to impose the following assumptions, mainly concerning the rates of convergence of the included sequences.

Assumption A4

i) The noise process $\varepsilon = (\varepsilon_t)_{t\geq 0}$ is an i.i.d. process, in particular the random variables $\varepsilon_{i\Delta}, i = 1, ..., n$, are i.i.d. with

$$E[\varepsilon_{i\Delta}] = 0, \quad E[\varepsilon_{i\Delta}^2] = \sigma_\varepsilon^2 < \infty$$

for all $0 \leq i\Delta \leq T$.

ii) The processes $X = (X_t)_{t\geq 0}$ and $\varepsilon = (\varepsilon_t)_{t\geq 0}$ are independent.

iii) The block length r fulfill

$$r \to \infty \text{ and } \Delta r \to 0 \text{ as } n \to \infty.$$

Moreover, the number of blocks m behaves like $m = \lfloor n/r \rfloor \to \infty$ as $n \to \infty$.

iv) The appearing parameters r, h, and Δ fulfill

$$n\Delta r h^5 \to 0,\ n(\Delta r)^2 h^{-3} \to 0,\ n\Delta r h \to \infty,$$
$$(\Delta r)^{1/2} h^{-2} \text{ and } n\Delta r^{-2} h^{-3} \to 0.$$

Now we are ready to state our main theorem concerning the asymptotic distribution of $\hat{b}_Y(x)$.

Theorem 2.19. *Under Assumptions A1-A4, provided that $\pi(x) > 0$, it holds that*

$$\sqrt{n\Delta h}\left(\hat{b}_Y(x) - b(x)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{||K||_2^2(\sigma^2(x) + \xi^2(x)Var(L(1)))}{\pi(x)}\right), \quad \text{as } n \to \infty.$$

It turns out that the key point for the derivation of this result is an analogous statement to Proposition 2.3. Under suitable assumptions, we have to bound the squared L^2-distance of small increments of X, but now in terms of $r\Delta$ instead of the original sampling frequency Δ.

Proposition 2.20. *Under Assumptions A1,i)-vi) and $\Delta \leq 1$, the following statements hold true*

$$1) \max_{1\leq j\leq m} E\left[(X_{jr\Delta} - X_{(j-1)r\Delta})^2\right] \lesssim r\Delta,$$

$$2) \max_{1\leq j\leq m} E\left[(\bar{X}_j - X_{(j-1)r\Delta})^2\right] \lesssim r\Delta,$$

$$3) \max_{1\leq j\leq m-1} E\left[(\bar{X}_{j+1} - \bar{X}_j)^2\right] \lesssim r\Delta.$$

Proof of Proposition 2.20. The first statement can be directly deduced by Proposition 2.3. For the second statement, observe that

$$
\begin{aligned}
&E[|\bar{X}_j - X_{(j-1)r\Delta}|] \leq \frac{1}{r}\sum_{i=1}^{r} E[|X_{((j-1)r+i)\Delta} - X_{(j-1)r\Delta}|] \\
&\leq \frac{1}{r}\sum_{i=1}^{r} \left(E[(X_{((j-1)r+i)\Delta} - X_{(j-1)r\Delta})^2]\right)^{1/2} \\
&\lesssim \frac{1}{r}\sum_{i=1}^{r} \Bigg(\Delta \int_{(j-1)r\Delta}^{((j-1)r+i)\Delta} E[b^2(X_s)]ds + \int_{(j-1)r\Delta}^{((j-1)r+i)\Delta} E[\sigma^2(X_s)]ds \\
&\qquad + Var(L(1)) \int_{(j-1)r\Delta}^{((j-1)r+i)\Delta} E[\xi^2(X_s)]ds\Bigg)^{1/2} \\
&\leq \frac{1}{r}\sum_{i=1}^{r} \left(\Delta^2 i E[b^2(X_0)] + i\Delta(||\sigma^2||_\infty + ||\xi^2||_\infty Var(L(1)))\right)^{1/2} \\
&\lesssim \frac{1}{r}\sum_{i=1}^{r} (i\Delta)^{1/2} \leq \frac{1}{r}\Delta^{1/2}\left(\sum_{i=1}^{r} i\right)^{1/2} \sqrt{r} \lesssim \frac{r^{3/2}\Delta^{1/2}}{r} = (r\Delta)^{1/2}.
\end{aligned}
$$

Now use Jensen´s inequality and the monotonicity of the function $t \to \sqrt{t}$ to finish the proof of statement 2).
Finally for 3), observe that

$$
\begin{aligned}
E[|\bar{X}_{j+1} - \bar{X}_j|] &\leq E[|\bar{X}_{j+1} - X_{jr\Delta}|] + E[|X_{jr\Delta} - X_{(j-1)r\Delta}|] \\
&\quad + E[|\bar{X}_j - X_{(j-1)r\Delta}|] \lesssim 3(r\Delta)^{1/2}
\end{aligned}
$$

due to 1) and 2). □

We are now able to proof our main statement, namely the asymptotic normality of the pre-averaged drift estimator.

Proof of Theorem 2.19. Recall that we observe a high-frequency sample

$$Y_{i\Delta} = X_{i\Delta} + \varepsilon_{i\Delta},\ i = 0, ..., n$$

consisting of the original diffusion process and contaminated by additional noise. Now define

$$\hat{b}_X(x) := \frac{\frac{1}{(m-1)h}\sum_{j=1}^{m-1}\left(X_{jr\Delta} - X_{(j-1)r\Delta}\right) K\left(\frac{X_{(j-1)r\Delta}-x}{h}\right)}{\frac{r\Delta}{(m-1)h}\sum_{j=1}^{m-1} K\left(\frac{X_{(j-1)r\Delta}-x}{h}\right)} := \frac{\hat{N}_X(x)}{\hat{D}_X(x)},$$

which is the drift estimator of $b(x)$ based on a high-frequency sample $\{X_{jr\Delta} := X_{j\delta}, j = 0, ..., m\}$. The new sample is a "thinned out version" with sampling frequency $\delta = \Delta r \to 0$ of the original sample. From an asymptotic point of view, this has no impact on the derivation of the asymptotic distribution since we are working in a high-frequency setting. Therefore, and under Assumptions A1-A4, this estimator is consistent and asymptotically normally distributed due to Theorems 2.6 and 2.7:

$$\sqrt{m\delta h}\left(\hat{b}_X(x) - b(x)\right) = \sqrt{n\Delta h}\left(\hat{b}_X(x) - b(x)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{||K||_2^2\tilde{\sigma}^2(x)}{\pi(x)}\right), \text{ as } n \to \infty.$$

Our aim is now to prove that

$$\hat{N}_X(x) - \hat{N}_Y(x) = o_P((n\Delta h)^{-1/2}) = o_P((m\delta h)^{-1/2}) \tag{2.9}$$

as well as

$$\hat{D}_X(x) - \hat{D}_Y(x) = o_P((n\Delta h)^{-1/2}) = o_P((m\delta h)^{-1/2}), \tag{2.10}$$

where analogously

$$\hat{b}_Y(x) = \frac{\frac{1}{(m-1)h}\sum_{j=1}^{m-1} K\left(\frac{\bar{Y}_j - x}{h}\right)\left(\bar{Y}_{j+1} - \bar{Y}_j\right)}{\frac{r\Delta}{(m-1)h}\sum_{j=1}^{m-1} K\left(\frac{\bar{Y}_j - x}{h}\right)} := \frac{\hat{N}_Y(x)}{\hat{D}_Y(x)}.$$

The proof of equations (2.9) and (2.10), together with the fact that all appearing estimators are bounded in probability, and are according to this of order $O_P(1)$, yields the desired result.

We will start with the derivation of (2.10) and evaluate the L^1-distance between $\hat{D}_X(x)$ and $\hat{D}_Y(x)$ to deduce the negligibility in probability by the Markov inequality. The key point will be Proposition 2.20 and the Lipschitz-continuity of K:

$$\begin{aligned} E[|\hat{D}_X(x) - \hat{D}_Y(x)|] &\leq \frac{r\Delta}{(m-1)h}\sum_{j=1}^{m-1} E\left[\left|K\left(\frac{\bar{Y}_j - x}{h}\right) - K\left(\frac{X_{(j-1)r\Delta} - x}{h}\right)\right|\right] \\ &\leq \frac{r\Delta||K'||_\infty}{(m-1)h^2}\sum_{j=1}^{m-1} E\left[|\bar{Y}_j - X_{(j-1)r\Delta}|\right] \\ &= \frac{r\Delta||K'||_\infty}{(m-1)h^2}\sum_{j=1}^{m-1} E\left[|\bar{X}_j - X_{(j-1)r\Delta} + \bar{\varepsilon}_j|\right]. \end{aligned}$$

Now we will focus on the above summands, which can be evaluated by the use of Proposition 2.20

$$
\begin{aligned}
&E\left[|\bar{X}_j - X_{(j-1)r\Delta} + \bar{\varepsilon}_j|\right] \leq E\left[|\bar{X}_j - X_{(j-1)r\Delta}|\right] + E\left[|\bar{\varepsilon}_j|\right] \\
&\lesssim (r\Delta)^{1/2} + \left(E\left[(\bar{\varepsilon}_j)^2\right]\right)^{1/2} = (r\Delta)^{1/2} + \left(E\left[\left(\frac{1}{r}\sum_{i=1}^{r} \varepsilon_{((j-1)r+i)\Delta}\right)^2\right]\right)^{1/2} \\
&= (r\Delta)^{1/2} + \frac{\sigma_\varepsilon}{\sqrt{r}}.
\end{aligned}
$$

Recall for the last equation that ε_j are i.i.d.
Now we can conclude that

$$
\begin{aligned}
E[|\hat{D}_X(x) - \hat{D}_Y(x)|] &\leq \frac{r\Delta ||K'||_\infty}{(m-1)h^2} \sum_{j=1}^{m-1} \left(E\left[|\bar{X}_j - X_{(j-1)r\Delta}|\right] + E\left[|\bar{\varepsilon}_j|\right]\right) \\
&\lesssim \frac{r\Delta}{(m-1)h^2} \sum_{j=1}^{m-1} \left((r\Delta)^{1/2} + \frac{\sigma_\varepsilon}{\sqrt{r}}\right) \\
&= O\left(\frac{(\Delta r)^{3/2}}{h^2} + \frac{\Delta r^{1/2}}{h^2}\right).
\end{aligned}
$$

To prove (2.10), both following terms have to converge to zero in probability:

$$
\sqrt{n\Delta h}\left(\hat{D}_X(x) - \hat{D}_Y(x)\right) = O_P\left(\frac{(n\Delta h)^{1/2}(\Delta r)^{3/2}}{h^2} + \frac{(n\Delta h)^{1/2}\Delta r^{1/2}}{h^2}\right) \overset{!}{=} o_P(1). \quad (2.11)
$$

To derive (2.11), both connections of the appearing parameters can be found in assumption A4, iv). Due to the Markov inequality, the difference of the denominators fulfills

$$
\hat{D}_X(x) - \hat{D}_Y(x) = o_P((n\Delta h)^{-1/2}), \text{ as } n \to \infty.
$$

The more interesting step of the proof is the treatment of the difference

$$
\hat{N}_X(x) - \hat{N}_Y(x).
$$

Our aim is to prove that this difference is also of order $o_P((n\Delta h)^{-1/2})$ as $n \to \infty$. For this purpose we decompose the considered difference as follows:

$$
\begin{aligned}
&\hat{N}_X(x) - \hat{N}_Y(x) \\
&= \frac{1}{(m-1)h} \sum_{j=1}^{m-1} \left(\left(\bar{Y}_{j+1} - \bar{Y}_j\right) K\left(\frac{\bar{Y}_j - x}{h}\right) - \left(X_{jr\Delta} - X_{(j-1)r\Delta}\right) K\left(\frac{X_{(j-1)r\Delta} - x}{h}\right) \right) \\
&= \frac{1}{(m-1)h} \sum_{j=1}^{m-1} \left(\left(\bar{Y}_{j+1} - \bar{Y}_j\right) \left(K\left(\frac{\bar{Y}_j - x}{h}\right) - K\left(\frac{X_{(j-1)r\Delta} - x}{h}\right) \right) \right) \\
&\quad + \frac{1}{(m-1)h} \sum_{j=1}^{m-1} \left(\left(\left(\bar{Y}_{j+1} - \bar{Y}_j\right) - \left(X_{jr\Delta} - X_{(j-1)r\Delta}\right) \right) K\left(\frac{X_{(j-1)r\Delta} - x}{h}\right) \right) \\
&:= A_n(x) + B_n(x).
\end{aligned}
$$

We will again determine the order of the L^1-distance of both terms $A_n(x)$ and $B_n(x)$. Again, by the use of the Markov inequality, we will deduce that both terms will converge as fast as required to zero. We will start with the first term $A_n(x)$ and use the Lipschitz-continuity of K as well as Proposition 2.20:

$$
\begin{aligned}
&E[|A_n(x)|] \leq \frac{1}{(m-1)h} \sum_{j=1}^{m-1} E\left[|\bar{Y}_{j+1} - \bar{Y}_j| \cdot \left| K\left(\frac{\bar{Y}_j - x}{h}\right) - K\left(\frac{X_{(j-1)r\Delta} - x}{h}\right) \right| \right] \\
&\leq \frac{||K'||_\infty}{(m-1)h^2} \sum_{j=1}^{m-1} E\left[|\bar{Y}_{j+1} - \bar{Y}_j| \cdot |\bar{Y}_j - X_{(j-1)r\Delta}| \right] \\
&\leq \frac{||K'||_\infty}{(m-1)h^2} \sum_{j=1}^{m-1} \left(E\left[\left(\bar{Y}_{j+1} - \bar{Y}_j\right)^2 \right] \right)^{1/2} \cdot \left(E\left[\left(\bar{Y}_j - X_{(j-1)r\Delta}\right)^2 \right] \right)^{1/2} \\
&= \frac{||K'||_\infty}{(m-1)h^2} \sum_{j=1}^{m-1} \left(E\left[\left((\bar{X}_{j+1} - \bar{X}_j) + (\bar{\varepsilon}_{j+1} - \bar{\varepsilon}_j)\right)^2 \right] \right)^{1/2} \cdot \left(E\left[\left(\bar{X}_j - X_{(j-1)r\Delta} + \bar{\varepsilon}_j\right)^2 \right] \right)^{1/2} \\
&= \frac{||K'||_\infty}{(m-1)h^2} \sum_{j=1}^{m-1} \Bigg(\left(E\left[(\bar{X}_{j+1} - \bar{X}_j)^2 + 2(\bar{X}_{j+1} - \bar{X}_j)(\bar{\varepsilon}_{j+1} - \bar{\varepsilon}_j) + (\bar{\varepsilon}_{j+1} - \bar{\varepsilon}_j)^2 \right] \right)^{1/2} \\
&\qquad \cdot \left(E\left[(\bar{X}_j - X_{(j-1)r\Delta})^2 + 2(\bar{X}_j - X_{(j-1)r\Delta})\bar{\varepsilon}_j + \bar{\varepsilon}_j^2 \right] \right)^{1/2} \Bigg) \\
&= \frac{||K'||_\infty}{(m-1)h^2} \sum_{j=1}^{m-1} \left(E\left[(\bar{X}_{j+1} - \bar{X}_j)^2 \right] + E\left[(\bar{\varepsilon}_{j+1} - \bar{\varepsilon}_j)^2 \right] \right)^{1/2} \\
&\qquad \cdot \left(E\left[(\bar{X}_j - X_{(j-1)r\Delta})^2 \right] + E[\bar{\varepsilon}_j^2] \right)^{1/2} \\
&\lesssim \frac{||K'||_\infty}{(m-1)h^2} \sum_{j=1}^{m-1} \left(r\Delta + \frac{1}{r} \right)^{1/2} \cdot \left(r\Delta + \frac{1}{r} \right)^{1/2} = \frac{r\Delta}{h^2} + \frac{1}{rh^2},
\end{aligned}
$$

where we used the fact that

$$E\left[(\bar{\varepsilon}_{j+1}-\bar{\varepsilon}_j)^2\right] \leq 2\left(E[\bar{\varepsilon}_{j+1}^2]+E[\bar{\varepsilon}_j^2]\right) = \frac{4\sigma_\varepsilon^2}{r}.$$

Hence, we conclude that

$$E[|A_n(x)|] = O\left(\frac{r\Delta}{h^2}+\frac{1}{rh^2}\right), \text{ as } n \to \infty.$$

Furthermore, to prove that the term $A_n(x)$ possesses the proper rate $o_P((n\Delta h)^{-1/2})$, we have to ensure that

$$\sqrt{n\Delta h}A_n(x) = O_P\left(\frac{(n\Delta h)^{1/2}r\Delta}{h^2}+\frac{(n\Delta h)^{1/2}}{rh^2}\right) \overset{!}{=} o_P(1).$$

Due to Assumption A4, iv), both constraints are fulfilled and we can deduce that $A_n(x)$ exhibits the needed order. Now we will continue with the derivation of the second term $B_n(x)$. We will use a detailed decomposition to derive the order of this term. The derivation will be rather tedious, but in the end we will see that this term possesses the needed order, too. At first decompose $B_n(x)$ as follows:

$$\begin{aligned}
B_n(x) &= \frac{1}{(m-1)h}\sum_{j=1}^{m-1}\left(\left((\bar{Y}_{j+1}-\bar{Y}_j)-\left(X_{jr\Delta}-X_{(j-1)r\Delta}\right)\right)K\left(\frac{X_{(j-1)r\Delta}-x}{h}\right)\right)\\
&= \frac{1}{(m-1)h}\sum_{j=1}^{m-1}\left(\left((\bar{X}_{j+1}-\bar{X}_j)-\left(X_{(j+1)r\Delta}-X_{jr\Delta}\right)\right)K\left(\frac{X_{(j-1)r\Delta}-x}{h}\right)\right)\\
&\quad+\frac{1}{(m-1)h}\sum_{j=1}^{m-1}\left(\left(\left(X_{(j+1)r\Delta}-X_{jr\Delta}\right)-\left(X_{jr\Delta}-X_{(j-1)r\Delta}\right)\right)K\left(\frac{X_{(j-1)r\Delta}-x}{h}\right)\right)\\
&\quad+\frac{1}{(m-1)h}\sum_{j=1}^{m-1}\left((\bar{\varepsilon}_{j+1}-\bar{\varepsilon}_j)K\left(\frac{X_{(j-1)r\Delta}-x}{h}\right)\right) := \sum_{k=1}^{3}B_{n,k}(x).
\end{aligned}$$

Now we will focus on each term separately and start with a further decomposition of the latter:

$$\begin{aligned}
B_{n,3}(x) &= \frac{1}{(m-1)h}\sum_{j=1}^{m-1}\left((\bar{\varepsilon}_{j+1}-\bar{\varepsilon}_j)K\left(\frac{X_{(j-1)r\Delta}-x}{h}\right)\right)\\
&= \frac{\bar{\varepsilon}_m}{(m-1)h}K\left(\frac{X_{(m-2)r\Delta}-x}{h}\right)-\frac{\bar{\varepsilon}_1}{(m-1)h}K\left(\frac{X_0-x}{h}\right)\\
&\quad+\frac{1}{(m-1)h}\sum_{j=2}^{m-1}\bar{\varepsilon}_j\left(K\left(\frac{X_{(j-2)r\Delta}-x}{h}\right)-K\left(\frac{X_{(j-1)r\Delta}-x}{h}\right)\right).
\end{aligned}$$

The corresponding L^1-distance can now easily be derived by the use of the boundedness of K as well as Proposition 2.20:

$$
\begin{aligned}
E[|B_{n,3}(x)|] \leq & \frac{1}{(m-1)h} E\left[\left|\bar{\varepsilon}_m \cdot K\left(\frac{X_{(m-2)r\Delta} - x}{h}\right)\right|\right] \\
& + \frac{1}{(m-1)h} E\left[\left|\bar{\varepsilon}_1 \cdot K\left(\frac{X_0 - x}{h}\right)\right|\right] \\
& + \frac{1}{(m-1)h} \sum_{j=2}^{m-1} E\left[|\bar{\varepsilon}_j| \cdot \left|K\left(\frac{X_{(j-2)r\Delta} - x}{h}\right) - K\left(\frac{X_{(j-1)r\Delta} - x}{h}\right)\right|\right] \\
\leq & \frac{2||K||_\infty}{(m-1)h} E[|\bar{\varepsilon}_m|] + \frac{||K'||_\infty}{(m-1)h^2} \sum_{j=2}^{m-1} E\left[|\bar{\varepsilon}_j| \cdot |X_{(j-2)r\Delta} - X_{(j-1)r\Delta}|\right] \\
\leq & \frac{2||K||_\infty}{(m-1)h} \left(E[\bar{\varepsilon}_m^2]\right)^{1/2} \\
& + \frac{||K'||_\infty}{(m-1)h^2} \sum_{j=2}^{m-1} \left(E[\bar{\varepsilon}_j^2]\right)^{1/2} \left(E[(X_{(j-2)r\Delta} - X_{(j-1)r\Delta})^2]\right)^{1/2} \\
\lesssim & \frac{1}{(m-1)hr^{1/2}} + \frac{1}{(m-1)h^2} \sum_{j=1}^{m-2} \frac{1}{r^{1/2}} (r\Delta)^{1/2} = O\left(\frac{1}{(m-1)hr^{1/2}} + \frac{\Delta^{1/2}}{h^2}\right).
\end{aligned}
$$

Again, we found two rates, which have to fulfill certain requirements. In particular, we have to assure that

$$
\sqrt{n\Delta h} B_{n,3}(x) = O_P\left(\frac{(n\Delta h)^{1/2}}{(m-1)hr^{1/2}} + \frac{(n\Delta h)^{1/2}\Delta^{1/2}}{h^2}\right) \overset{!}{=} o_P(1),
$$

which is guaranteed by Assumption A4, iv). Now we will focus on $B_{n,2}(x)$. We will perform a comparable decomposition as before and get three different terms as follows:

$$
\begin{aligned}
B_{n,2}(x) = & \frac{1}{(m-1)h} \sum_{j=1}^{m-1} \left((X_{(j+1)r\Delta} - X_{jr\Delta}) - (X_{jr\Delta} - X_{(j-1)r\Delta})\right) K\left(\frac{X_{(j-1)r\Delta} - x}{h}\right) \\
= & \frac{(X_{mr\Delta} - X_{(m-1)r\Delta})}{(m-1)h} K\left(\frac{X_{(m-2)r\Delta} - x}{h}\right) + \frac{(X_0 - X_{r\Delta})}{(m-1)h} K\left(\frac{X_0 - x}{h}\right) \\
& + \frac{1}{(m-1)h} \sum_{j=1}^{m-2} (X_{(j+1)r\Delta} - X_{jr\Delta}) \cdot \left(K\left(\frac{X_{(j-1)r\Delta} - x}{h}\right) - K\left(\frac{X_{jr\Delta} - x}{h}\right)\right).
\end{aligned}
$$

Now the L^1-distance can be bounded by

$$\begin{aligned} E[|B_{n,2}(x)|] &\leq \frac{||K||_\infty}{(m-1)h}\left(E[|X_{mr\Delta} - X_{(m-1)r\Delta}|] + E[|X_0 - X_{r\Delta}|]\right) \\ &\quad + \frac{||K'||_\infty}{(m-1)h^2}\sum_{j=1}^{m-2} E[|X_{(j+1)r\Delta} - X_{jr\Delta}| \cdot |X_{(j-1)r\Delta} - X_{jr\Delta}|] \\ &\lesssim \frac{(r\Delta)^{1/2}}{(m-1)h} + \frac{(r\Delta)}{h^2}. \end{aligned}$$

Finally, to guarantee that $B_{n,2}(x)$ converges fast enough to zero in probability, we have to ensure that

$$\sqrt{n\Delta h} B_{n,2}(x) = O_P\left(\frac{(n\Delta h)^{1/2}(r\Delta)^{1/2}}{(m-1)h} + \frac{(n\Delta h)^{1/2} r\Delta}{h^2}\right) \overset{!}{=} o_P(1).$$

Basic reformulations of Assumption A4, iv) ensure that the appropriate terms converge to zero as fast as it is needed.
Finally, we will focus on $B_{n,1}(x)$. As a first step, we will decompose it as follows:

$$\begin{aligned} B_{n,1}(x) &= \frac{1}{(m-1)h}\sum_{j=1}^{m-1}\left(\left((\bar{X}_{j+1} - \bar{X}_j) - (X_{(j+1)r\Delta} - X_{jr\Delta})\right) K\left(\frac{X_{(j-1)r\Delta} - x}{h}\right)\right) \\ &= \frac{(X_{r\Delta} - \bar{X}_1)}{(m-1)h} K\left(\frac{X_0 - x}{h}\right) + \frac{(\bar{X}_m - X_{(m-1)r\Delta})}{(m-1)h} K\left(\frac{X_{(m-1)r\Delta} - x}{h}\right) \\ &\quad + \frac{1}{(m-1)h}\sum_{j=1}^{m-2}\left(\bar{X}_{j+1} - X_{(j+1)r\Delta}\right)\left(K\left(\frac{X_{(j-1)r\Delta} - x}{h}\right) - K\left(\frac{X_{jr\Delta} - x}{h}\right)\right). \end{aligned}$$

Due to Proposition 2.20, we are allowed to proceed in the same way as for the treatment of $B_{n,2}(x)$. For the sake of brevity, we will restrict ourselves to the final result, namely

$$E[|B_{n,1}(x)|] \lesssim \frac{(r\Delta)^{1/2}}{(m-1)h} + \frac{(r\Delta)}{h^2}.$$

Hence,

$$B_{n,1}(x) = o_P((n\Delta h)^{-1/2}), \text{ as } n \to \infty,$$

which finishes the proof. □

Example 2.21. *In the noisy setting it would be interesting to derive an optimal rate of convergence under the assumptions on r, Δ and h, too.*
By letting

$$\Delta \sim n^{-\alpha},\ h \sim n^{-\beta},\ \textit{and } r \sim n^{\rho},$$

a variety of connections between α, β and ρ exists. It would be appropriate to make use of a simplex algorithm conducted in the previous section, too. Hence, we solve the following linear optimization problem. Let

$$G(\alpha, \beta) := \frac{1-\alpha-\beta}{2}$$

and maximize the function G with respect to

$$\begin{aligned}\alpha-\rho+5\beta>1,\ 2(\alpha-\rho)-3\beta>1,\ \alpha-\rho+\beta<1,\\ \alpha-\rho-4\beta>0,\ 4\alpha-3\rho-3\beta>1,\ 3\alpha-\rho-3\beta>1,\\ 3\alpha-3\beta-2\rho>1,\ \alpha+2\rho-3\beta>1,\ -\alpha+\beta+\rho<1,\\ 2\alpha-3\beta>1,\ -2\alpha+\beta+3\rho<1,\ \text{and } 3\alpha-2\rho-3\beta>1.\end{aligned}$$

The simplex algorithm yields the following result:

$$\alpha \approx 0.821, \quad \beta \approx 0.077,$$

which leads to an optimal rate under our assumptions of $n^{0.051}$. This rate is slower than in the case of non-noisy data, which is not surprising. We already know that in our setting $r\Delta$ denotes the sampling frequency of the new random sample $\{\bar{X}_j, j = 1, ..., m\}$ and should behave like Δ in the previous section without measurement errors. Therefore, we choose $r \sim n^{0.2}$ such that $r\Delta \sim n^{-0.62}$, which is approximately the optimal result for α in the previous section.

2.13 Drift estimation of an integrated jump diffusion process

In this section, we will extend our results from nonparametric drift estimation for Lévy driven jump diffusion models to the case of integrated processes. We will at first concretize what is meant by an integrated process. Consider a two-dimensional process $(X_t, V_t)_{t\geq 0}$ such that

$$\begin{aligned} dX_t &= V_t dt, \quad X_0 = 0 \\ dV_t &= b(V_t)dt + \sigma(V_t)dW_t + \xi(V_{t\_})dL_t, \quad V_0 \overset{\mathcal{D}}{=} \eta, \end{aligned} \tag{2.12}$$

where again $W = (W_t)_{t\geq 0}$ is a standard Brownian Motion and $L = (L_t)_{t\geq 0}$ is a centered Lévy process with finite variance $E[L^2(1)] = \int_{\mathbb{R}} y^2 \nu(dy) < \infty$ such that

$$dL_t = \int_{\mathbb{R}} z(\mu(dt, dz) - \nu(dz)dt).$$

W and L are independent and η is independent of both W and L. Hence, V is actually the same process as in our first section.

Usual estimation schemes for diffusion processes, as for example in our first considered model, are based on a sample of the original process. In contrast to the non-noisy setting, we are now assuming that we cannot observe the process V itself but rather a running integral of the process V. In particular, we only observe the first coordinate $X_t = \int_0^t V_s ds$ of our original bidimensional process at equidistant time points $k\Delta, k = 1, ..., n+2$, such that

$$T := (n+2)\Delta \to \infty \text{ and } \Delta \to 0.$$

We changed the original definition of T due to some technical reasons. From an asymptotic point of view, this renaming has obviously no impact on the rate of divergence of T. Such integrated processes appear quite often in engineering science as well as in physics. For example, Comte et al. (2009) refer to a model where V denotes the velocity of a particle and X represents its coordinate. Further models and application of such processes can be found in Ditlevsen and Sørensen (2004) as well as in Lefebvre (1997).

For such models, parametric inference has been conducted in some works; see for example Ditlevsen and Sørensen (2004) or Glotter (2000, 2006) as well as Glotter and Gobet (2005). But in general, this topic has not arisen much attention, although it is quite interesting and important for real data applications.

In the nonparametric framework, we are only aware of two works, where the coefficients of such models have been consistently estimated. Comte et al. (2009) use a model selection approach to construct adaptive nonparametric estimators of b and σ on a fixed compact interval. This work extends their approach for estimating ordinary univariate diffusions and was also pursued by Schmisser (2014) in the case of univariate jump diffusions; see Section 2.8. Nicolau (2007) uses kernel estimators for pointwise consistent estimation of $b(x)$ and $\sigma^2(x)$. Nicolau (2007) and Comte et al. (2009) are both concerned with continuous integrated diffusions, in particular the case where $\xi \equiv 0$. To the best of our knowledge, nonparametric pointwise inference for the drift function $b(x)$ in an integrated jump diffusion model has not been done in the literature before.

We will now concretize our estimation approach. Hence, consider the available data set $\{X_{k\Delta},\ k = 1, ..., n+2\}$ of the process X given by (2.12). As mentioned, the process V is not observable and has to be approximated. The idea behind our estimation approach relies on the following transformation. We set

$$\overline{V}_k := \frac{1}{\Delta}\left(X_{(k+1)\Delta} - X_{k\Delta}\right) = \frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta} V_s ds,\ 1 \leq k \leq n+1.$$

Based on the sample $\{\overline{V}_k, k = 1, ..., n+1\}$, we will now propose the drift estimator for the considered model as follows:

$$\hat{b}_V(x) := \frac{\hat{b}_{V,D}(x)}{\hat{b}_{V,N}(x)} := \frac{\frac{1}{n\Delta h}\sum_{k=0}^{n-1} K\left(\frac{\overline{V}_{k+1}-x}{h}\right)\left(\overline{V}_{k+2} - \overline{V}_{k+1}\right)}{\frac{1}{nh}\sum_{k=0}^{n-1} K\left(\frac{\overline{V}_{k+1}-x}{h}\right)},$$

where we shortly recall that K is a bounded, symmetric, and Lipschitz-continuous probability density function possessing a finite second moment.
The equivalent data set $\{\overline{V}_k, k = 1, ..., n+1\}$ should now act as an approximation of the unobservable jump diffusion V:

$$\overline{V}_k \approx V_{k\Delta}, \; k = 1, ..., n+1.$$

After evaluating the goodness of the above approximation, we will be able to deduce the desired asymptotic properties of our proposed drift estimator by the use of the results in the ordinary Lévy driven diffusion case.
At first, we will reconsider the context of the first section, where we worked with a stationary and exponentially β-mixing jump diffusion process fulfilling the ergodicity property. Hence, we will impose assumptions A1 and A2 and start with a very useful proposition acting as a key point for our proofs. The following proposition generalizes Lemma 7.1-7.3 in Comte et al. (2010) to the case of integrated jump diffusions.

Proposition 2.22. *Under assumptions A1 and A2, the following observations hold true*

a) $\overline{V}_k + \frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta}(u - k\Delta)dV_u = V_{(k+1)\Delta}$.

b) $Y_{k+1} := \frac{\overline{V}_{k+2} - \overline{V}_{k+1}}{\Delta} = \frac{1}{\Delta^2}\int_{(k+1)\Delta}^{(k+3)\Delta}\psi_{(k+1)\Delta}(u)dV_u$, *where*

$$\psi_{k\Delta}(u) := (u - k\Delta)1_{[k\Delta,(k+1)\Delta)}(u) + ((k+2)\Delta - u)1_{[(k+1)\Delta,(k+2)\Delta)}(u).$$

c) For $\Delta \leq 1$, *we can bound second moments of "small" increments of the process* V *as follows:*

$$E[(V_{t+\Delta} - V_t)^2] \lesssim \Delta.$$

d) To value the goodness of our used approximation, we state that

$$E[(V_{(k+1)\Delta} - \overline{V}_k)^2] \lesssim \Delta.$$

Proof of Proposition 2.22. We start by proving a), which is more or less an interchange of integrals according to

$$\begin{aligned}\overline{V}_k &= \frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta} V_s ds = \frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta}(V_{k\Delta} + V_s - V_{k\Delta})ds \\ &= \frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta}\left(V_{k\Delta} + \int_{k\Delta}^{s} dVu\right)ds = V_{k\Delta} + \frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta}\left(\int_{u}^{(k+1)\Delta} ds\right)dV_u \\ &= V_{k\Delta} + \frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta}((k+1)\Delta - u)dV_u = V_{(k+1)\Delta} + \frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta}(k\Delta - u)dV_u.\end{aligned}$$

By the use of a), we are able to deduce statement b) as follows:

$$
\begin{aligned}
&= \frac{1}{\Delta}\left(V_{(k+3)\Delta} - \frac{1}{\Delta}\int_{(k+2)\Delta}^{(k+3)\Delta}(u-(k+2)\Delta)dV_u - V_{(k+2)\Delta} + \frac{1}{\Delta}\int_{(k+1)\Delta}^{(k+2)\Delta}(u-(k+1)\Delta)dV_u\right)\\
&= \frac{1}{\Delta^2}\int_{(k+1)\Delta}^{(k+3)\Delta}\big((u-(k+1)\Delta)1_{[(k+1)\Delta,(k+2)\Delta)}(u) + ((k+3)\Delta - u)1_{[(k+2)\Delta,(k+3)\Delta)}(u)\big)\,dV_u\\
&= \frac{1}{\Delta^2}\int_{(k+1)\Delta}^{(k+3)\Delta}\psi_{(k+1)\Delta}(u)dV_u.
\end{aligned}
$$

Statement c) is a reformulation of Proposition 2.3 acting as a reminder and again being crucial for our purposes.
The proof of d) is based on c) as well as the Cauchy-Schwarz inequality and is derived as follows:

$$
\begin{aligned}
E[(V_{(k+1)\Delta} - \overline{V}_k)^2] &= \frac{1}{\Delta^2}E\left[\left(\int_{k\Delta}^{(k+1)\Delta}(V_{(k+1)\Delta} - V_s)ds\right)^2\right]\\
&\leq \frac{1}{\Delta^2}\int_{k\Delta}^{(k+1)\Delta}\Delta E[(V_{(k+1)\Delta} - V_s)^2]ds \lesssim \Delta.
\end{aligned}
$$

□

To derive the consistency of $\hat{b}_V(x)$, we will decompose the denominator $\hat{b}_{V,D}(x)$ and the numerator $\hat{b}_{V,N}(x)$ separately. We start with a useful regression type decomposition of Y_{k+1}:

$$
\begin{aligned}
\hat{b}_{V,D}(x) &= \frac{1}{nh}\sum_{k=0}^{n-1}K\left(\frac{\overline{V}_{k+1}-x}{h}\right)Y_{k+1}\\
&= \frac{1}{nh}\sum_{k=0}^{n-1}K\left(\frac{\overline{V}_{k+1}-x}{h}\right)\left(\frac{\overline{V}_{k+2}-\overline{V}_{k+1}}{\Delta}\right)\\
&= \frac{1}{nh}\sum_{k=0}^{n-1}K\left(\frac{V_{(k+2)\Delta}-x}{h}\right)b(V_{(k+2)\Delta})\\
&\quad + \frac{1}{nh}\sum_{k=0}^{n-1}b(V_{(k+2)\Delta})\left(K\left(\frac{\overline{V}_{k+1}-x}{h}\right) - K\left(\frac{V_{(k+2)\Delta}-x}{h}\right)\right)\\
&\quad + \frac{1}{nh}\sum_{k=0}^{n-1}K\left(\frac{\overline{V}_{k+1}-x}{h}\right)\frac{1}{\Delta^2}\int_{(k+1)\Delta}^{(k+3)\Delta}\psi_{(k+1)\Delta}(u)\big(b(V_u) - b(V_{(k+2)\Delta})\big)\,du\\
&\quad + \frac{1}{nh}\sum_{k=0}^{n-1}K\left(\frac{\overline{V}_{k+1}-x}{h}\right)\frac{1}{\Delta^2}\int_{(k+1)\Delta}^{(k+3)\Delta}\psi_{(k+1)\Delta}(u)\sigma(V_u)dW_u
\end{aligned}
$$

$$+\frac{1}{nh}\sum_{k=0}^{n-1}K\left(\frac{\overline{V}_{k+1}-x}{h}\right)\frac{1}{\Delta^2}\int_{(k+1)\Delta}^{(k+3)\Delta}\psi_{(k+1)\Delta}(u)\xi(V_{u-})dL_u$$
$$:=\frac{1}{nh}\sum_{k=0}^{n-1}K\left(\frac{V_{(k+2)\Delta}-x}{h}\right)b(V_{(k+2)\Delta})+R_\Delta^{(1)}+R_\Delta^{(2)}+Z_\Delta^{(1)}+Z_\Delta^{(2)},$$

where we have used the fact that

$$\frac{1}{\Delta^2}\int_{(k+1)\Delta}^{(k+3)\Delta}\psi_{(k+1)\Delta}(u)du=1.$$

The first term is responsible for receiving a consistent estimator of $\pi_V(x)b(x)$, where $\pi_V(x)$ denotes the stationary density of the process V. The following two terms act as remainder terms, which are negligible, since Δ is small. The last two terms are noise terms and also negligible, but possess the slowest rate of convergence and are responsible for the asymptotic distribution. These facts will be seen in the proof of the following Theorem.
We are now ready to formulate our main theorem within this section, namely the consistency of $\hat{b}_V(x)$.

Theorem 2.23. *Under assumptions A1 and A2, provided that $\pi_V(x)>0$, we find that*

$$\hat{b}_V(x)\xrightarrow{P}b(x),\quad \textit{as } n\to\infty.$$

Proof of Theorem 2.23. Using the regression type decomposition, we are able to deduce that, under the Lipschitz-continuity of the kernel K, the first remainder term $R_\Delta^{(1)}$ fulfills

$$\begin{aligned}E[|R_\Delta^{(1)}|]&\le\frac{1}{nh}\sum_{k=0}^{n-1}E\left[|b(V_{(k+2)\Delta})|\cdot\left|K\left(\frac{\overline{V}_{k+1}-x}{h}\right)-K\left(\frac{V_{(k+2)\Delta}-x}{h}\right)\right|\right]\\&\le\frac{||K'||_\infty}{nh^2}\sum_{k=0}^{n-1}E\left[|b(V_{(k+2)\Delta})|\cdot|\overline{V}_{k+1}-V_{(k+2)\Delta}|\right]\\&\le\frac{||K'||_\infty}{nh^2}\sum_{k=0}^{n-1}\left(E[b^2(V_{(k+2)\Delta})|]\right)^{1/2}\cdot\left(E[(\overline{V}_{k+1}-V_{(k+2)\Delta})^2]\right)^{1/2}\\&\lesssim\frac{||K'||_\infty\left(E[b^2(V_0)|]\right)^{1/2}}{nh^2}\sum_{k=0}^{n-1}\Delta^{1/2}\lesssim\Delta^{1/2}h^{-2}\to0,\text{ as }n\to\infty.\end{aligned}$$

Now we will focus on the second remainder term $R_\Delta^{(2)}$, which can be handled by the use of the Lipschitz-continuity of b as well as Proposition 3.2:

$$\begin{aligned}
E[|R_\Delta^{(2)}|] &\leq \frac{1}{nh}\sum_{k=0}^{n-1}\frac{1}{\Delta^2}\int_{(k+1)\Delta}^{(k+3)\Delta}\psi_{(k+1)\Delta}(u)E\left[\left|K\left(\frac{\overline{V}_{k+1}-x}{h}\right)\right|\cdot|b(V_u)-b(V_{(k+2)\Delta})|\right]du\\
&\leq \frac{||K||_\infty||b'||_\infty}{nh}\sum_{k=0}^{n-1}\frac{1}{\Delta^2}\int_{(k+1)\Delta}^{(k+3)\Delta}\psi_{(k+1)\Delta}(u)E\left[|V_u-V_{(k+2)\Delta}|\right]du\\
&\leq \frac{||K||_\infty||b'||_\infty}{nh}\sum_{k=0}^{n-1}\frac{1}{\Delta^2}\int_{(k+1)\Delta}^{(k+3)\Delta}\psi_{(k+1)\Delta}(u)\left(E\left[(V_u-V_{(k+2)\Delta})^2\right]\right)^{1/2}du\\
&\lesssim \frac{||K||_\infty||b'||_\infty}{nh}\sum_{k=0}^{n-1}\frac{1}{\Delta^2}\int_{(k+1)\Delta}^{(k+3)\Delta}\psi_{(k+1)\Delta}(u)\Delta^{1/2}du \lesssim \Delta^{1/2}h^{-1}\to 0,\ \text{as } n\to\infty.
\end{aligned}$$

The Brownian noise term $Z_\Delta^{(1)}$ can be handled in analogous manner compared to the first section, where we used independent increments of W to find out that the L^2-distance is bounded by

$$\begin{aligned}
E[(Z_\Delta^{(1)})^2] &= E\left[\left(\frac{1}{nh}\sum_{k=0}^{n-1}K\left(\frac{\overline{V}_{k+1}-x}{h}\right)\frac{1}{\Delta^2}\int_{(k+1)\Delta}^{(k+3)\Delta}\psi_{(k+1)\Delta}(u)\sigma(V_u)dW_u\right)^2\right]\\
&= \frac{1}{n^2h^2}\sum_{k=0}^{n-1}\frac{1}{\Delta^4}\int_{(k+1)\Delta}^{(k+3)\Delta}\psi^2_{(k+1)\Delta}(u)E\left[K^2\left(\frac{\overline{V}_{k+1}-x}{h}\right)\sigma^2(V_u)\right]du\\
&\leq \frac{||K^2||_\infty E[\sigma^2(V_0)]}{n^2h^2}\sum_{k=0}^{n-1}\frac{1}{\Delta^4}\int_{(k+1)\Delta}^{(k+3)\Delta}\psi^2_{(k+1)\Delta}(u)du\\
&= \frac{||K^2||_\infty E[\sigma^2(V_0)]}{n^2h^2}\\
&\quad\sum_{k=0}^{n-1}\frac{1}{\Delta^4}\left(\int_{(k+1)\Delta}^{(k+2)\Delta}(u-(k+1)\Delta)^2du+\int_{(k+2)\Delta}^{(k+3)\Delta}(u-(k+3)\Delta)^2du\right)\\
&= \frac{||K^2||_\infty E[\sigma^2(V_0)]}{n^2h^2}\sum_{k=0}^{n-1}\frac{1}{\Delta^4}\frac{2\Delta^3}{3}\lesssim\frac{1}{n\Delta h^2}\to 0,\ \text{as } n\to\infty.
\end{aligned}$$

Finally, the Lévy term $Z_\Delta^{(1)}$ can be treated using the same steps:

$$\begin{aligned}
E[(Z_\Delta^{(1)})^2] &= E\left[\left(\frac{1}{nh}\sum_{k=0}^{n-1}K\left(\frac{\overline{V}_{k+1}-x}{h}\right)\frac{1}{\Delta^2}\int_{(k+1)\Delta}^{(k+3)\Delta}\psi_{(k+1)\Delta}(u)\xi(V_{u-})dL_u\right)^2\right]\\
&\leq \frac{||K^2||_\infty E[\xi^2(V_0)]E[L^2(1)]}{n^2h^2}\sum_{k=0}^{n-1}\frac{1}{\Delta^4}\int_{(k+1)\Delta}^{(k+3)\Delta}\psi^2_{(k+1)\Delta}(u)du
\end{aligned}$$

$$= \frac{||K^2||_\infty E[\xi^2(V_0)]E[L^2(1)]}{n^2h^2} \sum_{k=0}^{n-1} \frac{1}{\Delta^4} \frac{2\Delta^3}{3} \lesssim \frac{1}{n\Delta h^2} \to 0, \text{ as } n \to \infty.$$

Using the Markov inequality we are ready to conclude that

$$\begin{aligned}\hat{b}_{V,D}(x) &= \frac{1}{nh}\sum_{k=0}^{n-1} K\left(\frac{V_{(k+2)\Delta} - x}{h}\right) b(V_{(k+2)\Delta}) + O_P(\Delta^{1/2}h^{-2}) + O_P\left(\frac{1}{\sqrt{n\Delta h^2}}\right) \\ &= \pi_V(x)b(x) + o_P(1), \text{ as } n \to \infty.\end{aligned}$$

The derivation of the denominator $\hat{b}_{V,N}(x)$ is rather simple and is based on comparable steps. The advantage is that we do not have any noise terms. We omit the proof for the sake of brevity and finally conclude that

$$\hat{b}_V(x) \xrightarrow{P} \frac{\pi_V(x)b(x)}{\pi_V(x)} = b(x), \quad \text{as } n \to \infty.$$

□

3 Using varying bandwidths in nonparametric regression and density estimation

3.1 Sample smoothing estimators in nonparametric density estimation

After introducing the general idea of using nonparametric regression estimators for time-continuous stochastic processes, we now want to focus on some methods how to improve the performance of this class of estimators. In general, the class of local polynomial estimators provide promising results in many arising practical situations. Higher order polynomials of degree $d \geq 1$ were firstly introduced by Fan (1992) in order to estimate probability densities as well as regression functions. The class of Nadaraya-Watson estimators acts as the special case $d = 0$ and has been considered since the 1960s; see for example Nadaraya (1965), Watson (1964) or Wand and Jones (1994). To concretize the definition of local polynomial estimators, we consider the following weighted least squares problem. Let $\{X_{k\Delta},\ k = 1, ..., n\}$ be a high-frequency sample of the considered Levy driven jump process and define

$$\hat{\beta} := (\hat{\beta}_0, ..., \hat{\beta}_p) = \text{argmin}_{\beta_0,...,\beta_p} \sum_{i=0}^{n-1} \left(\frac{X_{(i+1)\Delta} - X_{i\Delta}}{\Delta} - \sum_{k=0}^{p} \beta_j (X_{i\Delta} - x)^j \right)^2 K_h(X_{i\Delta} - x).$$

In view of local polynomial estimators, the above-introduced Nadaraya-Watson like estimator for the drift b is given by the case $p = 0$ and $\hat{\beta}_0 = \hat{b}(x)$. The case $p = 1$ is called local linear estimator for b and arises via $\hat{\beta}_0 = \hat{b}(x)$ again. Moreover, $\hat{\beta}_1$ is an estimator of $b'(x)$, see Fan (1992). In the context of scalar valued continuous diffusion processes ($\xi \equiv 0$), Bandi and Phillips (2002) focused on the use of such estimators in Section 3.3. The reason why we will only focus on the Nadaraya-Watson case is due to the fact that proofs for higher order polynomial estimators are rather long and tedious and, that it is possible that they produce negative values in the context of density estimation. But it should be mentioned that the higher effort will yield to a bias reduction; see again Fan (1992) for a detailed analysis. Nevertheless, the possibility to invoke higher order polynomials for the estimation of b is only one possibility to improve the performance of the proposed kernel based estimators. There is a plethora of possible extensions and improvements of this estimation technique, until now especially developed for nonparametric regression and density estimation. Due to the fact that any list would be incomplete, we only mention the books by Härdle et al. (2004) as well as Härdle (1994) providing a summary of many existing approaches for nonparametric curve estimation.

In this chapter, we want to focus on a special technique to reduce the asymptotic order of the bias, namely the use of variable bandwidths. The intuition of using this type of unequal smoothing is rather simple: in many practical situations, the available data is

not equidistantly located across the corresponding support. There are regions exhibiting scattered sparse data points as well as data-rich regions. This seems very important, especially for practical issues. Although we are working in a high-frequency scheme, where the time lag between observations tends to zero, we think that it is still worth focusing on varying bandwidths due to the fact that real data sets often consist of unequally spaced data points. Approaches to handle the problem of unequally spaced data have been widely studied, especially within the context of nonparametric density estimation. Particularly, we want to mention two important works, which act as a foundation for this chapter:
Based on n independent and identically distributed random variables $X_i,\ i = 1, ..., n$, possessing a density f, Abramson (1982) was one of the first authors who proposed the idea of using varying bandwidths of the form $h_n(X_i)$. In particular, he considered random bandwidths dependent on the observations X_i. He provides a so-called "square root law" arguing for the use of random bandwidths proportional to $f(X_i)^{-1/2}$. This choice ensures that the bias term is of order $o(h^2)$, which is faster compared to the classical fixed bandwidth case. He did not give the exact asymptotic proof, i.e. how fast the new rate of the bias exactly is. A very interesting proof of the exact rate of convergence of the bias, which will also play a role in our subsequent analysis, was proposed in Terrell and Scott (1992). They derived the asymptotic rates for the variance and the bias of the so-called "sample smoothing" density estimator

$$\hat{f}_1(x) := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_n(X_i)} K\left(\frac{x - X_i}{h_n(X_i)}\right)$$

for independent and identically distributed observations $X_1, ..., X_n$, based on a bandwidth function $h_n(X_i)$ dependent on both the sample size and the observations. Moreover, they assumed that K is a symmetric density supported on $[-1, 1]$. Therefore, $\hat{f}_1$ is an average of kernels with individually equipped scaling. Due to the fact that K is a probability density, $\hat{f}_1$ will be a "bona fide" estimator, namely $\hat{f}_1$ integrates to 1 and is non-negative. For the sake of completeness, we want to mention that the original version of this estimator, proposed by Abramson (1982), is of the form

$$\hat{f}_A(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{c(X_i)}{h_n} K\left(\frac{c(X_i)(x - X_i)}{h_n}\right),$$

where c is a scalar valued function such that $c(x) = \bar{f}(0)^{-1/2} \bar{f}(x)$ and $\bar{f}(x) = f(x) \vee \frac{1}{10} f(0)$, where Abramson assumed that $f(0) > 0$. This version is called the "clipped version" of Abramson´s estimator (cf. Terrell and Scott (1992)). Most authors only focused on the unclipped version for the sake of simplicity. We fill proceed analogously.

Remark 3.1. *Certainly, there are other possibilities of varying the bandwidth, for example the k-th nearest neighbor method, which was originally proposed by Loftsgaarden and Quesenberry (1965) as well as the so-called "balloon estimator", where the bandwidth h is* *dependent on the point of estimation and not on the available sample, see again Terrell and Scott(1992):*

$$\hat{f}_2(x) := \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_n(x)} K\left(\frac{x - X_i}{h_n(x)}\right).$$

It turns out that the asymptotic analysis of this class of estimators is rather simple, but the resulting estimator is not a "bona fide" estimator, because $\hat{f}_2$ is in general not a density anymore; see Terrell and Scott (1992) as well as Tukey and Tukey (1981).

Abramson´s square root law requires a "pilot estimate" for the unknown density f evaluated at X_i. This can, for example, be a kernel based estimator with fixed bandwidth, see Terrell and Scott (1992). For the pilot estimator as well as the resulting density estimator, the bandwidth selection plays an important role. We will not address this problem in our subsequent analysis and will rely on the existing methods like least squares cross-validation, which works very well in most situations, see for example Härdle (1994).
For the sake of completeness, we will state the pointwise asymptotic properties of the sample smoothing estimator $\hat{f}_2(x)$ below, which will play a central role in our following analysis. This proposition can be found in Terrell and Scott (1992) denoted as Theorem 2.

Proposition 3.2. *Let K be a symmetric probability density function being square integrable and possessing a finite second moment. Moreover, let $h_n(x) \to 0$ as well as $nh_n(x) \to \infty$ as $n \to \infty$ and let*

$$t_y(x) := \frac{x - y}{h_n(y)}$$

be strictly monotone in x. Then, for h and f twice continuously differentiable, we have

$$AVar(\hat{f}_2(y)) = \frac{f(y)||K||_2^2}{nh_n(y)}$$

as well as

$$ABias(\hat{f}_2(y))^2 = \frac{1}{4}\left(\left(f(y)h_n^2(y)\right)''\right)^2,$$

where the "A", stands, as usual, for "asymptotic" and denotes only the term exhibiting the slowest rate of convergence.

Recall that the expression of the variance remains unchanged compared to the classical fixed bandwidth case and, moreover, that the bias expression reveals Abramson´s square root law. In particular, in contrast to the classical fixed bandwidth case, the bandwidth function has now "moved under the differential operator", which results in the fact that, for $h_n(y) = h_n f(y)^{-1/2}$, this term is equal to zero:

$$ABias(\hat{f}_2(y)) = \frac{1}{2}\left(f(y)h_n f(y)^{-1}\right)'' \equiv 0,$$

provided that $f(y) > 0$.
The next term occurring in the bias is of order $O(h^4)$ (due to Silverman (1986)), which results in an optimal rate of convergence of the (asymptotic) mean squared error of order $O(n^{-8/9})$. This is considerably faster than $O(n^{-4/5})$, which is the optimal rate in the classical setting, in which K is a symmetric probability density function with finite second moment. Moreover, this faster rate is generally reserved for higher order kernels, which are, in turn, no probability density functions anymore. The use of higher order kernels also provides a possibility to achieve faster rates of the mean squared error; see for example Härdle (1994) for more details.

Remark 3.3. *Although the sample smoothing estimator $\hat{f}_2$ exhibits appealing theoretical properties, one should mention that this estimator suffers from the so-called "nonlocality phenomenon" as it was pointed in Terrell and Scott (1992). This means that even observations X_i, which lie "far away" from the point of estimation x, may have significant influence on the estimation procedure. This case will be suspended by the technical assumption on the function $t_y(x)$, in our theoretical analysis. In practical issues, this assumption may not be satisfied as it was stated in Terrell and Scott (1992) by a chosen example.*

3.2 Sample smoothing estimators in nonparametric regression models

Now we will turn to our main topic, namely the construction and asymptotic analysis of the Nadaraya-Watson sample smoothing estimator for regression functions.
The technique for the estimation of the drift in our previous analysis originates from discrete time series analysis and was originally proposed for the estimation of conditional expectations, for instance, in nonparametric regression models. We will now have a closer look at these models.
Let $(X_i, Y_i),\ i = 1, ..., n$, be identically distributed copies of the bivariate random vector (X, Y). Our main interest lies in the determination or quantification of the dependence structure between X and Y. We call X a (one-dimensional) covariate and Y the scalar-valued response and assume that they are connected by a nonparametric regression model

$$Y = m(X) + \varepsilon,$$

where m is an unknown regression function. m has to be specified by the available sample and ε is a centered random variable possessing a density f_ε and finite variance. Moreover, we assume that $E[\varepsilon|X] = 0$. The regression function m, evaluated at x, can now be specified as

$$m(x) = E[Y|X = x].$$

Based on the sample $(X_i, Y_i),\ i = 1, ..., n$, one intuitive and very simple method to estimate the value $m(x)$ nonparametrically is provided by the use of the already mentioned

Nadaraya-Watson estimator (cf. Nadaraya (1965), Watson (1964))

$$\hat{m}(x) = \frac{\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) Y_i}{\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)},$$

whose components have been introduced. This estimator has widely been studied in the regression context and its properties are well-known for independent and identically distributed as well as for weakly dependent (strong mixing) data. In fact, we have that (cf. Härdle (1994) or Cai (2001) in a slightly different setting)

$$\begin{aligned} E[\hat{m}(x)] &= \frac{h^2 \int_{\mathbb{R}} z^2 K(z)dz}{2}\left(m''(x) + \frac{2m'(x)f'(x)}{f(x)}\right) + o(h^2) \\ Var(\hat{m}(x)) &= \frac{\int_{\mathbb{R}} K^2(z)dzVar(Y|X=x)}{nh} + o\left(\frac{1}{nh}\right), \text{ as } n \to \infty \end{aligned}$$

under common smoothness assumptions like $m, f \in \mathcal{C}^2(\mathbb{R})$ on the regression function m and the marginal density f of X. The main goal of this chapter will now be to establish the pointwise asymptotic rates for the bias and the variance, as well as the derivation of the asymptotic distribution of the sample smoothing (adaptive) version of the Nadaraya-Watson estimator. Therefore, define:

$$\hat{m}_{ANW}(x) := \frac{\frac{1}{nh}\sum_{i=1}^{n} \frac{1}{h(X_i)} K\left(\frac{x-X_i}{h(X_i)}\right) Y_i}{\frac{1}{nh}\sum_{i=1}^{n} \frac{1}{h(X_i)} K\left(\frac{x-X_i}{h(X_i)}\right)}.$$

To the best of our knowledge, this estimator has only been studied in Demir and Töktamis (2010), where a short Monte Carlo study and also an empirical study have been considered. The entire asymptotic analysis has been left out.

3.3 Asymptotic properties of $\hat{m}_{ANW}(x)$

In this section, we will focus on the derivation of the asymptotic distribution of $\hat{m}_{ANW}(x)$. We will work with strong mixing (α-mixing) data and make use of some very interesting tools for the derivation of the proofs. This dependence structure occurs naturally in many existing and popular time series models; see for example Doukhan (1994), who provides an introduction into the concept of strong mixing properties for time series. The reason why this dependence structure is also interesting for the treatment of time-continuous processes is quite intuitive. In Masuda (2007), different assumptions on the coefficients of certain jump diffusion models were proposed, which ensure the ergodicity of the solution of the considered stochastic differential equations by establishing the β-mixing properties of such processes. In particular, β-mixing of a stochastic process is a stronger property

than α-mixing, in the sense that $\alpha_X(t) \leq \beta_X(t)$ for mixing coefficients α_X and β_X of a (stationary) stochastic process X.
We will now define what is meant by α-mixing and derive the asymptotic distribution of $\hat{m}_{ANW}(x)$ in the following.

Definition 3.4 (Doukhan (1994), Section 1.1)**.** *Let $(\Omega, \mathcal{A}, P)$ be a probability space with sub-σ-fields $\mathcal{B}$ and $\mathcal{C}$. Define the mixing coefficient α of $\mathcal{B}$ and $\mathcal{C}$ as*

$$\alpha = \alpha(\mathcal{B}, \mathcal{C}) := \sup_{B \in \mathcal{B}, C \in \mathcal{C}} |P(B \cap C) - P(B)P(C)|.$$

Now let $\{X_t\}_{t \in \mathbb{Z}}$ be a stochastic process. Furthermore, let $\mathcal{B}_t := \sigma(X_s, s \leq t)$ and $\mathcal{C}_{t,k} := \sigma(X_s, \geq t+k)$. The time series (or the stochastic process) $\{X_t\}_{t \in \mathbb{Z}}$ is called strongly mixing, if

$$\lim_{k \to \infty} \alpha(k) := \lim_{k \to \infty} \sup_{t \in \mathbb{Z}} \sup_{A \in \mathcal{B}_t, C \in \mathcal{C}_{t,k}} |P(A \cap C) - P(A)P(C)| = 0.$$

Example 3.5 (Linear autoregressive sequences)**.** *Let $\{X_t\}_{t \in \mathbb{N}}$ be a sequence of random variables satisfying*

$$X_{t+1} = aX_t + \varepsilon_{t+1},$$

where $\{\varepsilon_t\}_{t \in \mathbb{N}}$ is a sequence of i.i.d. random variables with $E[|\varepsilon_1|] < \infty$. Moreover, let ε_t possess a density which is almost everywhere positive on $\mathbb{R}$. Furthermore, assume that $E[|X_1|] < \infty$ and $|a| < 1$. Then it holds that $\{X_t\}_{t \in \mathbb{N}}$ is geometrically strong mixing, which means that

$$\alpha(k) = O(\gamma^k), \text{ for some } 0 < \gamma < 1.$$

Suppose that we observe n identically distributed copies (X_i, Y_i), $i = 1, ..., n$, of the random vector (X, Y), whose components are connected via a nonparametric regression model

$$Y_i = m(X_i) + \varepsilon_i,$$

where the random variables ε_i fulfill the above stated assumptions. In particular, they are centered, possess a density with respect to the Lebesgue measure, and have finite variance.
At first, we state the assumptions on the kernel function, which are comparable to assumption A2 in the first chapter.

Assumption B1

i) Let K be a bounded symmetric probability density function exhibiting a finite second moment (or a kernel of order two, as it is referred to in the literature).

ii) Let K additionally fulfill

$$\int_{\mathbb{R}} z^2 K^2(z) dz < \infty, \int_{\mathbb{R}} z^4 K(z) dz < \infty, \int_{\mathbb{R}} K^2(z) dz < \infty.$$

Finally, we impose the following assumptions for the pointwise estimation of $m(x)$ in our regression model and the appearing random variables:

Assumption B2

i) $t_x(y) = \frac{x-y}{h_n(y)}$ is strictly monotone in y.

ii) $h_n(x) := h_n\bar{h}(x) \to 0,\ nh_n(x) \to \infty$ as $n \to \infty$ and $h_n(x) > 0$.

iii) $h_n^2(x),\ f_X(x),\ \sigma^2(x) = Var(Y|X = x),\ m(x)$ are all twice continuously differentiable.

iv) $\exists\ H_1 \in \mathbb{R}:\ E[|\varepsilon_1\varepsilon_i||X_1 = x_1, X_i = x_2] \leq H_1\ \forall i \in \mathbb{N}$.

v) $\exists\ g_2 \in \mathbb{R}:\ f_{X_1,X_i}(x_1, x_2) \leq g_2\ \forall i \in \mathbb{N}$, where f_{X_1,X_i} indicates the joint density of X_1 and X_i.

vi) $\exists\ \delta > 2$ and $H_3 \in \mathbb{R}$ such that $\vartheta(u) := E[|Y|^\delta|X = u] < H_3$.

vii) $\exists\ 2 < \delta' < 4$ such that $\sum_{k=1}^{\infty} k\alpha(k)^{1-2/\delta'} < \infty$, where $\alpha(k)$ denotes the mixing coefficient of the sequence (X_i, Y_i).

viii) $\exists\ \{s_n\}_{n\in\mathbb{N}} \subseteq \mathbb{R}^+ :\ s_n \to \infty,\ s_n = o(\sqrt{nh_n(x)}),\ \sqrt{\frac{n}{h_n(x)}}\alpha(s_n) \to 0$ as $n \to \infty$.

ix) $\exists\ \delta^* :\ 4 > \delta^* > \delta > 2$ and $\exists\ H_4 \in \mathbb{R}$ such that $E[|Y|^{\delta^*}|X = u] \leq H_4 < \infty$. Moreover, $\alpha(k) = O(k^{-\theta^*})$, where $\theta^* \geq \frac{\delta^*\delta}{2(\delta^*-\delta)}$ and $n^{1/2-\delta/4}(h_n(x))^{\delta/\delta^*-1/2-\delta/4} = O(1)$ as $n \to \infty$.

Assumption B2, vii) is always fulfilled, if $\alpha(k)$ decays exponentially fast. Assumptions B2, iv)-ix) originate from Cai (2001), where the asymptotic behavior of a re-weighted version of the ordinary Nadaraya-Watson estimator is studied.

Remark 3.6. *By assuming that X and ε are independent and that ε_i are independent and normally distributed random variables, Assumptions B2, iv) and vi) are satisfied.*

Now we are ready to formulate our first important theorem in this section

Theorem 3.7. *Under Assumptions B1 and B2, provided that $f_X(x) > 0$ and $h_n(x) = C_{\bar{h}}n^{-1/5}$, we have that*

$$\sqrt{nh_n(x)}\left(\hat{m}_{ANW}(x) - m(x) - \mu_2(K)\left(\frac{h_n^2(x)m(x)}{2C_{\bar{h}}} + \frac{m'(x)(f_X(x)h_n^2(x))'}{f_X(x)}\right)\right)$$
$$\xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{||K||_2^2Var(Y|X = x)}{f_X(x)}\right),\ \textit{as } n \to \infty.$$

If the bandwidth function $h_n(x)$ fulfills $h_n(x) = o(n^{-1/5})$ as $n \to \infty$, then the bias is negligible and the asymptotic distribution is centered:

$$\sqrt{nh_n(x)}\,(\hat{m}_{ANW}(x) - m(x)) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{||K||_2^2 Var(Y|X = x)}{f_X(x)}\right), \text{ as } n \to \infty.$$

In order to establish the proof of this statement, we need three useful results from discrete time series analysis, which can all, for instance, be found in Billingsley (1995), Section 27. These results let us bound the covariance between two random variables which are connected via an α-mixing relation. Recall that we always work on the already mentioned probability space $(\Omega, \mathcal{F}, P)$.

Lemma 3.8 (Davydov´s inequality). *Let $X \in L^p(P)$, $Y \in L^q(P)$ with $p, q > 1$ and $\frac{1}{p} + \frac{1}{q} < 1$, then*

$$|Cov(X, Y)| \leq 2r\,(2\alpha(\sigma(X), \sigma(Y)))^{1/r}\,||X||_p||Y||_q,$$

where $\frac{1}{p} + \frac{1}{q} + \frac{1}{r} = 1$. Especially for $p = q > 2$ and $X \overset{\mathcal{D}}{=} Y$ we have

$$|Cov(X, Y)| \leq 2\left(1 - \frac{2}{p}\right)(2\alpha(\sigma(X), \sigma(Y)))^{1-2/p}\,||X||_p^2.$$

Lemma 3.9 (Billingsley´s inequality). *Let $\{X_t\}_{t\in\mathbb{Z}}$ be a stationary real-valued sequence of $\alpha-$mixing random variables. If Y is $\sigma(X_1, \ldots, X_n)$ measurable and bounded by C_Y and if Z is $\sigma(X_{n+k}, X_{n+k+1}, \ldots)$ measurable and bounded by C_Z, it holds that*

$$|Cov(Y, Z)| \leq 4C_Y C_Z \alpha(k).$$

For our purposes, we have to extend this lemma to complex valued random variables:

Corollary 3.10. *If $\{X_t\}_{t\in\mathbb{Z}}$ is a complex valued sequence of random variables and the conditions of the previous lemma remain the same, it holds that*

$$|Cov(Y, Z)| \leq 16C_Y C_Z \alpha(k).$$

Proof of Corollary 3.10. The proof is based on the decomposition of X into real and imaginary parts and, after that, the use of Lemma 3.9 for both parts. □

We see that we can bound the covariance of the considered random variables by the corresponding mixing coefficient. Moreover, and due to the definition of strong mixing, the random variables are nearly uncorrelated when the lag k is large enough. We are now ready to state the proof of Theorem 3.7

Proof of Theorem 3.7. We will start with the derivation of the asymptotic bias and, therefore, decompose $\hat{m}_{ANW}(x)$ according to the considered regression model as follows

$$\hat{m}_{ANW}(x) = m(x) + \frac{\hat{m}_1(x)}{\hat{f}_A(x)} + (nh_n(x))^{-1/2}\frac{\hat{m}_2(x)}{\hat{f}_A(x)},$$

where

$$\hat{m}_1(x) := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n(X_i)}K\left(\frac{x - X_i}{h_n(X_i)}\right)(m(X_i) - m(x))$$

as well as

$$\hat{m}_2(x) := \left(\frac{h_n(x)}{n}\right)^{1/2}\sum_{i=1}^{n}\frac{1}{h_n(X_i)}K\left(\frac{x - X_i}{h_n(X_i)}\right)\varepsilon_i.$$

Moreover, $\hat{f}_A(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n(X_i)}K\left(\frac{x-X_i}{h_n(X_i)}\right)$ denotes the denominator of $\hat{m}_{ANW}(x)$. To introduce an advantageous technique used to handle the appearing terms, we will focus on the expectation of the numerator $\hat{r}_A(x)$ of $\hat{m}_{ANW}(x)$ and, without loss of generality, evaluate it at $x = 0$. The following ideas are based on Terrell and Scott (1992), proof of Theorem 2, pp. 1262, where the properties of the smoothing sample density estimator for independent and identically distributed data are explored.

$$\begin{aligned}E[\hat{r}_A(0)] &= E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n(X_i)}K\left(\frac{-X_i}{h_n(X_i)}\right)Y_i\right] = E\left[\frac{1}{h_n(X)}K\left(\frac{X}{h_n(X)}\right)Y\right]\\ &= E\left[E\left[\frac{1}{h_n(X)}K\left(\frac{X}{h_n(X)}\right)Y\middle|X\right]\right] = E\left[\frac{1}{h_n(X)}K\left(\frac{X}{h_n(X)}\right)m(X)\right]\\ &= \int_{\mathbb{R}}\frac{1}{h_n(x)}K\left(\frac{x}{h_n(x)}\right)m(x)f_X(x)dx.\end{aligned}$$

Now we will evaluate the last integral using the substitution $\frac{x}{h_n(x)} = z$. Because of the assumptions on $h_n(x)$, we are allowed to differentiate this ratio with respect to x:

$$\frac{dz}{dx} = \frac{h_n(x) - xh'(x)}{h_n^2(x)} \Leftrightarrow dx = \frac{h_n^2(x)}{h_n(x) - xh_n'(x)}dz.$$

Moreover, we will use the monotonicity condition on $t_0(x) = \frac{x}{h_n(x)}$, which means that for every $z \in \mathbb{R}$ there is at most one $x = x(z) \in \mathbb{R}$ solving the equation $\frac{x}{h_n(x)} = z$. The latter fact is a direct consequence of the injectivity of $t_0(x)$.

Hence, we are now able to plug in this substitution into the above integral term and remark that the bounds of integration are not changed by the considered substitution. This yields:

$$\begin{aligned}&\int_{\mathbb{R}}\frac{1}{h_n(x)}K\left(\frac{x}{h_n(x)}\right)m(x)f_X(x)dx = \int_{\mathbb{R}}\frac{1}{h_n(x(z))}K(z)\frac{m(x(z))f_X(x(z))h_n^2(x(z))}{h_n(x(z)) - x(z)h_n'(x(z))}dz\\ &= \int_{\mathbb{R}}K(z)\frac{m(x(z))f_X(x(z))h_n(x(z))}{h_n(x(z)) - x(z)h_n'(x(z))}dz = \int_{\mathbb{R}}K(z)\frac{m(x(z))f_X(x(z))}{1 - h_n'(x(z))z}dz.\end{aligned}$$

We will now use the residue theorem to evaluate this term. Thus, consider the curve

$$\gamma_{r,x(z)} : [0,2\pi] \to \mathbb{C},\ t \to x(z) + re^{it}$$

with $r \in \mathbb{R}^+$ such that $0 \in B_r(x(z)) := \{y \in \mathbb{C} : |y - x(z)| < r\}$.
Obviously, $\gamma_{r,x(z)}$ defines a Jordan curve on the complex plane. Furthermore, the integrand is a holomorphic function and we can use the (generalized) residue theorem. Observe that the integrand has a pole of first order at $x(z)$:

$$\frac{1}{2\pi i}\int_{\gamma_{r,x(z)}} \frac{m(x)f_X(x)}{x - zh_n(x)}dx = 2\pi i\left(\frac{1}{2\pi i}\sum_{a\in\{x:\ x-zh_n(x)=0\}} Res_a\frac{m(x)f_X(x)}{x - zh_n(x)}\right)$$
$$= Res_{x(z)}\frac{m(x)f_X(x)}{x - zh_n(x)} = \frac{m(x)f_X(x)}{\frac{\partial}{\partial x}(x - zh_n(x))\big|_{x=x(z)}} = \frac{m(x)f_X(x)}{1 - zh_n'(x(z))}.$$

Therefore, it holds:

$$\int_{\mathbb{R}} K(z)\frac{m(x(z))f_X(x(z))}{1 - h_n'(x(z))z}dz = \int_{\mathbb{R}} K(z)\left(\frac{1}{2\pi i}\int_{\gamma_{r,x(z)}} \frac{m(x)f_X(x)}{x - zh_n(x)}dx\right)dz.$$

Now we will use a Taylor approximation for the integrand of the inner integral. First, observe that for

$$g_{f,h_n,m}(z) := \frac{m(x)f_X(x)}{x - zh_n(x)} = \frac{m(x)f_X(x)}{x - zh_n\bar{h}(x)}$$

an expansion around $z = 0$ gives us

$$g_{f,h_n,m}(z) = g_{f,h_n,m}(0) + g_{f,h_n,m}'(0)z + g_{f,h_n,m}''(0)\frac{z^2}{2} + o(h_n^2),$$

where the remainder term possesses this order, because $h_n(x) = h_n\bar{h}(x)$ and by the assumptions on K, especially the bounded moments of K up to order four (cf. Terrell and Scott P. 1262). We can use this representation for the evaluation of the integral term by the use of the generalized residue theorem for finite many residues as well as the moment assumptions on the kernel function K:

$$\int_{\mathbb{R}} K(z)\left(\frac{1}{2\pi i}\int_{\gamma_{r,x(z)}} \frac{m(x)f_X(x)}{x - zh_n(x)}dx\right)dz = \int_{\mathbb{R}} K(z)\left(\frac{1}{2\pi i}\int_{\gamma_{r,x(z)}} \frac{m(x)f_X(x)}{x}dx\right)dz$$
$$+\int_{\mathbb{R}} K(z)\left(\frac{1}{2\pi i}\int_{\gamma_{r,x(z)}} \frac{m(x)f_X(x)h_n(x)}{x^2}dx\ z\right)dz$$
$$+\int_{\mathbb{R}} K(z)\left(\frac{1}{2\pi i}\int_{\gamma_{r,x(z)}} \frac{m(x)f_X(x)h_n^2(x)}{x^3}dx\ z^2\right)dz + o(h_n^2).$$

For a convenient handling of these integrals, we use the well-known limit formula for higher order poles of meromorphic functions (cf. Fischer and Lieb (2008), Chapter VI, Theorem 4.2.):
If $a \in \mathbb{C}$ is a pole of order $k \in \mathbb{N}$, then the residue of a holomorphic function f at a is

$$Res_a f = \frac{1}{(k-1)!} \lim_{z \to a} \frac{\partial^{k-1}}{\partial z^{k-1}} \left((z-a)^k f(z) \right).$$

Here, the integrands have poles at $0 \in B_r(x(z))$ of orders one, two and three. Therefore, we can evaluate these integrals and get, under consideration of the assumptions taken on the kernel, for the asymptotic bias of the numerator:

$$E[\hat{r}_A(0)] = m(0)f_X(0) + \frac{\mu_2(K)(m(0)f_X(0)h_n^2(0))''}{2} + o(h_n^2),$$

where $\mu_2(K) := \int_{\mathbb{R}} z^2 K(z)dz < \infty$.
Using this technique together with assumption B2, ii) and iii), we find out that

$$\begin{aligned} E[\hat{m}_1(x)] &= \frac{\mu_2(K)}{2} \left(m''(x)f_X(x)h_n^2(x) + 2m'(x)(f_X(x)h_n^2(x))' \right) o(h_n^2) \\ &= \mu_2(K)f_X(x) \left(\frac{m''(x)h_n^2(x)}{2} + \frac{m'(x)(f_X(x)h_n^2(x))'}{f_X(x)} \right) + o(h_n^2) \\ &:= f_X(x)\bar{B}(x)\mu_2(K) + o(h_n^2). \end{aligned}$$

Our aim will be to use a law of large numbers for $\hat{m}_1$ in order to show that it converges asymptotically to its expectation. We therefore have to derive the variance of $\hat{m}_1$, which is slightly more complicated since we work with strong mixing data. Lemma 3.8 will act as a useful tool.

$$\begin{aligned} Var(\hat{m}_1(x)) &= Var \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_n(X_i)} K \left(\frac{x - X_i}{h_n(X_i)} \right) (m(X_i) - m(x)) \right) \\ &= \frac{1}{n} Var \left(\frac{1}{h_n(X_1)} K \left(\frac{x - X_1}{h_n(X_1)} \right) (m(X_1) - m(x)) \right) \\ &\quad + \frac{1}{n^2} \sum_{1 \le i \neq j \le n} Cov \left(C_i, C_j \right), \end{aligned}$$

where

$$C_i := \frac{1}{h_n(X_i)} K \left(\frac{x - X_i}{h_n(X_i)} \right) (m(X_i) - m(x)), \; i = 1, ..., n.$$

We will handle the terms above separately and start with the second one. Because of the stationarity of $\{X_i\}_{i=1,\dots,n}$, we can conclude that

$$\left|\frac{1}{n^2}\sum_{1\leq i\neq j\leq n} Cov(C_i, C_j)\right| = \left|\frac{2}{n}\sum_{i=2}^{n}\left(1-\frac{i-1}{n}\right)Cov(C_1, C_i)\right|$$
$$\leq \frac{2}{n}\sum_{i=2}^{n}\left|Cov(C_1, C_i)\right| \overset{Davydov}{\leq} \frac{2}{n}\sum_{i=2}^{n}\alpha(i-1)^{1-2/\delta}||C_1||_\delta^2, \text{ with } \delta \in (2,4).$$

At first we derive $||C_1||_\delta^2$:

$$E[C_1^\delta] = E\left[\left(\frac{1}{h_n(X_1)}K\left(\frac{x-X_1}{h_n(X_1)}\right)(m(X_1)-m(x))\right)^\delta\right]$$
$$= \int_{\mathbb{R}}\frac{1}{h_n^\delta(y)}K^\delta\left(\frac{x-y}{h_n(y)}\right)(m(y)-m(x))^\delta f_X(y)dy \text{ we set w.l.o.g. } x=0 \text{ and } y/h_n(y) := z$$
$$= \int_{\mathbb{R}} K^\delta(z)(m(y(z))-m(0))^\delta\frac{f_X(y(z))h_n^{1-\delta}(y(z))}{1-zh_n'(y(z))}dz$$
$$= \int_{\mathbb{R}} K^\delta(z)\frac{1}{2\pi i}\int_{\gamma_{r,y(z)}}\frac{(m(y)-m(0))^\delta f_X(y)h_n^{1-\delta}(y)}{y-zh_n(y)}dy\, dz.$$

After expanding the inner integrand in a Taylor series around $z=0$, we have for $h_n(y) = h_n\tilde{h}(y)$ that

$$E[C_1^\delta] \leq Ch_n^{2-\delta}.$$

When we apply these results, we find out that for the sum of the covariances one has

$$\left|\frac{1}{n^2}\sum_{1\leq i\neq j\leq n} Cov(C_i, C_j)\right| \leq \frac{2}{n}\sum_{i=2}^{n}\alpha(i-1)^{1-2/\delta}||C_1||_\delta^2$$
$$\leq \tilde{C}\cdot\frac{1}{n}\sum_{i=1}^{n-1}\alpha(i)^{1-2/\delta}(h_n^{2-\delta})^{2/\delta} = O\left(\frac{h_n^{(4-2\delta)/\delta}}{n}\right), \text{ as } n\to\infty,$$

where the last equation is justified by the summability assumption on the mixing coefficients

$$\sum_{k=1}^{\infty} k\alpha(k)^{1-2/\delta} < \infty.$$

We will now examine the first summand:

$$\begin{aligned}
&\frac{1}{n} Var\left(\frac{1}{h_n(X_1)} K\left(\frac{x-X_1}{h_n(X_1)}\right)(m(X_1)-m(x))\right)\\
&\leq \frac{1}{n} E\left[\left(\frac{1}{h_n(X_1)} K\left(\frac{x-X_1}{h_n(X_1)}\right)(m(X_1)-m(x))\right)^2\right]\\
&\overset{w.l.o.g. x=0}{=} \frac{1}{n}\int_{\mathbb{R}} \frac{1}{h_n^2(y)} K^2\left(\frac{y}{h_n(y)}\right)(m(y)-m(0))^2 f_X(y)dy\\
&\overset{y/h_n(y)=z}{=} \frac{1}{n}\int_{\mathbb{R}} K^2(z)\frac{f_X(y(z))\frac{(m(y(z))-m(0))^2}{h_n(y(z))}}{1-zh_n'(y(z))}dz\\
&= \frac{1}{n}\int_{\mathbb{R}} K^2(z)\frac{1}{2\pi i}\int_{\gamma_{r,y(z)}} \frac{\frac{f_X(y)(m(y)-m(0))^2}{h_n(y)}}{y-zh_n(y)}dy\ dz.
\end{aligned}$$

Expanding the inner integrand around $z=0$ yields:

$$\begin{aligned}
&\frac{1}{n}\int_{\mathbb{R}} K^2(z)\frac{1}{2\pi i}\int_{\gamma_{r,y(z)}} \frac{\frac{f_X(y)(m(y)-m(0))^2}{h_n(y)}}{y-zh_n(y)}dy\ dz\\
&\doteq \frac{1}{n}\int_{\mathbb{R}} K^2(z)\frac{1}{2\pi i}\int_{\gamma_{r,y(z)}} \frac{f_X(y)(m(y)-m(0))^2}{yh_n(y)}dy\ dz\\
&\quad + \frac{1}{n}\int_{\mathbb{R}} K^2(z)z\frac{1}{2\pi i}\int_{\gamma_{r,y(z)}} \frac{f_X(y)(m(y)-m(0))^2}{y^2}dy\ dz\\
&+ \frac{1}{n}\int_{\mathbb{R}} K^2(z)\frac{z^2}{2}\frac{1}{2\pi i}\int_{\gamma_{r,y(z)}} \frac{h_n(y)f_X(y)(m(y)-m(0))^2}{y^3}dy\ dz\\
&= \frac{1}{n}||K||_2^2 \underbrace{\left(\frac{(m(y)-m(0))^2 f_X(y)}{h_n(y)}\right)\Bigg|_{y=0}}_{=0} + \frac{1}{n}\int_{\mathbb{R}} K^2(z)zdz \underbrace{\frac{\partial}{\partial y}\left(f_X(y)(m(y)-m(0))^2\right)\Bigg|_{y=0}}_{=0}\\
&+ \frac{1}{n}\int_{\mathbb{R}} K^2(z)\frac{z^2}{2}\frac{\partial^2}{\partial^2 y}\left(h_n(y)f_X(y)(m(y)-m(0))^2\right)\Bigg|_{y=0}\\
&= \frac{h_n}{n}\int_{\mathbb{R}} K^2(z)\frac{z^2}{2}dz \cdot f_X(0)m'(0)\tilde{h}(0) = O\left(\frac{h_n}{n}\right), \text{ as } n\to\infty.
\end{aligned}$$

Therefore, we can now conclude that:

$$Var\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n(X_i)}K\left(\frac{x-X_i}{h_n(X_i)}\right)(m(X_i)-m(x))\right) = O\left(\frac{h_n}{n}\right) + O\left(\frac{h_n^{(4-2\delta)/\delta}}{n}\right),\ \delta\in(2,4).$$

Using this result as well as the law of large numbers in connection with Chebyshev´s inequality, we can deduce that

$$\sqrt{nh_n(x)}(\hat{m}_1(x) - f_X(x)\tilde{B}(x)\mu_2(K)) \xrightarrow{P} 0 \text{ as } n\to\infty.$$

Using exact analogous techniques, we are able to derive the (weak) consistency of the denominator of $\hat{m}_{ANW}(x)$, too. The rates are the same as in the fixed bandwidth case and we can conclude that

$$\sqrt{nh_n(x)}\left(\frac{\hat{m}_1(x)}{\hat{f}_A(x)} - \tilde{B}(x)\mu_2(K)\right) \xrightarrow{P} 0 \text{ as } n \to \infty.$$

We will now turn to the examination of

$$\hat{m}_2(x) = \left(\frac{h_n(x)}{n}\right)^{1/2} \sum_{i=1}^{n} \frac{1}{h_n(X_i)} K\left(\frac{x - X_i}{h_n(X_i)}\right) \varepsilon_i.$$

As we will see, this term is responsible for the asymptotic distribution. Recall that $\hat{m}_2(x)$ is centered, because $E[\varepsilon_i|X_i] = 0$ holds true. We will now focus on the variance of this term. The result of this computation is the variance of the asymptotic distribution.

$$\begin{aligned}
&Var(\hat{m}_2(x)) = \frac{h_n(x)}{n} Var\left(\sum_{i=1}^{n} \frac{1}{h_n(X_i)} K\left(\frac{x - X_i}{h_n(X_i)}\right) \varepsilon_i\right)\\
&= h_n(x) Var\left(\frac{1}{h_n(X_1)} K\left(\frac{x - X_1}{h_n(X_1)}\right) \varepsilon_1\right)\\
&\quad + 2h_n(x) \sum_{i=1}^{n} \left(1 - \frac{i-1}{n}\right) Cov\left(\frac{1}{h_n(X_1)} K\left(\frac{x - X_1}{h_n(X_1)}\right) \varepsilon_1, \frac{1}{h_n(X_i)} K\left(\frac{x - X_i}{h_n(X_i)}\right) \varepsilon_i\right)\\
&=: I(x) + II(x)
\end{aligned}$$

We will handle these terms separately and start with $I(x)$:

$$\begin{aligned}
I(x) &= h_n(x) E\left[\left(\frac{1}{h_n(X_1)} K\left(\frac{x - X_1}{h_n(X_1)}\right) \varepsilon_1\right)^2\right]\\
&= h_n(x) \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{e^2 f_{X,\varepsilon}(y, e)}{h_n^2(y)} K^2\left(\frac{x - y}{h_n(y)}\right) dyde\\
&= h_n(x) \int_{\mathbb{R}} \frac{\sigma^2(y) f_X(y)}{h_n^2(y)} K^2\left(\frac{x - y}{h_n(y)}\right) dy, \text{ where } \sigma^2(y) := E[\varepsilon^2|X = y].
\end{aligned}$$

We will abbreviate the derivation of this integral, because the procedure remains the same as before. Using the residue theorem twice as well as a Taylor expansion yields:

$$I(x) = \sigma^2(x) f_X(x) ||K||_2^2 + O(h_n), \text{ as } n \to \infty.$$

For the derivation of $II(x)$, we choose a sequence $\{a_n\}_{n\in\mathbb{N}} \subseteq \mathbb{R}^+$ such that

$$a_n \to \infty \text{ as well as } a_n h_n(x) = a_n h_n \tilde{h}(x) \to 0 \text{ as } n \to \infty.$$

Furthermore, we introduce the abbreviation

$$\xi_i := \frac{h_n(x)^{1/2}\varepsilon_i}{h_n(X_i)} K\left(\frac{x - X_i}{h_n(X_i)}\right).$$

Using this assumptions we can conclude:

$$|II(x)| = 2\left|\sum_{i=2}^{n}\left(1 - \frac{i-1}{n}\right) Cov(\xi_1, \xi_i)\right| \leq 2\underbrace{\sum_{i=2}^{\lfloor a_n \rfloor} |Cov(\xi_1, \xi_i)|}_{:=II_1(x)} + 2\underbrace{\sum_{i=\lfloor a_n \rfloor+1}^{n} |Cov(\xi_1, \xi_i)|}_{:=II_2(x)}.$$

Now we use assumptions B2, iv) and v):

B2, iv) $\exists H_1 \in \mathbb{R}: \; E[|\varepsilon_1 \varepsilon_i||X_1, X_i] \leq H_1 \; \forall i \in \mathbb{N}$

B2, v) $\exists g_2 \in \mathbb{R}: \; f_{X_1,X_i}(x_1, x_2) \leq g_2 \; \forall i \in \mathbb{N}.$

Ad $II_1(x)$:

$$\begin{aligned}
II_1(x) &= \sum_{i=2}^{\lfloor a_n \rfloor} |Cov(\xi_1, \xi_i)| = \sum_{i=2}^{\lfloor a_n \rfloor} |E[\xi_1 \xi_i]| \\
&\leq h_n(x) \sum_{i=2}^{\lfloor a_n \rfloor} E\left[\frac{1}{h_n(X_1)} K\left(\frac{x - X_1}{h_n(X_1)}\right) \frac{1}{h_n(X_i)} K\left(\frac{x - X_i}{h_n(X_i)}\right) E[|\varepsilon_1 \varepsilon_i||X_1, X_i]\right] \\
&\leq H_1 h_n(x) \sum_{i=2}^{\lfloor a_n \rfloor} E\left[\frac{1}{h_n(X_1)} K\left(\frac{x - X_1}{h_n(X_1)}\right) \frac{1}{h_n(X_i)} K\left(\frac{x - X_i}{h_n(X_i)}\right)\right] \\
&\leq H_1 g_2^2 h_n(x) \sum_{i=2}^{\lfloor a_n \rfloor} \underbrace{\left(\int_{\mathbb{R}} \frac{1}{h_n(y)} K\left(\frac{x - y}{h_n(y)}\right) dy\right)^2}_{=O(1)} \\
&\leq \tilde{M} h_n(x) O(a_n) = O(h_n(x) a_n) = o(1), \text{ as } n \to \infty.
\end{aligned}$$

To handle the second term $II_2(x)$, we again make use of Davydov´s inequality:

$$|II_2(x)| \leq \sum_{i=a_n+1}^{n} |Cov(\xi_1, \xi_i)| \leq C \sum_{i=a_n+1}^{n} \alpha(i-1)^{1-2/\delta} (E[|\xi_1|^\delta])^{2/\delta}, \; \delta > 2.$$

First, observe that, under the assumption that $\vartheta(y) := E[|\varepsilon|^\delta | X = y] < H_3 < \infty$, it holds that

$$\begin{aligned}E[|\xi_1|^\delta] &= h_n(x)^{\delta/2} E\left[\left(\frac{1}{h_n(X_1)} K\left(\frac{x-X_1}{h_n(X_1)}\right) |\varepsilon_1|\right)^\delta\right] \\ &= h_n(x)^{\delta/2} E\left[\frac{1}{h_n^\delta(X_1)} K^\delta\left(\frac{x-X_1}{h_n(X_1)}\right) \vartheta(X_1)\right] \\ &= h_n(x)^{\delta/2} \int_{\mathbb{R}} \frac{1}{h_n^\delta(y)} K^\delta\left(\frac{x-y}{h_n(y)}\right) \vartheta(y) f_X(y) dy.\end{aligned}$$

After several computations we have

$$E[|\xi_1|^\delta] \leq C_1 \cdot h_n(x)^{1-\delta/2}.$$

Therefore:

$$\begin{aligned}|II_2(x)| \leq C \sum_{i=a_n+1}^{n} \alpha(i-1)^{1-2/\delta} (C_1 h_n(x)^{1-\delta/2})^{2/\delta} &= C_2 h_n(x)^{2/\delta-1} \sum_{i=a_n}^{n-1} \alpha(i)^{1-2/\delta} \\ = C_2 h_n(x)^{2/\delta-1} \sum_{i=a_n}^{n-1} i^{-1} i\alpha(i)^{1-2/\delta} &\leq C_2 h_n(x)^{2/\delta-1} a_n^{-1} \sum_{i=a_n}^{n-1} i\alpha(i)^{1-2/\delta}.\end{aligned}$$

Now choose a_n such that

$$h_n(x)^{2/\delta-1} a_n^{-1} = O(1) \Leftrightarrow h_n(x)^{1-2/\delta} a_n = O(1),$$

so that the assumptions $a_n \to \infty$ and $a_n h_n(x) \to 0$ are still valid. For example, choose $a_n = L h_n^{2/\delta-1}$, $L > 0$ and recall that $\delta > 2$. Now make use of the assumption that $\sum_{i=1}^\infty i\alpha(i)^{1-2/\delta} < \infty$ to conclude that

$$\sum_{i=\lfloor a_n \rfloor}^{n} i\alpha(i)^{1-2/\delta} \to 0 \text{ as } n \to \infty.$$

Thus, we have established an expression for the variance of $\hat{m}_1(x)$. Under the previous assumptions, the covariances are vanishing asymptotically and it holds that

$$Var(\hat{m}_2(x)) \to \sigma^2(x) f_X(x) ||K||_2^2 \text{ as } n \to \infty.$$

Let us now shortly summarize what we have shown until now:

$$\begin{aligned}&\sqrt{nh_n(x)}(\hat{m}_{ANW}(x) - m(x) - \tilde{B}(x)\mu_2(K)) \\ &= \sqrt{nh_n(x)}\left(\frac{\hat{m}_1(x)}{\hat{f}_A(x)} - \tilde{B}(x)\mu_2(K)\right) + \sqrt{nh_n(x)}(nh_n(x))^{-1/2}\frac{\hat{m}_2(x)}{\hat{f}_A(x)} \\ &= \frac{n^{-1/2}\sum_{i=1}^n h_n(x)^{1/2} \frac{1}{h_n(X_i)} K\left(\frac{x-X_i}{h_n(X_i)}\right)\varepsilon_i}{\hat{f}_A(x)} + o_P(1) = \frac{n^{-1/2}\sum_{i=1}^n \xi_i}{\hat{f}_A(x)} + o_P(1).\end{aligned}$$

Finally, we will show the asymptotic normality of $n^{-1/2}\sum_{i=1}^{n}\xi_i$. The sequence $\{\xi_i\}_{i=1,\dots,n}$ is α-mixing and, therefore, the classical Lindeberg theorem cannot be applied to this sequence. Instead, we will use a so-called "Large-Block"- and "Small-Block"-technique, which enables us to derive the asymptotic distribution of $n^{-1/2}\sum_{i=1}^{n}\xi_i$. This technique is commonly attributed to Bernstein (1927), sometimes referred to as Bernstein´s blocking. Other authors refer to Markov when dealing with this method. We do not want to take part in this discussion and will only focus on the main idea behind this approach. In the context of nonparametric density and regression estimation, this technique has played a major role in some articles; we refer to Cai (2001), where the blocking technique is used for the derivation of the asymptotic normality of a re-weighted version of the ordinary Nadaraya-Watson estimator in a regression context.

Now we will introduce the main idea. Consider the sum $\xi_1 + \ldots + \xi_n$ and divide it in $2k_n + 1$ alternating blocks of length r_n and s_n -the "big blocks" and the "small blocks"- as well as a residual block which length is smaller than $r_n + s_n$. Observe that, under these conditions, $k_n = \lfloor n/(r_n + s_n)\rfloor$. Now choose s_n so small that the small blocks are negligible in probability (they are of order $o_P(1)$) but even so large that the big blocks are asymptotically independent. Then we apply the Lindeberg theorem to the big blocks and the proof is completed.

At first, define the big blocks:

$$U_{ni} := \xi_{(i-1)(r_n+s_n)+1} + \ldots + \xi_{(i-1)(r_n+s_n)+r_n}, \text{ for } 1 \le i \le k_n.$$

Next, define the small blocks:

$$V_{ni} := \xi_{(i-1)(r_n+s_n)+r_n+1} + \ldots + \xi_{(i-1)(r_n+s_n)+r_n+s_n}, \text{ for } 1 \le i \le k_n.$$

Finally, define the residual block:

$$W_n := \xi_{k_n(r_n+s_n)+1} + \ldots + \xi_n.$$

We are now able to divide $n^{-1/2}\sum_{i=1}^{n}\xi_i$ into three parts:

$$n^{-1/2}\sum_{i=1}^{n}\xi_i = n^{-1/2}\sum_{i=1}^{k_n}U_{ni} + n^{-1/2}\sum_{i=1}^{k_n}V_{ni} + n^{-1/2}W_n.$$

Let us now verify that $E[(\sum_{i=1}^{k_n}V_{ni})^2] = o(n)$ as $n \to \infty$. If this order holds true we can conclude that the small blocks are negligible in probability by the use of Chebyshev´s inequality. Recall that the random variables ξ_i are centered and, hence, the second moment of the sum denotes its variance. Let us now make assumptions about the length s_n of the small blocks (cf. Cai (2001)):

$$s_n \to \infty,\ s_n = o((nh_n(x))^{1/2}),\ \left(\frac{n}{h_n(x)}\right)^{1/2}\alpha(s_n) \to 0,\ \text{as } n \to \infty.$$

Therefore, it exists a sequence β_n such that

$$\beta_n \to \infty,\ \beta_n s_n = o((nh_n(x))^{1/2}),\ \beta_n \left(\frac{n}{h_n(x)}\right)^{1/2} \alpha(s_n) \to 0,\ \text{as } n \to \infty.$$

Then define the length r_n of the big blocks as $r_n := \lfloor \frac{\sqrt{nh_n(x)}}{\beta_n} \rfloor$. It follows that

$$s_n/r_n \to 0,\ r_n/n \to 0,\ n/r_n\alpha(s_n) \to 0, r_n/\sqrt{nh_n(x)} \to 0,\ \text{as } n \to \infty.$$

Now we are ready to cope with the small blocks. Observe that V_{ni} are centered for all $1 \leq i \leq k_n$, too.

$$E[(\sum_{i=1}^{k_n} V_{ni})^2] = \sum_{i=1}^{k_n} Var(V_{ni}) + \sum_{1\leq i\neq j\leq k_n} Cov(V_{ni}, V_{nj}).$$

The variance of V_{ni} can be derived by the help of the previous findings as follows:

$$\begin{aligned}
Var(V_{nj}) &= Var\left(\sum_{i=(j-1)(r_n+s_n)+r_n+1}^{j(r_n+s_n)} h_n(x)^{1/2}\frac{1}{h_n(X_i)}K\left(\frac{x-X_i}{h_n(X_i)}\right)\varepsilon_i\right)\\
&= h_n(x) \sum_{i=(j-1)(r_n+s_n)+r_n+1}^{j(r_n+s_n)} Var\left(\frac{1}{h_n(X_i)}K\left(\frac{x-X_i}{h_n(X_i)}\right)\varepsilon_i\right)\\
&\quad + h_n(x) \sum_{(j-1)(r_n+s_n)+r_n+1\leq i\neq k\leq j(r_n+s_n)}\\
&\qquad Cov\left(\frac{1}{h_n(X_i)}K\left(\frac{x-X_i}{h_n(X_i)}\right)\varepsilon_i, \frac{1}{h_n(X_k)}K\left(\frac{x-X_k}{h_n(X_k)}\right)\varepsilon_k\right)\\
&= h_n(x)s_n Var\left(\frac{1}{h_n(X_1)}K\left(\frac{x-X_1}{h_n(X_1)}\right)\varepsilon_1\right)\\
&\quad + 2h_n(x)s_n\sum_{i=2}^{s_n}\left(1-\frac{i}{s_n}\right)Cov\left(\frac{1}{h_n(X_1)}K\left(\frac{x-X_1}{h_n(X_1)}\right)\varepsilon_1, \frac{1}{h_n(X_i)}K\left(\frac{x-X_i}{h_n(X_i)}\right)\varepsilon_i\right)\\
&= s_n(\sigma^2(x)f_X(x)||K||_2^2 + o(1)).
\end{aligned}$$

By the use of the stationarity of V_{ni}, we can conclude that

$$\begin{aligned}
\sum_{i=1}^{k_n} Var(V_{ni}) &\leq k_n s_n(\sigma^2(x)f_X(x)||K||_2^2 + o(1))\\
&\overset{k_n=\lfloor\frac{n}{r_n+s_n}\rfloor}{\leq} \frac{ns_n}{s_n+r_n}(\sigma^2(x)f_X(x)||K||_2^2 + o(1))\\
&= \frac{n\frac{s_n}{r_n}}{\frac{s_n}{r_n}+1}(\sigma^2(x)f_X(x)||K||_2^2 + o(1)) = o(n),\ \text{as } n \to \infty,
\end{aligned}$$

because $s_n/r_n \to 0$ as $n \to \infty$.
Now we will focus on the covariances. This procedure is rather simple, because the main work has been done already.

$$\sum_{1\le i\neq j\le k_n} Cov(V_{ni},V_{nj}) = \sum_{1\le i\neq j\le k_n} Cov\left(\sum_{l=i(r_n+s_n)+r_n+1}^{i(r_n+s_n)+s_n} \xi_l, \sum_{m=j(r_n+s_n)+r_n+1}^{j(r_n+s_n)+s_n} \xi_m\right)$$
$$\overset{\text{index shift}}{=} \sum_{1\le i\neq j\le k_n}\sum_{m=1}^{s_n}\sum_{l=1}^{s_n} Cov\left(\xi_{i(r_n+s_n)+r_n+m}, \xi_{j(r_n+s_n)+r_n+l}\right).$$

The lag between two of the ξ_i‘s amounts at least r_n (this is the case, where two consecutive small blocks and the rightmost- and leftmost-lying random variables inside both blocks are considered). Thus, we will consider all covariances of random variables, whose lag is at least r_n:

$$\left|\sum_{1\le i\neq j\le k_n} Cov(V_{ni},V_{nj})\right|$$
$$\le 2h_n(x)\sum_{l=1}^{n-r_n}\sum_{m=l+r_n}^{n}\left|Cov\left(\frac{1}{h_n(X_l)}K\left(\frac{x-X_l}{h_n(X_l)}\right)\varepsilon_l, \frac{1}{h_n(X_m)}K\left(\frac{x-X_m}{h_n(X_m)}\right)\varepsilon_m\right)\right|$$
$$\overset{\text{index shift}}{=:} 2h_n(x)\sum_{l=1}^{n-r_n}\sum_{m=r_n}^{n-l}|Cov(\xi_l',\xi_m')| \le 2nh_n(x)\sum_{m=r_n}^{n-1}|Cov(\xi_1',\xi_{m+1}')|$$
$$\le 2nh_n(x)\sum_{m=1}^{n-1}|Cov(\xi_1',\xi_{m+1}')| = o(n) \text{ as } n\to\infty,$$

where the last equality comes from the fact that the covariances are asymptotically vanishing, as we have already seen.
Therefore:

$$E\left[\left(\sum_{i=1}^{k_n} V_{ni}\right)^2\right] = o(n) \Leftrightarrow \frac{1}{n}E\left[\left(\sum_{i=1}^{k_n} V_{ni}\right)^2\right] = o(1), \text{ as } n\to\infty.$$

Now we are ready to make sure that the residual block W_n converges to zero as $n \to \infty$ in probability, too:

$$\frac{1}{n}E[W_n^2] = \frac{h_n(x)}{n}E\left[\left(\sum_{i=k_n(r_n+s_n)+1}^{n}\xi_i\right)^2\right]$$
$$= \frac{h_n(x)}{n}\cdot(n-k_n(r_n+s_n))Var(\xi_1) + \frac{h_n(x)}{n}\sum_{k_n(r_n+s_n)+1\le i\neq j\le n} Cov(\xi_i',\xi_j')$$

$$\begin{aligned} &\leq \frac{h_n(x)}{n} \cdot (n - k_n(r_n + s_n))Var(\xi_1) + 2h_n(x) \sum_{j=1}^{n-1} |Cov(\xi_1', \xi_j')| \\ &\leq \frac{h_n(x)}{n} \cdot (r_n + s_n) + o(1) = o(1) \text{ as } n \to \infty \end{aligned}$$

since $n - k_n(r_n + s_n) \leq r_n + s_n$, and $r_n/n \to 0$ as $n \to \infty$.
Now we can summarize what we have shown and what is left to be established:

$$\begin{aligned} n^{-1/2} \sum_{i=1}^{n} \xi_i &= n^{-1/2} \sum_{i=1}^{k_n} U_{ni} + n^{-1/2} \sum_{i=1}^{k_n} V_{ni} + n^{-1/2} W_n \\ &= n^{-1/2} \sum_{i=1}^{k_n} U_{ni} + o_P(1) \overset{!}{\xrightarrow{\mathcal{D}}} \mathcal{N}(0, \sigma^2(x) f_X(x) ||K||_2^2). \end{aligned}$$

For the sake of brevity, we will shorten the proof and will only focus on the idea behind it. At first, recall that the random variables U_{ni} are not independent. Now consider random variables $U_{n1}', \dots, U_{nk_n}'$, which have the same distribution as U_{ni} for all i and n, and which additionally are independent. We will use Lemma 3.9 and, in particular, Corollary 3.10 to show that $n^{-1/2} \sum_{i=1}^{k_n} U_{ni}$ is asymptotically normally distributed, if and only if $n^{-1/2} \sum_{i=1}^{k_n} U_{ni}'$ is. Consider therefore the random variables $Y_U := e^{\frac{itU_{n1}}{\tau(x)\sqrt{n}}}$ and $Z_U = e^{\frac{itU_{n2}}{\tau(x)\sqrt{n}}}$, where $\tau^2(x) := \sigma^2(x) f_X(x) ||K||_2^2$. Observe that $|Y_U|, |Z_U| \leq 1$ and conclude that:

$$\begin{aligned} &\left| E\left[[\exp\left(\frac{it(U_{n1} + U_{n2})}{\tau(x)\sqrt{n}} \right)] \right] - E\left[\exp\left(\frac{it(U_{n1}' + U_{n2}\prime)}{\tau(x)\sqrt{n}} \right) \right] \right| \\ &= \left| E\left[\exp\left(\frac{it(U_{n1} + U_{n2})}{\tau(x)\sqrt{n}} \right) \right] - E\left[\exp\left(\frac{itU_{n1}'}{\tau(x)\sqrt{n}} \right) \right] E\left[\exp\left(\frac{itU_{n2}'}{\tau(x)\sqrt{n}} \right) \right] \right| \\ &= \left| E\left[\exp\left(\frac{it(U_{n1} + U_{n2})}{\tau(x)\sqrt{n}} \right) \right] - E\left[\exp\left(\frac{itU_{n1}}{\tau(x)\sqrt{n}} \right) \right] E\left[\exp\left(\frac{itU_{n2}}{\tau(x)\sqrt{n}} \right) \right] \right| \\ &= |Cov(Y, Z)| \leq 16\alpha(s_n). \end{aligned}$$

Via induction, we find out that

$$\begin{aligned} &\left| E\left[\exp\left(\frac{it \sum_{i=1}^{k_n} U_{ni}}{\tau(x)\sqrt{n}} \right) \right] - \prod_{i=1}^{k_n} E\left[\exp\left(\frac{itU_{ni}'}{\tau(x)\sqrt{n}} \right) \right] \right| \leq 16(k_n - 1)\alpha(s_n) \\ &\leq 16 k_n \alpha(s_n) \leq 16 \frac{n\alpha(s_n)}{r_n + s_n} = 16 \frac{n/r_n \cdot \alpha(s_n)}{1 + s_n/r_n} = o(1), \text{ as } n \to \infty. \end{aligned}$$

We see that the characteristic function of $\frac{it \sum_{i=1}^{k_n} U_{ni}}{\tau(x)\sqrt{n}}$ converges to $e^{-t^2/2}$, if and only if $\frac{it \sum_{i=1}^{k_n} U_{ni}'}{\tau(x)\sqrt{n}}$ does.

Now we shorten the proof and indicate that the asymptotic variance of $\sum_{i=1}^{k_n} U'_{ni}$ is given by

$$Var\left(\sum_{i=1}^{k_n} U'_{ni}\right) \to \tau(x)^2 = \sigma^2(x) f_X(x) ||K||_2^2, \text{ as } n \to \infty.$$

Moreover, one can show that by Assumption B2, ix) (see also Cai (2001)), the independent and identically distributed random variables U'_{nj} fulfill the Lindeberg condition such that finally

$$n^{-1/2} \sum_{j=1}^{k_n} U'_{nj} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \tau^2(x)), \text{ as } n \to \infty.$$

Therefore, we have finished our proof. □

3.4 Asymptotic mean squared error

A natural question is now, how to use these results to actually improve -even in an asymptotical sense- the rate of convergence of the considered Nadaraya-Watson estimator. For this purpose, we will now state the rates for the asymptotic mean squared error and compare them to the classical case, where $h(X_i) \equiv h$.

Lemma 3.11. *Under assumptions B1 and B2, provided that $f_X(x) > 0$, we have the following representation of the asymptotic mean squared error ("MSE") of $\hat{m}_{ANW}(x)$*

$$\begin{aligned} AMSE(\hat{m}_{ANW}(x)) &= AVar(\hat{m}_{ANW}(x)) + (ABias(\hat{m}_{ANW}(x)))^2 \\ &= \frac{1}{nh(x)} \frac{Var(Y|X=x)}{f_X(x)} ||K||_2^2 + \frac{\mu_2^2(K)}{4} \left(h^2(x) m''(x) + 2 \frac{m'(x)(f_X(x)h^2(x))'}{f_X(x)} \right)^2 \\ &= \frac{1}{nh} \frac{Var(Y|X=x)}{\bar{h}(x) f_X(x)} ||K||_2^2 + \frac{h^4 \mu_2^2(K)}{4} \left(\bar{h}^2(x) m''(x) + 2 \frac{m'(x)(f_X(x)\bar{h}^2(x))'}{f_X(x)} \right)^2, \end{aligned}$$

where we used $h(x) = h\bar{h}(x)$.

In contrast to the result of the adaptive version, we recall the result in the classical case such that

$$\hat{m}(x) = \frac{\frac{1}{nh}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) Y_i}{\frac{1}{nh}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)}.$$

Under appropriate smoothness assumptions on m and f_X, the AMSE of $\hat{m}(x)$ has the form

$$AMSE(\hat{m}(x)) = \frac{1}{nh} \frac{Var(Y|X=x)}{f_X(x)} ||K||_2^2 + \frac{h^4 \mu_2^2(K)}{4} \left(m''(x) + 2 \frac{m'(x) f'_X(x)}{f_X(x)} \right)^2.$$

It is notable that the terms are almost the same for the adaptive and the classical Nadaraya-Watson estimator. The only difference, which impact should now be examined, is that the bandwidth function has moved under the differential operator.
Now focus again on the asymptotic bias of $\hat{m}_{ANW}(x)$:

$$
\begin{aligned}
Bias(\hat{m}_{ANW}(x))^2 &= \frac{h^4\mu_2^2(K)}{4}\left(\bar{h}^2(x)m''(x) + 2\frac{m'(x)(f_X(x)\bar{h}^2(x))'}{f_X(x)}\right)^2 \\
&= \frac{h^4\mu_2^2(K)}{4f_X^2(x)}\left(\bar{h}^2(x)f_X(x)m''(x) + 2m'(x)(f_X(x)\bar{h}^2(x))'\right)^2 \\
&= \frac{h^4\mu_2^2(K)}{4f_X^2(x)}\left(\bar{h}^2(x)(2m'(x)f_X'(x) + m''(x)f_X(x)) + 2(\bar{h}^2(x))'m'(x)f_X(x)\right)^2.
\end{aligned}
$$

Our aim is now to choose $\bar{h}(x)$ such that the term inside the brackets vanishes. For this purpose, we suppose that $m' \neq 0$. The concerning term is an ordinary differential equation with variable coefficients of order one. We can solve this equation and get

$$
\tilde{h}_{opt}(x) = \begin{cases} C_1(f_X(x))^{-1/2}(m'(x))^{-1/4}, & \text{for } m'(x) > 0 \\ C_1(f_X(x))^{-1/2}(-m'(x))^{-1/4}, & \text{for } m'(x) < 0. \end{cases}
$$

Therefore, we can formulate the following corollary as a direct consequence of this choice of the bandwidth function:

Corollary 3.12. *By choosing $h(x) = h\bar{h}(x) = h_n\tilde{h}_{opt}(x)$, it follows that*

$$
ABIAS(\hat{m}_{ANW}(x)) = o(h_n^2), \text{ as } n \to \infty.
$$

This rate of convergence is generally reserved for kernels of higher orders. For simulation purposes, it is necessary to construct the optimal bandwidth $h_{opt}(x)$ via pilot estimators of the appearing unknown values, namely $m'(x)$ and $f_X(x)$. For nonparametric density estimation, where Abramson´s square root law yields an optimal choice of the bandwidth function $h(x)$ according to

$$
h(X_i) = hf_X^{-1/2}(X_i),
$$

Silverman (1986, Section 5.3.1) proposes a strategy for implementing this bandwidth based on a pilot estimate $\hat{f}_P(x)$ such that $\hat{f}_P(X_i) > 0$ for all $i = 1, ..., n$. For example, $\hat{f}_P(x)$ can be an ordinary density estimator with fixed bandwidth.
We propose a comparable strategy, but in our case, we have to use pilot estimates for two unknown quantities. Hence, choose a pilot estimate $\hat{f}_P(x)$ at first such that $\hat{f}_P(X_i) > 0$ for all $i = 1, ..., n$. Then, choose an estimator for the derivative of $m(x)$, for example

$$
\hat{m}'(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial x}W_{ni}(x)Y_i,
$$

where

$$W_{ni}(x) = \frac{\frac{1}{nh}K\left(\frac{x-X_i}{h}\right)}{\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)}$$

denotes the Nadaraya-Watson weight with fixed bandwidth. After that, estimate $\bar{h}_{opt}(x)$by

$$\hat{\bar{h}}_{opt}(x) := (C|\hat{m}'(x)|^{1/2}\hat{f}_P(x))^{-1/2},$$

where C denotes a real and positive constant. In performed simulations by Silverman (1986) as well as Terrell and Scott (1992), C was chosen as the geometric mean g of $\hat{f}_P(X_i)$:

$$\log(g) = \frac{1}{n}\sum_{i=1}^{n}\log(\hat{f}_P(X_i)).$$

Demir and Töktamis (2010) used various methods for choosing the constant C, for example a weighted average, the harmonic mean, and the arithmetic mean of $\hat{f}_P(X_i)$. As already mentioned, they did not focus on the asymptotic analysis and, hence, their simulations are based on Abramson´s square root law. In this section, we have seen that this square root law has to be adapted in regression models, such that the optimal bandwidth possesses a different form.

3.5 Adaptive Nadaraya-Watson like estimators for stochastic differential equations

After this excursion in the discrete time series context, we will now again focus on stochastic differential equations. Our aim is to adapt the findings of the previous section to the nonparametric drift estimation in diffusion models. We will only focus on ordinary diffusions without jumps. Hence, reconsider a stochastic process $X = (X_t)_{t\geq 0}$ fulfilling the stochastic differential equation

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t, \;\; X_0 \overset{\mathcal{D}}{=} \eta.$$

Again, we observe a high-frequency sample $\{X_{k\Delta}, k = 0, ..., n\}$ and aim to estimate the unknown drift function b. In analogy of the previous section, we define the adaptive version of the nonparametric drift estimator $\hat{b}(x)$ according to

$$\hat{b}_A(x) := \frac{\frac{1}{n}\sum_{i=0}^{n-1}\frac{1}{h(X_{i\Delta})}K\left(\frac{X_{i\Delta}-x}{h(X_{i\Delta})}\right)\frac{(X_{(i+1)\Delta}-X_{i\Delta})}{\Delta}}{\frac{1}{n}\sum_{i=0}^{n-1}\frac{1}{h(X_{i\Delta})}K\left(\frac{X_{i\Delta}-x}{h(X_{i\Delta})}\right)}.$$

We will not focus on the nonparametric estimation of the diffusion coefficient here, but an analogous adaptive version of this estimator is given by

$$\hat{\sigma}^2(x) := \frac{\frac{1}{n}\sum_{i=0}^{n-1} \frac{1}{h(X_{i\Delta})} K\left(\frac{X_{i\Delta}-x}{h(X_{i\Delta})}\right) \frac{(X_{(i+1)\Delta}-X_{i\Delta})^2}{\Delta}}{\frac{1}{n}\sum_{i=0}^{n-1} \frac{1}{h(X_{i\Delta})} K\left(\frac{X_{i\Delta}-x}{h(X_{i\Delta})}\right)}.$$

We will specify the asymptotic properties of the adaptive drift estimator.

Lemma 3.13. *Under assumptions A1, A2 iii), B1 and B2 i), provided that $\pi(x) > 0$, $\xi \equiv 0$, and $b, \sigma, \pi \in \mathcal{C}^2(\mathbb{R})$, we find that*

$$i.)\ ABias(\hat{b}_A(x)) = \frac{\mu_2(K)}{2\pi(x)}\Big(h^2(x)(b''(x)\pi(x) + 2b'(x)\pi'(x)) + 2(h^2(x))'b'(x)\pi(x)\Big),$$

$$ii.)\ AVar(\hat{b}_A(x)) = \frac{\sigma^2(x)||K||_2^2}{n\Delta h(x)\pi(x)},\ \textit{as } n \to \infty.$$

In particular, the consistency of $\hat{b}_A(x)$ can be deduced by this representation of the mean squared error.

Corollary 3.14. *Under the assumptions of Lemma 3.13, $\hat{b}_A(x)$ is a pointwise weak consistent estimator of $b(x)$.*

4 Boundary bias correction methods

In the following chapter, we will face a problem occurring in nonparametric estimation approaches, which are based on symmetrical kernels. We will start with an intuitive illustration of the so-called “boundary bias” problem and will introduce some methods avoiding the bias near endpoints of the support of the underlying stationary density corresponding to the considered stochastic process X. Our main interest will lie in the exploration of certain properties of the class of asymmetric kernel estimators, which were originally proposed by Chen (1999, 2000). After introducing this approach in the context of density estimation, we will also focus on the regression case. Moreover, we show how to use asymmetric kernels in the context of estimation of diffusion models and, finally, we will leave the one-dimensional setting and focus on multivariate diffusions. We will present a non-negative multiplicative bias correction method for the estimation of multivariate densities and use this method afterwards to construct nonparametric estimators for the drift vector of a multivariate diffusion process.

4.1 Boundary bias of nonparametric kernel estimators

In this section, we will introduce the boundary bias problem of nonparametric kernel based estimators. For illustration purposes, we will focus on a discrete sample $\{X_i\}_{i=1,...,n}$ of independent and identically distributed random variables with density f. Moreover, we assume that the support of f fulfills

$$\operatorname{supp}(f) = [0, \infty)$$

and, hence,

$$X_i \geq 0, \text{ a.s. for every } i = 1, ..., n.$$

We suppose that f is unknown and twice continuously differentiable and we want to estimate it at a certain point $x \in \mathbb{R}^+$ by the use of the ordinary kernel based estimator

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i),$$

where h denotes the bandwidth again and $K_h(x) = \frac{1}{h} K(x/h)$ is the already introduced symmetric kernel function. For technical reasons, we assume here that K has bounded support $[-1, 1]$. For the sake of our estimation, we will distinguish between two cases, namely points “near” the origin and those lying in the interior region of the support of f. To concretize this idea, we state the following definition.

Definition 4.1. *A point $x \in \mathbb{R}^+$ is called an interior point, if $x > h$. In contrast, it is called a boundary point, if it exists a $\rho \in [0, 1]$, such that $x = \rho h$.*

We are now ready to determine the bias of $\hat{f}(x)$ under our used assumptions.

$$\begin{aligned} E[\hat{f}(x)] &= \int_0^1 K_h(x-y)f(y)dy = -\int_{x/h}^{(x-1)/h} K(u)f(x-uh)du \\ &= \int_{(x-1)/h}^{x/h} K(u)\left(f(x) - uhf'(x) + \frac{(uh)^2}{2}f''(x)\right)du + o(h^2) \text{ as } n \to \infty. \end{aligned}$$

If x lies in the interior region, then

$$x/h > 1 \text{ as well as } (x-1)/h < -1$$

for n large enough and we get the usual result that

$$E[\hat{f}(x)] = f(x) + O(h^2), \text{ as } n \to \infty$$

and, hence, the boundedness of supp(f) does not influence the rate of convergence. But, if, in contrast, x lies in the boundary region, then

$$x = \rho h \leq 1 - h$$

for n large enough and thus

$$\begin{aligned} E[\hat{f}(x)] &= \int_{(x-1)/h}^{x/h} K(u)\left(f(x) - uhf'(x) + \frac{(uh)^2}{2}f''(x)\right)du + o(h^2) \\ &= f(x)\int_{-1}^{\rho} K(u)du - hf'(x)\int_{-1}^{\rho} uK(u)du + \frac{f''(x)h^2}{2}\int_{-1}^{\rho} K(u)u^2du + o(h^2), \end{aligned}$$

as $n \to \infty$.
For the variance of $\hat{f}(x)$ we get

$$Var(\hat{f}(x)) = \frac{f(x)}{nh}\int_{-1}^{\rho} K^2(t)dt + o((nh)^{-1}) \text{ as } n \to \infty$$

by the use of an analogous argumentation. We conclude that the rate of the variance is not influenced through the bounded support of f, whereas the bias is highly affected. For example, we get that

$$E[\hat{f}(0)] = f(0)/2 + O(h), \text{ as } n \to \infty,$$

by the symmetry of K. Provided that $f(0) \neq 0$, we see that $\hat{f}$ produces a bias near the origin, which is not even asymptotically vanishing. In general, we see that the true density is underestimated inside this region. There are many approaches to avoid this effect, for example:

- the reflection method; see Cline and Hart (1991) and Schuster (1985) among others

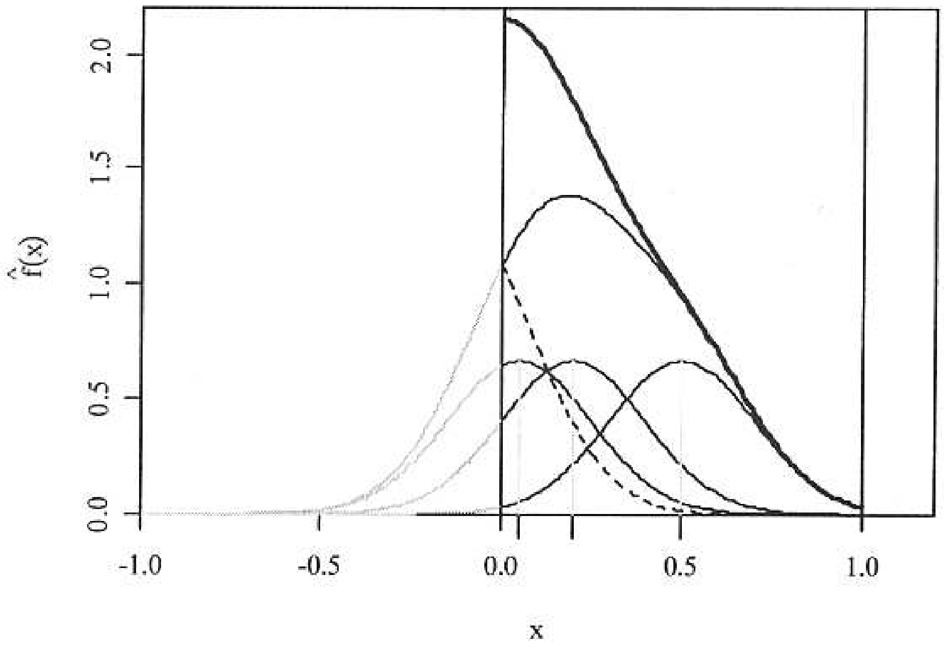

Figure 3: Plot of the true density in a dark solid line and the resulting kernel estimator based on three observations in a dark thin line. The dashed line denotes the reflected estimated density. Source: Schmid and Trede (2006): Finanzmarktstatistik, p. 106.

- the use of boundary kernels; see Jones (1993) and Zhang and Karunamuni (2000) among others
- a transformation method; see Karunamuni and Alberts (2005)
- a pseudo-data method; see Cowling and Hall (1996)
- convolution power based estimators; see Comte and Genon-Catalot (2012).

An approach which has attracted some attention in recent literature is the one by Chen, firstly used for the estimation of compact supported densities in 1999 and for the estimation of positive supported densities in 2000. We will now introduce his ideas. Consider therefore a random sample $\{X_i\}_{i=1,\ldots,n}$ of independent and identically distributed random variables and assume that every X_i has a density f with respect to the Lebesgue measure, which is supported on the positive real line. Consider now the density $G(p,\gamma)$ of a Gamma-distributed random variable with parameter p and γ, which is defined as

$$G(p,\gamma)(u) = \frac{u^{p-1}e^{-u/\gamma}}{\gamma^p\Gamma(p)} \cdot 1_{\{u\geq 0\}},\ p,\gamma > 0.$$

Now choose $p = x/h + 1$ and $\gamma = h$ such that

$$G(x/h+1, h)(u) = \frac{u^{x/h}e^{-u/h}}{h^{x/h+1}\Gamma(x/h+1)},\ u \geq 0.$$

Suppose again that f is twice continuously differentiable. Chen (2000) has now proposed the use of

$$\hat{f}_G(x) = n^{-1}\sum_{i=1}^{n} G(x/h+1, h)(X_i),$$

where h denotes the bandwidth such that $h \to 0$ and $nh \to \infty$ as $n \to \infty$. For illustration purposes, we will now shortly explore the explicit forms of the pointwise bias and variance of this estimator. The crucial observation is the following basic identity:

$$E[\hat{f}_G(x)] = E[G(x/h+1, h)(X_1)] = \int_0^\infty G(x/h+1, h)(y)f(y)dy = E[f(\xi_x)],$$

where $\xi_x \overset{\mathcal{D}}{=} G(x/h+1, h)$. Using the well-known moment properties of the Gamma distribution, we see that

$$E[\xi_x] = (x/h+1)h = x + h,\ Var(\xi_x) = (x/h+1)h^2 = xh + h^2.$$

By assuming that f is twice continuously differentiable and by using a Taylor expansion, we find out that the bias can be derived as

$$\begin{aligned}
&E[\hat{f}_G(x)] = E[f(\xi_x)] = E[f(E[\xi_x] + \xi_x - E[\xi_x])] \\
&= f(E[\xi_x]) + \frac{1}{2}f''(x)Var(\xi_x) + o(h),\ \text{as } n \to \infty \\
&= f(x) + h(f'(x) + \frac{1}{2}xf''(x)) + o(h) \to f(x)\ ,\quad \text{as } n \to \infty.
\end{aligned}$$

The bias of $\hat{f}_G(x)$ is independent of the location of x of order $O(h)$ as $n \to \infty$. The variance can be established in the following way; see Chen (2000):

$$Var(\hat{f}_G(x)) = n^{-1}Var(G(x/h+1, h)(X_1)) = n^{-1}E[G^2(x/h+1, h)(X_1)] + O(n^{-1}).$$

Now let $\eta_x \sim G(2x/h+1, h)$ and $B_h(x) := \frac{h^{-1}\Gamma(2x/h+1)}{2^{2x/h+1}\Gamma^2(x/h+1)}$, then it holds that

$$E[G^2(x/h+1, h)(X_1)] = B_h(x)E[f(\eta_x)].$$

Moreover, let $R(z) := \sqrt{2\pi}e^{-z}z^{z+1/2}/\Gamma(z+1)$ for $z \geq 0$. Hence

$$hB_h(x) = \frac{h^{1/2}x^{-1/2}R^2(x/h)}{2\sqrt{\pi}R(2x/h)}.$$

Due to Brown and Chen (1999), we get by monotonicity arguments that

$$B_h(x) \doteq \begin{cases} \frac{h^{-1/2}x^{-1/2}}{2\sqrt{\pi}}, & \text{as } x/h \to \infty \\ \frac{\Gamma(2\kappa+1)h^{-1}}{2^{1+2\kappa}\Gamma^2(\kappa+1)}, & \text{as } x/h \to \kappa > 0. \end{cases}$$

Therefore, we finally get that

$$Var(\hat{f}_G(x)) \doteq \begin{cases} \frac{n^{-1}h^{-1/2}x^{-1/2}f(x)}{2\sqrt{\pi}}, & \text{as } x/h \to \infty \\ \frac{\Gamma(2\kappa+1)h^{-1}n^{-1}f(x)}{2^{1+2\kappa}\Gamma^2(\kappa+1)}, & \text{as } x/h \to \kappa > 0. \end{cases}$$

Now we can summarize these results and state the following lemma

Lemma 4.2. *[Chen (2000), Section 2 and 3] Let $h = h_n$ be a positive sequence such that $h \to 0,\ nh \to \infty$ as $n \to \infty$. Moreover, let f be an unknown probability density, which is twice continuously differentiable as well as supported on the positive real line. Then,*

$$MSE(\hat{f}_G(x)) = \begin{cases} h^2\left(f'(x) + \frac{xf''(x)}{2}\right)^2 + \frac{f(x)}{nh^{1/2}x^{1/2}2\sqrt{\pi}} & , \textit{ as } x/h \to \infty \\ h^2\left(f'(x) + \frac{xf''(x)}{2}\right)^2 + \frac{\Gamma(2\kappa+1)f(x)}{nh2^{1+2\kappa}\Gamma^2(\kappa+1)} & , \textit{ as } x/h \to \kappa > 0. \end{cases}$$

Corollary 4.3. *Under the assumptions of Lemma (4.2), $\hat{f}_G(x)$ is a weak consistent estimator of $f(x)$ for $x \in \mathbb{R}^+$.*

We see that the bias of the Gamma kernel estimator is of order $O(h)$, which is slower compared to the rate for the classical kernel estimator ($O(h^2)$). Nevertheless, this class of estimators possesses many attractive properties, which make it worth taking a closer look at them. At first, the smoothing changes in an adaptive manner, which means that for each x, the shape of the kernel function varies and, therefore, an individual smoothing takes place. This opposes the classical case, where the amount of smoothing does not change. Moreover, one can show that $\hat{f}_G(x)$ reaches the optimal rate of convergence of the mean integrated squared error in the class of kernels of order two. Finally, the variance is inversely proportional to the location of the design point x and, hence, as x moves far from the origin, the variance shrinks.

Asymmetric kernel estimators are not restricted to the Gamma distribution. Although this concept is quite new in the literature, there are several choices of proper distributions available, like the use of Modified Gamma kernels (Chen (2000)), Log Normal and Birnbaum Saunders kernels (Jin and Kawczak (2003)), as well as Inverse Gaussian and Reciprocal Inverse Gaussian kernels (Scaillet (2004)). The exploration of asymptotic properties of all the mentioned choices rely on distribution-specific properties like moment relations and are, hence, diversified approaches. A relatively new paper by Hirukawa and Sakudo (2015) provides a remedy for this aspect. They introduce a family of asymmetric

Table 1: Univariate Asymmetric Kernels

Kernel	**Explicit Form**
G	$K_{G,h,x}(z) = \frac{z^{x/h}\exp(-z/h)}{h^{x/h+1}\Gamma(x/h+1)}$
MG	$K_{MG,h,x}(z) = \frac{z^{\rho_h(x)-1}\exp(-z/h)}{h^{\rho_h(x)}\Gamma(\rho_h(x))}$,
	where $\rho_h(x) = \frac{x}{h}1(x \geq 2h) + (\frac{x^2}{4h^2} + 1)1(x < 2h)$
IG	$K_{IG,h,x}(z) = \frac{1}{\sqrt{2\pi h z^3}} \exp\left[-\frac{1}{2hx}\left(z/x - 2 + x/z\right)\right]$
RIG	$K_{RIG,h,x}(z) = \frac{1}{\sqrt{2\pi h z}} \exp\left[-\frac{(x-h)}{2h}\left(z/(x-h) - 2 + (x-h)/z\right)\right]$
LN	$K_{LN,h,x}(z) = \frac{1}{z\sqrt{2\pi h z}} \exp\left(-\frac{(\log(z)-\log(x))^2}{2h}\right)$
BS	$K_{BS,h,x}(z) = \frac{1}{2x\sqrt{2\pi h}}\left((x/z)^{1/2} + (x/z)^{3/2}\right)$
	$\cdot \exp\left(-\frac{1}{2h}(z/x - 2 + x/z)\right)$
NM	$K_{NM,h,x}(z) = \frac{2z^{\alpha-1}\exp\left(-(z/(\beta\Gamma(\alpha/2)/\Gamma((\alpha+1)/2)))^2\right)}{(\beta\Gamma(\alpha/2)/\Gamma((\alpha+1)/2))^{\alpha}\Gamma(\alpha/2)}$,
	where $(\alpha,\beta) = \begin{cases} \left(\frac{x}{h}, x\right), & \text{for } x \geq 2h \\ \left(\frac{x^2}{4h^2}+1, \frac{x^2}{4h}+h\right), & \text{for } x < 2h \end{cases}$
B	$K_{B,h,x}(z) = \frac{z^{x/h}(1-z)^{(1-x)/h}}{B(x/h+1,(1-x)/h+1)}$

kernels for which bias and variance approximations can be employed by general assumptions on the whole set of possible kernel functions. Examples of distributions contained in this family are the Modified Gamma kernels (see also Chen (2000)) and the Nakagami-m-kernels. For an overview see Table 1, where several choices of possible kernel functions are displayed. Furthermore, in the context of estimating a density with compact support, Chen (1999) suggested the use of Beta-kernels. We will return to this class of estimators later on, when we focus on nonparametric estimation of copula densities, which are naturally supported on the unit hyper cube $[0,1]^d,\ d \geq 2$.

Referring to our context, it is mentionable that asymmetric kernels were also used for the nonparametric estimation of the coefficients of an ordinary diffusion process driven by a Brownian motion; see Gospodinov and Hirukawa (2012). After introducing two multiplicative bias corrected density estimators in the following section, we will propose the use of those estimators in the context of multivariate diffusion models.

4.2 Multivariate density estimation via asymmetric kernels

In this section we will at first introduce a product Gamma kernel based estimator for multivariate densities. Afterwards, we will present two non-negative multiplicative bias

correction ("MBC") methods, which improve the rate of convergence of the product kernel estimator. Finally, asymmetric kernels are used for the estimation of the drift of a multivariate diffusion process.
As we already mentioned, the boundary bias problem of classical symmetric kernel estimators has been dealt within the literature by several works. Although this problem plays even a more significant role in higher dimensions, due to the fact that the boundary region increases as the dimension does, there are only a few papers which provide boundary correction methods for higher dimensions. In the context of density estimation, we are only aware of Müller and Stadtmüller (1999) who extended the boundary kernel approach from the univariate to the multivariate case and Bouezmarni and Rombouts (2010) who investigated an asymmetric product kernel based estimator. We will now focus on the latter one and state their main results and will afterwards introduce a bias corrected version of this estimator. The following explanations are mainly listed in Funke and Kawka (2015), where the already mentioned MBC technique for multivariate density estimation was introduced.
Given a random sample $\{(X_{i1}, ..., X_{id})\}_{i=1,...,n}$ of independent and identically distributed random vectors with positively supported marginals, Bouezmarni and Rombouts (2010) estimate the unknown joint density function f using a product kernel estimator

$$\hat{f}_{m,h}(x) = \hat{f}_{m,h}(x_1, ..., x_d) := \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{d} K_{m,h_j,x_j}(X_{ij}),$$

where $h := (h_1, ..., h_d)$ denotes the vector of bandwidths and m= G or MG.
Based on this approach, they derived asymptotic approximations of bias and variance and, moreover, determined the optimal rate of convergence of the asymptotic mean integrated squared error ("AMISE"). For the sake of simplicity we cite here Theorem 1 of Bouezmarni and Rombouts (2010) only for the case

$$x_j/h_j \to \infty \text{ as } n \to \infty \text{ for all } j = 1, ..., d.$$

Theorem 4.4 (Bouezmarni and Rombouts (2010), Theorem 1, p.141)**.** *Let f be the joint density of the random vector $(X_1, ..., X_d)$ which is twice continuously differentiable in all components such that*

$$\int_{\mathbb{R}^d} \left(\sum_{j=1}^{d} \frac{\partial f(x)}{\partial x_j} \right)^2 dx < \infty$$

and

$$\int_{\mathbb{R}^d} \left(\sum_{j=1}^{d} x_j \frac{\partial^2 f(x)}{\partial^2 x_j} \right)^2 dx < \infty$$

Let the bandwidths h_j, $j = 1, ..., d$, fulfill

$$h_j \to 0 \text{ and } n^{-1} \prod_{j=1}^{d} h_j^{-1/2} \to 0 \text{ as } n \to \infty.$$

Then the optimal bandwidths, which minimize the AIMSE, are given by

$$h_{j,opt} := c_j n^{-2/(d+4)}.$$

Furthermore, the optimal $AIMSE_{opt}$ can be in this case decomposed into

$$AIMSE_{opt} = \left(\int_{\mathbb{R}^d} \left(\sum_{j=1}^{d} c_j B_j(x) \right)^2 dx + \prod_{j=1}^{d} c_j^{-1/2} \int_{\mathbb{R}^d} V(x) dx \right) n^{-4/(d+4)},$$

where

$$B_j(x) := \frac{\partial f(x)}{\partial x_j} + \frac{x_j}{2} \frac{\partial^2 f(x)}{\partial^2 x_j}$$

and

$$V(x) := \left(2\sqrt{\pi}\right)^{-d} f(x) \prod_{j=1}^{d} x_j^{-1/2}.$$

We see, that the optimal AIMSE is of order $O(n^{-4/(d+4)})$, which is the same rate as for the classical symmetric kernel based estimator; see for example Silverman (1986). In contrast to this approach, the Gamma kernel estimator avoids the boundary problem and does not place weights outside the support of f.
Moreover, we recognize the well-known "curse of dimensionality": the rate of convergence gets slower, when the dimension of the random vector X increases. Further, it can be shown that the variance in the interior region is smaller than in the boundary region; see the Appendix in Bouezmarni and Rombouts (2010).

4.3 MBC techniques for multivariate density estimation

In this section, we will now introduce two MBC techniques, originally proposed by Jones et al. (1995) and Terrell and Scott (1980) for univariate kernel density estimation based on symmetric kernels. In the context of asymmetric kernel estimation, both techniques were extended inside the univariate setting to the use of Beta kernels (Hirukawa (2010)) as well as Gamma and Generalized Gamma kernels (Hirukawa and Sakudo (2014, 2015)). We will now extend the results to the multivariate case and start with the formulation of the multivariate Jones, Linton and Nielsen ("MV-JLN") estimator as well as the multivariate Terrell Scott estimator ("MV-TS"). Consider therefore a random sample

$\{X_{i1},...,X_{id}\}_{i=1,...,n}$ of independent and identically distributed random vectors with joint density f. The MBC technique proposed by Jones et al. (1995) is based on the identity

$$f(x) = \hat{f}_m(x) \left(\frac{f(x)}{\hat{f}_m(x)} \right).$$

In this sense, the unknown density f at a vector $x = (x_1,...,x_d) \in (\mathbb{R}^+)^d$ is estimated via

$$\hat{f}_{JLN,m,h}(x) := \hat{f}_{m,h}(x) \frac{1}{n} \sum_{i=1}^{n} \frac{\prod_{j=1}^{d} K_{m,h_j,x_j}(X_{ij})}{\hat{f}_{m,h}(X_i)},$$

where $X_i := (X_{i1},...,X_{id})$ and $m =$ G, MG, IG, RIG, LN, BS, NM or B; see Table 1. The second MBC technique, published by Terrell and Scott (1980), is based on a geometric extrapolation originating from numerical mathematics. Hence, let $c_d \equiv c \in (0,1)$ be a constant depending on the dimension d. The unknown density function f is now estimated by

$$\hat{f}_{TS,m,h}(x) := \left(\hat{f}_{m,h}(x)\right)^{\frac{1}{1-c}} \left(\hat{f}_{m,h/c}(x)\right)^{-\frac{c}{1-c}}.$$

In contrast to other boundary correction methods like boundary kernels or local linear estimators, both estimators only produce positive values for f.
We will now develop the asymptotic properties of our estimators. Due to notational limitations and for the sake of brevity, we will only focus on the case where $h_1 = ... = h_d \equiv h$ and also where all components of the vector x lie in the interior region or at the boundary of the support. Results for more involved locations as well as unequal bandwidths can be examined in a straight forward manner. In contrast to Bouezmarni and Rombouts (2010), we have to strengthen the assumptions on the unknown joint density f. Roughly speaking, f has to be four times continuously differentiable in each component. Hence, we assume that

Assumption C1

i) f has four continuous and bounded partial derivatives.

ii) The bandwidth h fulfills $h \to 0$ as well as $nh^{d(r_j+1/2)+2} \to 0$ as $n \to \infty$, where

$$r_j = \begin{cases} 1/2, & \text{for } j = \text{ G, MG, RIG, NM, B} \\ 1, & \text{for } j = \text{ LN, BS,} \\ 3/2, & \text{for } j = \text{ IG.} \end{cases}$$

Assumption C1, i) is usually needed for estimation via fourth order product kernels. Assumption C2, ii) is used to control the convergence order of remainder terms appearing in certain Taylor expansions within the proofs. Now state the bias and variance approximations of both MBC methods dependent on the location of x.

Theorem 4.5. *Let $x := (x_1, ..., x_d)$ be a given design point such that $f(x) > 0$. Under Assumptions 1 and 2, the bias of the MV-TS MBC estimator is given by*

$$Bias\left(\hat{f}_{TS,m,h}(x)\right) = \frac{h^2}{c}\left[\frac{a_{1,m}^2(x; f)}{f(x)} - a_{2,m}(x; f)\right] + o(h^2) := h^2\frac{p_m(x; f)}{c} + o(h^2),$$

where $a_{j,m}(x; f) := a_{j,m}(x_1, ..., x_d; f)$, $j = 1, 2$ are functions depending on the choice of the kernel and their explicit forms are given in Tables 2 and 3. The variance can be approximated by

$$Var\left(\hat{f}_{TS,m,h}(x)\right) = \begin{cases} n^{-1}h^{-d/2}f(x)\nu_m(x)\lambda_d(c) + o(n^{-1}h^{-d/2}), & \text{for an interior vector } x, \\ O\left(\left(nh^{d(r_j+1/2)}\right)^{-1}\right), & \text{for a boundary vector } x, \end{cases}$$

where

$$\lambda_d(c) := \frac{(1+c)^{d/2}\left(1 + c^{(d+4)/2}\right) - (2c)^{(d+2)/2}}{(1-c)^2(1+c)^{d/2}}$$

and

$$\nu_m(x) := \nu_m(x_1, ..., x_d) = (2\sqrt{\pi})^{-d}\prod_{j=1}^{d} x_j^{-r_j}.$$

Theorem 4.6. *Let x be a given design point such that $f(x) > 0$. Under Assumptions 1 and 2, the bias of the MV-JLN MBC estimator is given by*

$$Bias\left(\hat{f}_{JLN,m,h}(x)\right) = -h^2 f(x)a_{1,m}(x; g) + o(h^2) := q_m(x; f)h^2 + o(h^2),$$

where $a_{1,m}(x; g)$ is defined by replacing f in the definition of $a_{1,m}(x; f)$ by

$$g := g(x; f) := \frac{a_{1,m}(x; f)}{f(x)}.$$

Furthermore, the variance can be approximated by

$$Var\left(\hat{f}_{JLN,m,h}(x)\right) = \begin{cases} n^{-1}h^{-d/2}f(x)\nu_m(x) + o(n^{-1}h^{-d/2}), & \text{for an interior vector } x, \\ O\left(\left(nh^{d(r_j+1/2)}\right)^{-1}\right), & \text{for a boundary vector } x. \end{cases}$$

The rate of the mean squared error ("MSE") can now easily be derived using Theorems 4.5 and 4.6. Therefore, let $x = (x_1, ..., x_d)$ be a vector of interior points, then:

$$\text{MSE}\left(\hat{f}_{TS,m,h}(x)\right) = \frac{h^4 p_m^2(x; f)}{c^2} + n^{-1}h^{-d/2}f(x)\nu_m(x)\lambda_d(c) + o(h^4 + n^{-1}h^{-d/2})$$

as well as

$$\text{MSE}\left(\hat{f}_{JLN,m,h}(x)\right) = h^4 q_m^2(x; f) + n^{-1}h^{-d/2}f(x)\nu_m(x) + o(h^4 + n^{-1}h^{-d/2}).$$

Table 2: Explicit formulas for $a_{1,m}(x;f)$

Kernel	$a_{1,m}(x;f)$
G	$\sum_{j=1}^{d}\left(\frac{\partial f(x)}{\partial x_j}+\frac{x_j}{2}\frac{\partial^2 f(x)}{\partial^2 x_j}\right)$
MG	$\sum_{j=1}^{d}\left(\frac{x_j}{2}\frac{\partial^2 f(x)}{\partial^2 x_j}1(x_j>2h)+\xi_h(x_j)\frac{\partial f(x)}{\partial x_j}1(x_j\leq 2h)\right)$, where $\xi_h(x_j):=\rho_h(x_j)-\frac{x_j}{h}=O(1)$
IG	$\frac{1}{2}\sum_{j=1}^{d}x_j^3\frac{\partial^2 f(x)}{\partial^2 x_j}$
RIG	$\frac{1}{2}\sum_{j=1}^{d}x_j\frac{\partial^2 f(x)}{\partial^2 x_j}$
LN	$\frac{1}{2}\sum_{j=1}^{d}x_j\left(\frac{\partial f(x)}{\partial x_j}+x_j\frac{\partial^2 f(x)}{\partial^2 x_j}\right)$
BS	$\frac{1}{2}\sum_{j=1}^{d}x_j\left(\frac{\partial f(x)}{\partial x_j}+\frac{\partial^2 f(x)}{\partial^2 x_j}\right)$
NM	$\sum_{j=1}^{d}\left(\frac{x_j}{4}\frac{\partial^2 f(x)}{\partial^2 x_j}1(x_j>2h)+\xi_h(x_j)\frac{\partial f(x)}{\partial x_j}1(x_j\leq 2h)\right)$,
B	$\sum_{j=1}^{d}\left((1-2x_j)\frac{\partial f(x)}{\partial x_j}+\frac{1}{2}x_j(1-x_j)\frac{\partial^2 f(x)}{\partial^2 x_j}\right)$

This yields to the optimal smoothing parameters

$$h_{opt,TS,m}=n^{-2/(8+d)}\left(\frac{f(x)\nu_m(x)\lambda_d(c)c^2 d}{8p_m^2(x;f)}\right)^{2/(8+d)} \tag{4.13}$$

and

$$h_{opt,JLN,m}=n^{-2/(8+d)}\left(\frac{f(x)\nu_m(x)d}{8q_m^2(x;f)}\right)^{2/(8+d)}. \tag{4.14}$$

These parameters lead us now to the optimal rate of convergence of the MSE for an interior vector x:

$$MSE_{opt}\left(\hat{f}_{TS,m,h}(x)\right)\sim n^{-8/(8+d)}\frac{(d+8)}{(8^{8/d}d)^{d/(8+d)}}\gamma_d(c)p_m^{2d/(8+d)}(x;f)\left(\nu_m(x)f(x)\right)^{8/(8+d)}$$

and

$$MSE_{opt}\left(\hat{f}_{JLN,m,h}(x)\right)\sim n^{-8/(8+d)}\frac{(d+8)}{(8^{8/d}d)^{d/(8+d)}}q_m^{2d/(8+d)}(x;f)\left(\nu_m(x)f(x)\right)^{8/(8+d)},$$

where

$$\gamma_d(c):=\frac{\lambda_d(c)^{8/(8+d)}}{c^{2d/(8+d)}}=\left(\frac{(1+c)^{d/2}(1+c^{(4+d)/2})-(2c)^{(2+d)/2}}{c^{d/4}(1+c)^{d/2}(1-c)^2}\right)^{8/(8+d)}. \tag{4.15}$$

Table 3: Explicit formulas for $a_{2,m}(x;f)$

Kernel	$a_{2,m}(x;f)$
G	$\sum_{j=1}^{d}\left(\frac{\partial^2 f(x)}{\partial^2 x_j}+\frac{5}{6}x_j\frac{\partial^3 f(x)}{\partial^3 x_j}+\frac{1}{8}x_j\frac{\partial^4 f(x)}{\partial^4 x_j}\right)$
	$+\sum_{i\neq j}\left(\frac{\partial^2 f(x)}{\partial x_j\partial x_i}+\frac{x_i}{2}\frac{\partial^3 f(x)}{\partial^2 x_i\partial x_j}+\frac{x_i x_j}{4}\frac{\partial^4 f(x)}{\partial^2 x_i\partial^2 x_j}\right)$
MG	$\sum_{j=1}^{d}\left(\frac{x_j}{3}\frac{\partial^3 f(x)}{\partial^3 x_j}+\frac{x_j^2}{8}\frac{\partial^4 f(x)}{\partial^4 x_j}\right)1(x_j>2h)$
	$+\frac{1}{2}\sum_{j=1}^{d}\left(\xi_h^2(x_j)+\xi_h(x_j)+\frac{x_j}{h}\right)\frac{\partial^2 f(x)}{\partial^2 x_j}1(x_j\leq 2h)$
	$+\sum_{i\neq j}\frac{x_i x_j}{4}\frac{\partial^4 f(x)}{\partial^2 x_j\partial^2 x_i}1(x_i>2h;x_j>2h)$
	$+\sum_{i\neq j}\xi_h(x_j)\xi_h(x_i)\frac{\partial^2 f(x)}{\partial x_j\partial x_i}1(x_j\leq 2h;x_i\leq 2h)$
IG	$\frac{1}{2}\sum_{j=1}^{d}\left(x_j^5\frac{\partial^3 f(x)}{\partial^3 x_j}+\frac{x_j^6}{4}\frac{\partial^4 f(x)}{\partial^4 x_j}\right)+\sum_{i\neq j}\frac{x_i^3 x_j^3}{4}\frac{\partial^4 f(x)}{\partial^2 x_j\partial^2 x_i}$
RIG	$\frac{1}{2}\sum_{j=1}^{d}\left(\frac{\partial^2 f(x)}{\partial^2 x_j}+x_j\frac{\partial^3 f(x)}{\partial^3 x_j}+\frac{x_j^2}{4}\frac{\partial^4 f(x)}{\partial^4 x_j}\right)+\sum_{i\neq j}\frac{x_i x_j}{4}\frac{\partial^4 f(x)}{\partial^2 x_j\partial^2 x_i}$
LN	$\sum_{j=1}^{d}\left(\frac{x_j}{8}\frac{\partial f(x)}{\partial x_j}+\frac{7x_j}{8}\frac{\partial^2 f(x)}{\partial^2 x_j}+\frac{3x_j^3}{4}\frac{\partial^3 f(x)}{\partial^3 x_j}+\frac{x_j^4}{8}\frac{\partial^4 f(x)}{\partial^4 x_j}\right)$
	$+\sum_{i\neq j}x_i x_j\left(\frac{\partial^2 f(x)}{\partial x_j\partial x_i}+2x_i\frac{\partial^3 f(x)}{\partial^2 x_i\partial x_j}+\frac{x_i x_j}{4}\frac{\partial^4 f(x)}{\partial^2 x_i\partial^2 x_j}\right)$
BS	$\sum_{j=1}^{d}\left(\frac{3x_j^2}{4}\frac{\partial^2 f(x)}{\partial^2 x_j}+\frac{3x_j^3}{4}\frac{\partial^3 f(x)}{\partial^3 x_j}+\frac{x_j^4}{8}\frac{\partial^4 f(x)}{\partial^4 x_j}\right)$
	$+\sum_{i\neq j}x_i x_j\left(\frac{\partial^2 f(x)}{\partial x_j\partial x_i}+\frac{x_i x_j}{4}\frac{\partial^4 f(x)}{\partial^2 x_i\partial^2 x_j}\right)+\frac{x_i^2}{2}\frac{\partial^3 f(x)}{\partial^2 x_i\partial x_j}$
NM	$\frac{1}{8}\sum_{j=1}^{d}\left(\frac{\partial^2 f(x)}{\partial^2 x_j}+\frac{x_j}{3}\frac{\partial^3 f(x)}{\partial^3 x_j}+\frac{x_j^2}{4}\frac{\partial^4 f(x)}{\partial^4 x_j}\right)1(x_j>2h)$
	$+\sum_{i\neq j}\frac{x_i x_j}{16}\frac{\partial^4 f(x)}{\partial^2 x_j\partial^2 x_i}1(x_i>2h;x_j>2h)$
	$+\sum_{j=1}^{d}\left(\left(\xi_h(x_j)+\frac{x_j}{h}\right)^2\frac{\Gamma\left(\frac{\xi_h(x_j)+\frac{x_j}{h}}{2}\right)\Gamma\left(\frac{\xi_h(x_j)+\frac{x_j}{h}}{2}+1\right)}{\Gamma^2\left(\frac{\xi_h(x_j)+\frac{x_j}{h}+1}{2}\right)}\right.$
	$\left.-\frac{2x_j\xi_h(x_j)}{h}+\left(\frac{x_j}{h}\right)^2\right)1(x_j\leq 2h)$
	$+\sum_{i\neq j}\xi_h(x_j)\xi_h(x_i)\frac{\partial^2 f(x)}{\partial x_j\partial x_i}1(x_j\leq 2h;x_i\leq 2h)$
B	$\sum_{j=1}^{d}\left(-2(1-2x_j)\frac{\partial f(x)}{\partial x_j}+\frac{1}{2}(11x_j^2-11x_j+2)\frac{\partial^2 f(x)}{\partial^2 x_j}\right)$
	$+\sum_{j=1}^{d}\left(\frac{5}{6}x_j(1-x_j)(1-2x_j)\frac{\partial^3 f(x)}{\partial^3 x_j}+\frac{1}{8}x_j^2(1-x_j)^2\frac{\partial^4 f(x)}{\partial^4 x_j}\right)$
	$+\sum_{i\neq j}\left((1-2x_i)(1-2x_j)\frac{\partial^2 f(x)}{\partial x_j\partial x_i}+\frac{1}{2}(1-2x_i)x_j(1-x_j)\frac{\partial^3 f(x)}{\partial x_i\partial^2 x_j}\right)$
	$+\sum_{i\neq j}\frac{1}{4}x_i(1-x_i)x_j(1-x_j)\frac{\partial^4 f(x)}{\partial^2 x_i\partial^2 x_j}$

Remark 4.7. *Observe that Assumption 2 guarantees that the smoothing parameter h converges slower than $O(n^{-1/((r_j+1/2)d+2)})$. As we can see, the optimal bandwidth parameter h_{opt} is of order $O(n^{-2/(8+d)})$ for both estimators. This means that our requirement is fulfilled for the G, MG, RIG, NM and B kernel only in the case $d < 4$. For other kernels, h_{opt} does not fulfill this assumption. From a practical point of view, this should not be a major problem, due to the fact that the bandwidth in finite sample examples is often chosen by a data driven method and nonparametric density estimation suffers extremely from the curse of dimensionality in dimensions higher than 3.*

Similarly, we can derive the rates for boundary vectors. The MSE of both estimators is in this case of order $O(h^4 + (nh^{d(r_j+1/2)})^{-1})$ and the optimal bandwidth parameter h^*_{opt} consequently fulfills $h^*_{opt} = O(n^{-1/(4+d(r_j+1/2))})$, which leads to an optimal mean squared error for boundary vectors of order

$$MSE^* \left(\hat{f}_{JLN,m,h}(x)\right) = MSE^* \left(\hat{f}_{TS,m,h}(x)\right) = O(n^{-4/((4+d(r_j+1/2))})).$$

It is reasonable to include also a global performance criterion such as the mean integrated squared error ("MISE"). As suggested in Chen (2000), a trimming argument yields that the unwanted slower rates near the origin do not affect the global performance. One can easily use this argument in each direction due to the product form of the chosen kernel; see Bouezmarni and Rombouts (2010). Thus, we have the following rates for the MISEs of the proposed estimators:

$$MISE \left(\hat{f}_{TS,m,h}(x)\right) = \frac{h^4}{c^2} \int_0^\infty p_m^2(x)dx + \frac{\lambda_d(c)}{nh^{d/2}} \int_0^\infty f(x)\nu_m(x)dx + o(h^4 + n^{-1}h^{-d/2})$$

as well as

$$MISE \left(\hat{f}_{JLN,m,h}(x)\right) = h^4 \int_0^\infty q_m^2(x)dx + \frac{1}{nh^{d/2}} \int_0^\infty f(x)\nu_m(x)dx + o(h^4 + n^{-1}h^{-d/2}),$$

provided that all appearing integrals exist.

The optimal smoothing parameters for the MISE are, hence, given by

$$\bar{h}_{opt,TS,m} = n^{-2/(8+d)} \left(\frac{\lambda_d(c)c^2 d \int_0^\infty f(x)\nu_m(x)dx}{8 \int_0^\infty p_m^2(x;f)} \right)^{2/(8+d)}$$

and

$$\bar{h}_{opt,JLN,m} = n^{-2/(8+d)} \left(\frac{d \int_0^\infty f(x)\nu_m(x)dx}{8 \int_0^\infty q_m^2(x;f)dx} \right)^{2/(8+d)}.$$

Consequently, the optimal MISEs are of order

$$MISE_{opt}(\hat{f}_{TS,m,h}(x)) \\ \sim n^{-8/(8+d)} \frac{(d+8)}{(8^{8/d}d)^{d/(8+d)}} \gamma_d(c) \left(\int_0^\infty p_m^2(x;f)dx \right)^{2/(8+d)} \left(\int_0^\infty \nu_m(x)f(x)dx \right)^{8/(8+d)}$$

and

$$MISE_{opt}(\hat{f}_{JLN,m,h}(x))$$
$$\sim n^{-8/(8+d)} \frac{(d+8)}{(8^{8/d}d)^{d/(8+d)}} \left(\int_0^\infty q_m^2(x;f)dx \right)^{2/(8+d)} \left(\int_0^\infty \nu_m(x)f(x) \right)^{8/(8+d)} .$$

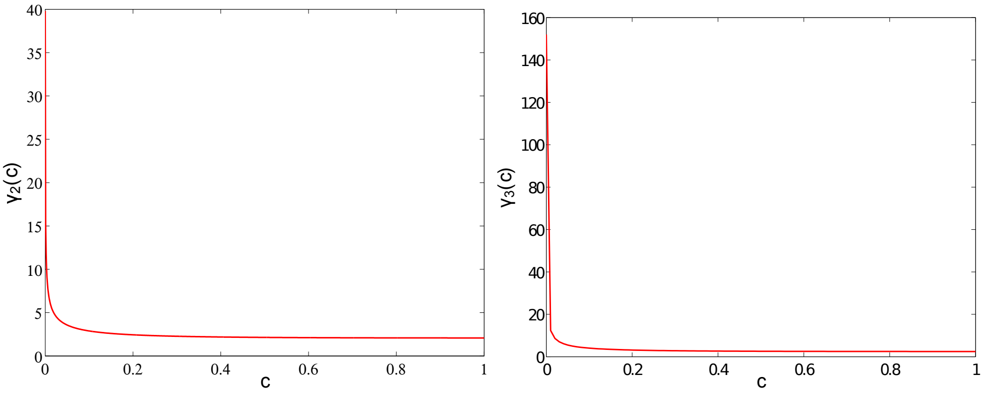

Figure 4: Plots of the functions $\gamma_d(c)$ (cf. (4.15)) for $c \in (0,1)$. Left γ_2, right γ_3.

Plots of the functions γ_2 and γ_3 are displayed in Figure 4. We see that γ_2 as well as γ_3 are strictly decreasing on the interval $(0,1]$. We decided to present both plots here to show that monotonicity property not only holds in the two-dimensional case. For the simulations in Funke and Kawka (2015), c was chosen according to $c = 0.7$, because there is not much difference between $\gamma_2(0.7) \approx 2.0974$ and $\lim_{c\to 1} \gamma_2(c) \approx 2.0814$.

Remark 4.8. *We can easily transmit the results when bias and variance have to be established at a generic vector $x \in (\mathbb{R}^d)^+$. As we have seen, the bias remains unchanged and is uniformly of order $O(h^2)$ over the whole support. Moreover, the variance exhibits the following order:*

$$Var\left(\hat{f}_{JLN,m,h}(x)\right) = Var\left(\hat{f}_{TS,m,h}(x)\right) = O\left(n^{-1}\prod_{l=1}^{d} h^{-(1/2+r_j 1_l)}\right),$$

which depends on the location of the components and, furthermore, where $1_l := 1(x_l/h \to \kappa_l > 0)$ is a function indicating whether a component lies in the boundary region.

Proof of Theorem 4.5. The proof largely follows Hirukawa (2010) and can also be found in Funke and Kawka (2015). For the sake of brevity, we will only focus on the proof for the use of Gamma kernels. Moreover, in Hirukawa and Sakudo (2012), the univariate case

is studied in more detail compared to Hirukawa and Sakudo (2014).
For the bias of the MV-TS estimator, we start with a multivariate Taylor expansion of f up to order 4. Therefore, let $Y_i,\ i=1,...,d$, be independent and Gamma-distributed random variables such that $Y_i \sim G(x_i/h+1,h)$ with mean $\mu_i = x_i + h$. Using the smoothness assumption on the unknown density f as well as the fact that

$$E[(Y_i-x_i)^r] = O(h^3), \text{ as } n \to \infty \text{ for } r \geq 5,$$

(see Lemma B2 in Gospodinov and Hirukawa (2007)), we can derive that

$$\begin{aligned}
&I_h(x) := E[\hat{f}_{G,h}(x)] \\
&= \int_{\mathbb{R}^+} \cdots \int_{\mathbb{R}^+} \prod_{j=1}^{d} K_{G,h,x_j}(y_j) f(y_1,...,y_d) dy_1...dy_d = E[f(Y_1,...,Y_d)] \\
&= f(x_1,...,x_d) + \sum_{j=1}^{d} E[Y_j - x_j] \frac{\partial f(x_1,...,x_d)}{\partial x_j} + \sum_{j=1}^{d} \frac{E[(Y_j-x_j)^2]}{2} \frac{\partial^2 f(x_1,...,x_d)}{\partial^2 x_j} \\
&\quad + \sum_{i\neq j} E[Y_i - x_i] E[Y_j - x_j] \frac{\partial^2 f(x_1,..,x_d)}{\partial x_i \partial x_j} + \sum_{j=1}^{d} \frac{E[(Y_j-x_j)^3]}{6} \frac{\partial^3 f(x_1,..,x_d)}{\partial^3 x_j} \\
&\quad + \sum_{i\neq j} \frac{E[(Y_j-x_j)]E[(Y_i-x_i)^2]}{2} \frac{\partial^3 f(x_1,..,x_d)}{\partial x_j \partial^2 x_i} + \sum_{j=1}^{d} \frac{E[(Y_j-x_j)^4]}{24} \frac{\partial^4 f(x_1,..,x_d)}{\partial^4 x_j} \\
&\quad + \sum_{i\neq j} \frac{E[(Y_j-x_j)^2]E[(Y_i-x_i)^2]}{4} \frac{\partial^4 f(x_1,..,x_d)}{\partial^2 x_j \partial^2 x_i} + o(h^2) \\
&= f(x_1,...,x_d) + h\left(\sum_{j=1}^{d} \left(\frac{\partial f(x_1,...,x_d)}{\partial x_j} + \frac{x_j}{2} \frac{\partial^2 f(x_1,...,x_d)}{\partial^2 x_j} \right) \right) \\
&\quad + h^2 \Bigg(\sum_{j=1}^{d} \left(\frac{\partial^2 f(x_1,...,x_d)}{\partial^2 x_j} + \frac{5x_j}{6} \frac{\partial^3 f(x_1,...,x_d)}{\partial^3 x_j} + \frac{x_j}{8} \frac{\partial^4 f(x_1,...,x_d)}{\partial^4 x_j} \right) \\
&\quad + \sum_{i\neq j} \left(\frac{\partial^2 f(x_1,..,x_d)}{\partial x_i \partial x_j} + \frac{x_i x_j}{4} \frac{\partial^4 f(x_1,...,x_d)}{\partial^2 x_j \partial^2 x_i} + \frac{x_i}{2} \frac{\partial^3 f(x_1,...,x_d)}{\partial^2 x_i \partial x_j} \right) \Bigg) + o(h^2) \\
&:= f(x_1,...,x_d) \left(1 + h \frac{a_{1,G}(x;f)}{f(x)} + h^2 \frac{a_{2,G}(x;f)}{f(x)} + o(h^2) \right),
\end{aligned}$$

where all appearing Landau symbols describe the asymptotical order as $n \to \infty$, which is also true for the following derivations in this proof.
We remark that the main difference between the univariate and the multivariate case is the appearance of cross derivatives. One important tool is the independence of the random variables $Y_j,\ j=1,...,d$, in order to derive the mixed moments. Moreover, not all families

of asymmetric kernels are exactly centered at x, i.e. the Gamma kernels. Hence, in such cases additional cross terms appear.
The terms $I_h(x)$ and $I_{h/c}(x)$ can be handled in an analogous manner. Thus, taking the logarithm on both sides and expanding the logarithm by a Taylor expansion yields

$$\begin{aligned}&\log(I_h(x)) =: \log(f(x)) + \log(1+\tilde{a}(x))\\&= \log(f(x)) + h\frac{a_{1,G}(x;f)}{f(x)} + h^2\frac{a_{2,G}(x;f)}{f(x)} - h^2\frac{a_{1,G}^2(x;f)}{2f^2(x)} + o(h^2)\\&= \log(f(x)) + h\frac{a_{1,G}(x;f)}{f(x)} + \frac{h^2}{f^2(x)}\left(f(x)a_{2,G}(x;f) - \frac{a_{1,G}^2(x;f)}{2}\right) + o(h^2).\end{aligned}$$

In addition, for $\log(I_{h/c}(x))$ we have that

$$\log(I_{h/c}(x)) = \log(f(x)) + \frac{ha_{1,G}(x;f)}{cf(x)} + \frac{h^2}{c^2f^2(x)}\left(a_{2,G}(x;f) - \frac{a_{1,G}^2(x;f)}{2}\right) + o(h^2)$$

and, consequently,

$$\begin{aligned}&\frac{1}{1-c}\log(I_h(x)) - \frac{c}{1-c}\log(I_{h/c}(x))\\&= \log(f(x)) - \frac{h^2}{c}\left(\frac{a_{2,G}(x;f)f(x) - a_{1,G}^2(x;f)/2}{f^2(x)}\right) + o(h^2).\end{aligned}$$

In a first order approximation of the exponential function, the following statement holds:

$$\begin{aligned}(I_h(x))^{\frac{1}{1-c}}\left(I_{h/c}(x)\right)^{\frac{-c}{1-c}} &= f(x)\left(1 + \frac{h^2}{c}\left(\frac{a_{1,G}^2(x;f)}{2f^2(x)} - \frac{a_{2,G}(x;f)}{f(x)}\right) + o(h^2)\right)\\&= f(x) + \frac{h^2}{c}\left(\frac{a_{1,G}(x;f)}{2f(x)} - a_{2,G}(x;f)\right) + o(h^2).\end{aligned}$$

Now let $Z := \hat{f}_{G,h}(x) - I_h(x)$ and $W := \hat{f}_{G,h/c}(x) - I_{h/c}(x)$. As we will see in the derivation of the variance, $E[Z^2]$, $E[W^2]$, and $E[ZW]$ are at most of order $O(n^{-1}b^{-d})$ as $n \to \infty$. Using C1, ii), we can deduce that

$$O(n^{-1}b^{-d}) = o(h^2).$$

By a first order Taylor expansion of the function $\tilde{f}(x) := (1+x)^a$, $a > 0$, around zero, we find that

$$\begin{aligned}&\hat{f}_{TS,G,h}(x) = (I_h(x))^{\frac{1}{1-c}}\left(1 + \frac{Z}{I_h(x)}\right)^{\frac{1}{1-c}}\left(I_{h/c}(x)\right)^{\frac{-c}{1-c}}\left(1 + \frac{W}{I_{h/c}(x)}\right)^{\frac{-c}{1-c}}\\&= (I_h(x))^{\frac{1}{1-c}}\left(I_{h/c}(x)\right)^{\frac{-c}{1-c}}\left(1 + \frac{Z}{(1-c)I_h(x)}\right)\left(1 - \frac{cW}{(1-c)I_{h/c}(x)}\right)\\&= (I_h(x))^{\frac{1}{1-c}}\left(I_{h/c}(x)\right)^{\frac{-c}{1-c}} + \frac{Z}{1-c}\left(\frac{I_h(x)}{I_{h/c}(x)}\right)^{\frac{c}{1-c}} - \frac{cW}{1-c}\left(\frac{I_h(x)}{I_{h/c}(x)}\right)^{\frac{1}{1-c}} + O(n^{-1}h^{-d}).\end{aligned} \tag{4.16}$$

Using the approximations of the expectations $I_h(x)$ and $I_{h/c}(x)$, we see that they are asymptotically equivalent in terms of their order, which means that

$$I_h(x) = I_{h/c}(x) + O(h).$$

Using this property together with Assumption C1, ii) as well as the fact that Z and W are centered, we conclude that

$$\begin{aligned} E[\hat{f}_{TS,G,h}(x)] &= (I_h(x))^{\frac{1}{1-c}} \left(I_{h/c}(x)\right)^{\frac{-c}{1-c}} + O(n^{-1}h^{-d}) \\ &= f(x) + \frac{h^2}{c(1-c)} \left(\frac{a_{1,G}^2(x;f)}{2f(x)} - a_{2,G}(x;f) \right) + o(h^2), \text{ as } n \to \infty. \end{aligned}$$

Making use of (4.16), the variance can be decomposed as follows:

$$\begin{aligned} &Var\left(\hat{f}_{TS,G,h}(x)\right) = E\left[\left(\frac{Z}{1-c} - \frac{cW}{1-c} \right)^2 \right] + O(n^{-1}) \\ &= \frac{1}{(1-c)^2} \left(Var(\hat{f}_{G,h}(x)) - 2cCov(\hat{f}_{G,h}(x), \hat{f}_{G,h/c}(x)) + c^2 Var(\hat{f}_{G,h/c}(x)) \right) + O(n^{-1}). \end{aligned}$$

We will restrict ourselves only to the case where x is a vector including interior components x_j for all $j = 1, ..., d$. Using the results in Bouezmarni and Rombouts (2010), we can conclude that

$$Var(\hat{f}_{G,h}(x)) = n^{-1}h^{-d/2}(2\sqrt{\pi})^{-d} f(x_1, ..., x_d) \prod_{j=1}^{d} x_j^{-1/2} + o(n^{-1}h^{-d/2}),$$

$$Var(\hat{f}_{G,h/c}(x)) = n^{-1}h^{-d/2}c^{d/2}(2\sqrt{\pi})^{-d} f(x_1, ..., x_d) \prod_{j=1}^{d} x_j^{-1/2} + o(n^{-1}h^{-d/2}),$$

and

$$\begin{aligned} Cov(\hat{f}_{G,h}(x), \hat{f}_{G,h/c}(x)) &= n^{-1} E\left[\prod_{j=1}^{d} K_{G,h,x_j}(X_{1j}) \prod_{k=1}^{d} K_{G,h/c,x_k}(X_{1k}) \right] + O(n^{-1}) \\ &= n^{-1} \prod_{j=1}^{d} C_h(x_j)(f(x_1, ..., x_d) + o(1)) + O(n^{-1}), \end{aligned}$$

where

$$C_h(x_j) := \frac{\Gamma\left(\frac{x_j(1+c)}{h} + 1\right) \left(\frac{h}{1+c}\right)^{\frac{x_j(1+c)}{h}+1}}{\Gamma\left(\frac{x_j}{h}+1\right) h^{\frac{x_j}{h}+1} \Gamma\left(\frac{cx_j}{h}+1\right) \left(\frac{h}{c}\right)^{\frac{cx_j}{h}+1}}.$$

Using equation (6) in Hirukawa (2010) and its adjacent derivations for the function $C_h(x_j)$, the covariance can be approximated by

$$Cov(\hat{f}_{G,h}(x), \hat{f}_{G,h/c}(x)) = n^{-1}h^{-d/2}\left(\frac{2c}{1+c}\right)^{d/2} f(x_1,...,x_d)\prod_{j=1}^{d} x_j^{-1/2} + o(n^{-1}h^{-d/2}).$$

Bringing all three parts of the variance of the MV-TS estimator together, we are finally able to deduce that

$$\begin{aligned} Var\left(\hat{f}_{TS,G,h}(x)\right) &\doteq \frac{1}{(1-c)^2}\left(n^{-1}h^{-d/2}\frac{f(x_1,...,x_d)}{(2\sqrt{\pi})^d\prod_{j=1}^{d} x_j^{1/2}}\left(1-2c\left(\frac{2c}{1+c}\right)^{d/2}+c^{2+d/2}\right)\right) \\ &= n^{-1}h^{-d/2}\frac{f(x_1,...,x_d)}{(2\sqrt{\pi})^d\prod_{j=1}^{d} x_j^{1/2}}\left(\frac{(1+c)^{d/2}(1+c^{(d+4)/2})-(2c)^{(d+2)/2}}{(1-c)^2(1+c)^{d/2}}\right) \\ &:= n^{-1}h^{-d/2}\frac{f(x_1,...,x_d)\lambda_d(c)}{(2\sqrt{\pi})^d\prod_{j=1}^{d} x_j^{1/2}}. \end{aligned}$$

□

Now we will state the proof of the bias and variance approximation for the MV-JLN estimator. Again, the proof is based on the results in Hirukawa (2010) and can be found in Funke and Kawka (2015).

Proof of Theorem 4.6. We will start with the derivation of the bias term. For this purpose, we decompose the considered estimator as follows:

$$\hat{f}_{JLN}(x) = \hat{f}_G(x)\hat{\alpha}(x) = f(x)\left(1+\frac{\hat{f}_G(x)-f(x)}{f(x)}\right)((\hat{\alpha}(x)-1)+1),$$

where

$$\hat{\alpha}(x) = \hat{\alpha}(x_1,...,x_d) := \frac{1}{n}\sum_{i=1}^{n}\frac{\prod_{j=1}^{d} K_{G,h,x_j,}(X_{ij})}{\hat{f}_G(X_i)}.$$

The expectation of the proposed estimator can therefore be decomposed as

$$\begin{aligned} &E[\hat{f}_{JLN}(x)] = f(x) \\ &\quad + f(x)\left(E\left[\frac{\hat{f}_G(x)-f(x)}{f(x)}\right] + E[(\hat{\alpha}(x)-1)]E\left[\frac{(\hat{f}_G(x)-f(x))}{f(x)}(\hat{\alpha}(x)-1)\right]\right) \\ &:= f(x) + f(x)(I + II + III). \end{aligned}$$

We will handle the terms inside the brackets separately and start with term II. Using a geometric expansion around $f(x)$, $\hat{\alpha}(x)$ can be approximated by the following:

$$\hat{\alpha}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{\prod_{j=1}^{d} K_{G,h,x_j}(X_{ij})}{f(X_i)}\Bigg(1 - \frac{\hat{f}_G(X_i) - f(X_i)}{f(X_i)} + \left(\frac{\hat{f}_G(X_i) - f(X_i)}{f(X_i)}\right)^2\Bigg) + o(h^2 + (nh^{d/2})^{-1}), \text{ as } n \to \infty.$$

Notice that the order of the remainder term is $o(h^2+(nh^{d/2})^{-1}) = o(h^2)$ due to Assumption C1, ii). Moreover, the order of the remainder term can be determined by the bias and variance approximation of the uncorrected product estimator examined in Bouezmarni and Rombouts (2010). Now we will evaluate the first moment conditioned on $X_i := (X_{i1}, ..., X_{id})$

$$E\left[\frac{\hat{f}_G(X_i) - f(X_i)}{f(X_i)}\middle| X_i\right] = E\left[\frac{f(Y_i) - f(X_i)}{f(X_i)}\middle| X_i\right],$$

where $Y_i := (Y_{i1}, ..., Y_{id})$ is a random vector such that $Y_{ij} \overset{\mathcal{D}}{=} G(X_{ij}/h + 1, h)$. A Taylor expansion around $(\mu_{i1}, ..., \mu_{id}) := (X_{i1} + h, ..., X_{id} + h)$ yields

$$\begin{aligned}
E[f(Y_i)|X_i] &= E\Bigg[f(\mu_{i1}, ..., \mu_{id}) + \frac{1}{2}\sum_{j=1}^{d}(Y_{ij} - \mu_{ij})^2\frac{\partial^2 f}{\partial^2 x_j} + \frac{1}{6}\sum_{j=1}^{d}(Y_{ij} - \mu_{ij})^3\frac{\partial^3 f}{\partial^3 x_j} \\
&+ \frac{1}{24}\sum_{j=1}^{d}(Y_{ij} - \mu_{ij})^4\frac{\partial^4 f}{\partial^4 x_j} + \sum_{j\neq k}^{d}(Y_{ij} - \mu_{ij})^2(Y_{ik} - \mu_{ik})^2\frac{\partial^4 f}{\partial^2 x_j \partial^2 x_k}\Bigg| X_i\Bigg] + o(h^2) \\
&= f(\mu_{i1}, ..., \mu_{id}) + \frac{1}{2}\sum_{j=1}^{d} Var(Y_{ij}|X_i)\frac{\partial^2 f}{\partial^2 x_j} + \frac{1}{6}\sum_{j=1}^{d} E[(Y_{ij} - \mu_j)^3|X_i]\frac{\partial^3 f}{\partial^3 x_j} \\
&+ \frac{1}{24}\sum_{j=1}^{d} E[(Y_{ij} - \mu_{ij})^4|X_i]\frac{\partial^4 f}{\partial^4 x_j} + \sum_{j\neq k}^{d} Var(Y_{ij}|X_i)Var(Y_{ik}|X_i)\frac{\partial^4 f}{\partial^2 x_j \partial^2 x_k} + o(h^2) \\
&= f(\mu_{i1}, ..., \mu_{id}) + \frac{1}{2}\sum_{j=1}^{d}(X_{ij}h + h^2)\frac{\partial^2 f}{\partial^2 x_j} + \frac{1}{6}\sum_{j=1}^{d}(2h^2 X_{ij} + 8h^3)\frac{\partial^3 f}{\partial^3 x_j} \\
&+ \frac{1}{24}\sum_{j=1}^{d} 3h^2\frac{\partial^4 f}{\partial^4 x_j} + \sum_{j\neq k}^{d}(X_{ij}h + h^2)(X_{ik}h + h^2)\frac{\partial^4 f}{\partial^2 x_j \partial^2 x_k} + o(h^2).
\end{aligned}$$

Note that the other terms vanish due to the independence of the Gamma distributed random variables Y_{ij} for $j = 1, ..., d$. Now expand these terms again, but now around the point of interest $X_i = (X_{i1}, ..., X_{id})$:

$$\begin{aligned}
&f(\mu_{i1}, ..., \mu_{id}) + \frac{1}{2}\sum_{j=1}^{d}(X_{ij}h + h^2)\frac{\partial^2 f}{\partial^2 x_j} + \frac{1}{6}\sum_{j=1}^{d}(2h^2 X_{ij} + 8h^3)\frac{\partial^3 f}{\partial^3 x_j} \\
&\quad + \frac{1}{24}\sum_{j=1}^{d}3h^2\frac{\partial^4 f}{\partial^4 x_j} + \sum_{j\neq k}^{d}(X_{ij}h + h^2)(X_{ik}h + h^2)\frac{\partial^4 f}{\partial^2 x_j \partial^2 x_k} + o(h^2) \\
&= f(X_{i1}, ..., X_{id}) + h\sum_{j=1}^{d}\frac{\partial f}{\partial x_j} + \frac{h^2}{2}\sum_{j=1}^{d}\frac{\partial^2 f}{\partial^2 x_j} + h^2\sum_{j\neq k}^{d}\frac{\partial^2 f}{\partial x_j \partial x_k} \\
&\quad + \frac{1}{2}\sum_{j=1}^{d}(X_{ij}h + h^2)\left(\frac{\partial^2 f}{\partial^2 x_j} + h\frac{\partial^3 f}{\partial^3 x_j}\right) + \frac{1}{6}\sum_{j=1}^{d}2h^2 X_{ij}\frac{\partial^3 f}{\partial^3 x_j} \\
&\quad + \frac{h^2}{8}\sum_{j=1}^{d}\frac{\partial^4 f}{\partial^4 x_j} + h^2\sum_{j\neq k}^{d}X_{ij}X_{ik}\frac{\partial^4 f}{\partial^2 x_j \partial^2 x_k} + o(h^2) \\
&= f(X_i) + h\left(\sum_{j=1}^{d}\frac{\partial f}{\partial x_j} + \frac{1}{2}X_{ij}\frac{\partial^2 f}{\partial^2 x_j}\right) \\
&\quad + h^2\left(\sum_{j=1}^{d}\frac{\partial^2 f}{\partial^2 x_j} + \frac{5X_{ij}}{6}\frac{\partial^3 f}{\partial^3 x_j} + \frac{1}{8}\frac{\partial^4 f}{\partial^4 x_j} + \sum_{j\neq k}^{d}\frac{\partial^2 f}{\partial x_j \partial x_k} + X_{ij}X_{ik}\frac{\partial^4 f}{\partial^2 x_j \partial^2 x_k}\right) + o(h^2).
\end{aligned}$$

Therefore, we can approximate the first conditional moment of the examined difference as

$$\begin{aligned}
E\left[\frac{\hat{f}_G(X_i) - f(X_i)}{f(X_i)}\middle| X_i\right] &:= h\frac{a_1(X_i)}{f(X_i)} + h^2\frac{a_2(X_i)}{f(X_i)} + o(h^2) \\
&:= hg_1(X_i) + h^2 g_2(X_i) + o(h^2), \text{ as } n \to \infty,
\end{aligned}$$

where the functions g_1 and g_2 are defined as

$$g_1(x) = \frac{a_{1,G}(x)}{f(x)} = \frac{\sum_{j=1}^{d}\left(\frac{\partial f}{\partial x_j} + \frac{1}{2}x_j\frac{\partial^2 f}{\partial^2 x_j}\right)}{f(x)}$$

as well as

$$g_2(x) = \frac{a_{2,G}(x)}{f(x)} = \frac{\sum_{j=1}^{d}\left(\frac{\partial^2 f}{\partial^2 x_j} + \frac{5x_j}{6}\frac{\partial^3 f}{\partial^3 x_j} + \frac{1}{8}\frac{\partial^4 f}{\partial^4 x_j}\right) + \sum_{j\neq k}^{d}\left(\frac{\partial^2 f}{\partial x_j \partial x_k} + x_j x_k\frac{\partial^4 f}{\partial^2 x_j \partial^2 x_k}\right)}{f(x)}.$$

Using the usual bias and variance decomposition and under the already known order of the variance, we can conclude that

$$E\left[\left(\frac{\hat{f}_G(X_i) - f(X_i)}{f(X_i)}\right)^2 \middle| X_i\right] = h^2 g_1^2(X_i) + o(h^2), \text{ as } n \to \infty.$$

Conditioned on X_i, the i-th summand of $\hat{\alpha}(x)$ can be approximated by

$$\frac{\prod_{j=1}^{d} K_{G,h,x_j}(X_{ij})}{f(X_i)} \left(1 - hg_1(X_i) - h^2(h_g^2(X_i) - g_2(X_i))\right) + o(h^2), \text{ as } n \to \infty.$$

Now we are able to derive the unconditioned expectation of $\hat{\alpha}(x)$:

$$\begin{aligned}
&E[\hat{\alpha}(x)] = E[E[\hat{\alpha}(x)|X_i]] \\
&= E\left[\frac{\prod_{j=1}^{d} K_{G,h,x_j}(X_{ij})}{f(X_i)}\right] - hE\left[\frac{g_1(X_i)\prod_{j=1}^{d} K_{G,h,x_j}(X_{ij})}{f(X_i)}\right] \\
&\quad + h^2 E\left[\frac{(g_2(X_i) - g_1^2(X_i))\prod_{j=1}^{d} K_{G,h,x_j}(X_{ij})}{f(X_i)}\right] + o(h^2) \\
&= \int_0^\infty \dots \int_0^\infty \frac{\prod_{j=1}^{d} K_{G,h,x_j}(y_j)}{f(y_1, ..., y_d)} f(y_1, ..., y_d) dy_1 ... dy_d - hE[g_1(Z)] \\
&\quad + h^2 E[(g_2(Z) - g_1^2(Z))] + o(h^2), \text{ as } n \to \infty, \\
&= 1 - hE[g_1(Z)] + h^2 E[(g_2(Z) - g_1^2(Z))] + o(h^2), \text{ as } n \to \infty,
\end{aligned}$$

where $Z := (Z_1, ..., Z_d)$ denotes a random vector consisting of independent Gamma distributed margins with $Z_k \sim G(x_k/h+1, h)$, $k = 1, ..., d$. Now we expand both summands by a Taylor expansion. At first, we expand g_1 and g_2 around $(\nu_1, ..., \nu_d) := (x_1 + h, ..., x_d + h)$ and, afterwards, around $x = (x_1, ..., x_d)$ to obtain

$$E[g_1(Z)] = g_1(x) + h\sum_{j=1}^{d}\left(\frac{\partial g_1}{\partial x_j} + \frac{x_j}{2}\frac{\partial^2 g_1}{\partial^2 x_j}\right) + o(h), \text{ as } n \to \infty$$

as well as

$$E[(g_2(Z) - g_1^2(Z))] = g_2(x) - g_1^2(x) + O(h), \text{ as } n \to \infty.$$

Thus, the term II can finally be approximated by

$$\begin{aligned}
&II = E[\hat{\alpha}(x) - 1] \\
&= -hg_1(x) - h^2\left(\sum_{j=1}^{d}\left(\frac{\partial g_1}{\partial x_j} + \frac{x_j}{2}\frac{\partial^2 g_1}{\partial^2 x_j}\right) + g_2(x) - g_1^2(x)\right) + o(h^2), \text{ as } n \to \infty.
\end{aligned}$$

Using the above results, term I is of the following form

$$I = E\left[\frac{(\hat{f}_G(x) - f(x))}{f(x)}\right] = hg_1(x) + h^2 g_2(x) + o(h^2), \text{ as } n \to \infty.$$

Finally, by the use of the Cauchy-Schwarz inequality and the derived results for terms I and II, we have for the last term III that

$$III = E\left[\frac{(\hat{f}_G(x) - f(x))}{f(x)}(\hat{\alpha}(x) - 1)\right] = -h^2 g_1^2(x) + o(h^2), \text{ as } n \to \infty.$$

Assembling all results, we have for the bias of our proposed estimator that

$$\begin{aligned}
&E[\hat{f}_{JLN}(x)] = f(x) + f(x)\Bigg(hg_1(x) + h^2 g_2(x)\\
&\quad - hg_1(x) - h^2\left(\sum_{j=1}^{d}\left(\frac{\partial g_1}{\partial x_j} + \frac{x_j}{2}\frac{\partial^2 g_1}{\partial^2 x_j}\right) + g_2(x) - g_1^2(x)\right) - h^2 g_1^2(x)\Bigg) + o(h^2)\\
&= f(x) - h^2 f(x)\sum_{j=1}^{d}\left(\frac{\partial g_1}{\partial x_j} + \frac{x_j}{2}\frac{\partial^2 g_1}{\partial^2 x_j}\right) + o(h^2), \text{ as } n \to \infty,\\
&=: f(x) - h^2 f(x) a_{1,G}(x; g) + o(h^2), \text{ as } n \to \infty.
\end{aligned}$$

We will now turn to the variance of the MV-JLN estimator. At first, we observe that, under the consistency of the product Gamma kernel based estimator and a geometric expansion, the following asymptotic representation of $\hat{f}_{JLN}(x)$ holds true.

$$\begin{aligned}
\hat{f}_{JLN}(x) &= \hat{f}_G(x)\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\prod_{j=1}^{d} K_{G,h,x_j}(X_{ij})}{\hat{f}_G(X_i)}\right)\\
&= f(x)\left(1 + \frac{\hat{f}_G(x) - f(x)}{f(x)}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\prod_{j=1}^{d} K_{G,h,x_j}(X_{ij})}{f(X_i)}\right.\\
&\quad \times \left.\left(1 - \frac{\hat{f}_G(X_i) - f(X_i)}{f(X_i)} + o_P\left(\frac{\hat{f}_G(X_i) - f(X_i)}{f(X_i)}\right)\right)\right)\\
&= f(x)\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\prod_{j=1}^{d} K_{G,h,x_j}(X_{ij})}{f(X_i)}\right)\left(2 - \frac{\hat{f}_G(X_i)}{f(X_i)}\right) + o_P((nh^{d/2})^{-1})\\
&:= f(x)\frac{1}{n}\sum_{l=1}^{n}\zeta(X_l) + o_P((nh^{d/2})^{-1}),
\end{aligned}$$

where

$$\begin{aligned}
\zeta(u) := \zeta(u_1, ..., u_d) &= 2\frac{\prod_{j=1}^{d} K_{G,h,x_j}(u_j)}{f(u_1, ..., u_d)} - \frac{1}{n}\sum_{i=1}^{n}\frac{\prod_{j=1}^{d} K_G(X_{ij}, h)(u_j) K_G(x_j, h)(X_{ij})}{f^2(X_{i1}, ..., X_{id})}\\
&:= \zeta_1(X_l) - \zeta_2(X_l).
\end{aligned}$$

We will now approximate the average ζ_2 for a given data vector

$$X_i := (X_{i1}, ..., X_{id}) \neq (0, ..., 0)$$

as follows:

$$\begin{aligned}
\zeta_2(X_{i1}, ..., X_{id}) = \zeta_2(X_i) &\doteq \psi_2(X_i) \\
&:= E\left[\frac{1}{n}\sum_{j=1}^{n} \frac{\prod_{k=1}^{d} K_{G,h,X_{jk}}(X_{ik}) K_{G,h,x_k}(X_{jk})}{f^2(X_j)} \middle| X_i\right] \\
&= E\left[\frac{\prod_{k=1}^{d} K_{G,h,X_{jk}}(X_{ik}) K_{G,h,x_k}(X_{jk})}{f^2(X_j)} \middle| X_i\right] \\
&= \int_{(\mathbb{R}^+)^d} \frac{\prod_{k=1}^{d} K_{G,h,y_k}(X_{ik}) K_{G,h,x_k}(y_k)}{f^2(y_1, ...y_d)} f(y_1, ..., y_d) dy_1...dy_d \\
&:= \int_{(\mathbb{R}^+)^d} \psi_G(y_1, ..., y_d) \prod_{k=1}^{d} K_{G,h,y_k}(X_{ik}) dy_1...dy_d \\
&= \int_{(\mathbb{R}^+)^d} \psi_G(y_1, ..., y_d) \prod_{k=1}^{d} \frac{X_{ik}^{y_k/h} \exp(-X_{ik}/h)}{h^{y_k/h+1}\Gamma(y_k/h+1)} dy_1...dy_d, \qquad (4.17)
\end{aligned}$$

where

$$\psi_G(y_1, ..., y_d) := \frac{\prod_{k=1}^{d} K_{G,h,x_k}(y_k)}{f(y_1, ..., y_d)}.$$

Let us now approximate the integrands using Stirling´s approximation of the Gamma function. Due to the fact that $h \to 0$ as $n \to \infty$, we will only focus on first and second order expansions and omit the remainder terms converging with a faster rate. Now approximate the following term for every $j = 1, ..., d$:

$$\frac{u_j^{y_j/h} \exp(-u_j/h)}{h^{y_j/h+1}\Gamma(y_j/h+1)} = \exp\left(\frac{y_j}{h}\log(u_j) - \frac{u_j}{h} - \log(y_j) - \frac{y_j}{h}\log(h) - \log\left(\Gamma\left(\frac{y_j}{h}\right)\right)\right).$$

Make then use of Stirling´s series for the Gamma function; see Wrench (1968). It holds that

$$\log\left(\Gamma\left(\frac{y_j}{h}\right)\right) = (\frac{y_j}{h}-\frac{1}{2})\log(y_j)-(\frac{y_j}{h}-\frac{1}{2})\log(h)-\frac{y_j}{h}+\log(\sqrt{2\pi})+\frac{h}{12y_j}+O(h^3), \text{ as } n \to \infty.$$

Now we are able to approximate the integral term 4.17 according to

$$\begin{aligned}
&\int_{(\mathbb{R}^+)^d} \psi_G(y_1, ..., y_d) \prod_{k=1}^{d} \frac{X_{ik}^{y_k/h} \exp(-X_{ik}/h)}{h^{y_k/h+1}\Gamma(y_k/h+1)} dy_1...dy_d \\
&= \int_{(\mathbb{R}^+)^d} \psi_G(y_1, ..., y_d) \prod_{k=1}^{d} \frac{1}{\sqrt{y_k h 2\pi}} \\
&\quad \times \exp\left(\frac{y_k}{h}\log\left(\frac{X_{ik}}{y_k}\right) - \frac{(X_{ik}-y_k)}{h} - \frac{h}{12y_k} + O(h^3)\right) dy_1...dy_d.
\end{aligned}$$

We will further approximate this integral and utilize the substitution

$$w_k = \frac{X_{ik}-y_k}{h^{1/2}},\ k = 1, .., d.$$

Using a first order Taylor Expansion of the logarithm and the exponential function we have

$$\begin{aligned}
&\int_{-\infty}^{X_{i1}/h^{1/2}} ... \int_{-\infty}^{X_{id}/h^{1/2}} \psi_G(X_{i1} - h^{1/2}w_1, ..., X_{id} - h^{1/2}w_d) \prod_{k=1}^{d} \frac{(2\pi)^{-1/2}}{(X_{ik} - h^{1/2})^{1/2}} \\
&\times \exp\Bigg(\left(\frac{X_{ik} - h^{1/2}w_k}{h}\right) \log\left(\frac{X_{ik}}{X_{ik} - h^{1/2}w_k}\right) \\
&\qquad - h^{-1/2}w_k - \frac{h}{12(X_{ik} - h^{1/2}w_k)} + O(h^3)\Bigg) dw_1...dw_d \\
&= \int_{-\infty}^{X_{i1}/h^{1/2}} ... \int_{-\infty}^{X_{id}/h^{1/2}} \psi_G(X_{i1} - h^{1/2}w_1, ..., X_{id} - h^{1/2}w_d) \prod_{k=1}^{d} \frac{(2\pi)^{-1/2}}{(X_{ik} - h^{1/2})^{1/2}} \\
&\times \exp\left(-\frac{w_k^2}{2u_{ik}}\right)\left(1 - \frac{h^{1/2}w_k^3}{6X_{ik}^2} - \frac{hw_k^4}{12X_{ik}^3} - \frac{h}{12X_{ik}} + \frac{hw_k^6}{72X_{ik}^4} + O(h^{3/2})\right) dw_1...dw_d. \quad (4.18)
\end{aligned}$$

Now make use of the final substitution

$$v_k = \frac{w_k}{X_{ik}^{1/2}},\ k = 1, ..., d,$$

and denote, as usual, by ϕ the density and by Φ the cumulative distribution function of a standard normal distributed random variable. It holds that 4.18 can be expressed as

$$
\begin{aligned}
&\int_{-\infty}^{\sqrt{X_{ik}/h}} \dots \int_{-\infty}^{\sqrt{X_{id}/h}} \psi_G(X_{i1} - (hX_{i1})^{1/2}v_1, ..., X_{id} - (hX_{id})^{1/2}v_d) \\
&\qquad \prod_{k=1}^{d} \sqrt{\left(\frac{X_{ik}}{X_{ik} - (X_{ik}h)^{1/2}v_k}\right)\phi(v_k)} \\
&\times \left(1 - \frac{h^{1/2}v_k^3}{6\sqrt{X_{ik}}} - \frac{hv_k^4}{12X_{ik}} - \frac{h}{12X_{ik}} + \frac{hv_k^6}{72X_{ik}} + O(h^{3/2})\right) dv_1...dv_d \\
&\doteq \psi_G(X_{i1}, ..., X_{id}) \times \prod_{k=1}^{d} \begin{cases} 1 & \text{, if } X_{ik}/h \to \infty,\ a.s. \\ \Phi(\kappa_k^{1/2}) & \text{, if } X_{ik}/h \to \kappa_k;\ a.s. \end{cases} ;\ \text{as } n \to \infty \\
&= \prod_{k=1}^{d} \frac{K_{G,h,x_k}(X_{ik})}{f(X_{i1}, ..., X_{id})} \times \begin{cases} 1 & \text{, if } X_{ik}/h \to \infty,\ a.s. \\ \Phi(\kappa_k^{1/2}) & \text{, if } X_{ik}/h \to \kappa_k;\ a.s. \end{cases} ;\ \text{as } n \to \infty.
\end{aligned}
$$

To conclude the rate of the variance, we have to determine the second moment of the function ζ, where we use the following abbreviation:

$$
\rho_k = \begin{cases} 1 & \text{, if } X_{ik}/h \to \infty,\ a.s. \\ 2 - \Phi(\kappa_k^{1/2}) & \text{, if } X_{ik}/h \to \kappa_k;\ a.s. \end{cases} ;\ \text{as } n \to \infty.
$$

Now observe that, by adapting the trimming argument of Chen (1999) to the multivariate case (see Bouezmarni and Rombouts (2010), equation 7b), it holds that

$$
\begin{aligned}
&E[\zeta^2(X_{11}, ..., X_{1d})] = E\left[\left(\prod_{k=1}^{d} \rho_k \frac{K_{G,h,x_k}(X_{1k})}{f(X_1)}\right)^2\right] \\
&= \prod_{k=1}^{d} A(x_k, h) E[f^{-1}(Y_1, ..., Y_d)] = f^{-1}(x_1, ..., x_d) \prod_{j=1}^{d} A(x_j, h) + O(h),
\end{aligned}
$$

where $Y_i, i = 1, ..., d$ are independent and Gamma distributed such that $Y_i \sim G(2x_i/h + 1, h/2)$ and

$$
A(x_j, h) := \frac{h^{-1}\Gamma(2x_j/h + 1)}{2^{2x_j/h+1}\Gamma^2(x_j/h + 1)} = \begin{cases} \frac{h^{-1/2}}{2\sqrt{\pi x_j}} & \text{, if } x_j/h \to \infty \\ \frac{h^{-1}\Gamma(2\kappa_j+1)}{2^{2\kappa_j+1}\Gamma^2(\kappa_j+1)} & \text{, if } x_j/h \to \kappa_j. \end{cases}
$$

Therefore, we finally have for an interior vector $x = (x_1, ..., x_d)$ that

$$
\begin{aligned}
Var(\hat{f}_{JLN}(x)) &= \frac{f^2(x)E[\zeta^2(X_1)]}{n} \\
&= \frac{f(x)}{nh^{d/2}} \prod_{k=1}^{d} \frac{1}{2\sqrt{\pi x_k}} + o((nh^{d/2})^{-1}), \text{ as } n \to \infty.
\end{aligned}
$$

□

4.4 Bandwidth selection for MBC estimators

In this section, we shortly want to focus on the problem of selecting a proper bandwidth for the proposed two MBC estimators. We will restrict ourselves to the case $d = 2$, in order to keep the notations and formulas as simple as possible. According to (4.13) and (4.14), the optimal bandwidths $h_{opt,TS,m}$ and $h_{opt,JLN,m}$ for the MSE are proportional to $n^{-1/5}$. Hence, an intuitive and rather simple choice is given by letting

$$h_{opt,TS,m} = h_{opt,JLN,m} = C_{opt} n^{-1/5},$$

where C_{opt} is a generic constant, which has to be chosen by the practitioner.
The rule of thumb by Scott for multivariate nonparametric density estimation (see Silverman (1986)) is given by

$$h_j = \hat{\sigma}_j n^{-1/6}, \ j = 1, 2,$$

where $\hat{\sigma}_j$ denotes the standard deviation of the j-th component of the random vector (X_1, X_2). This leads to an analogous rule of thumb in our context, which will be used in the following section. Due to the fact that comparable selection procedures are highly computable, a rule of thumb provides a very simple and fast way to select a suitable bandwidth for checking the performance of a proposed estimator, although it only minimizes theoretically the pointwise M(I)SE.
When invoking the integrated squared error ("ISE") for a general density estimator $\hat{f}(x)$

$$\text{ISE}(h) := \int (\hat{f}(x) - f(x))^2 \, dx = \int \hat{f}^2(x) \, dx - 2 \int \hat{f}(x) f(x) \, dx + \int f^2(x) \, dx,$$

where $h = (h_1, h_2)$, as global performance criterion, a widely used approach for minimizing this error is the so called least squares cross-validation ("LSCV") method. Since the last term of the ISE is independent of h, we are only interested in finding a minimum of the remaining two terms. The integral in the middle still depends on the unknown density function f, but we can represent this expression as

$$\int \hat{f}(x) f(x) \, dx = E[\hat{f}(X)].$$

The LSCV selected bandwidth h is hence defined as

$$\mathrm{LSCV}(h) = \mathrm{argmin}_{h_1,h_2}\left(\int \hat{f}^2_{g,m,h}(x)\,\mathrm{d}x - 2\hat{E}[\hat{f}_{g,m,h}(X_1,X_2)]\right),$$

where $\hat{E}$ denotes the expectation with respect to the empirical distribution and $g =$TS or JLN. The last expression is determined via the so-called "leave-one-out" estimator, which we define as

$$\begin{aligned}\hat{E}[\hat{f}_{TS,m,h}(X_1,X_2)] &:= \frac{1}{n}\sum_{i=1}^{n}\hat{f}_{TS,m,h,-i}(X_{i1},X_{i2})\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\left(\sum_{\substack{j=1\\ j\neq i}}^{n} K_{m,h_1,X_{i1}}(X_{j1})K_{m,h_2,X_{i2}}(X_{j2})\right)^{\frac{1}{1-c}}\\ &\quad\left(\sum_{\substack{j=1\\ j\neq i}}^{n} K_{m,h_1/c,X_{i1}}(X_{j1})K_{m,h_2/c,X_{i2}}(X_{j2})\right)^{-\frac{c}{1-c}}\end{aligned}$$

for the TS estimator and as

$$\begin{aligned}&\hat{E}[\hat{f}_{JLN,m,h}(X_1,X_2)] = \frac{1}{n}\sum_{i=1}^{n}\hat{f}_{JLN,m,h,-i}(X_{i1},X_{i2})\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\left(\sum_{\substack{j=1\\ j\neq i}}^{n} K_{m,h_1,X_{i1}}(X_{j1})K_{m,h_2,X_{i2}}(X_{j2})\sum_{\substack{j=1\\ j\neq i}}^{n}\frac{K_{m,h_1,X_{i1}}(X_{j1})K_{m,h2,X_{i2}}(X_{j2})}{\sum_{k=1}^{n}K_{m,h_1,X_{j1}}(X_{k1})K_{m,h_2,X_{j2}}(X_{k2})}\right)\end{aligned}$$

for the JLN estimator. The integral is approximated via a proper discretization. For further details, we refer to Funke and Kawka (2015), where this criterion is used for selecting the bandwidth in a real data example.

4.5 Finite sample performance

In this section, we shortly want to present some finite sample results for the proposed MBC estimators. The following results originate from Funke and Kawka (2015), where a variety of additional performances were made. For illustration purposes, we will only focus on selected results and refer the interested reader to the aforementioned article for further details. Hence, let

$$g(x) = \left(\frac{2x}{3}\right)^{1/2}\exp\left(-\,(2x/3)^{3/2}\right)$$

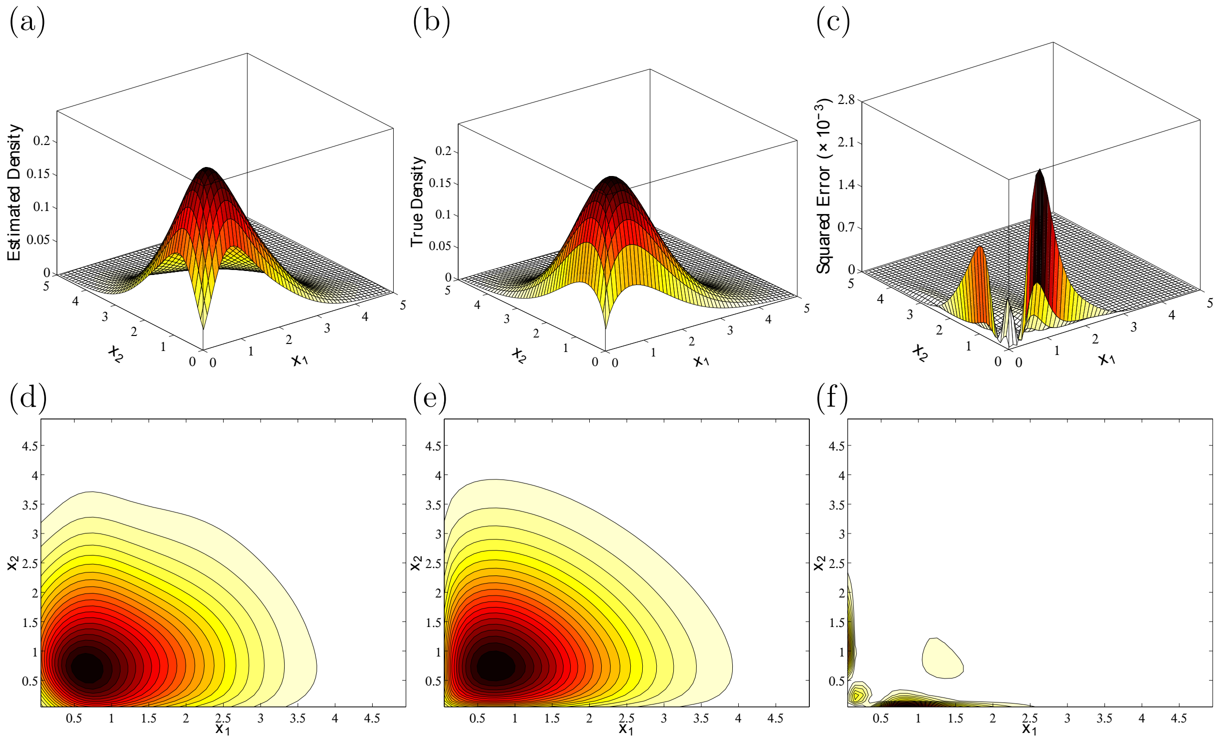

Figure 5: Surface and contour plots of estimated Weibull $(1.5, 1.5)$ density. (a),(d): JLN-estimated density with G kernel. (b),(e): True density. (c),(f): Squared error.

be the univariate density of a Weibull distribution with parameters $(1.5, 1.5)$ and, moreover, let

$$f(x, y) = g(x)g(y)$$

be the joint density of two independent Weibull $(1.5, 1.5)$ distributed random variables. We simulated 1000 data sets of $n = 250$ and $n = 500$ data points and performed a kernel estimation via the Gamma and the Modified Gamma kernels. Furthermore, we measured the performance in terms of the mean root integrated squared error ("MRISE"), which is defined by

$$\text{MRISE}\left(\hat{f}_{g,m,h}\right) = E\left[\left(\int\left(\hat{f}_{g,m,h}(x) - f(x)\right)^2 \mathrm{d}x\right)^{1/2}\right],$$

as well as the integrated absolute bias ("IAB")

$$\text{IAB}\left(\hat{f}_{g,m,h}\right) = \int\left|E\left[\hat{f}_{g,m,h}(x)\right] - f(x)\right| \mathrm{d}x,$$

where $g =$ TS, JLN or C. Here, C denotes the classical product kernel estimator, based on asymmetric kernels. Moreover, the expected values are approximated by the sample

Table 4: Simulation results: MRISE, Standard deviation (SD), and IAB for the estimation of $f(x,y) = g(x)g(y)$, where g is either the density of a Gamma(1.5,1) or a Weibull(1.5,1.5) random variable

Distribution:		Gamma		Weibull	
	Kernel:	G	MG	G	MG
Panel 1: $n = 250$					
TS	MRISE	0.0654	0.0688	0.0697	0.0701
	SD	0.0125	0.0111	0.0151	0.0119
	IAB	0.0851	0.1098	0.1052	0.1172
JLN	MRISE	0.0629	0.0638	0.0662	0.0676
	SD	0.0121	0.0085	0.0145	0.0103
	IAB	0.1220	0.0923	0.1345	0.0968
C	MRISE	0.0680	0.0653	0.0729	0.0689
	SD	0.0114	0.0106	0.0130	0.0123
	IAB	0.1354	0.1366	0.1440	0.1448
Panel 2: $n = 500$					
TS	MRISE	0.0564	0.0570	0.0578	0.0566
	SD	0.0093	0.0068	0.0114	0.0084
	IAB	0.0752	0.0939	0.0895	0.0992
JLN	MRISE	0.0541	0.0556	0.0547	0.0559
	SD	0.0090	0.0058	0.0111	0.0073
	IAB	0.1010	0.0820	0.1119	0.0827
C	MRISE	0.0574	0.0538	0.0606	0.0554
	SD	0.0086	0.0079	0.0104	0.0093
	IAB	0.1133	0.1116	0.1199	0.1168

mean. The integrals are approximated on an equidistant grid of 100^2 points over the square $(0,5) \times (0,5)$, because the probability mass of the given distributions is nearly zero outside this square.

The bandwidths are chosen according to the mentioned rule of thumb

$$h_1 = \sigma(X_1)n^{-1/5} \text{ and } h_2 = \sigma(X_2)n^{-1/5}.$$

The results can be found in Table 4, which originates from Funke and Kawka (2015). It turns out that the MBC methods work quite well and, moreover, especially in terms of the IAB, a significant improvement is obtained. In particular, when estimating the joint density of two independent Gamma$(1.5, 1)$ distributed random variables via the TS estimator, the IAB for the Gamma kernel is 0.0851 whereas the IAB for the classical product kernel is 0.1354. For further discussions and additional examples we refer to Funke and Kawka (2015).

4.6 Nonparametric inference for multivariate diffusions

In order to create a link to the first part of this thesis, we will now again focus on time-continuous stochastic processes, in particular on multivariate diffusions. We will see that comparable approximations of the unknown components of the appearing drift vector exist, as it was the case for univariate diffusions.
Hence, focus on the following multivariate diffusion model

$$d\boldsymbol{X}_t = \boldsymbol{b}(\boldsymbol{X}_t)dt + \boldsymbol{\sigma}(\boldsymbol{X}_t)d\boldsymbol{W}_t,\ \boldsymbol{X}_0 \overset{\mathcal{D}}{=} \eta, \tag{4.19}$$

where

$$\boldsymbol{b} : \mathbb{R}^d \to \mathbb{R}^d$$

is a d-dimensional drift vector and

$$\boldsymbol{\sigma} : \mathbb{R}^d \to \mathbb{R}^{d\times d}$$

is a dispersion matrix. Moreover, $\boldsymbol{W} = \left(W^{(1)}, ..., W^{(d)}\right)$ is a d-dimensional Brownian motion independent of η. In our subsequent analysis, we will impose assumptions, which ensure that (4.19) possesses a stationary solution equipped with a time invariant probability measure $\Gamma(dx)$ such that $\eta \in L^2(\Gamma(dx))$. Based on a high-frequency sample, we will introduce a nonparametric estimator of $\boldsymbol{b}(x)$ for a generic vector $x := (x_1, ..., x_d) \in \mathbb{R}^d$. Afterwards, we will present an estimator based on the already introduced MBC technique via asymmetric kernels.
Nonparametric inference for multivariate diffusion models has been considered, for instance, in Bandi and Moloche (2008) and in Schmisser (2013). Bandi and Moloche (2008) focused on non-stationary processes, which are at least Harris-recurrent. They developed an asymptotic theory for product kernel based estimators of $\boldsymbol{b}$ and $\boldsymbol{\sigma}\boldsymbol{\sigma}^T$. Nevertheless, we will impose alternative assumptions and will work in a stationarity framework. In contrast to the kernel based approach, Schmisser (2013) used a penalized least squares approach based on model selection, which was already introduced in the first part of this thesis. In particular, the approach of Comte et al. (2007) was adapted to the multivariate setting.
The justification of using kernel based estimators in the multivariate setting, too, lies in the fact that $\boldsymbol{b}$ and $\boldsymbol{\sigma}\boldsymbol{\sigma}^T$ can also be recovered via infinitesimal conditional moments. In fact, following Karatzas and Shreve (1996), Chapter 5, pp. 281, we have that

$$b_i(x) = \frac{1}{\Delta}E\left[X_{t+\Delta}^{(i)} - X_t^{(i)}\middle| X_t = x\right] + O(\Delta),\ \text{as } \Delta \to 0$$

as well as

$$a_{ik}(x) = \frac{1}{\Delta} E\left[\left(X_{t+\Delta}^{(i)} - X_t^{(i)}\right)\left(X_{t+\Delta}^{(k)} - X_t^{(k)}\right)\middle| X_t = x\right] + O(\Delta), \text{ as } \Delta \to 0,$$

where $X_t := (X_t^{(1)}, ..., X_t^{(d)})$ and

$$a_{ik}(x) := \sum_{l=1}^{d} \sigma_{il}(x)\sigma_{lk}(x).$$

These approximations suggest the use of multivariate regression techniques in analogy to the univariate case.
Therefore, let us impose that we observe a high-frequency sample

$$\boldsymbol{X}_0, \boldsymbol{X}_\Delta, ..., \boldsymbol{X}_{n\Delta}$$

of d-dimensional random vectors, where we assume that $\Delta \equiv \Delta_n \to 0$ and $n\Delta := T \to \infty$ as $n \to \infty$. Hence, we will work in a double asymptotics scheme, which was one of the crucial assumptions in the first part of this thesis, too.
Let us, therefore, propose the drift vector estimator $\hat{\boldsymbol{b}}$ according to

$$\begin{aligned}\hat{\boldsymbol{b}}(x) &:= \frac{\frac{1}{nh^d}\sum_{i=0}^{n-1} K\left(\frac{\boldsymbol{X}_{i\Delta}-x}{h}\right)\left(\boldsymbol{X}_{(i+1)\Delta} - \boldsymbol{X}_{i\Delta}\right)}{\frac{\Delta}{nh^d}\sum_{i=0}^{n-1} K\left(\frac{\boldsymbol{X}_{i\Delta}-x}{h}\right)} \\ &:= \frac{\frac{1}{nh^d}\sum_{i=0}^{n-1}\prod_{j=1}^{d} k\left(\frac{X_{i\Delta}^{(j)}-x^{(j)}}{h}\right)\left(\boldsymbol{X}_{(i+1)\Delta} - \boldsymbol{X}_{i\Delta}\right)}{\frac{\Delta}{nh^d}\sum_{i=0}^{n-1}\prod_{j=1}^{d} k\left(\frac{X_{i\Delta}^{(j)}-x^{(j)}}{h}\right)},\end{aligned}$$

where $x = (x^{(1)}, ..., x^{(d)})$, $h \equiv h_n$ denotes the bandwidth and

$$\boldsymbol{X}_{i\Delta} := \left(X_{i\Delta}^{(1)}, ..., X_{i\Delta}^{(d)}\right)^T.$$

Now let $\langle x, y\rangle := \sum_{l=1}^{d} x_l y_l$ be the standard scalar product on $\mathbb{R}^d$ and $||x||$ its associated norm. For convenience, we will choose the Euclidean as the associated norm. Moreover, $||A||$ denotes a matrix norm for $A \in \mathbb{R}^{d\times d}$. Now we will make the following assumptions on the considered model:

Assumption D1

i) The drift vector $\boldsymbol{b}$ and the diffusion matrix $\boldsymbol{\sigma}$ are globally Lipschitz-continuous.

ii) The drift vector $\boldsymbol{b}$ is elastic:

$$\exists\ M \in \mathbb{R}^+ :\ \forall x,\ ||x|| > M :\ \langle \boldsymbol{b}(x), x\rangle \lesssim -||x||^2.$$

iii) The matrix $\boldsymbol{A}(x) := \boldsymbol{\sigma}^T(x)\boldsymbol{\sigma}(x) = \{a_{ij}(x)\}_{1\leq i,j\leq d}$ fulfills

$$0 < \min_{1\leq i,j\leq d} a_{ij}(x) \leq a_0.$$

Moreover, let

$$Tr(\boldsymbol{A}(x)) = \sum_{l=1}^{d} a_{ll}(x) \leq \sigma_0^2\ \ \forall\ x \in \mathbb{R}^d.$$

iv) For the matrix $\boldsymbol{A}$, there are constants $\lambda_-,\ \lambda_+ > 0$ such that for all $x \in \mathbb{R}^d$

$$\lambda_-||x||^2 \leq \langle \boldsymbol{A}(x), x\rangle \leq \lambda_+||x||^2.$$

Assumption D1, i) guarantees the existence and uniqueness of a solution of (4.19); see Karatzas and Shreve (1996), Theorem 2.9. Assumptions D1, ii) and D1, iii) ensure that the solution is endowed with an invariant measure $\Gamma(dx)$, which is additionally absolutely continuous with respect to the Lebesgue measure; see Schmisser (2013). Hence, a stationary density $\pi(x)$ of the process exists.

v) To ensure that X is stationary, we further assume that

$$\eta \sim \pi(x)dx.$$

According to Pardoux and Veretennikov (2001) and assumptions $D1, i) - iv)$, the process X is exponentially β-mixing, too. Therefore, X admits the ergodicity property that for all g such that $\int_{\mathbb{R}^d} |g(y)|\Gamma(dy) < \infty$ we can deduce that

$$\frac{1}{T}\int_0^T g(\boldsymbol{X}_s)ds \xrightarrow{P} \int_{\mathbb{R}^d} g(y)\pi(y)dy,\ \text{as } T \to \infty.$$

In analogy to the kernel specific assumptions of the first part of this thesis, we have to ensure that comparable assumptions also hold true for the product kernel.

Assumption D2

i) For $x \in \mathbb{R}^d$ let $K(x) = \prod_{j=1}^{d} k(x_j)$, where the univariate kernel k is a symmetric, bounded, differentiable, and Lipschitz-continuous probability density function.

ii) Let the product kernel K fulfill

$$\int_{\mathbb{R}^d} z^2 K(z)dz < \infty \text{ and } \int_{\mathbb{R}^d} K^2(z)dz < \infty.$$

iii) Let the sampling frequency and the bandwidth be coupled according to

$$\frac{\Delta^{1/2}}{h^{2d}} \to 0 \text{ as well as } \frac{1}{n\Delta h^{2d}} \to 0 \text{ as } n \to \infty.$$

We briefly remark that K is also bounded and Lipschitz-continuous by the use of assumption D1, i). In particular, let $x, y \in \mathbb{R}^d$, then we can conclude that

$$\begin{aligned} |K(x) - K(y)| &= \left| \prod_{j=1}^{d} k(x_j) - \prod_{j=1}^{d} k(y_j) \right| \leq ||k||_{\infty}^{d-1} \sum_{j=1}^{d} |k(x_j) - k(y_j)| \\ &\leq ||k||_{\infty}^{d-1} L_k \sum_{j=1}^{d} |x_j - y_j| := L_K ||x - y||. \end{aligned}$$

The following theorem states the consistency of $\hat{\boldsymbol{b}}(x)$ under assumptions $D1$ and $D2$. As already mentioned, Bandi and Moloche (2008) derived the consistency and asymptotic normality of $\hat{\boldsymbol{b}}(x)$ in a non-stationary framework under alternative assumptions. Nevertheless, we think that it is worth mentioning this theorem, because it leads us, afterwards, to the use of the introduced MBC techniques for improving this class of estimators under the stationarity assumption.

Theorem 4.9. *Under assumptions $D1$ and $D2$, provided that $\pi(x) > 0$, we have that*

$$\hat{\boldsymbol{b}}(x) \xrightarrow{P} \boldsymbol{b}(x), \text{ as } n \to \infty.$$

A useful proposition, which has already been established for Lévy driven univariate diffusions in Proposition 2.3, is also useful in the present context. It is stated in Schmisser (2013) and can also be found in Glotter (2000) for the univariate case.

Proposition 4.10 (Proposition 5, Schmisser (2013))**.** *Let $\boldsymbol{X} = (\boldsymbol{X}_t)_{t\geq 0}$ be the solution of* (4.19)*. Under assumption $D1$, provided that $\Delta \leq 1$, we have that*

$$E\left[\sup_{|s-t|\leq\Delta} ||\boldsymbol{b}(\boldsymbol{X}_s) - \boldsymbol{b}(\boldsymbol{X}_t)|| \right] \lesssim \Delta^{1/2}.$$

Proof of Theorem 4.9. We will only give a sketch of the proof and start with a decomposition as follows:

$$\frac{\boldsymbol{X}_{(i+1)\Delta}-\boldsymbol{X}_{i\Delta}}{\Delta}=\boldsymbol{b}(\boldsymbol{X}_{i\Delta})+\frac{1}{\Delta}\int_{i\Delta}^{(i+1)\Delta}\boldsymbol{\sigma}(\boldsymbol{X}_s)d\boldsymbol{W}_s+\frac{1}{\Delta}\int_{i\Delta}^{(i+1)\Delta}\left(\boldsymbol{b}(\boldsymbol{X}_s)-\boldsymbol{b}(\boldsymbol{X}_{i\Delta})\right)ds$$

The second summand is a noise term whereas the third summand is a remainder one. Furthermore, all appearing integrals are understood coordinatewisely.
Using the boundedness of K, we find that

$$\begin{aligned}&E\left[\left|\left|\frac{1}{nh^d}\sum_{i=0}^{n-1}K\left(\frac{\boldsymbol{X}_{i\Delta}-x}{h}\right)\frac{1}{\Delta}\int_{i\Delta}^{(i+1)\Delta}\left(\boldsymbol{b}(\boldsymbol{X}_s)-\boldsymbol{b}(\boldsymbol{X}_{i\Delta})\right)ds\right|\right|\right]\\&\leq\frac{||k||_\infty^d}{n\Delta h^d}\sum_{i=0}^{n-1}\int_{i\Delta}^{(i+1)\Delta}E\left[||\boldsymbol{b}(\boldsymbol{X}_s)-\boldsymbol{b}(\boldsymbol{X}_{i\Delta})||\right]ds\lesssim\frac{\Delta^{1/2}}{h^{2d}}.\end{aligned}$$

For the noise term, define the random vector

$$\frac{1}{n\Delta h^d}\sum_{i=0}^{n-1}K\left(\frac{\boldsymbol{X}_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta}\boldsymbol{\sigma}(\boldsymbol{X}_s)d\boldsymbol{W}_s=:\boldsymbol{Z}:=\left(Z^{(1)},...,Z^{(d)}\right)^T,$$

where

$$Z^{(j)}:=\frac{1}{n\Delta h^d}\sum_{i=0}^{n-1}K\left(\frac{\boldsymbol{X}_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta}\sum_{l=1}^{d}\sigma_{jl}(\boldsymbol{X}_s)dW_s^{(l)}.$$

Using the independence of the components of W, the Itô-isometry as well as the boundedness of the trace of the matrix $\boldsymbol{A}$, we can conclude that

$$\begin{aligned}E[Z^TZ]&=\frac{1}{n^2\Delta^2h^{2d}}\sum_{i=0}^{n-1}E\left[K^2\left(\frac{\boldsymbol{X}_{i\Delta}-x}{h}\right)\int_{i\Delta}^{(i+1)\Delta}\sum_{l=1}^{d}a_{ll}(\boldsymbol{X}_s)ds\right]\\&\leq\frac{||k^2||_\infty^d\sigma_0^2}{n\Delta h^{2d}}.\end{aligned}$$

Using assumption $D1$, we are able to deduce that both terms are negligible. Moreover, we are able to finish the proof of the consistency of $\hat{\boldsymbol{b}}(x)$ due to the following reasons. Firstly, the denominator is a consistent estimate of the stationary density $\pi(x)$, which can be deduced in an analogous manner to the univariate case; see Bandi and Moloche (2008). Moreover,

$$\begin{aligned}&\frac{1}{nh^d}\sum_{i=0}^{n-1}K\left(\frac{\boldsymbol{X}_{i\Delta}-x}{h}\right)\boldsymbol{b}(\boldsymbol{X}_{i\Delta})\\&=\frac{1}{nh^d}\sum_{i=0}^{n-1}\prod_{j=1}^{d}k\left(\frac{X_{i\Delta}^{(j)}-x^{(j)}}{h}\right)\boldsymbol{b}(\boldsymbol{X}_{i\Delta})\xrightarrow{P}\boldsymbol{b}(x)\pi(x),\text{ as }n\to\infty.\end{aligned}$$

The latter statement can be deduced by standardly used arguments which were taken into account in the first part of this thesis. □

To get into account of the usage of asymmetric kernels in this context, we will at first focus on Gospodinov and Hirukawa (2012), who investigated a univariate diffusion model

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t,\ X_0 \overset{\mathcal{D}}{=} \eta,$$

where b and σ are unknown and W is a Brownian motion independent of η. This process is assumed to be almost surely positive and, furthermore, that a high-frequency sample $\{X_{i\Delta}\}_{i=1,\dots,n}$ is observed. In an analogous manner to Bandi and Phillips (2003), the unknown drift b is estimated at a point x via

$$\hat{b}_G(x) := \frac{\frac{1}{n\Delta}\sum_{i=1}^{n-1} G(x/h+1,h)(X_{i\Delta})(X_{(i+1)\Delta} - X_{i\Delta})}{\frac{1}{n}\sum_{i=1}^{n-1} G(x/h+1,h)(X_{i\Delta})},$$

where $G(x/h+1,h)(u)$ denotes the already introduced density of a Gamma distribution with parameters $p = x/h+1$ and $\gamma = h$. They show in Theorem 1 and also in Corollary 1 that, under the assumption that it exists a stationary solution X of the considered stochastic differential equation, the asymptotic distribution of $\hat{b}_G(x)$ can be derived in dependency of the location of x. For an interior point x, the asymptotic distribution is given by

$$\sqrt{n\Delta h^{1/2}}\left(\hat{b}_G(x) - b(x) - h\left(b'(x)\left(1+\frac{x\pi'(x)}{\pi(x)}\right) + \frac{xb''(x)}{2}\right)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{\sigma^2(x)}{2\sqrt{\pi x}\pi(x)}\right)$$

as $n \to \infty$ and $h = O((n\Delta)^{-2/5})$.

For a boundary point $x = \rho h$, $0 < \rho \leq 1$, the bias is negligible due to the choice of h and the asymptotic distribution can be deduced as

$$\sqrt{n\Delta h}\left(\hat{b}_G(x) - b(x)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{\Gamma(2\rho+1)\sigma^2(x)}{2^{2\rho+1}\Gamma^2(\rho+1)\pi(x)}\right),\ \text{as } n \to \infty.$$

This work acts as a motivation for us to consider multivariate diffusions, where the components are almost surely positively supported. Under the upper sampling scheme and in view of Bandi and Moloche (2008) as well as Bouezmarni and Rombouts (2010), a natural estimator of $\hat{\boldsymbol{b}}_G(x)$ is given by

$$\hat{\boldsymbol{b}}_G(x) := \frac{\frac{1}{n}\sum_{i=1}^{n-1}\prod_{l=1}^{d} G(x_l/h+1,h)(X_{i\Delta}^{(l)})(\boldsymbol{X}_{(i+1)\Delta} - \boldsymbol{X}_{i\Delta})}{\frac{\Delta}{n}\sum_{i=1}^{n-1}\prod_{l=1}^{d} G(x_l/h+1,h)(X_{i\Delta}^{(l)})},$$

where $h_1 = ... = h_d \equiv h$ for the sake of simplicity.

In view of the previous chapter, a further improvement of this estimator can be established by using the proposed MBC techniques. In order to propose adequate estimators, let us, in a first step, use the following identity

$$\boldsymbol{b}(x) = \begin{pmatrix} b_1(x) \\ \vdots \\ b_d(x) \end{pmatrix} = \begin{pmatrix} \hat{b}_1(x)\frac{b_1(x)}{\hat{b}_1(x)} \\ \vdots \\ \hat{b}_d(x)\frac{b_d(x)}{\hat{b}_d(x)} \end{pmatrix}.$$

In view of the JLN technique, a possible estimator of the j-th component of $\boldsymbol{b}$ is given by

$$\hat{b}_{j,G,JLN}(x) := \hat{b}_{G,j}(x) \sum_{i=1}^{n-1} \frac{\left(X_{(i+1)\Delta}^{(j)} - X_{i\Delta}^{(j)}\right) \prod_{l=1}^{d} G(x_l/h+1,h)(X_{i\Delta}^{(l)})}{\Delta \hat{b}_{G,j}(\boldsymbol{X}_{i\Delta}) \sum_{m=1}^{n-1} \prod_{k=1}^{d} G(x_k/h+1,h)(X_{m\Delta}^{(k)})},$$

where

$$\hat{b}_{G,j}(x) := \frac{\sum_{i=1}^{n-1} \left(X_{(i+1)\Delta}^{(j)} - X_{i\Delta}^{(j)}\right) \prod_{l=1}^{d} G(x_l/h+1,h)(X_{i\Delta}^{(l)})}{\Delta \sum_{i=1}^{n-1} \prod_{l=1}^{d} G(x_l/h+1,h)(X_{i\Delta}^{(l)})}, \ j = 1, ..., d.$$

Finally, we suggest the use of

$$\hat{\boldsymbol{b}}(x) = \begin{pmatrix} \hat{b}_{1,G,JLN}(x) \\ \vdots \\ \hat{b}_{d,G,JLN}(x) \end{pmatrix}.$$

An analogous construction can be executed for the derivation of the Terrell Scott estimator.
However, it seems to be quite complicated to derive the asymptotic properties of these estimators. Nevertheless, we conjecture that, due to the findings for the i.i.d. case, those estimators are also consistent when dealing with stationary and exponential β-mixing data.

5 Nonparametric estimation of copula densities

In this section, we will introduce the concept of copula functions and present methods for the estimation of unknown copula densities. Copulas are multivariate distribution functions, which are able to cover the whole dependence structure of random vectors. Especially, they are able to capture non-linear dependence between the coordinates and are, thus, a more useful tool than correlations. We will start with a short presentation of the most important properties of copulas, in particular, the theorem of Sklar will play a central role. Subsequently, we will introduce two methods for the nonparametric estimation of such densities. At first, we will have a look at Beta product kernel based estimators, which were introduced in the previous section. Afterwards, we will make use of a multivariate version of the Weierstrass approximation theorem, which states that every continuous function f defined on a compact interval $[a, b]^d$, can be uniformly approximated via a sequence of multidimensional polynomials. In our case, we have $a = 0$ and $b = 1$ and choose the Bernstein polynomials as approximating sequence. They have appealing properties, which will be explored later on. Afterwards, we will turn to the estimation of joint densities and conditional densities via the corresponding copula representation. Finally, we will describe how conditional moments can be estimated via a Bernstein polynomial based copula estimator. This allows us to present a substantially different way of estimating the coefficients in a diffusion model and to create a link to the first part of this thesis.

5.1 Copula functions and multivariate distributions

In view of the previous section, the estimation of unknown copulas and their densities is highly influenced by the boundary bias effect. The support of copula densities is naturally given by the unit hyper cube $[0, 1]^d$, where d denotes the dimension of the underlying data set. Hence, when estimating a multivariate compact supported density, the boundary region increases exponentially in terms of d. Thus, the adequate choice of boundary corrected kernels is of major interest in higher dimensions.
As toy-examples of such functions, we will now introduce the concept of copulas. A standard reference in this context is “An introduction to Copulas” by Nelsen (1999), where the following introductory definitions, lemmas, and theorems are taken from.

Definition 5.1. *A multivariate distribution function $C(x_1, ..., x_d)$, such that the marginal distributions are uniformly distributed on the unit interval $(0, 1)$, is called a copula.*

Lemma 5.2. *Let $C(x_1, ..., x_d)$ be a copula and let $F_1, ..., F_d$ be univariate distribution functions, then*

$$F(x_1, ..., x_d) = C(F_1(x_1), ..., F_d(x_d))$$

is a multivariate distribution function possessing the margins $F_1, ..., F_d$.

The upper lemma describes the close relation between copulas and multivariate distribution functions. Probably the most famous theorem in this context is the converse statement of Lemma 5.2, which goes back to Sklar (1959).

Theorem 5.3 (Sklar´s theorem). *Let F be a d-dimensional distribution function possessing the margins $F_1, ..., F_d$. Then, a copula C exists such that*

$$F(x_1, ..., x_d) = C(F_1(x_1), ..., F_d(x_d)).$$

Moreover, if $F_1, ..., F_d$ are continuous, C is unique.

Another interesting property is the existence of upper and lower bounds for every copula C given by the so-called Fréchet-Hoeffding bounds.

Proposition 5.4. *For every d-dimensional Copula C, it holds that*

$$\max\left\{\sum_{j=1}^{d} u_j + 1 - d;\ 0\right\} \leq C(u_1, ..., u_d) \leq \min\{u_1; ...; u_d\}.$$

The upper bound is a copula, too, whereas the lower is a copula iff $d = 2$.

An additional useful property is the fact that copulas are invariant under strictly increasing transformations.

Lemma 5.5. *Let $X_1, ..., X_d$ be random variables possessing the copula C and margins $F_1, ..., F_d$. Let*

$$T_k : \ D_k \to \mathbb{R},\ k = 1, ..., d,$$

be strictly increasing functions, each defined on the range D_k of the random variable X_k. Let $\tilde{X}_k := T_k(X_k)$ with corresponding margins $\tilde{F}_k$, then the joint distribution $\tilde{F}$ is given by

$$\tilde{F}(x_1, ..., x_d) = C(\tilde{F}_1(x_1), ..., \tilde{F}_d(x_d)).$$

The copula will only be slightly different, if T_k is strictly decreasing.
We will now state some important examples of copulas, which can be found in Schmidt (2007):

Example 5.6. *i) Using Sklar´s theorem, we are easily able to deduce that*

$$C(u_1, ..., u_d) = \prod_{j=1}^{d} u_j,$$

if and only if the random variables $X_1, ..., X_d$ are independent.

ii) As already mentioned, the Fréchet-Hoeffding upper bound

$$M(u_1, ..., u_d) := \min\{u_1, ..., u_d\}$$

is also a copula, which is often referred to as the comonotonicity copula. This copula describes perfect positive dependence. Indeed, for a random variable X_1 and strictly increasing functions $T_k, k = 2, ..., d$, define $X_k = T_k(X_1)$. Then, we have that

$$C_{X_1,...,X_d}(u_1, ..., u_d) = M(u_1, ..., u_d).$$

iii) In the case $d = 2$, the Fréchet-Hoeffding lower bound is a copula, too. It is often referred to as the countermonotonicity copula and is given by

$$W(u_1, u_2) := \max\{u_1 + u_2 - 1;\ 0\}.$$

It describes perfect negative dependence and can, analogously to ii), be constructed by setting $T_1(X_1) := X_2$ for a strictly decreasing function T_1 defined on the range of X_1.

iv) Sklar´s theorem allows us to construct copulas for given margins $F_1, ..., F_d$ and the joint distribution F via the representation

$$C(u_1, ..., u_d) = F(F_1^{-1}(u_1), ..., F_d^{-1}(u_d)) \text{ and the relation } F_k(F_k^{-1}(y)) \geq y.$$

Recall that the generalized inverse $F_k^{-1}(y)$ is defined as

$$F_k^{-1}(y) := \inf\{x \in \mathbb{R} :\ F(x) \geq y\}.$$

Hence, let $X_1, ..., X_d$ be standard normally distributed random variables, which are also jointly normally distributed with mean vector $\boldsymbol{0}$ and correlation matrix Σ. The Gaussian copula $C_{\Sigma,G}$ is then defined as

$$C_{\Sigma,G}(u_1, ..., u_d) = \Phi_\Sigma(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_d)),$$

where Φ_Σ denotes the multivariate and Φ the univariate normal distribution function.

v) Another famous example of a distributional construction of a copula is given by the Student copula

$$C_{\nu,\Sigma,t}(u_1, ..., u_d) := t_{\nu,\Sigma}(t_\nu^{-1}(u_1), ..., t_\nu^{-1}(u_d)),$$

where $t_{\nu,\Sigma}$ describes the cumulative distribution function of the multivariate Student distribution with parameter ν and correlation matrix Σ. Moreover, t_ν is the univariate distribution function of the Student distribution with ν degrees.

vi) The class of Archimedean copulas is widely used, because it provides a possibility to construct several copulas having desired properties for certain purposes. We will omit the exact definition and refer the interested reader to Nelsen (1999), Chapter 4. Instead, we will state three different examples of bivariate Archimedean copulas, which are defined in terms of one parameter.
The Gumbel copula is defined as

$$C_{\theta,Gu}(u_1, u_2) = \exp\left(-\left((-\ln u_1)^\theta + (-\ln u_u)^\theta\right)^{\frac{1}{\theta}}\right),$$

where $\theta \geq 1$. In terms of θ, this family of copulas interpolates between the independence ($\theta = 1$) and the comonotonicity ($\theta \to \infty$) copula. For values $\theta > 1$, these copulas exhibit a tail dependence corresponding to high profits in both components.
The Clayton copula

$$C_{\theta,Cl}(u_1, u_2) = \left(\max\{u_1^{-\theta} + u_2^{-\theta} - 1;\ 0\}\right)^{-\frac{1}{\theta}},$$

where $\theta \in [-1, \infty)\backslash\{0\}$ interpolates between the countermonotonicity ($\theta = -1$), the independence ($\theta \to 0$), and the comonotonicity ($\theta \to \infty$) copulas. Thus, this Copula family is able -comparable to the Gumbel copula family- to capture various dependence structures.
The last example of an Archimedean copula is the Frank copula

$$C_{\theta,Fr}(u_1, u_2) := -\frac{1}{\theta}\ln\left(1 + \frac{\left(e^{-\theta u_1} - 1\right)\cdot\left(e^{-\theta u_2} - 1\right)}{e^{-\theta} - 1}\right),$$

where $\theta \in \mathbb{R}\backslash\{0\}$. This copula possesses a bounded density and will be of interest in the following section.

For our purposes, it suffices to finish this introductory section by some remarks concerning the question why correlation is not able to capture general dependency.
There are several different concepts of measuring dependency of bivariate random vectors; see Schmidt (2007), Section 4. Linear correlation as a measure of dependence is only reasonable when dealing with, for example, normally distributed random variables.
But in general, besides the fact that correlation is invariant under linear transformation, it can change under general transformations. See the following example.

Example 5.7. *Let $X_1 \sim Exp(2)$ be an exponentially distributed random variable with parameter 2, then we can deduce that*

$$E[X_1] = \frac{1}{2} \text{ and } Var(X_1) = \frac{3}{4}.$$

Now let $X_2 = X_1^2$, then it obviously holds that

$$E[X_2] = 1 \text{ and } Var(X_2) = 11.$$

The correlation coefficient ρ can now be determined by

$$\rho(X_1, X_2) = \frac{Cov(X_1, X_2)}{\sqrt{Var(X_1)} \cdot \sqrt{Var(X_2)}} = \frac{2.5}{\frac{\sqrt{33}}{2}} \approx 0.87.$$

Although X_1 and $X_2 = X_1^2$ are perfectly dependent, the correlation coefficient is not equal to 1. The corresponding copula is, in view of Lemma 5.5, invariant under the transformation.

Moreover, let $X_3 = X_1^4$, then

$$\rho(X_2, X_3) = \frac{Cov(X_2, X_3)}{\sqrt{Var(X_2)} \cdot \sqrt{Var(X_3)}} = \frac{348}{\sqrt{220176}} \approx 0.74.$$

Even when the pair (X_2, X_3) can be derived by squaring the components of (X_1, X_2), the correlation coefficient changes.

5.2 Nonparametric estimation of copulas

In this section, we will briefly describe how estimators for unknown copula functions are constructed.

Hence, let $\{X_{i1}, ..., X_{id}\},\ i = 1, ..., n$, be a sample of identically distributed random vectors, which are not necessarily independent. Recall that, by Sklar´s theorem, a copula C exists such that

$$C(u_1, ..., u_d) = F(F_1^{-1}(u_1), ..., F_d^{-1}(u_d))\ u_k \in [0,1]^d.$$

When no additional information, concerning the margins, is available, a natural nonparametric choice for an estimator of F_k is the empirical distribution function

$$\hat{F}_k(x) := \frac{1}{n} \sum_{j=1}^{n} 1(X_{jk} \leq x),\ k = 1, ..., d,$$

which yields to a consistent asymptotically normally distributed estimator. Moreover, the joint distribution function F can be estimated via the joint empirical distribution function according to

$$\hat{F}(x_1, ..., x_d) := \frac{1}{n} \sum_{j=1}^{n} 1(X_{j1} \leq x_1, ..., X_{jd} \leq x_d).$$

Thus, a natural estimator of the corresponding copula is given by its empirical counterpart:

$$\hat{C}(u_1, ..., u_d) := \frac{1}{n} \sum_{j=1}^{n} 1(X_{j1} \leq \hat{F}_1^{-1}(u_1), ..., X_{jd} \leq \hat{F}_1^{-1}(u_d)) = \hat{F}(\hat{F}_1^{-1}(u_1), ..., \hat{F}_1^{-1}(u_d)), \tag{5.20}$$

which has been introduced in Deheuvels (1979) and has further been studied in Fermanian et al. (2004) as well as Doukhan et al. (2004). This method is fully nonparametric and no additional smoothing parameters have to be chosen. Moreover, $\hat{C}$ is a consistent estimator of C and the empirical copula process

$$\sqrt{n}(\hat{C}(u) - C(u))$$

converges weakly to a Gaussian process by Donsker´s theorem; see Fermanian et al. (2004) or Bücher and Volgushev (2013). One disadvantage is the fact that $\hat{C}$ is not differentiable by construction. In our subsequent analysis, we are interested in estimators for the copula density c, which is defined as the derivative of C with respect to $(u_1, ..., u_d)$:

$$c(u_1, ..., u_d) := \frac{\partial^d}{\partial u_1 ... \partial u_d} C(u_1, ..., u_d), \text{ at every } u = (u_1, ..., u_d) \in [0, 1]^d.$$

Therefore, an alternative method is provided by using smoothed estimators such that the joint distribution function is estimated by

$$\hat{F}(x_1, ..., x_d) = \frac{1}{n} \sum_{j=1}^{n} \prod_{k=1}^{d} K\left(\frac{X_{jk} - x_k}{h}\right),$$

where $K(z) := \int_{-\infty}^{z} k(y)dy$ is the distribution function corresponding to the probability measure $k(z)dz$. Note that the bandwidth h only occurs in the argument of K which opposes the density estimation case. The usual product kernel based estimator for probability densities can easily be derived by differentiation of $\hat{F}$ with respect to $(x_1, ..., x_d)$. Furthermore, the margins can be consistently estimated by the use of ordinary kernel based estimators.

5.3 Copula density estimation and the boundary bias effect

In this section, we want to state possible estimation approaches for estimating copula densities nonparametrically and additionally by incorporating of their bounded support. Let $(X_1, ..., X_d)$ be a random vector with marginal distribution functions F_k and corresponding copula C. Due to the fact that

$$(F_1(X_1), ..., F_d(X_d)) \sim C$$

by Sklar´s theorem, a natural estimator of the copula density c is given by

$$\hat{c}_K(u) := \frac{\partial^d}{\partial u_1 ... \partial u_d} \hat{C}_K(u),\ u = (u_1, ..., u_d) \in [0,1]^d,$$

where $\hat{C}_K$ denotes a kernel based estimator for the unknown copula C based on the observations

$$(\hat{F}_1(X_{i1}), ..., \hat{F}_d(X_{id})),\ i = 1, ..., n.$$

For the sake of simplicity, we restrict ourselves to the case $d = 2$ and define

$$\hat{c}_K(u_1, u_2) := \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{u_1 - \hat{F}_1(X_{i1})}{h}\right) K\left(\frac{u_1 - \hat{F}_2(X_{i2})}{h}\right),$$

where K is a symmetrical univariate probability density, whose support is $[-1, 1]$ and

$$\hat{F}_1(X_{i1}) = \frac{1}{n+1} \sum_{j=1}^{n} 1(X_{j1} \leq X_{i1}) \in \left\{\frac{1}{n+1}, ..., \frac{n}{n+1}\right\}$$

denotes the rescaled rank of X_{i1}. The scaling by $\frac{1}{n+1}$ has technical reason and should avoid estimation problems in the corners. This estimator has been investigated by Gijbels and Mielniczuk (1990) as well as Behnen et al. (1985). Note that $\hat{c}_K$ is based on so-called pseudo-observations $(\hat{F}_1(X_{i1}), \hat{F}_2(X_{i2}))$.

By assuming that c is twice continuously differentiable, this estimator is consistent at every interior point $(u, v) \in (0, 1)$, but in contrast it is biased at any corner and on the interior of the borders; see Charpentier et al. (2006), page 12. In fact, Charpentier et al. (2006) derived

$$E[\hat{c}_K(0, 0)] = \frac{1}{4} c(0, 0) + O(h),\ \text{as } n \to \infty$$

and

$$E[\hat{c}_K(0, v)] = \frac{1}{2} c(0, v) + O(h),\ v \in (0, 1) \text{ as } n \to \infty.$$

This is the already known boundary bias effect for multivariate compact supported probability densities. To face this problem, several techniques have been considered. Most of them are already presented in the previous section. Figure 6 shows the mirror image method, which consists of the reflection of the sample points with respect to all edges and corners. In the sequel analysis, we will present two methods introducing alternative approaches for estimating unknown copula densities. The first one uses Beta kernels, which were already used in the previous chapter as a starting point for the proposed correction methods. The other method is based on Bernstein polynomials and will lead us to a regression estimator.

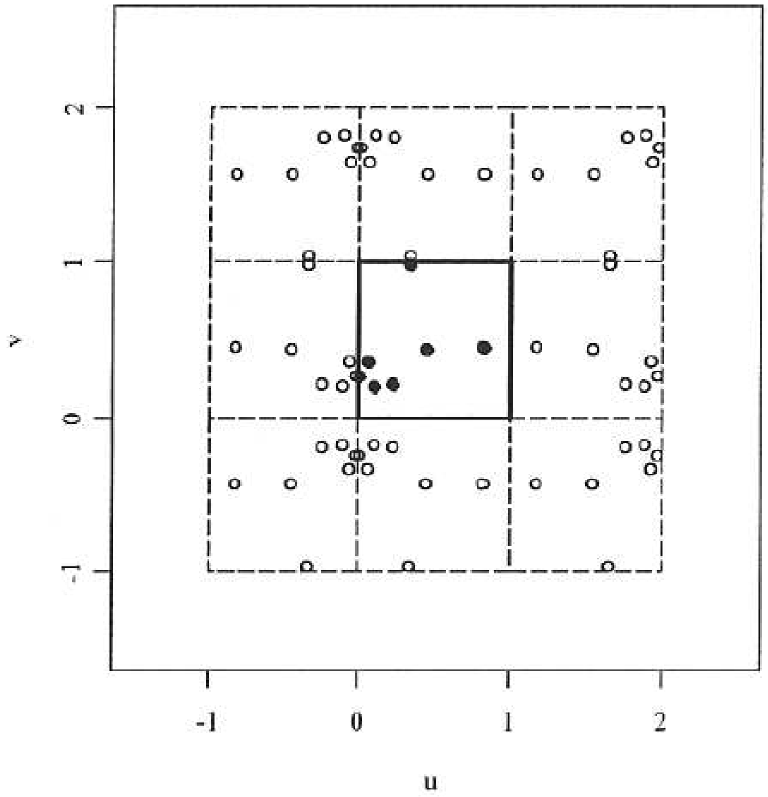

Figure 6: Plot of seven data points inside the unit square $[0,1]^2$ in black and their reflected pseudo counterparts, which are black-rimmed. Source: Schmid and Trede (2006): Finanzmarktstatistik, p. 106.

5.4 Beta kernel based copula density estimation

We will now introduce the beta kernel based copula density estimator, for which the asymptotic mean squared error as well as the asymptotic normality for strongly mixing data are determined. To the best of our knowledge, this has not been investigated in the literature before. We are only aware of an article by Bouezmarni et al. (2010), where the asymptotic normality and the consistency of the Bernstein density copula estimator has been derived under weakly dependent data.
Hence, suppose that $\{(X_{i1},...,X_{id})\},\ i=1,...,n$, is a sample of $d-$dimensional random vectors, which are α-mixing with joint cumulative distribution function F and joint probability density function f. Moreover, let

$$C(F_1(u_1),...,F_d(u_d)) := F(u_1,...,u_d)$$

be the corresponding copula function C possessing a density c. Further, assume that the mixing coefficient $\alpha(k)$ fulfills

$$\alpha(k) \leq \rho^k, \ \rho \in (0,1).$$

We will focus on the nonparametric estimation of the copula density c of the random vector X. For this purpose, we use an estimator based on Beta kernels

$$\hat{c}(u_1, ..., u_d) := \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{d} K_B \left(U_{ij}, \frac{u_j}{h} + 1, \frac{1-u_j}{h} + 1 \right),$$

where $K_B \left(x, \frac{x}{h} + 1, \frac{1-x}{h} + 1 \right)$ denotes the Beta density with parameters

$$(a, b) = \left(\frac{x}{h} + 1, \frac{1-x}{h} + 1 \right),$$

h is a bandwidth, and

$$U_i := (U_{i1}, ..., U_{id}) = (F_1(X_{i1}), ..., F_d(X_{id})) \in [0,1]^d.$$

Note that the random vector U_i is distributed according to

$$U_i = (U_{i1}, ..., U_{id}) \sim C.$$

We will investigate the asymptotic bias, the variance, as well as the asymptotic normality.

Lemma 5.8. *Suppose that c is twice continuously differentiable and that the function*

$$g_i := f_{U_1, U_i} - f_{U_1} f_{U_i}$$

is uniformly bounded for all i by a non-random constant. Let $h \to 0$ and $n^{-1} h^{-d/2} \to 0$ as $n \to \infty$. For an interior point $u \in (0,1)^d$, it holds that

$$E[\hat{c}(u)] = c(u) + h \left(\sum_{j=1}^{d} (1 - 2u_j) \frac{\partial c(u)}{\partial u_j} + \frac{1}{2} u_j (1 - u_j) \frac{\partial^2 c(u)}{\partial^2 u_j} \right) + o(h), \ \textit{as } n \to \infty.$$

Moreover, for an interior point $u \in (0,1)^d$, it holds that

$$Var(\hat{c}(u)) = \frac{c(u)}{n h^{d/2} (4\pi)^{d/2} \prod_{j=1}^{d} \sqrt{u_j (1 - u_j)}} + o((n h^{d/2})^{-1}), \ \textit{as } n \to \infty.$$

Proof of Lemma 5.8. We will start with the derivation of the bias term:

$$E[\hat{c}(u)] = E\left[\prod_{j=1}^{d} K_B\left(U_{1j}, \frac{u_j}{h}+1, \frac{1-u_j}{h}+1\right)\right]$$

$$= \int_{[0,1]^d} \prod_{j=1}^{d} K_B\left(y_j, \frac{u_j}{h}+1, \frac{1-u_j}{h}+1\right) c(y_1, ..., y_d) d(y_1, ..., y_d) = E[c(Y)],$$

where $Y = (Y_1, ..., Y_d)$ consists of independent and Beta distributed random variables with

$$Y_j \overset{\mathcal{D}}{=} B\left(\frac{u_j}{h}+1, \frac{1-u_j}{h}+1\right).$$

Now we will make use of a Taylor-Expansion around $\mu := (\mu_1, ..., \mu_d)$, where $\mu_i := E[Y_i]$:

$$c(y_1, .., , y_d) = c(\mu) + \sum_{j=1}^{d}(y_j - \mu_j)\frac{\partial c(\mu)}{\partial u_j} + \frac{1}{2}\sum_{j=1}^{d}(y_j - \mu_j)^2\frac{\partial^2 c(\mu)}{\partial^2 u_j}$$

$$+ \frac{1}{2}\sum_{1\leq j\neq l\leq d}(y_j - \mu_j)(y_l - \mu_l)\frac{\partial c(\mu)}{\partial u_j}\frac{\partial c(\mu)}{\partial u_l} + o(h^2), \text{ as } n \to \infty.$$

Taking the expectation on both sides yields to

$$E[c(Y)] = c(\mu_1, ..., \mu_d) + \frac{1}{2}\sum_{j=1}^{d} Var(Y_j)\frac{\partial^2 c(\mu)}{\partial^2 u_j} + o(h), \text{ as } n \to \infty.$$

Note that $Var(Y_j) = hu_j(1-u_j)$ and make use of a Taylor-Expansion around $u = (u_1, ..., u_d)$ such that

$$c(\mu_1, ..., \mu_d) = c(u_1, ..., u_d) + \sum_{j=1}^{d}(\mu_j - u_j)\frac{\partial c(u)}{\partial u_j} + \frac{h}{2}\sum_{j=1}^{d} u_j(1-u_j)\frac{\partial^2 c(u)}{\partial^2 u_j} + o(h)$$

$$= c(u) + \sum_{j=1}^{d}(1-2u_j)\frac{\partial c(u)}{\partial u_j} + \frac{h}{2}\sum_{j=1}^{d} u_j(1-u_j)\frac{\partial^2 c(u)}{\partial^2 u_j} + o(h), \text{ as } n \to \infty.$$

We will now derive the variance as follows:

$$
\begin{aligned}
&Var(\hat{c}(u)) = Var\left(\frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d} K_B\left(U_{ij}, \frac{u_j}{h}+1, \frac{1-u_j}{h}+1\right)\right)\\
&= \frac{1}{n}Var\left(\prod_{j=1}^{d} K_B\left(U_{1j}, \frac{u_j}{h}+1, \frac{1-u_j}{h}+1\right)\right)\\
&\quad + \frac{2}{n}\sum_{i=1}^{n-1}(1-\frac{i}{n})\\
&Cov\left(\prod_{j=1}^{d} K_B\left(U_{1j}, \frac{u_j}{h}+1, \frac{1-u_j}{h}+1\right), \prod_{j=1}^{d} K_B\left(U_{i+1j}, \frac{u_j}{h}+1, \frac{1-u_j}{h}+1\right)\right)\\
&=: A + B
\end{aligned}
$$

Using the Davydov-inequality for α-mixing time series, we observe that

$$
\begin{aligned}
&\left|Cov\left(\prod_{j=1}^{d} K_B\left(U_{1j}, \frac{u_j}{h}+1, \frac{1-u_j}{h}+1\right), \prod_{j=1}^{d} K_B\left(U_{i+1j}, \frac{u_j}{h}+1, \frac{1-u_j}{h}+1\right)\right)\right|\\
&\le 4\alpha(i)\left|\left|\prod_{j=1}^{d} K_B\left(U_{1j}, \frac{u_j}{h}+1, \frac{1-u_j}{h}+1\right)\right|\right|_{\infty}^{2}\\
&\overset{(*)}{\le} 4\alpha(i)O((h^{-d/2})) \le C\rho^{i}h^{-d},
\end{aligned}
$$

where inequality $(*)$ is derived via Stirling´s formula and from the fact that x is the mode of the Beta kernels; see Chen(2000), p. 88.
Furthermore, because $||g_i||_\infty \le C_2$, observe that

$$
\begin{aligned}
&\left|Cov\left(\prod_{j=1}^{d} K_B\left(U_{1j}, \frac{u_j}{h}+1, \frac{1-u_j}{h}+1\right), \prod_{j=1}^{d} K_B\left(U_{i+1j}, \frac{u_j}{h}+1, \frac{1-u_j}{h}+1\right)\right)\right|\\
&\le ||g_{i+1}||_{\infty}\\
&\int_{[0,1]^d}\prod_{j=1}^{d} K_B\left(s_{1j}, \frac{u_j}{h}+1, \frac{1-u_j}{h}+1\right)\int_{[0,1]^d}\prod_{j=1}^{d} K_B\left(\tilde{s}_{i+1j}, \frac{u_j}{h}+1, \frac{1-u_j}{h}+1\right)dsd\tilde{s}\\
&\le C_2
\end{aligned}
$$

Now conclude that the covariances are bounded by

$$|B| \leq$$

$$\leq \frac{2}{n}\sum_{i=1}^{\lfloor T_n \rfloor}\left|Cov\left(\prod_{j=1}^{d} K_B\left(U_{1j}, \frac{u_j}{h}+1, \frac{1-u_j}{h}+1\right), \prod_{j=1}^{d} K_B\left(U_{i+1j}, \frac{u_j}{h}+1, \frac{1-u_j}{h}+1\right)\right)\right|$$

$$+\frac{2}{n}\sum_{i=\lfloor T_n \rfloor+1}^{n-1}$$

$$\left|Cov\left(\prod_{j=1}^{d} K_B\left(U_{1j}, \frac{u_j}{h}+1, \frac{1-u_j}{h}+1\right), \prod_{j=1}^{d} K_B\left(U_{i+1j}, \frac{u_j}{h}+1, \frac{1-u_j}{h}+1\right)\right)\right|$$

$$\leq \frac{2C_2T_n}{n}+\frac{8C}{nh^d}\sum_{i=\lfloor T_n \rfloor+1}^{n-1}\rho^i,$$

where T_n is a sequence such that $T_n \to \infty$ as $n \to \infty$ and $T_n = O(h^{-\varkappa})$, $\varkappa \in (0, d/2)$. Hence, we can deduce that $B = o(n^{-1}h^{-d/2})$, because $\sum_{i=\lfloor T_n \rfloor+1}^{n-1}\rho^i = o(1)$ as $n \to \infty$. Now examine the first term A:

$$\begin{aligned} A &= \frac{1}{n}Var\left(\prod_{j=1}^{d} K_B\left(U_{1j}, \frac{u_j}{h}+1, \frac{1-u_j}{h}+1\right)\right) \\ &= \frac{1}{n}\left(E\left[\left(\prod_{j=1}^{d} K_B\left(U_{1j}, \frac{u_j}{h}+1, \frac{1-u_j}{h}+1\right)\right)^2\right]\right)+O(n^{-1}) \\ &:= \frac{1}{n}\prod_{j=1}^{d} A_j(u_j;h)E[c(Z)]+O(n^{-1}), \end{aligned}$$

where $Z = (Z_1, ..., Z_d)$ is a random vector of independent random variables such that $Z_j \overset{\mathcal{D}}{=} B\left(\frac{2u_j}{h}+1, \frac{2(1-u_j)}{h}+1\right)$ and

$$A_j(u_j;h) := \frac{B\left(\frac{2u_j}{h}+1, \frac{2(1-u_j)}{h}+1\right)}{\left(B\left(\frac{u_j}{h}+1, \frac{(1-u_j)}{h}+1\right)\right)^2},$$

where $B(\alpha, \beta)$ denotes the Beta-function. The asymptotic behavior of $A_j(u_j;h)$ as $h \to 0$ was investigated in Chen (2000), p. 86:

$$A_j(u_j;h) = \begin{cases} \frac{1}{2\sqrt{\pi u_j(1-u_j)h}}, & \text{if } u_j/h \to \infty \text{ and } (1-u_j)/h \to \infty, \\ \frac{\Gamma(2\kappa+1)}{2^{1+2\kappa}\Gamma^2(\kappa+1)}, & \text{if } u_j/h \to \kappa \text{ or } (1-u_j)/h \to \kappa, \end{cases}$$

as $h \to 0$ and κ is a positive constant.
Finally, a first order Taylor-expansion yields to

$$E[c(Z)] = c(u) + O(h), \text{ as } n \to \infty.$$

Therefore, we can conclude that, for an interior point u,

$$Var(c(u)) = \frac{c(u)}{nh^{d/2}\sqrt{(4\pi)^d \prod_{j=1}^d u_j(1-u_j)}} + o(n^{-1}h^{-d/2}), \text{ as } n \to \infty.$$

□

The variance at a generic point $u \in (0,1)^d$ is proportional according to

$$Var\left(\hat{c}(u)\right) = O\left(n^{-1}\prod_{l=1}^{d} h^{-(1/2+1/2\cdot 1_l)}\right),$$

which depends on the location of the components. Moreover, $1_l := 1(x_l/h \to \kappa_l > 0)$ is a function indicating whether a component lies in the boundary region; see also the proposed MBC techniques in section 4.3.
We will now derive the asymptotic normality of the proposed estimator. For this purpose, we make use of the big-block and small-block technique, which we already used for the adaptive version of the Nadaraya-Watson estimator.

Proposition 5.9. *Suppose that c is twice continuously differentiable and suppose that X_i is an α-mixing time series of d-dimensional random vectors with mixing coefficients*

$$\alpha(k) \leq \rho^k, \ \rho \in (0,1).$$

If $h = O(n^{-2/(d+4)})$ it holds that

$$\sqrt{nh^{d/2}}(\hat{c}(u) - E[\hat{c}(u)]) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2(u)), \ \textit{as } n \to \infty$$

for $u \in (0,1)^d$ and where $\sigma^2(u) := \frac{c(u)}{\sqrt{(4\pi)^d \prod_{j=1}^d u_j(1-u_j)}}$. If $h = o(n^{-2/(d+4)})$, then

$$\sqrt{nh^{d/2}}(\hat{c}(u) - c(u)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2(u)), \ \textit{as } n \to \infty.$$

Proof. The second statement is a direct consequence of the first one by considering the fact that

$$Bias(\hat{c}(u)) = O(h), \text{ as } n \to \infty.$$

Hence, we only state proof for the first statement. Now define

$$
\begin{aligned}
& n^{1/2}h^{d/4}\left(\frac{\hat{c}(u)-E[\hat{c}(u)]}{\sqrt{\sigma^2(u)}}\right) \\
&= \frac{\frac{n^{1/2}h^{d/4}}{n}\sum_{i=1}^{n}\left(\prod_{j=1}^{d}K_B\left(U_{ij},\frac{u_j}{h}+1,\frac{1-u_j}{h}+1\right)-E\left[\prod_{j=1}^{d}K_B\left(U_{ij},\frac{u_j}{h}+1,\frac{1-u_j}{h}+1\right)\right]\right)}{\sqrt{\sigma^2(u)}} \\
&:= n^{-1/2}h^{d/4}\sum_{i=1}^{n}Y_i := n^{-1/2}\xi_n(u).
\end{aligned}
$$

We will make use of the aforementioned big-block and small-block technique. Therefore, decompose the sum in alternate big and small blocks and a remainder block.
Let $i = 1, ..., r$ and consider the big blocks

$$V_i := h^{d/4}(Y_{(i-1)(p+q)+1} + ... + Y_{ip+(i-1)q}),$$

the small blocks

$$V_i^* := h^{d/4}(Y_{ip+(i-1)q+1} + ... + Y_{i(p+q)}),$$

and finally the remainder block: for $r(p+q) \le n \le r(p+q+1)$:

$$R := \sum_{i=r(p+q)}^{n} Y_i.$$

Therefore, we decompose the sum in

$$\xi_n(u) = \sum_{i=1}^{r} V_i + \sum_{i=1}^{r} V_i^* + R$$

At first, we show that

$$n^{-1/2}(\sum_{i=1}^{r} V_i^* + R) \xrightarrow{P} 0, \text{ as } n \to \infty.$$

For this purpose, it suffices to show that

$$\frac{1}{n}Var\left(\sum_{i=1}^{r} V_i^*\right) = o(1) \text{ and } \frac{1}{n}Var\left(\sum_{i=r(p+q)}^{n} Y_i\right) = o(1), \text{ as } n \to \infty.$$

Now choose $r = O(n^a)$, $p = O(n^{1-a})$, and $q = O(n^c)$, $a \in (0,1)$, $c \in (0, 1-a)$ and start with the small blocks:

$$Var\left(\sum_{i=1}^{r} V_i^*\right) = \sum_{i=1}^{r} Var(V_i^*) + \sum_{1\le i\ne j\le r} Cov(V_i^*, V_j^*) := I + II.$$

We will start with I. Under the stationary condition and the previous computations concerning the covariances, it holds that

$$\begin{aligned}
I &= \sum_{i=1}^{r} Var\left(\sum_{l=ip+(i-1)q+1}^{i(p+q)} h^{d/4}Y_l\right) \\
&= \sum_{i=1}^{r}\left(\sum_{l=ip+(i-1)q+1}^{i(p+q)} Var(h^{d/4}Y_l) + \sum_{ip+(i-1)q+1\le l\neq m\le i(p+q)} Cov(h^{d/4}Y_l, h^{d/4}Y_m)\right) \\
&= \sum_{i=1}^{r}\left(qh^{d/2}Var(Y_1) + h^{d/2}\sum_{ip+(i-1)q+1\le l\neq m\le i(p+q)} Cov(Y_l, Y_m)\right) \\
&= r\left(q + h^{d/2}\sum_{p+1\le l\neq m\le p+q} Cov(Y_l, Y_m)\right) \\
&= O(rq) + o(r) = O(n^{a+c}) + o(n^a), \text{ as } n\to\infty.
\end{aligned}$$

For the second summand II, we again make use of the knowledge that the covariances are asymptotically negligible:

$$\begin{aligned}
&\sum_{1\le i\neq j\le r} Cov(V_i^*, V_j^*) \\
&= \sum_{1\le i\neq j\le r} h^{d/2}\sum_{l=ip+(i-1)q+1}^{i(p+q)}\ \sum_{m=jp+(j-1)q+1}^{j(p+q)} Cov(Y_l, Y_m) = o(n), \text{ as } n\to\infty.
\end{aligned}$$

Therefore, the small blocks converge to zero in probability:

$$\frac{1}{n}Var\left(\sum_{i=1}^{r} V_i^*\right) = O(n^{a+c-1}) + o(n^{a-1}) + o(1) \to 0, \text{ as } n\to\infty.$$

The remainder terms can analogously be treated. We omit the proof and approximate $\xi_n(u)$ by

$$n^{-1/2}\xi_n = n^{-1/2}\sum_{i=1}^{r} V_i + o_P(1).$$

Now use Lemma 3.9 to conclude that

$$\begin{aligned}
&\left|E\left[\exp\left(it\sum_{i=1}^{r} V_i\right)\right] - \prod_{i=1}^{r} E[\exp(itV_i)]\right| \le 16(r-1)\alpha(q+1) \\
&\le 16r\alpha(q+1) = 16n^a\rho^{q+1} = 16n^a\rho^{n^c+1} = 16\rho n^a\exp(\log(\rho)n^c) = o(1),\ n\to\infty
\end{aligned}$$

and apply Lyapunov´s central limit theorem for $\delta = 1$:

$$\begin{aligned}
&\frac{\sum_{i=1}^{r} E[|V_i|^3]}{(rVar(V_1))^{3/2}} \leq ||V_i||_\infty (rVar(V_1))^{-1/2} \\
&\leq h^{d/4} p ||Y_1||_\infty \left(r \left(\sum_{j=1}^{p} Var(h^{d/4} Y_j) + h^{d/2} \sum_{1 \leq j \neq l \leq p} Cov(Y_j, Y_l) \right) \right)^{-1/2} \\
&\lesssim h^{d/4} p h^{-d/2} (n^a n^{1-a})^{-1/2} = O\left(n^{\frac{d+2}{d+4} - a} \right).
\end{aligned}$$

Now choose a such that $0 < \frac{d+2}{d+4} < a < 1$. For this choice, the Lyapunov condition is fulfilled and it holds that

$$n^{-1/2} \sum_{i=1}^{r} V_i \xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \text{ as } n \to \infty.$$

□

5.5 Bernstein copula density estimator

In this section, we will briefly introduce the concept of Bernstein copulas and corresponding estimators for copula densities. The Bernstein copula was defined by Sancetta and Satchell (2004) and is used for the nonparametric estimation of a d-dimensional copula C. To motivate the estimation procedure, we will focus on a multivariate version of the uniform approximation of a continuous function by a sequence of Bernstein polynomials.

Proposition 5.10. *Let $f : [0,1]^d \to \mathbb{R}$ be a continuous function and define*

$$B_{f,k_1,\dots,k_d}(x_1,..,x_d) := \sum_{\nu_1=0}^{k_1} \dots \sum_{\nu_d=0}^{k_d} f(\nu_1/k_1, \dots, \nu_d/k_d) \prod_{j=1}^{d} \binom{k_j}{\nu_j} x_j^{\nu_j} (1-x_j)^{n_j - \nu_j}.$$

Now the following approximation holds for all $(x_1, \dots, x_d) \in [0,1]^d$:

$$f(x_1, \dots, x_d) = \lim_{k_1 \to \infty, \dots, k_d \to \infty} B_{f,k_1,\dots,k_d}(x_1,..,x_d).$$

Due to the fact that every copula C is continuous (i.e. C is Lipschitz-continuous; see Nelsen (1999)), we can make use of this approximation as a starting point for the construction of a nonparametric estimator for C. Moreover, because we now smooth the estimator by the use of Bernstein polynomials

$$P_{k_j,\nu_j}(x_j) = \binom{k_j}{\nu_j} x_j^{\nu_j} (1-x_j)^{k_j - \nu_j},$$

this leads us to an alternative estimator of a copula density c by differentiation. Now let $X = (X_1, ..., X_d)$ be a random vector possessing the unknown copula C. By substitution of C via its empirical counterpart $\hat{C}$ (see (5.20)), we derive the empirical Bernstein copula according to

$$\hat{C}_B(u_1, ..., u_d) = \sum_{\nu_1=0}^{k_1} ... \sum_{\nu_d=0}^{k_d} C_n(\nu_1/k_1, ..., \nu_d/k_d) \prod_{j=1}^{d} P_{k_j,\nu_j}(u_j).$$

By assuming that C possesses a density c, we derive the Bernstein copula density estimator of c_B by differentiation according to

$$\hat{c}_B(u_1, ..., u_d) = \sum_{\nu_1=0}^{k_1} ... \sum_{\nu_d=0}^{k_d} C_n(\nu_1/k_1, ..., \nu_d/k_d) \prod_{j=1}^{d} P'_{k_j,\nu_j}(u_j),$$

where

$$P'_{k_j,\nu_j}(u_j) = k_j(P_{k_j-1,\nu_j-1}(u_j) - P_{k_j,\nu_j-1}(u_j))$$

by the well-known properties of the Bernstein polynomials; see Lorentz (1986) for an appealing survey of certain properties of this class of polynomials.
Using the upper identity of the derivative of a Bernstein polynomial, we are able to derive an alternative form of the Bernstein copula density estimator by setting $k_1 = ... = k_d \equiv k$ as follows:

$$\hat{c}_B(u_1, ..., u_d) = \frac{k^d}{n} \sum_{i=1}^{n} \sum_{\nu_1=0}^{k-1} ... \sum_{\nu_d=0}^{k-1} 1(S_i \in B_\nu) \prod_{j=1}^{d} P_{k-1,\nu_j}(u_j),$$

where

$$S_i := (\hat{F}_1(X_{i1}), ..., \hat{F}_d(X_{id}))$$

and

$$B_v := \left[\frac{\nu_1}{k}, \frac{\nu_1 + 1}{k}\right] \times ... \times \left[\frac{\nu_d}{k}, \frac{\nu_d + 1}{k}\right].$$

The Bernstein copula estimator has been studied by Janssen et al. (2012), where the strong consistency and the asymptotic normality for i.i.d. d-dimensional random vectors $X_i,\ i = 1, ..., n$, were derived. The Bernstein copula density estimator has been studied by Bouezmarni et al. (2010) in the case of α-mixing data under the assumption that the marginal distributions are known. They derived the consistency and the asymptotic normality, too. In fact, we will present the derived rates for the asymptotic bias and variance below.

Proposition 5.11 (Proposition 1 and 2, Bouezmarni et al. (2010)). *Let $X = \{(X_{i1}, ..., X_{id})\}$, $i = 1, ..., n$, be a sample of n observations of α-mixing random vectors with unknown corresponding copula C, copula density c, joint distribution F, and joint density f. Assume that the mixing coefficients fulfill*

$$\alpha(k) \leq \rho^k,\ \rho \in (0, 1).$$

Moreover, assume that the copula density is twice continuously differentiable and fix a vector $u = (u_1, ..., u_d) \in (0,1)^d$. *If* $k \to \infty$ *we have for known marginals* $F_k,\ k = 1, ..., d$, *that*

$$E[\hat{c}_B(u)] = c(u) + k^{-1}\frac{1}{2}\sum_{j=1}^{d}\left(\frac{\partial c}{\partial u_j}(1-2u_j) + \frac{\partial^2 c}{\partial^2 u_j}u_j(1-u_j)\right) + o(k^{-1}).$$

For the derivation of the asymptotic variance, we further assume that

$$||g_j||_\infty := ||f_{X_1,X_j} - f_{X_1}f_{X_j}||_\infty \leq C,$$

where f_{X_1,X_j} *denotes the joint density of the random vectors* X_1 *and* X_j, $j = 2, ..., n$. *In addition, let the regularization parameter* k *fulfill* $n^{-1}k^{d/2} \to 0$ *as* $n \to \infty$. *Then, we have that*

$$Var(\hat{c}_B(u)) = n^{-1}k^{d/2}\frac{c(u)}{(4\pi)^{d/2}\prod_{j=1}^{d}(u_j(1-u_j))^{1/2}} + o(n^{-1}k^{d/2}),\ \text{as } n \to \infty.$$

Comparing the asymptotic properties of the Beta kernel and the Bernstein polynomial based estimator, we see that the variance is actually the same, whereas the bias differs only in a slightly manner. Nevertheless, the rates coincide for both components.
In a recent paper, Janssen et al. (2014) derived the asymptotic properties of the Bernstein copula density estimator, where the marginals are estimated via their empirical counterparts. In the case of i.i.d. data, they derived the asymptotic distribution and the exact representation of the bias term. For α-mixing data, the corresponding asymptotic properties are not derived, yet.

5.6 Nonparametric estimation approaches via copula based representations

In this section, we will describe how the presented copula density estimators can be used for the estimation of joint densities, conditional densities, and conditional expectations. For this purpose, consider the theorem of Sklar again, which states that

$$C(F_1(x_1), ..., F_d(x_d)) = F(x_1, ..., x_d)$$

for a d-dimensional random vector X possessing the multivariate joint distribution function F with marginals F_k and a corresponding copula C. Now let C be twice differentiable such that the joint density f can be represented by

$$f(x_1, ..., x_d) = c(F_1(x_1), ..., F_d(x_d))\prod_{j=1}^{d} f_j(x_j),$$

where f_j denotes the marginal density of the j-th component of X. The nonparametric estimation of joint densities suffers from the curse of dimensionality. This effect describes that the rate of convergence decreases as the dimension d increases. Liebscher (2005) used the upper representation of f to construct a semiparametric estimator $\hat{f}$. Liebscher assumes that the corresponding copula density c is known to belong to a parametric class of copulas and is, hence, known up to a finite dimensional parameter ϑ. Due to the fact that the unknown parameter ϑ can be $\sqrt{n}$-consistently estimated, the proposed estimator

$$\hat{f}(x_1, ..., x_d) = c(\hat{F}_1(x_1), ..., \hat{F}_d(x_d); \hat{\vartheta}) \prod_{j=1}^{d} \hat{f}(x_j)$$

overcomes the curse of dimensionality. Nevertheless, in many practical situations, the knowledge of a corresponding parametric family of copulas cannot be guaranteed. A recent work by Dette et al. (2014) illustrates with rather simple examples, which kind of impact a misspecification of the copula family has.

Now consider the case $d = 2$ and a bivariate random vector (Y, X). The conditional density $f_{Y|X=x}(y, x)$ can be represented by

$$f_{Y|X=x}(y, x) = \frac{f_{Y,X}(x, y)}{f_X(x)} = c(G(y), F(x))g(y),$$

where $Y \sim G$ and $X \sim F$. Moreover, c denotes the copula density as well as f and g denote the marginal densities. In Faugeras (2009), this representation is used as a starting point for the construction of a nonparametric estimator of the conditional density. Faugeras (2009) called this the "quantile-copula approach" and estimated the copula density by means of symmetric kernels, which cause an additional bias near the boundaries. This was also mentioned in Faugeras (2009) in Remark 4. Hence, in order to overcome this drawback, we propose an alternative estimator $\hat{f}_{Y|X=x}(y, x)$ based on Bernstein polynomials in view of the previous findings. Therefore, define the Bernstein polynomial based conditional density estimator $\hat{f}_{Y|X=x}(y, x)$ as follows:

$$\begin{aligned}\hat{f}_{Y|X=x}(y, x) &= \hat{c}_B(\hat{G}(y), \hat{F}(x))\hat{g}(y)\\ &:= \sum_{k=0}^{m}\sum_{l=0}^{m} C_n\left(\frac{k}{m}, \frac{l}{m}\right) P'_{mk}(\hat{G}(y))P'_{ml}(\hat{F}(x))\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{y - Y_i}{h}\right),\end{aligned}$$

where we chose an ordinary density estimator for the estimation of g. This estimator can also be replaced by a Gamma kernel estimator or a Beta kernel estimator, whether the support is bounded or not. This copula based representation possesses some appealing advantages over the ordinary representation

$$f_{Y|X=x}(y, x) = \frac{f_{Y,X}(y, x)}{f_X(x)}.$$

In fact, the quotient-shaped plug-in estimator suffers from the fact that it may shows an explosive behavior in numerical implementations as the denominator too small. Moreover, from a theoretical point of view, the marginal density has to be bounded away from zero at x to guarantee that the quotient is well-defined. Furthermore, the upper estimator overcomes these problems and is free of boundary bias, too.

We will omit the exact proof of the asymptotic properties and only remark that the consistency as well as the derivation of the asymptotic distribution are based on the following decomposition:

$$\begin{aligned}
&\hat{f}_{Y|X=x}(y,x) - f_{Y|X=x}(y,x) = \hat{c}_B(\hat{G}(y),\hat{F}(x))\hat{g}(y) - c(G(y),F(x))g(y) \\
&= (\hat{g}(y)-g(y))\hat{c}_B(\hat{G}(y),\hat{F}(x)) + g(y)(\hat{c}_B(\hat{G}(y),\hat{F}(x)) - c(G(y),F(x))) \\
&= (\hat{g}(y)-g(y))(\hat{c}_B(\hat{G}(y),\hat{F}(x)) - \hat{c}_B(G(y),F(x))) \\
&\quad + (\hat{g}(y)-g(y))(\hat{c}_B(G(y),F(x)) - c(G(y),F(x))) + (\hat{g}(y)-g(y))c(G(y),F(x)) \\
&\quad + g(y)(\hat{c}_B(\hat{G}(y),\hat{F}(x)) - \hat{c}_B(G(y),F(x))) + g(y)(\hat{c}_B(G(y),F(x)) - c(G(y),F(x))) \\
&:= \sum_{i=1}^{5} c_{B,i}(y,x).
\end{aligned}$$

Applying the consistency result of Janssen et al. (2014) to the terms $c_{B,2}(y,x)$ and $c_{B,5}(y,x)$ as well as the well-known rates of the ordinary kernel density estimator to the terms $c_{B,1}(y,x)$, $c_{B,2}(y,x)$, and $c_{B,3}(y,x)$, yields to the desired consistency as

$$m \to \infty,\ h \to 0,\ nm^{-1} \to \infty,\ nh \to \infty, \text{ as } n \to \infty.$$

Moreover, for the derivation of this result, the continuity of the Bernstein polynomials as well as the consistency of the empirical distribution function are used for the terms $c_{B,1}(y,x)$ and $c_{B,4}(y,x)$. Furthermore, under additional smoothness assumptions on c and g, the asymptotic normality can be derived, if m increases not too fast. In fact, this means that $m = o(n^{1/2})$; see Janssen et al. (2014) for further details.

To illustrate the upper results, we present the following figure. We simulated $n = 100$ independent copies of the two-dimensional random vector (X,Y) such that the marginals are given by

$$X_i \sim \mathcal{N}(0,1) \text{ and } Y_i \sim \mathcal{N}(0,1), \quad \forall\, i = 1,...,n.$$

Moreover, let the corresponding copula of the random vector (X,Y) be given by the Frank copula

$$C_{\theta,Fr}(u,v) = \frac{-1}{\theta}\log\left(1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right)$$

with parameter $\theta = 5.7$, which corresponds to Kendalls τ of 0.5. We estimated the conditional density $f_{Y|X=x}(y,x)$ via two different ways. At first, we used the ordinary quotient-shaped approach. It turns out, that this estimator is sensitive in such a way that it is unbounded in some areas, where the denominator gets too small. In contrast, the Bernstein polynomial based estimator does not suffer from this effect, although its resulting estimate is not really smooth anymore. The latter fact rises due to the construction via a smoothed version of the empirical copula function. Nevertheless, it can be seen that the copula based estimator provides better results in this case.

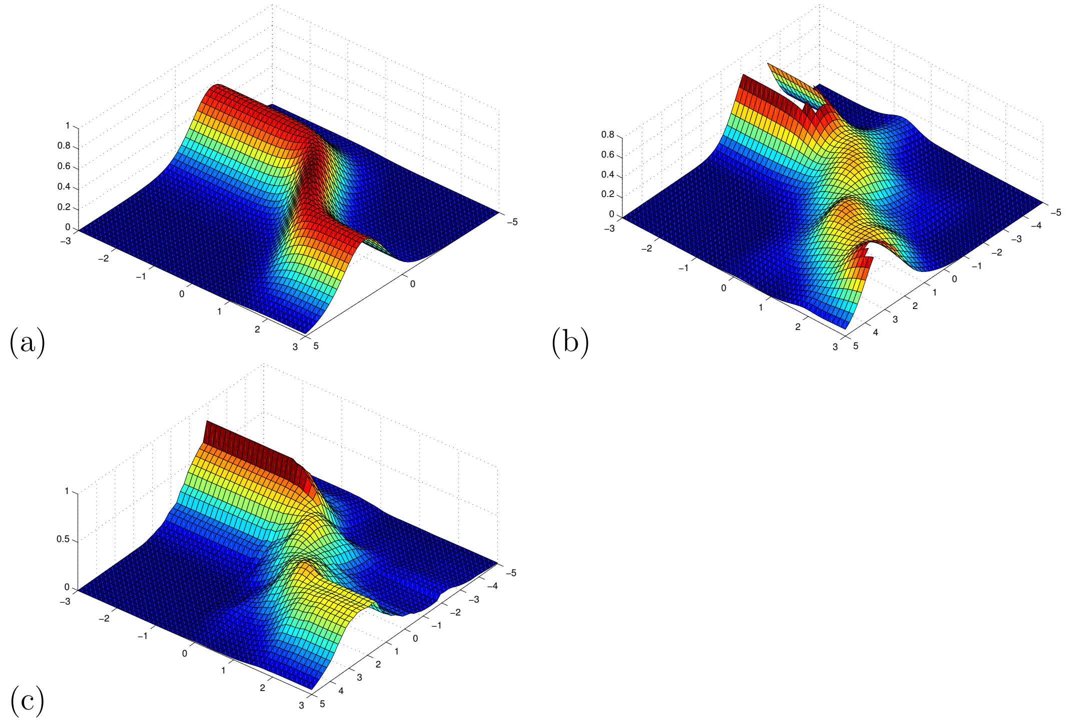

Figure 7: Surface plots of estimated conditional density $f_{Y|X=x}(y,x)$. (a): True density ,(b): Ordinary symmetrical kernel based estimator. (c): Bernstein polynomial based estimator.

Finally, we will focus on the estimation of conditional expectations relying on the corresponding copula representation. By taking the upper findings into account, the conditional moment of a random variable Y given $X = x \in \mathbb{R}$ can be represented as follows:

$$\begin{aligned} m(x) := E[Y|X = x] &= \int_{\mathbb{R}} y f_{Y|X=x}(y,x) dy \\ &= \int_{\mathbb{R}} y c(G(y), F(x)) g(y) dy = E[Y c(G(Y), F(x))]. \end{aligned}$$

In view of Liebscher (2005), Noh et al. (2013) proposed a regression estimator $\hat{m}(x)$ based on i.i.d. random variables according to this representation as

$$\hat{m}(x) = \int_{\mathbb{R}} yc(\hat{G}(y), \hat{F}(x), \hat{\vartheta})d\hat{G}(y) = \frac{1}{n}\sum_{i=1}^{n} Y_i c(\hat{G}(Y_i), \hat{F}(x), \hat{\vartheta}),$$

where $\hat{\vartheta}$ denotes the maximum likelihood estimator based on the pseudo observations $(\hat{G}(Y_i), \hat{F}(X_i))$. Noh et al. (2013) derived a stochastic expansion of this estimator and, thus, deduced the consistency as well as the asymptotic normality of $\hat{m}(x)$ under common regularity assumptions on c and $E[|Y|] < \infty$.

Example 5.12. *If* $(G(Y), F(X)) \sim$ *Gaussian copula with parameter* ρ_1*,* $Y \overset{\mathcal{D}}{=} \mathcal{N}(\mu_Y, \sigma_Y^2)$*, and* $X \overset{\mathcal{D}}{=} \mathcal{N}(\mu_X, \sigma_X^2)$*, then*

$$m(x) = E[Y|X = x] = \mu_Y + \sigma_Y \rho_1 \Phi^{-1}(F(x)).$$

For further examples see Noh et al. (2013), Section 5.

In the already mentioned article by Dette et al. (2014), it is illustrated that this estimator reveals extremely dissatisfying results, provided that the parametric copula family is not known a priori.

This observation motivates the construction of a fully nonparametric estimator $\hat{m}(x)$ of $m(x)$ according to

$$\hat{m}(x) := \frac{1}{n}\sum_{i=1}^{n} Y_i \hat{c}_B(\hat{G}(Y_i), \hat{F}(x)),$$

where c_B denotes either the Bernstein polynomial based or the Beta kernel based copula density estimator.

The derivation of the asymptotic properties is sophisticated due to the twofold estimation of the unknown distribution of Y and the unknown copula density c.

To create a link to the first part of this thesis, it would be interesting, if the class of these estimators could be used for the meaningful estimation of coefficients of diffusion processes by the approximation via conditional expectations. Hence, when observing a high-frequency sample

$$X_0, X_\Delta, ..., X_{n\Delta}$$

of a univariate ordinary diffusion driven by a Brownian motion, a potential estimator of the drift function would be given by

$$\hat{b}_C(x) = \frac{1}{n\Delta}\sum_{i=0}^{n-1} Y_{(i+1)\Delta} \hat{c}_B\left(\hat{F}_Y(Y_{(i+1)\Delta}), \hat{F}_X(x)\right),$$

where

$$Y_{(i+1)\Delta} := X_{(i+1)\Delta} - X_{i\Delta}$$

and $\hat{F}_Y(y) := \frac{1}{n}\sum_{i=0}^{n-1} 1(Y_{i\Delta} \leq y)$ denotes the empirical distribution function of the increments of the process X.
Due to the fact that, even in the discrete time analysis, only the results for i.i.d. data are available, we think that a generalization towards continuous-time processes requires a lot of additional effort. Especially, the generalization to dependent data should be a challenging task.

6 Conclusion

In this section, we shortly want to summarize the topics and results of the present thesis. The main subject of the first part was the nonparametric estimation of the drift function in a jump diffusion model. In particular, the process under investigation was driven by a pure jump Lévy process and a Brownian motion, which were independent. Motivated by approximations of the infinitesimal generator, we constructed a pointwise kernel estimator of the unknown drift function via techniques which originate from classical regression analysis in discrete time. Our findings are coherent to those of Bandi and Phillips (2003) as well as Bandi and Nguyen (2003), where analogous estimators were investigated in a diffusion model and a jump diffusion model where the driving process possesses finite activity. In particular, we derived the weak consistency and the asymptotic normality of the proposed estimator. In view of the mentioned articles, a possible extension would be given by more general drivers, such that the solution of the considered stochastic differential equation is still a semimartingale.
Moreover, we analyzed the case where one only observes a noise-contaminated sample. Using the pre-averaging approach, firstly proposed in Podolskij and Vetter (2006), we were able to handle this data adequately and derived the corresponding asymptotic properties. In view of Comte et al. (2010) as well as Kanaya and Kristensen (2015), the use of the proposed kernel estimators and the incorporation of noisy data in the context of stochastic volatility models would be interesting.

In the second part, we investigated several bias reduction techniques. The main aspect was the use of asymmetric kernels (see Chen (1999) and (2000)) in the context of nonparametric density and regression estimation. Our main contribution was the formulation of a multiplicative bias corrected multivariate nonparametric density estimator based on asymmetric kernels. This estimator overcomes the boundary bias effect and, moreover, possesses an optimal rate of convergence of the MSE which is proportional to $n^{-8/(8+d)}$, where d denotes the dimension of the underlying data set. This rate is considerably faster than $n^{-4/(4+d)}$, which can be derived for the classical product kernel based estimator using asymmetric kernels. We quantified the performance of the new proposed estimator in a short finite sample Monte Carlo study, which originates from Funke and Kawka (2015). In view of these considerations and additionally motivated by Gospodinov and Hirukawa (2012), the use of asymmetric kernels for the estimation of jump diffusion models and multivariate diffusions, would be an interesting further topic. We think that, from a practical point of view, asymmetric kernels should be invoked when estimating positive economic processes of interest, which possess a natural boundary at the origin.

In our last part, we focused on copulas and their nonparametrically guided estimation. Using Sklar´s theorem, an alternative product-shaped representation of conditional expectations involving the corresponding copula density can be derived. In view of this findings, an alternative estimator for conditional expectations can also be defined. According to this representation, it would be interesting, if these estimators can be used for the estimation of the coefficients of diffusion models, too.

Finally, we conclude that this thesis generalizes existing ideas for the nonparametric estimation of diffusions due to more general drivers. Moreover, it presents a couple of intentions how to expedite the presented techniques for the reasonable estimation of continuous-time stochastic processes by invoking data-specific properties like positivity or an accumulation of non-negative data near the origin.

7 References

Abramson, I.S. (1982). On Bandwidth Variation in Kernel Estimates- A Square Root Law. *The Annals of Statistics.* **10**, 4. 1217-1223.

Bandi, F.M., Phillips, P.C.B (2003). Fully Nonparametric Estimation of Scalar Diffusion Models. *Econometrica.* **71**. 241-283.

Bandi, F.M., Nguyen T. (2003). On the Functional Estimation of Jump-Diffusion Models. *Journal of Econometrics.* **116**. 293-328.

Bandi, F.M., Moloche, G. (2008). On the Functional Estimation of Multivariate Diffusion Processes. Under revision.

Bandi, F.M., Phillips, P.C.B (2009). Nonstationary Continuous-Time Processes. Handbook of Financial Econometrics. Elsevier Science.

Barndorff-Nielsen, O.E., Mikosch, T., Resnick, S. (Eds.) (2001). Lévy Processes: Theory and Applications. Birkhäuser, Basel.

Barndorff-Nielsen, O.E.,Shephard, N., Winkel, M. (2006). Limit Theorems for Multipower Variation in the Presence of Jumps. *Stochastic Processes and Their Applications.* **116**, 5. 796-806.

Behnen, K., Hušková, Neuhaus, G. (1985). Rank Estimators of Scores for Testing Independence. *Statistics and Decision.* **3**, 239-262.

Bernstein, S.N. (1927). Sur l´Extension du Théorème Limite du Calcul des Probabilités aux Sommes de Quantités Dependantes. *Mathematische Annalen.* **97**, 1-59.

Billingsley, P. (1995). Probability & Measure. 3rd Edition. John Wiley & Sons, New York.

Bouezmarni, T., Rombouts J.V.K. (2010). Nonparametric Density Estimation for Multivariate Bounded Data. *Journal of Statistical Planning and Inference.* **140**, No. 1 . 139-152.

Bouezmarni, T., Rombouts J.V.K., Taamouti, A. (2010). Asymptotic Properties of the Bernstein Density Copula Estimator for α-Mixing Data. *Journal of Multivariate Analysis.* **101**, 1-20.

Bücher, A., Volgushev, S. (2013). Empirical and Sequential Empirical Copula Processes Under Serial Dependence. *Journal of Multivariate Analysis.* **119**. 61-70.

Burman, P., Chow, E., Nolan D. (1994). A Cross-Validatory Method for Dependent Data. *Biometrika.* **81**. 351-358.

Cai, Z. (2001). Weighted Nadaraya-Watson Regression Estimation. *Statistics and Probability Letters.* **51**. 299-318.

CARR, P., GEMAN,H., MADAN, D.B., YOR, M. (2002). The Fine Structure of Asset Returns: An Empirical Investigation. *Journal of Business.* **75**, No. 2. 305-332.

CHARPENTIER, A., FERMANIAN, J.D., SCAILLET, O. (2006). The Estimation of Copulas: Theory and Practice. *In: Copulas: From Theory to Applications in Finance.* Risk Books, London. 35-62.

CHEN, S.X. (1999). Beta Kernel Estimators for Density Functions. *Computational Statistics & Data Analysis.* **31**. 131-145.

CHEN, S.X. (2000). Probability Density Function Estimation Using Gamma Kernels. *Annals of the Institute of Statistical Mathematics.* **52**. 471-480.

CHU, C.K., MARRON, J.S. (1991). Comparison of Two Bandwidth Selectors with Dependent Errors. *Annals of the Institute of Statistical Mathematics.* **19**. 1906-1918.

CLINE, D.B.H., HART, J.D. (1991). Kernel Estimation of Densities of Discontinuous Derivatives. *Statistics.* **22**. 69-84.

COMTE, F., GENON-CATALOT, V., ROZENHOLC, Y. (2007). Penalized Nonparametric Mean Square Estimation of the Coefficients of Diffusion Processes. *Bernoulli.* **13**, No. 2. 514-543.

COMTE, F., GENON-CATALOT, V., ROZENHOLC, Y. (2009). Nonparametric Adaptive Estimation for Integrated Diffusions. *Stochastic Processes and Their Applications.* **119**, 3. 811-834.

COMTE, F., GENON-CATALOT, V., ROZENHOLC, Y. (2010). Nonparametric Estimation for a Stochastic Volatility Model. *Finance and Stochastics.* **14**, No. 1. 49-80.

COMTE, F., GENON-CATALOT, V. (2012). Convolution Power Kernels for Density Estimation. *Journal of Statistical Planning and Inference.* **142**, 1698-1715.

CONT, R., TANKOV, P. (2004). Financial Modeling with Jump Processes. Chapman & Hall/ CRC Financial Mathematics Series, London.

COWLING, A., HALL, P. (1996). On Pseudo Data Methods for Removing Boundary Effects in Kernel Density Estimation. *Journal of the Royal Statistical Society Series B 58.* 551-563.

DEHEUVELS, P. (1979). La Fonction de Dépendance Empirique et ses Propriétés. *Académie Royale de Belgique. Bulletin de la Classe des Sciences.* **65**, 274-292.

DEMIR, S., TOKTAMIS, Ö. (2010). On the Adaptive Nadaraya-Watson Kernel Regression Estimators. *Hacettepe Journal of Mathematics and Statistics.* **39**, 3. 429-437.

DETTE, H., VAN HECKE, R., VOLGUSHEV, S. (2014). Some Comments on Copula-Based Regression. *Journal of the American Statistical Association.* **109**, 507. 1319-1324.

DITLEVSEN, S., SØRENSEN, M. (2004). Inference for Observations of Integrated Diffusion

Processes. *Scandinavian Journal of Statistics.* **31**, 417-429.

Doukhan, P. (1994). Mixing: Properties and Examples. Lecture Notes in Statistics. Springer, Berlin.

Doukhan, P., Fermanian, J.D., Lang, G. (2004). Copulas of a Vector-valued Stationary Weakly Dependent Process. Working Paper CREST.

Faugeras, O.P. (2009). A Quantile-Copula Approach to Conditional Density Estimation. *Journal of Multivariate Analysis.* **100**, 2083-2099.

Fermanian, J.D., Radulovic, D., Wegkamp, M. (2004). Weak Convergence of Empirical Copula Processes. *Bernoulli.* **10**, 847-860.

Fischer, W., Lieb, I. (2008). Funktionentheorie: Komplexe Analysis in Einer Veränderlichen. 9th Edition. Vieweg + Teubner, Wiesbaden.

Florens-Zmirous, D. (1993). On Estimating the Diffusion Coefficient from Discrete Observations. *Journal of Applied Probability.* **30**. 790-804.

Funke, B., Kawka, R. (2015). Nonparametric Density Estimation for Multivariate Bounded Data Using two Non-Negative Multiplicative Bias Correction Methods. To appear in *Computational Statistics & Data Analysis.* doi:10.2016/j.csda.2015.07.006.

Gijbels, I., Mielniczuk, J. (1990). Estimating the Density of a Copula Function. *Communications in Statistics: Theory and Methods.* **19**, 445-464.

Glotter, A. (2000). Discrete Sampling of an Integrated Diffusion Process and Parameter Estimation of the Diffusion Coefficient. *ESAIM: Probability and Statistics.* **4**. 205-227.

Glotter, A. (2006). Parameter Estimation for a Discretely Observed Integrated Process. *Scandinavian Journal of Statistics.* **33**. 83-104.

Glotter, A., Gobet, E. (2008). LAMN Property for Hidden Processes: The Case of Integrated Diffusions. *Annales de l´Institut Henri Poincaré, Probabilités et Statistiques.* **44**, No.1. 104-128.

Gospodinov, N., Hirukawa, M. (2007). Time Series Nonparametric Regression using Asymmetric Kernels with an Application to Estimation of Scalar Diffusion Processes. Unpublished manuscript available online at
http://alcor.concordia.ca/˜gospodin/research/npdiff.pdf.

Gospodinov, N., Hirukawa, M. (2012). Nonparametric Estimation of Scalar Diffusion Models of Interest Rates Using Asymmetric Kernels. *Journal of Empirical Finance.* **19**, 595-609.

Greenwood, P.E., Heckman, N., Lee, W., Wefelmeyer, W. (2015). Preaveraged kernel estimators for the drift function of a diffusion process in the presence of microstructure noise. Preprint.

Härdle, W., Marron, J.S. (1985). Optimal Bandwidth Selection in Nonparametric Regression Function Estimation. *The Annals of Statistics.* **13**, 4. 1465-1481.

Härdle, W. (1994). Applied Nonparametric Regression. New Rochelle, Cambridge.

Härdle, W., Müller, M., Sperlich, S., Werwatz, A. (2004). Nonparametric and Semiparametric Models. Springer, Berlin.

Hirukawa, M. (2009). Nonparametric Multiplicative Bias Correction for Kernel-type Density Estimation on the Unit Interval. *Computational Statistics & Data Analysis.* **54**, 2. 473-495.

Hirukawa, M., Sakudo, M. (2012). Nonparametric Multiplicative Bias Correction for Kernel-type Density Estimation using Positive Data. *DBJ Discussion Paper Series.* No. 1209.

Hirukawa, M., Sakudo, M. (2014). Nonnegative Bias Reduction Methods for Density Estimation using Asymmetric Kernels. *Computational Statistics & Data Analysis.* **75**, 112-123.

Hirukawa, M., Sakudo, M. (2015). Family of the Generalized Gamma Kernels: A Generator of Asymmetric Kernels for Nonnegative Data. *Journal of Nonparametric Statistics.* **27**, 1. 41-63.

Jacod, J., Li, Y., Mykland, P.A., Podolskij, M., Vetter, M. (2009). Microstructure Noise in the Continuous Case: The Pre-Averaging Approach. *Stochastic Processes and Their Applications.* **119**, 2249-2276.

Janssen, P., Swanepoel, J., Veraverbeke, N. (2012). Large Sample Behavior of the Bernstein Copula Estimator. *Journal of Statistical Planning and Inference.* **142**. 1189-1197.

Janssen, P., Swanepoel, J., Veraverbeke, N. (2014). A Note on the Asymptotic Behavior of the Bernstein Estimator of the Copula Density. *Journal of Multivariate Analysis.* **124**. 480-487.

Jin, X., Kawczak, J. (2003). Birnbaum-Saunders and Lognormal Kernel Estimators for Modeling Duration in High-Frequency Financial Data. *Annals of Economics and Finance.* **4**. 103-124.

Johannes, M. (2004). The Statistical and Economic Role of Jumps in Continuous-Time Interest Rate Models. *The Journal of Finance.* **59**, 1. 227-260.

Jones, M.C. (1993). Simple Boundary Correction for Kernel Density Estimation. *Statistics and Computing.* **3**. 135-146.

Jones, M.C., Linton, O., Nielsen, J.P. (1995). A Simple Bias Reduction Method for Density Estimation. *Biometrika.* **82**. 327-338.

JONES, C.S. (2003). Nonlinear Mean Reversion in the Short-Term Interest Rate. *Review of Financial Studies.* **16**. 793-843.

KALLSEN, J. (2006). A Didactic Note on Affine Stochastic Volatility Models. In: From Stochastic Calculus to Mathematical Finance, eds. Kabanov, Y., Liptser, R., Stoyanov, J. Springer, Berlin, 343-368.

KANAYA, S., KRISTENSEN, D. (2015). Estimation of Stochastic Volatility Models by Nonparametric Filtering. The Institute for Fiscal Studies, Department of Economics, UCL. Cemmap Working Paper CEW 09/15.

KARATZAS, I., SHREVE, S.E. (1996). Brownian Motion and Stochastic Calculus. 2nd Edition. Springer, Berlin.

KARUNAMUNI, R.J., ALBERTS, T. (2005). On Boundary Correction in Kernel Density Estimation. *Statistical Methodology.* **3**, Issue 3. 191-212.

KESSLER, M., LINDNER, A., SØRENSEN, M. (2013). Statistical Methods for Stochastic Differential Equations. Taylor & Francis Group. London.

LEE, W. (2014). Kernel Estimation of the Drift Coefficient of a Diffusion Process in the Presence of Measurement Error. Master Thesis, submitted at the University of British Columbia (Vancouver). Available online at
https://circle.ubc.ca/bitstream/handle/2429/46990/ubc_2014_september_lee_wooyong.pdf

LIEBSCHER, E. (2005). Semiparametric Density Estimators Using Copulas. *Communications in Statistics: Theory and Methods.* **34**. 59-71.

LOFTSGAARDEN, D.O., QUESENBERRY, C.P. (1965). A Nonparametric Estimate of a Multivariate Density Function. *The Annals of Mathematical Statistics.* **36**. 1049-1051.

LORENTZ, G.G. (1986). Bernstein Polynomials. AMS Chelsea Publishing, AMS. Providence, Rhode Island.

MANCINI, C., RENÒ, R. (2011). Threshold Estimation of Markov Models with Jumps and Interest Rate Modeling. *Journal of Econometrics.* **160**, 1. 77-92.

MASUDA, H. (2007). Ergodicity and Exponential β-Mixing Bounds for Multidimensional Diffusions with Jumps. *Stochastic Processes and their Applications.* **117**, 1. 35-56.

MÜLLER, H., STADTMÜLLER, U. (1999). Multivariate Boundary Kernels and a Continuous Least Square Principle. *Journal of the Royal Statistical Society.* Series B 61. 439-458.

NADARAYA, E.A. (1965). On Nonparametric Estimates of Density Functions and Regression Curves. *Theory of Probability and its Applications.* **10**. 186-190.

NELSEN, R.B. (1999). An Introduction to Copulas. Springer, Berlin.

NICOLAU, N. (2007). Nonparametric Estimation of Second-Order Stochastic Differential Equations. *Econometric Theory.* **23**, No.5. 880-898.

Noh, H., El Ghouch, A., Bouezmarni, T. (2013). Copula-Based Regression and Inference. *Journal of the American Statistical Association.* **108**, 502. 676-688.

Papapantoleon, A. (2008). An introduction to Lévy Processes with Applications in Finance. Lecture notes, available online at http://arxiv.org/pdf/0804.0482.pdf.

Pardoux, E., Veretennikov, A.Y. (2001). On the Poisson Equation and Diffusion Approximation. *The Annals of Probability.* **29**, 3. 1061-1085.

Podolskij, M., Vetter, M. (2006). Estimation of Volatility Functionals in the Simultaneous Presence of Microstructure Noise and Jumps. Technical Report, Ruhr-Universität Bochum.

Protter, P. E. (2005). Stochastic Integration and Differential Equations. 3rd Edition. Springer, Berlin.

Sancetta, A., Satchell, S. (2004). The Bernstein Copula and its Applications to Modeling and Approximations of Multivariate Distributions. *Econometric Theory.* **20**, 535-562.

Sato, K.I. (1999). Lévy Processes and Infinitely Divisible Distributions. Cambridge Studies in Advanced Mathematics.

Scaillet, O. (2004). Density Estimation Using Inverse and Reciprocal Inverse Gaussian Kernels. *Journal of Nonparametric Statistics.* **16**, 217-226.

Scott, D.W., Terrell, G.R. (1992). Variable Kernel Density Estimation. *The Annals of Statistics.* **20**, 3. 1236-1265.

Schmid, F., Trede, M. (2006). Finanzmarktstatistik. 2nd Edition. Springer, Berlin.

Schmidt, T. (2007). Coping with Copulas. *In: Copulas: From Theory to Applications in Finance.* Risk Books, London. 3-34.

Schmisser, E. (2013). Penalized Nonparametric Drift Estimation for a Multidimensional Diffusion Process. *Statistics.* **47**, No. 1. 61-84.

Schmisser, E. (2014). Nonparametric Adaptive Estimation of the Drift for a Jump Diffusion Process. *Stochastic Processes and Their Applications.* **124**, No. 1. 883-914.

Schuster, E.F. (1985). Incorporating Support Constraints into Nonparametric Estimators of Densities. *Communications in Statistics. Part A-Theory and Methods.* **14**. 1123-1136.

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. Chapman & Hall, New York.

Sklar, A. (1959). Fonction de Répartition à n Dimensions et Leurs Marges. *Publications de l´Institut de Statistique de l´Université de Paris.* **8**, 229-231.

Stanton, R. (1997). A Nonparametric Model of Term Structure Dynamics and the Market Price of Interest Rate Risk. *Journal of Finance.* **52**, 5. 1973-2002.

Terrell, G.R., Scott, D.W. (1980). On Improving Convergence Rates for Non-Negative Kernel Density Estimation. *The Annals of Statistics.* **8**. 1160-1163.

Tukey, P.A., Tukey, J.W. (1981). Data-Driven View Selection: Agglomeration and Sharpening. *Interpreting Multivariate Data.* John Wiley & Sons, New York. 215-243.

Vieu, P. (1993). Bandwidth Selection for Kernel Regression: A Survey. *In: Computer Intensive Methods in Statistics and Computing, eds. W. Härdle et al.* 134-149.

Wand, M.P., Jones, M.C. (1995). Kernel Smoothing. Chapman & Hall, CRC Monographs on Statistics & Applied Probability, New York.

Watson, G.S. (1964). Smooth Regression Analysis. *Sankhya.* **26**, 15. 175-184.

Wrench, J.W. (1968). Concerning Two Series for the Gamma Function. *Mathematics of Computation.* **22**. 617-626.

Zhou, B. (1996). High-Frequency Data and Volatility in Foreign-Exchange Rates. *Journal of Business and Economic Statistics.* **14**, 45-52.

Zhang, S., Karunamuni, R.J. (2000). On Nonparametric Density Estimation at the Boundary. *Nonparametric Statistics.* **12**, 197-221.

Zhang, L., Mykland, P.A., Aït-Sahalia, Y. (2005). A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High-Frequency Data. *Journal of the American Statistical Association.* **100**, 1394-1419.