Consistency and Bandwidth Selection for Dependent Data in Non-Parametric Functional Data Analysis

Von der Fakultät Mathematik und Physik
der Universität Stuttgart
zur Erlangung der Würde eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
genehmigte Abhandlung

von
Simon Peter Müller
geboren in Tettnang

Hauptberichter:	Priv.-Doz. Dr. J. Dippon
Mitberichter:	Prof. Dr. I. Steinwart
Tag der Einreichung:	26.7.2011
Tag der mündlichen Prüfung:	27.9.2011

Institut für Stochastik und Anwendungen
2011

ACKNOWLEDGMENTS

During my time at the Institute of Stochastic and Applications as a research and teaching assistant I drank thousands of cups of coffee, gave hundreds of tutorials, organised several lectures, i. e. probability theory, mathematical statistics, optimisation and biostatistics, designed certification exams for them, was a recording clerk of countless oral examinations, and wrote a book. This work was, at times, challenging, exhausting, and sometimes even frustrating, but well, finally I made it. This book would not have been possible without the great support and the highly appreciated scientific freedom that was granted me by my doctoral advisor PD Dr. J. Dippon. Special thanks to him! Furthermore, I would like to thank Prof. Dr. I. Steinwart for the co-examination of my thesis.

Apart from that, there are some more people whom I owe a debt of gratitude. In the first place, Dr. Fritz and Prof. Alscher, who supported me with their medical knowledge and for pushing our medical project. And secondly, Stefan Winter, who introduced me to the position of a teaching assistant. Furthermore, I want to thank all the other former and present colleagues. Thanks for the great time!

Furthermore, I want to give special thanks to my wife Alexandra. She was the person who supported me all the time and a profuse sorry for getting all the frustrations.

Last but not least, I would like to thank my family for their unceasing support.

Die moderne Geschichte ist der Dialog zwischen zwei Männern:
einer, der an Gott glaubt, ein anderer, der Gott zu sein glaubt.

— Gómez Dávila

DEUTSCHE ZUSAMMENFASSUNG

In der vorliegenden Dissertation betrachten wir Aspekte der nichtparametrischen funktionalen Datenanalyse. Es wird der funktionale Zusammenhang zweier Zufallsvariablen, einer erklärenden Zufallsvariablen X und einer abhängigen Zufallsvariablen Y, untersucht. Dabei bezieht sich der Begriff funktional in funktionaler Datenanalyse auf den Ursprung der erklärenden Zufallsvariablen X. Bei dieser wird angenommen, dass sie aus einem Funktionenraum E stammt. Die abhängige Zufallsvariable Y sei dagegen reellwertig.

Neben der Einführung in die nichtparametrische funktionale Datenanalyse in Kapitel 1 beinhaltet diese Dissertation drei weitere Kapitel, deren Inhalt in den nachfolgenden drei Absätzen zusammengefasst ist.

In Kapitel 2 betrachten wir die funktionale nichtparametrische Regression für α-mischende Daten $((X_i, Y_i))_{i=1}^n$. Dabei ist man an einer Schätzung der unbekannten Regressionsfunktion $m(x) := E[Y|X = x]$ interessiert. Im Gegensatz zur parametrischen Regression machen wir keine Annahmen über die Gestalt von $m(x)$, wir setzen lediglich gewisse Regularitätsannahmen voraus. Eine Methode zur Schätzung der Regressionsfunktion $m(x)$ ist der k-Nächste Nachbarn Kernschätzer. Der k-NN Kernschätzer gehört zu den lokalen Mittelungsschätzern. Bei diesem Verfahren bildet man ein gewichtetes Mittel über die abhängigen Zufallsvariablen Y_i, die den k nächsten Nachbarn des Elementes x zugeordnet sind, um damit eine Schätzung von $m(x)$ zu erhalten. Wir werden beweisen, dass der k-NN Schätzer für α-mischende Daten punktweise konsistent ist, und wir geben, unter zwei sich unterscheidenden Voraussetzungen an den Kovarianzterm, jeweils die Konvergenzraten an.

Zu guter Letzt geben wir einen Ausblick, wie man die Anfälligkeit des k-NN Kernschätzers gegenüber Ausreißern vermeiden kann. Wir umreißen dabei, wie man diesen robusten k-NN Schätzer konstruiert und zu einer Konsistenzaussage gelangt.

In Kapitel 3 befassen wir uns mit der gleichmäßigen Konvergenz von Kernschätzern auf einer kompakten Menge S_E verschiedener bedingter Größen, wie dem bedingten Erwartungswert, der bedingten Verteilungsfunktion und der bedingten Dichtefunktion für α-mischende Daten. Wie bereits im zweiten Kapitel setzen wir für diese drei bedingten Größen lediglich gewisse Regularitätsannahmen voraus. In den Beweisen für die Konvergenzraten der verschiedenen bedingten Größen stellt sich heraus, dass ein Zusammenhang zwischen der Überdeckungszahl von S_E und der Art der Abhängigkeit der Daten vorliegt. Besitzt S_E eine exponentiell wachsende Überdeckungszahl, so ist es mit den uns bekannten Mitteln nicht möglich, gleichmäßige Konvergenzraten für allgemein α-mischende Zufallsvariablen zu erhalten. Für Funktionenräume mit derartiger Eigenschaft von kompakten Teilmengen müssen wir uns auf geometrisch α-mischende Zufallsvariablen beschränken. Bei Mengen S_E mit polynomial wachsenden Überdeckungszahlen erhält man Resultate auch für arithmetisch α-mischende Zufallsvariablen.

Des Weiteren präsentieren wir Resultate für den Kernschätzer der Regressionsfunktion, bei denen man unter zusätzlichen Voraussetzungen ähnliche Konvergenzraten erhält wie für unabhängige Daten. Mit leicht modifizierten Voraussetzungen erhält man für die Kernschätzer der bedingten Verteilungs- und Dichtefunktion ähnliche Aussagen. Dies führen wir aber in dieser Arbeit nicht aus. Darüber hinaus geben wir für den Kernschätzer der Regressionsfunktion eine mögliche Beweisidee, um für α-mischende Daten die Konsistenz der Kreuzvalidierung als Bandbreitenwahl zu erhalten.

Im abschließenden Kapitel 4 beschäftigen wir uns mit einem lokalen datenabhängigen Verfahren der Bandbreitenwahl für den Kernschätzer der Regressionsfunktion. Als naheliegendes Maß für die Genauigkeit der Schätzung und somit der Güte der Bandbreitenwahl bietet sich der punktweise L_2–Fehler an. Da die Regressionsfunktion $m(\cdot)$ unbekannt ist, ist dieser jedoch nicht bestimmbar und es ist notwendig, eine geeignete Approximation zu finden. In der Literatur werden hierzu verschiedene Methoden eingesetzt, wie z. B. Kreuzvalidierung oder verschiedene Bootstrap-Methoden. Wir haben in unserer Arbeit ein Bootstrap-Verfahren aufgegriffen und dieses auf den Fall der funktionalen nichtparametrischen Regression übertragen. Hierzu beweisen wir, dass unsere Methode asymptotisch gegen den zu approximierenden L_2–Fehler konvergiert und wir vergleichen unser Verfahren anschließend auf simulierten und realen Datensätzen mit einer lokalen und globalen Version der Kreuzvalidierung. Die simulierten Daten sind derart konstruiert, dass verschiedene Stufen zwischen homogen und heterogen angenommen werden. Bei den homogenen Daten erreichen, wie erwartet, die globale und die lokale Methode eine ähnliche Genauigkeit. Bei immer stärker werdender Heterogenität der Daten hingegen, schneide das lokale Verfahren gegenüber der globalen deutlich besser ab. Zudem konnten wir in allen Beispielen feststellen, dass die Bootstrap-Methode zu einer höheren oder gleich guten Genauigkeit führt wie die lokale Kreuzvalidierung. Der Vorteil des Bootstrap-Verfahrens gegenüber der Kreuzvalidierung ist, dass man mit wenig Mehraufwand Konfidenzbänder berechnen kann. Man muss allerdings eine höhere Rechenzeit in Kauf nehmen, da man für das Bootstrapping-Verfahren eine Pilot-Kernschätzung benötigt.

CONTENTS

1

INTRODUCTION TO NON-PARAMETRIC FUNCTIONAL DATA ANALYSIS

1.1 REGRESSION ANALYSIS

Let (E, d) be a semi-metric space and (X, Y) be a pair of random variables valued in the measurable space $(E \times \mathbb{R}, \mathcal{E}_d \otimes \mathcal{B}(\mathbb{R}))$, where $\mathcal{B}(\mathbb{R})$ is the Borel σ–algebra and $\mathcal{E}_d$ is the σ–algebra generated by the topology of E which is defined by the semi-metric d. Any random variable considered in this work is defined on the same probability space, namely $(\Omega, \mathcal{A}, P)$.

In regression analysis, one is interested in how the response variable Y depends on the observation X. The problem herein is to find a measurable function $f : E \to \mathbb{R}$, such that $f(X)$ is a *good approximation*, in some sense, of Y. Since $|f(X) - Y|$ is a $\mathbb{R}$–valued random variable, the L_p–risk is used to measure the accuracy

$$E\left[|Y - f(X)|^p\right],$$

for some $p \in \mathbb{N}$. In this work we consider the case $p = 2$. The advantage of the L_2–risk is that the solution can be explicitly calculated and the minimisation of the L_2–risk leads to estimates that can be computed quickly, see Györfi et al. [33, p. 2 or p. 158]. Therefore, we are interested in a measurable function $f : E \to \mathbb{R}$ such that this function minimises the mean squared error,

$$E\left[(Y - m(X))^2\right] = \min_{\substack{f:E\to\mathbb{R} \\ \text{measurable}}} E\left[(Y - f(X))^2\right]. \tag{1.1}$$

The *regression function*

$$m(x) = E\left[Y \mid X = x\right] \tag{1.2}$$

is the explicit solution of the minimisation problem in (1.1).

1.2 DESCRIPTION OF THE DATA AND RANDOM DESIGN

Let $(X_i, Y_i)_{i=1}^n$ be n pairs identically distributed as (X, Y). At the beginning, let us start with some notation which we will use throughout this work.

Definition 1.2.1 *We denote by a lower case letter x a non-random element of a functional semi-metric space (E, d) and by a capital letter X a functional E-valued random variable.*

The word functional in non-parametric functional data analysis is linked with the nature of the observation X, namely that it lives in an infinite-dimensional space E. We identify these elements x and X as functions $x : T \to \mathbb{R}$ and $X : T \to \mathbb{R}$, where T is a subset of $\mathbb{R}^p$ for some $p \in \mathbb{N}$. If we speak of curves, we have the one-dimensional case in mind when $T \subset \mathbb{R}$, for instance in the analysis of time series. Another example is image analysis, where the colour gradient is examined ($T \subset \mathbb{R}^2$), or the colour gradient of a 3-d image ($T \subset \mathbb{R}^3$).

The problem we examine is called regression estimation with random design. Random design means that the observation is made at a random element X and not at a fixed element chosen by the user. The estimate of the regression function can then be characterised as follows: the statistician observes some response value Y_i of an unknown measurable function $m(\cdot)$ at some random function X_i with an additive random error ε_i and he wants to recover $m(X_i)$, the true value of the function at these observation. In this model the data $(X_i, Y_i)_{i=1}^n$ can be rewritten as

$$Y_i = m(X_i) + \varepsilon_i. \tag{1.3}$$

It is assumed that the additive random error ε_i depends on the observation X_i and satisfies $E[\varepsilon_i | X_i] = 0$. For a more detailed description of the difference of the random and the fixed design, we refer to Györfi et al. [33, p. 15].

1.3 PARAMETRIC VERSUS NON-PARAMETRIC REGRESSION

In this analysis the notion of non-parametric is motivated by the space which we assume the regression function belongs to. In the case of a parametric model the statistician assumes that the structure of the regression function is known. For example, one assumes that the regression function is linear. If the model is well-chosen, an advantage of a parametric model is that the practitioner gets good results for small sample sizes, otherwise the parametric model performs badly. Another handicap one has in the multivariate case where it is difficult to visualise the data and it therefore is that may be difficult to choose a suitable model. Even in the univariate case this is sometimes difficult, see for instance the illustrative example given by Györfi et al. [33, p. 10 et seq.], where they use a regression function that is composed of different parametric models. This inflexibility of the parametric model leads to non-parametric regression estimates, where the statistician does not assume that the regression function can be described by a finite number of parameters.

Let us now present an example for a parametric model and a non-parametric function regression model.

Examples:

1. *Parametric functional model (see e.g. [30, p. 9] or [59]):*
   Let H be a Hilbert space, X a H-valued random variable and assume that the regression function $m(x)$ is a linear and continuous function, $m : H \to \mathbb{R}$. By

Riesz representation theorem, there exists a unique element $h \in H$ such that $m(\cdot) = \langle \cdot, h \rangle_H$. The linear regression model may then be expressed as

$$Y_i = \langle X_i, h \rangle_H + \varepsilon_i.$$

2. *Non-parametric functional model:*
Let H be a Hilbert space, X a H-valued random variable and assume that the regression function $m(x)$ is continuous. This model may be expressed as

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

Bosq [6] gives an good introduction into functional data analysis for linear processes in function spaces, also Ramsay and Silverman treat functional linear regression in [59] or [60]. The non-parametric functional model was examined in the monograph by Ferraty and Vieu [30].

In this work, we examine the non-parametric functional regression model, more precisely, we assume that the regression function is of one of the following two types:

Definition 1.3.1 *The regression function is of continuity-type, if*

$$m \in C(E) := \left\{ f : E \to \mathbb{R} \middle| f \text{ is continuous} \right\}.$$

Definition 1.3.2 *The regression function is of Hölder-type, if*

$$m \in L^{\beta}(E) := \left\{ f : E \to \mathbb{R} \middle| f \text{ is Hölder continuous with parameter } \beta \right\}$$

with $\beta > 0$.

These assumptions may be be replaced by the following condition,

$$\lim_{h \to 0} \frac{1}{\mu(B(x,h))} \int_{B(x,h)} |m(\omega) - m(x)| d\mu(\omega) = 0 \tag{1.4}$$

where $B(x,h)$ is a closed ball centred at x with radius h, as Dabo-Niang and Rhomari [18] did in their work. This assumption covers a wider class of regression functions $m(\cdot)$ than we use, e.g. $m = 1_{[0,1] \cap \mathbb{Q}}$ and μ as the Lebesgue measure satisfies the condition in (1.4), but is obviously not continuous. For further discussion of this example, we refer to *Remarque 1* in [18]. Another discussion on this assumption can be found in [13].

1.4 REGRESSION ESTIMATION, CONSISTENCY, AND RATE OF CONVERGENCE

In practise, the distribution of the pair (X, Y) is unknown and so is the regression function. Because of this, the regression function is estimated by a data set of random variables $(X_i, Y_i)_{i=1}^n$ which is identically distributed as (X, Y). We denote the estimate by $\hat{m}(x) := \hat{m}(x; (X_1, Y_1), \ldots, (X_n, Y_n)) : E \to \mathbb{R}$, which is assumed to be a measurable function of the data. Commonly the estimate $\hat{m}(x)$ will not be equal to the true regression function. Because of this a measurement of accuracy is needed. In the literature following distinct error criteria are used (see Györfi et al. [33, p. 3]):

- the pointwise error

  $$|\hat{m}(x) - m(x)|$$

  for $x \in E$,

- the supremum norm error

  $$\sup_{x \in S_E} |\hat{m}(x) - m(x)|,$$

  where $S_E \subset E$ is a totally bounded set, and

- the pointwise L_p–error,

  $$E\left[|\hat{m}(x) - m(x)|^p\right]$$

  for $p \in \mathbb{N}$ and $x \in E$.

Next, we present the type of convergence that we use in this work for defining consistency. We will see later in the proofs in Chapter 2, 3, and 4, the almost complete convergence is in some sense easier to state than the almost sure one. Furthermore, the almost complete convergence implies the almost sure convergence and the convergence in probability. (For proofs see Ferraty and Vieu [25, p. 229 et seq.].)

Definition 1.4.1 (Ferraty and Vieu [30], p. 228) *Let (Z_n) be a sequence of random variables. Then (Z_n) converges almost completely to a random variable Z, if and only if*

$$\forall \varepsilon > 0 : \sum_{n=1}^{\infty} P\left(|Z_n - Z| > \varepsilon\right) < \infty,$$

in short: $\lim_{n\to\infty} Z_n = Z$ *almost completely.*

The following definition is presented to introduce the notion of the almost complete convergence rate, which was first introduced by Ferraty and Vieu.

Definition 1.4.2 (Ferraty and Vieu [30], p. 230) *Let (Z_n) be a sequence of random variables and (u_n) a positive decreasing sequence converging to zero. Then the rate of almost complete convergence of (Z_n) to Z is said to be of order (u_n) , if and only if*

$$\exists \varepsilon_0 > 0 : \sum_{n=1}^{\infty} P\left(|Z_n - Z| > \varepsilon_0 u_n\right) < \infty,$$

in short: $Z_n - Z = \mathcal{O}_{a.co.}(u_n)$.

1.5 CONSTRUCTION OF THE NON-PARAMETRIC REGRESSION ESTIMATE

Györfi et al. [33, p. 18] describe four paradigms for non-parametric regression, namely *local averaging*, *local modelling*, *global modelling* and *penalised modelling*. We restrict ourselves to the examination of local averaging. Recall, that the data can be written as in (1.3). By this, the fact that a function x is close, in some sense, to X_i should imply that the estimate $\hat{m}(x)$ is close to the response Y_i that is associated to the observation X_i. Such an estimate is given as

$$\hat{m}(x) = \sum_{i=1}^{n} Y_i W_{n,i}(x), \tag{1.5}$$

for $x \in E$ and the weight function $W_{n,i}(x) \in [0,1]$ depends on the data. We assume for this weight function that $W_{n,i}(x)$ is close to 0 if X_i is far away from x. We examine in this work the *Naradaya-Watson kernel estimate* and the *k-Nearest Neighbour kernel estimate* (k-NN kernel estimate). In Györfi et al. [33, p. 19] one can additionally find the *partitioning estimate* as an example of local averaging.

Nadaraya-Watson Kernel Estimate

This type of estimate was first proposed by Nadaraya [53] and Watson [70], so this estimate is called the Nadaraya-Watson kernel estimate. For $\mathbb{R}^p$-valued observations X this was extensively examined by Györfi et al. [33].

The weight function in (1.5) for this type of estimate is defined as

$$W_{n,i}(x) := \frac{K\left(h_n^{-1} d(x, X_i)\right)}{\sum_{i=1}^{n} K\left(h_n^{-1} d(x, X_i)\right)}, \tag{1.6}$$

where $K : \mathbb{R} \to \mathbb{R}^+$ is a kernel function, d the semi-metric of the function space E and h_n is a strictly positive decreasing sequence. We get then for the kernel estimate

$$\hat{m}(x) = \sum_{i=1}^{n} Y_i \frac{K\left(h_n^{-1} d(x, X_i)\right)}{\sum_{i=1}^{n} K\left(h_n^{-1} d(x, X_i)\right)}, \quad \text{if} \quad \sum_{j=1}^{n} K\left(h_n^{-1} d(x, X_i)\right) \neq 0, \tag{1.7}$$

otherwise $\hat{m}(x) = 0$ and $x \in E$. Hereafter, any reference to a kernel estimate should be understood as a Nadaraya-Watson kernel estimate.

k-Nearest Neighbour Kernel Estimate

The k-NN kernel estimate differs from the Nadaraya-Watson kernel estimate in how the smoothing parameter is chosen. The bandwidth is chosen here as the radius of a ball with centre x such that k data points X_i are within the ball. More precisely,

$$H_{n,k} := d(x, X_{(k)}), \tag{1.8}$$

where the set $(X_{(i)}, Y_{(i)})_{i=1}^n$ is the re-indexed set $(X_i, Y_i))_{i=1}^n$ such that

$$d(x, X_{(1)}) \leqslant d(x, X_{(2)}) \leqslant \ldots \leqslant d(x, X_{(n)}).$$

By this definition (1.8) of the bandwidth, we find that $H_{n,k}$ is a positive real-valued random variable depending on the data $(X_i, Y_i)_{i=1}^n$. The following theorem, proven by Cover and Hart [15], shows that the choice of bandwidth in (1.8) is a sequence converging to zero under some conditions.

Theorem 1.5.1 (Cover and Hart [15]) *Denote by μ the probability measure of X. Let $(E, d) = (\mathbb{R}^p, d)$ with a metric d, for $x \in \operatorname{supp}(\mu)$, and $\lim_{n\to\infty} k/n = 0$ we have*

$$\lim_{n\to\infty} H_{n,k} = \lim_{n\to\infty} d(x, X_{(k)}) = 0$$

with probability 1.

A proof of Theorem 1.5.1 can be found in the monograph of Devroye et al. [21, p. 63]. There it is given for $(\mathbb{R}^p, d)$ and for independent data, but it may be extended to a general metric but separable space, see for instance [13]. Then the k-NN kernel estimate is defined as

$$\hat{m}_{k\text{-NN}}(x) = \sum_{i=1}^n Y_i \frac{K\left(H_{n,k}^{-1} d(x, X_i)\right)}{\sum_{i=1}^n K\left(H_{n,k}^{-1} d(x, X_i)\right)}, \quad \text{if} \quad \sum_{j=1}^n K\left(H_{n,k}^{-1} d(x, X_i)\right) \neq 0,$$

otherwise $\hat{m}_{k\text{-NN}}(x) = 0$ and $x \in E$.

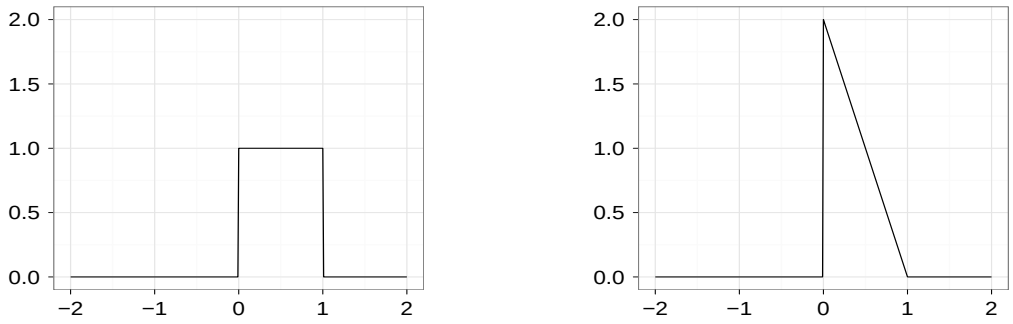
In the next section we treat the kernel function K more precisely.

The Kernel Function and Some of its Properties

In contrast to the one-dimensional regression analysis, we have in the functional and multivariate consideration only a positive input to the kernel function because we are considering asymmetric kernel functions. We assume that the asymmetrical kernel function has its peak at zero and decreases monotonically as the input increases. This assumption ensures that if the function of interest x is close to X_i the response value Y_i plays in the estimate of $\hat{m}(x)$ a more important role as a Y_j which observation X_j is far from x. Figure 1.1 shows some typical kernel functions.

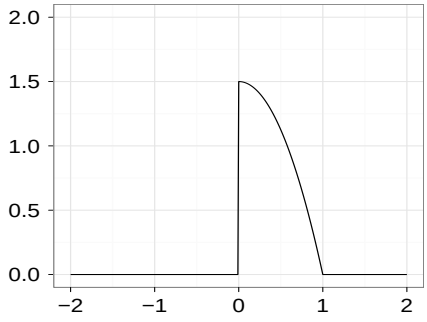
Moreover, as can be seen in (1.7), the kernel estimate depends on the parameter h_n. This *smoothing parameter* or *bandwidth* controls the width of this asymmetric kernel function and, therefore, how many data points X_i are considered for the prediction of the regression function at x. If the amount of data grows we assume that $h_n \to 0$.

In the following, we specify the kernel function K more precisely.

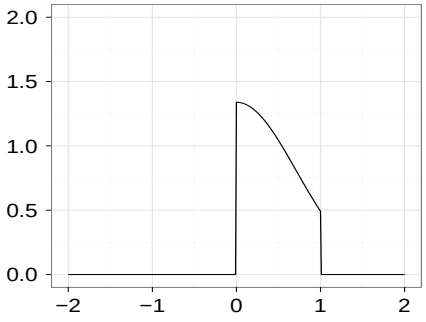


(a) Box kernel

(b) Triangle kernel



(c) Quadratic kernel



(d) Truncated Gaussian kernel

Figure 1.1: Four typical kernel functions.

Definition 1.5.1 (Ferraty and Vieu [30], p. 42) *We consider two types of kernel functions.*

- *A function* $K : \mathbb{R} \mapsto \mathbb{R}^+$ *such that* $\int_{\mathbb{R}} K(u)du = 1$ *is called a kernel of discontinuous-type if there exist two constants* $0 < C_1 < C_2 < \infty$ *such that*

$$\forall u \in \mathbb{R} : \; C_1 1_{[0,1]}(u) \leqslant K(u) \leqslant C_2 1_{[0,1]}(u).$$

- *A function* $K : \mathbb{R} \mapsto \mathbb{R}^+$ *such that* $\int_{\mathbb{R}} K(u)du = 1$ *is called a kernel of continuous-type if its support is* $[0, 1]$, K *is differentiable in* $[0, 1]$, $K(1) = 0$, *and there exists two constants* $-\infty < C_1 < C_2 < 0$ *such that*

$$\forall u \in [0, 1] : \; C_1 \leqslant K'(u) \leqslant C_2.$$

The box or the truncated Gaussian kernel function are two examples of discontinuous-type kernel functions and the triangle or the quadratic kernel function are two examples of continuous-type kernel functions. For these two types of kernel functions we present some theoretical advances as we will use them throughout in this dissertation. For the proof of these lemmas we refer to the monograph of Ferraty and Vieu [30, p. 43 et seq.].

Lemma 1.5.1 (Ferraty and Vieu [30], p. 43) *Assume that* K *is a kernel function of discontinuous-type, then there are two constants* $C_3, C_4 \in (0, \infty)$ *such that*

$$C_3 P(d(x, X) \leqslant h_n) \leqslant E\left[K\left(h_n^{-1} d(x, X)\right)\right] \leqslant C_4 P(d(x, X) \leqslant h_n).$$

Next, for the continuous-type kernel functions we get, with an additional assumption, the same result as for discontinuous-type kernel functions.

Lemma 1.5.2 (Ferraty and Vieu [30], p. 44) *Let* X *be an* E*-valued random variable and assume* K *is a continuous-type kernel function, and there are two constants* $C_5 > 0$ *and* $\varepsilon_0 > 0$ *such that we have*

$$\forall \varepsilon < \varepsilon_0 : \int_0^{\varepsilon} P(d(x, X) \leqslant u)\, du > C_5 \varepsilon P(d(x, X) \leqslant \varepsilon),$$

where $P(d(x, X) \leqslant \cdot)$ *is the probability distribution function of* X*. Then we have for small* h_n *and for* $C_6, C_7 \in \mathbb{R}^+$*,*

$$C_6 P(d(x, X) \leqslant h_n) \leqslant E\left[K\left(h_n^{-1} d(x, X)\right)\right] \leqslant C_7 P(d(x, X) \leqslant h_n).$$

As we will see, the *small ball probability*,

$$F_x(h) := P(d(x, X) \leqslant h), \tag{1.9}$$

plays a crucial role in functional data analysis. The index of the small ball probability $F_x(h)$ shall emphasise that this concentration function depends on the non-random element $x \in E$.

1.6 SMALL BALL PROBABILITY

On infinite-dimensional spaces, we have no default measure, unlike the Lebesgue measure in a finite-dimensional space. Therefore a *density-free* approach was developed. Because of this circumstance, the problem is deferred to the examination of the small ball probability $F_x(h)$. This function plays a role similar to the density function in the finite-dimensional case. Both the density function and the small ball probability are measures of the concentration of the random variable. Because of this behaviour of $F_x(h)$, it has an affect on

- the rate of convergence (see Ferraty and Vieu [30, p. 80]).
- the choice of the optimal bandwidth h_n (see Rachdi and Vieu [58]).
- or the asymptotic evaluation of the L_p–error (see Delsol [20] or Ferraty et al. [25]).

We will give a short overview of this large and current field of research here. The results presented here are taken from the monograph by Ferraty and Vieu [30], Chapter 13, the paper by Ferraty et al. [25], Delsol [20], and the monograph by Bogachev [4].

In the case of independent data, the kernel estimate $\hat{m}(x)$, introduced in (1.7), converges to the regression function $m(x)$ with rate

$$\hat{m}(x) - m(x) = \mathcal{O}\left(h_n^{\beta}\right) + \mathcal{O}_{a.co.}\left(\sqrt{\frac{\log n}{n F_x(h_n)}}\right), \tag{1.10}$$

for $x \in E$ and β is the Hölder constant. For conditions see Theorem 6.11 [30, p. 80]. As can be seen in (1.10), the rate of convergence is governed by two parts. Here, we have to consider for the choice of the bandwidth h_n that there is a trade-off between the first and the second term. By the first term one wants to choose a fast decaying smoothing parameter h_n, but in such a case, the second term blows up, as $F_x(h_n) \to 0$ for $h_n \to 0$. The bandwidth h_n also has to fulfil the condition $nF_x(h_n) \to \infty$. Therefore the concentration of the random variable X determines how to choose the sequence h_n. If the data $X_1, \ldots, X_n$ is dispersed, we get a slow rate. On the other hand, for concentrated data, we have a more efficient rate. Before giving some examples for $F_x(h)$, we will discuss the link of functional data analysis to the finite-dimensional case.

Let d be the standard Euclidian metric in $E = \mathbb{R}^p$, X be a random variable whose probability distribution function is absolutely continuous with respect to the Lebesgue measure. Assume that the density function f is continuous and strictly positive for all $x \in \mathbb{R}^p$, then the small ball probability is expressed as

$$F_x(\varepsilon) = C\varepsilon^p + o\left(\varepsilon^p\right)$$

for some $C > 0$, see Lemma 13.13 [30, p. 219]. Then the almost complete convergence rate of the kernel regression estimate is expressed as

$$\hat{m}(x) - m(x) = \mathcal{O}_{a.co.}\left(\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+p}}\right), \tag{1.11}$$

for $x \in \mathbb{R}^p$. For kernel estimates of the non-parametric regression function, Stone [66] proved that this rate is optimal. Therefore, Ferraty and Vieu ansatz to functional data analysis includes the finite-dimensional approaches. However, the rate for the kernel estimate given in (1.10) is just an upper bound, the optimality having not yet proven.

Next, the definition of two types of small ball probabilities is presented. We write $f(\varepsilon) \sim Cg(\varepsilon)$, iff $\left|\frac{f(\varepsilon)}{g(\varepsilon)} - C\right| \to 0$ for $\varepsilon \to 0$.

Definition 1.6.1 (Ferraty and Vieu, [30], p. 207 and p. 209) *Let (E, d) be a semi-metric space, X be an E-valued random variable, and $x \in E$ fixed.*

- *X is considered of fractal-type with order $\tau > 0$ with respect to d, if there exists a positive and finite constant C such that*

$$F_x(\varepsilon) \sim C\varepsilon^{\tau} \quad \text{for} \quad \varepsilon \to 0. \tag{1.12}$$

- X *is considered of exponential-type with order* (τ_1, τ_2), $\tau_1, \tau_2 \geqslant 0$, *with respect to* d, *if there exists a positive and finite constant* C *such that*

$$F_x(\varepsilon) \sim C \exp\left(-\frac{1}{\varepsilon^{\tau_1}} \log\left(\frac{1}{\varepsilon}\right)^{\tau_2}\right) \text{ for } \varepsilon \to 0. \tag{1.13}$$

The constants in the definition may depend on x.

The fractal-type random variable was introduced by Ferraty and Vieu [28]. They transferred the idea of fractal-dimensions from applications in physics (Pesin [56]), to functional data analysis. In the paper of Ferraty et al. [22] such fractal-type processes were examined for functionally dependent data in the case of a non-parametric functional model. Moreover, they prove the uniform convergence on a compact set of the non-parametric regression kernel estimate for a random variable of that type.

If we have a E-valued random variables X of fractal-type, similar convergence rates are obtained as for $\mathbb{R}^p$-valued random variables. For independent data distributed as a fractal-type random variable X with order τ we have the rate

$$\hat{m}(x) - m(x) = \mathcal{O}_{a.co.}\left(\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+\tau}}\right), \tag{1.14}$$

for $x \in E$. In comparison to (1.11), we have for fractal-type random variables similar rates (1.14) as for $\mathbb{R}^p$-valued random variables. Unfortunately, most E-valued random variables are of the exponential-type. Dabo-Niang and Rhomari [18] extended the fractal-type ansatz of Ferraty and Vieu [28] to exponential-type random variables, such as for example the following.

Example for an Exponential-Type Random Variable, Ferraty et al. [24] and Bogachev [4]

Let P^W be the Wiener measure on the space $C([0,1], \mathbb{R})$ that is equipped with supremum norm

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

(For a definition of Wiener measure we refer to Bogachev [4, p. 42 and p. 54], Definition 2.2.1 and Example 2.3.11 therein.) Then we have for small centred balls ($\varepsilon > 0$)

$$P^W(x \in C([0,1], \mathbb{R}) : \|x\|_\infty < \varepsilon) \sim \frac{4}{\pi} \exp\left(-\frac{\pi^2}{8\varepsilon^2}\right), \tag{1.15}$$

see Bogachev [4, p. 187]. In accordance with Bogachev [4, p. 61] Theorem 2.4.5 we get the following Reproducing Kernel Hilbert Space

$$H := \left\{x \in C([0,T], \mathbb{R}) : P^W \ll P_x^W \wedge P^W \gg P_x^W\right\},$$

where $P_x^W(\cdot) := P^W(\cdot - x)$ is the translated measure of P^W and $P^W \ll P_x^W$ means that P_x^W is absolutely continuous with respect to P^W. By this, the above result in (1.15) can be extended and we get for $\varepsilon > 0$

$$\forall\, \tilde{x} \in H \;:\; P^W(x \in H :\; \|x - \tilde{x}\|_\infty < \varepsilon) \sim C_{\tilde{x}} \frac{4}{\pi} \exp\left(-\frac{\pi^2}{8\varepsilon^2}\right).$$

Let $(B_t)_{t\in[0,1]}$ be a Brownian motion with $B_0 := 0$ and let $S := (S_t)_{t\in[0,1]}$ with $S_0 := 0$, the Ornstein-Uhlenbeck process, which is defined as the solution of the stochastic differential equation

$$dS_t = -\frac{1}{2} S_t\, dt + dB_t \quad \forall\, t \in (0,1].$$

The Ornstein-Uhlenbeck process has a probability measure that is absolutely continuous with respect to the Wiener measure P^W, so that we have for $\varepsilon > 0$

$$\forall\, \tilde{x} \in H \;:\; P(S \in B(\tilde{x}, \varepsilon)) \sim C_{\tilde{x}} \frac{4}{\pi} \exp\left(-\frac{\pi^2}{8\varepsilon^2}\right).$$

Therefore, S is of exponential-type with order $(\tau_1, \tau_2) = (2, 0)$.

More examples can be found in Bogachev [4, Chapter 4.10 p. 197 et seqq.], Ferraty et al. [24], in the references given in the monograph by Ferraty and Vieu [25, p. 209 et seq.], or in the paper by Van der Vaart and Van Zanten [69].

It is a disadvantage of exponential-type random variables that the rate of convergence of the regression function estimation is only of order $(\log n)^{-t}$ for some $t > 0$. In the case that E is a separable Hilbert space, this disadvantage can be overcome by choosing a semi-metric d adapted to the functional variable X. A statistician is able to transform X to a random variable of fractal-type by using a projection-based semi-metric, as for instance functional principal component analysis, Fourier basis, wavelets, see Lemma 13.6 [30, p. 213]. This idea also effects dimension reduction in the finite-dimensional non-parametric regression. If one uses a projection-based semi-metric instead of a metric (see Ferraty and Vieu [30, p. 221] Lemma 13.15 and Proposition 13.16 therein), it is possible to get, with some additional assumptions, a faster rate than with respect to the Euclidian metric, as in

$$\hat{m}(x) - m(x) = \mathcal{O}\left(\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+p}}\right).$$

To apply this projection-based semi-metric to non-parametric multivariate regression the absolute continuity with respect to the Lebesgue measure of the projected part of the random variable X has to be assumed. For more references see Delsol [20] or Ferraty and Vieu [30, p. 210].

Furthermore, as it can be seen in the definition (1.9) of the small ball probability $F_x(h)$, the choice of the semi-metric d plays an important role. What follows is a brief discussion of this issue. For functional data analysis the choice of the semi-metric is still an open field of research. A recent publication on the choice of the semi-metric in functional data analysis is the paper of Ferraty and Vieu [32].

1.7 ASPECTS OF UNIFORM CONVERGENCE IN FUNCTIONAL SPACES

For the examination of almost complete convergence in the supremum norm, namely let

$$\sup_{x \in S_E} |\hat{m}(x) - m(x)|,$$

with $S_E \subset E$ be compact, that is to say there has to exist for all $\varepsilon > 0$ finitely many balls with radius ε such that these balls cover S_E. This covering property is not only needed for the convergence in the supremum norm. It is also used in proving the optimality of methods for bandwidth estimation, cross-validation, see Rachdi and Vieu [58] or Benhenni et al. [3], or bootstrapping Chapter 4, or for building confidence intervals by bootstrapping Ferraty and Vieu [27]. More details on this topic are given in Chapter 3.

1.8 MODELLING OF WEAK DEPENDENCE OF RANDOM VARIABLES

This section gives a short introduction to the concept of α-mixing. This type of dependence of random variables was first introduced by Rosenblatt [63]. There are some other ways of modelling the dependence of a sequence of random variables in the case of mixing, see for example the survey of Bradley [7], or for a deeper study [8], [9], or [10].

To start with, some notations are introduced. Let (X_n) be a sequence of random variables on the probability space $(\Omega, \mathcal{A}, P)$, which takes values in the measurable space $(\tilde{\Omega}, \tilde{\mathcal{A}})$. Denote $\mathcal{A}_j^k$, $-\infty \leqslant j \leqslant k \leqslant \infty$, the σ-algebra, which is generated by the random variables $\{X_j, \ldots, X_k\}$.

Definition 1.8.1 *The strong mixing coefficient of a sequence (X_n) of random variables is defined as*

$$\alpha(n) = \sup_k \sup_{A \in \mathcal{A}_{-\infty}^k} \sup_{B \in \mathcal{A}_{k+n}^{\infty}} |P(A \cap B) - P(A)\,P(B)|.$$

The sequence (X_n) is called α-mixing (or strong mixing), if

$$\lim_{n \to \infty} \alpha(n) = 0.$$

Depending on the rate of convergence of $\alpha(n)$ one considers two cases.

Definition 1.8.2 *A sequence (X_n) is called arithmetic α-mixing (or algebraic) with rate $b > 0$ if*

$$\exists C > 0 : \ \alpha(n) \leqslant C n^{-b}.$$

The sequence is called geometric α-mixing if

$$\exists b > 0, \ C > 0 : \ \alpha(n) \leqslant \exp\left(-C n^b\right).$$

For mixing in the functional context, we refer to the monograph of Ferraty and Vieu [25, p. 155], especially Proposition 10.3 and 10.4.

1.9 SUMMARY OF THIS THESIS

Besides the introduction this dissertation contains three more chapters. In the following paragraphs we will give a short summary for each of them.

In Chapter 2 we examine non-parametric regression for α-mixing functional data. A method for estimating the regression function $m(x)$ is the k-nearest neighbour kernel estimate. We prove that the k-NN kernel estimate is pointwise almost complete consistent for α-mixing data and we present, for two different assumptions on the covariance term, the almost complete convergence rate. The results are obtained on the one hand by using results of the functional kernel estimate, where a deterministic bandwidth sequence is used, and on the other hand by applying lemmas from Bradley [5, p. 18] and Burba et al. [11].

Finally, we give an outline on how to avoid the drawback of susceptibility of the k-NN kernel estimate to outliers. We adumbrate on how to construct such a robust kernel estimate and on how get almost complete convergence.

Chapter 3 is focused on uniform convergence rates on a compact set S_E of non-parametric estimates for α-mixing random variables of various conditional quantities, such as the conditional expectation, the conditional distribution function, and the conditional density function. It turns out in our proofs that there is a link between the covering number of the set S_E and the type of α-mixing. Indeed, there are many functional spaces on which a compact set has a covering number that grows exponentially. For such sets S_E it is not possible to get uniform almost complete rates for general α-mixing random variable, there we have to restrict on geometric α-mixing random variables. Instead, if the covering number grows polynomially, we get almost complete rates for general α-mixing random variables. Furthermore, we present two results for the kernel estimate of the regression function, where we get with some additional conditions similar rates as in the independent case. With slightly modified assumptions, not listed in this thesis, we get similar results for the kernel estimate of the conditional distribution function and the conditional density function. Moreover, we comment on the uniform almost complete rate for the estimate of the non-parametric regression function and outline how to possibly prove the validity of a cross-validation bandwidth selection procedure for α-mixing functional data.

In the last Chapter 4 we discuss the issue of a local adaptive bandwidth selection procedure for the kernel estimate of the regression function. Here, an obvious measure for the optimality of the parameter selection is the pointwise mean squared error. As the regression function $m(\cdot)$ is unknown, we cannot calculate it. In the literature different approximation methods as cross-validation or bootstrapping are presented. We pick up a bootstrap method for approximating this pointwise mean squared error for non-parametric functional regression. We prove that our approximation converges against the true error and afterwards we compare our method on simulated and real world data with a global and local version of a cross-validation method. The simulated data is constructed such that we have different nuances between homogenous and heterogenous data. The results differ then in the following way. On the one hand if the data is more homogenous, global and local methods perform similarly, on the other hand if the data gets more heterogenous,

the local methods outperform the global bandwidth selection procedure more and more. In addition, we notice that in all examples the bootstrap method performs better or equal than the local cross-validation procedure. Moreover, it is possible to calculate confidence intervals from the bootstrapped data. As we need a pilot kernel estimate for bootstrapping, more calculation time is needed for that bootstrap procedure.

2

K-NN KERNEL ESTIMATE FOR NON-PARAMETRIC FUNCTIONAL REGRESSION IN TIME SERIES ANALYSIS

2.1 INTRODUCTION

In this chapter we examine the functional k-nearest neighbours, shortly k-NN, non-parametric regression estimate in case of α-mixing data. The classical non-parametric regression estimate introduced in (1.5), Section 1.1, depends on a real-valued non-random bandwidth sequence h_n. On the contrary, the smoothing parameter of the k-NN regression estimate depends on the numbers of neighbours at the point of interest at which we want to make a prediction. In cases where data is sparse, the k-NN kernel estimate has a significant advantage over the classical kernel estimate. The k-NN kernel estimate is also automatically able to take into account the local structure of the data. This advantage, however, may turn into a disadvantage. If there is an outlier in the dataset, the local prediction may be bad. To avoid this, a robust non-parametric regression ansatz may be chosen (for references on this topic see Section 2.6). Selecting the bandwidth depending on the data turns the bandwidth into a random variable. Hence we are no longer able to use the same techniques in the consistency proofs as in the case of a non-random bandwidth sequence.

The k-NN kernel estimate is a widely studied estimate if the explanatory variable is an element of a finite-dimensional space, see Györfi et al. [33]. In the functional case with real-valued response, two different approaches for the k-NN regression estimation exist. The first one, published by Laloë [45], examines a k-NN kernel estimate when the functional variable is an element of a separable Hilbert space. For that case Laloë establishes a weak consistency result. However, his ansatz is not completely functional. Laloë's strategy is to reduce the dimension of the input variable by using a projection onto a finite-dimensional subspace and then applying multivariate techniques on the projected data. The second result, from Burba et al. [11], is based on a pure functional approach instead. Burba et al. examine the problem on a semi-metric functional space. They proved almost complete convergence and rates for independent data. Furthermore, Burba et al. extended a lemma that we will also use in our proofs. This lemma originates from Collomb [14]. We will cite it in Section 2.4 and make some additional comments on it. Additionally, the k-NN kernel estimate is examined for classification in infinite dimension by Cérou and Guyader [13] and there exists a convergence result for the k-NN regression estimate when the response is an element of a Hilbert space (see Lian [47]).

In the case of a finite-dimensional explanatory variable, the k-NN kernel estimate for α-mixing random variables is treated by Tran [67] and Lu and Cheng [48]. Both results are based on Collomb's [14] results. We combined their idea with Burba et alii's [11] results to prove consistency and the rate.

This chapter is organised as follows. In Section 2.2 we present the k-NN kernel estimate. Afterwards, we introduce the assumptions and the main result, the almost complete convergence and the convergence rate. In Section 2.4, some technical auxiliary results are deployed and in Section 2.5, we show the proofs of our main result. In the end, we outline some applications and discuss how to get a robust k-NN kernel estimate.

2.2 METHOD AND ASSUMPTIONS

Let $(X_i, Y_i)_{i=1}^n$ be n pairs identically distributed as (X, Y), the latter being a random pair with values in the measurable space $(E \times \mathbb{R}, \mathcal{E}_d \otimes \mathcal{B})$. Here (E, d) is a semi-metric space and $\mathcal{E}_d$ is the σ–algebra generated by the topology of E that is defined by the semi-metric d, and $\mathcal{B}$ is the Borel σ–algebra. In order to characterise the model of dependence, we use the notion of α-mixing.

We examine the k-NN kernel estimate that is defined for $x \in E$ as

$$\hat{m}_{k\text{-NN}}(x) = \sum_{i=1}^{n} Y_i \frac{K\left(H_{n,k}^{-1} d(x, X_i)\right)}{\sum_{i=1}^{n} K\left(H_{n,k}^{-1} d(x, X_i)\right)}, \quad \text{if} \quad \sum_{j=1}^{n} K\left(H_{n,k}^{-1} d(x, X_i)\right) \neq 0, \tag{2.1}$$

otherwise $\hat{m}_{k\text{-NN}}(x) = 0$. $K : \mathbb{R} \to \mathbb{R}^+$ is a kernel function and $H_{n,k}$ is the bandwidth that is defined as

$$H_{n,k} := d(x, X_{(k)}), \tag{2.2}$$

where the sequence $(X_{(i)}, Y_{(i)})_{i=1}^n$ is the re-indexed sequence $(X_i, Y_i)_{i=1}^n$ such that

$$d(x, X_{(1)}) \leqslant d(x, X_{(2)}) \leqslant \ldots \leqslant d(x, X_{(n)}).$$

From now on, when we refer to the bandwidth of the k-NN kernel estimate, we mean the number of neighbours k we are considering.

To prove the almost complete convergence of the k-NN kernel estimate, we need some results of the Nadaraya-Watson kernel estimate. Hereafter, the notion kernel estimate will refer to the Nadaraya-Watson kernel estimate. Let $x \in E$, then

$$\hat{m}(x) = \sum_{i=1}^{n} Y_i \frac{K\left(h_n^{-1} d(x, X_i)\right)}{\sum_{i=1}^{n} K\left(h_n^{-1} d(x, X_i)\right)}, \quad \text{if} \quad \sum_{j=1}^{n} K\left(h_n^{-1} d(x, X_i)\right) \neq 0, \tag{2.3}$$

otherwise $\hat{m}(x) = 0$. K is a kernel function and $h := h_n$ is a non-random bandwidth.

Prior to the presentation of our main results, we outline the assumptions.

Condition on the small ball probability

(F) Let $x \in E$. Assume that the probability of observing the functional random variable X around x is strictly positive, that means

$$\forall \varepsilon > 0 : \; F_x(\varepsilon) := P\left(d(x, X) \leqslant \varepsilon\right) > 0.$$

Condition on the kernel function K

(K) Assume that the kernel function K is of continuous- or of discontinuous-type. Furthermore, assume for continuous-type kernel functions following technical assumption

$$\exists C > 0 \; \exists \varepsilon_0 > 0 \; \forall \; 0 < \varepsilon < \varepsilon_0 : \; \int_0^{\varepsilon} F(u)du > C\varepsilon F(\varepsilon).$$

Condition on the response variable Y

(M) Assume that the conditional moments of Y are bounded,

$$\forall m \in \mathbb{N} : \; E\left[|Y|^m | X = x\right] < \sigma_m(x) < \infty,$$

with $\sigma_m(\cdot)$ continuous at x.

Condition on the mixing coefficient

(A) Assume that the sequence (X_i, Y_i) is arithmetic α-mixing (or algebraic),

$$\exists C > 0 : \; \alpha(n) \leqslant Cn^{-b}$$

for some $C > 0$ and rate $b > 0$, which is defined more exactly in the theorems.

Condition on the covariance terms

The terms of covariance, which are a measure of dependence, are here denoted by

$$s_{n,1}(x) = \sum_{i,j=1}^{n} |Cov(\Delta_i(x), \Delta_j(x))| \text{ and}$$
$$s_{n,2}(x) = \sum_{i,j=1}^{n} |Cov(Y_i\Delta_i(x), Y_j\Delta_j(x))|,$$

where

$$\Delta_i(x) := \frac{K(h^{-1}d(x, X_i))}{E\left[K(h^{-1}d(x, X_1))\right]}.$$

Note that we can split for example $s_{n,2}(x)$ as

$$s_{n,2}(x) = \underbrace{\sum_{i=1}^{n} Var\left[Y_i\Delta_i(x)\right]}_{I} + \underbrace{\sum_{\substack{i,j=1 \\ j\neq i}}^{n} |Cov(Y_i\Delta_i(x), Y_j\Delta_j(x))|}_{II}. \tag{2.4}$$

Term II in (2.4) is a measure of the dependence of the random variables. We want to remark, if the X_i are α-mixing then also the $\Delta_i(x)$ are α-mixing, see e.g. Lemma 10.3 in [30, p. 155].

(D) Assume for the covariance term $s_n(x) := \max\{s_{n,1}(x), s_{n,2}(x)\}$ that there exists a $\theta > 2$ such that

$$s_n^{-(b+1)} = o\left(n^{-\theta}\right),$$

where b is the rate of the mixing coefficient.

Condition on the bandwidth

(B) Assume for the sequence of bandwidths $k := k_n$ that there exists a $\gamma \in (0,1)$ such that

$$k \sim n^{\gamma}.$$

Condition (B) is not more restrictive than in the independent case. However, for their consistency result Burba et al. [11] need the following two conditions,

$$\frac{k}{n} \to 0 \text{ and } \frac{\log n}{k} \to 0 \text{ as } n \to \infty,$$

so k must exceed logarithmic order. As Lian comments in [47], in most cases in the functional context the small ball probability is of exponential-type. Hence the convergence speed is logarithmic, no matter if the number of neighbours k increases logarithmically or polynomially. For example, if we have for the small ball probability

$$F_x(h) \sim \exp\left(-\frac{1}{h^{\tau}}\right), \text{ then } F^{-1}\left(\frac{k}{n}\right) \sim \left(\frac{1}{\log\left(\frac{k}{n}\right)}\right)^{\tau},$$

where $F_x^{-1}(y) := \inf\{h | F_x(h) \geqslant y\}$ (see [47]). It can be easily seen that the order of k is less important for such small probabilities.

Condition on the distribution and joint distribution function

(D1) This condition is on the distribution of two distinct pairs (X_i, Y_i) and (X_j, Y_j). We assume that

$$\forall i \neq j: \ E\left[Y_i Y_j | X_i X_j\right] \leqslant C < \infty,$$

and the joint distribution functions $P\left(X_i \in B(x,h),\ X_j \in B(x,h)\right)$ satisfy

$$\exists \varepsilon_1 \in (0,1]: \ 0 < G_x(h) = \mathcal{O}\left(F_x(h)^{1+\varepsilon_1}\right),$$

where

$$G_x(h) := \max_{i,j \in \{1,\ldots,n\}, i \neq j} P\left(X_i \in B(x,h),\ X_j \in B(x,h)\right).$$

Condition (D1) is, as Ferraty and Vieu [30, p. 163] in Note 11.2 describe, not too restrictive. For example, if we choose $E = \mathbb{R}^p$, then $\varepsilon_1 = 1$ as soon as each pair of random variables (X_i, X_j) has a bounded density $f_{i,j}$ with respect to the Lebesgue measure.

Next, we formulate a more general condition on the joint distribution function.

(D2) Define $\chi(x,h) := \max\left\{1, \frac{G_x(h)}{F_x(h)^2}\right\}$ and $s = 1/(b+1)$ with b as the rate of the mixing coefficient. Then assume that

$$\frac{\log(n)\chi(x,h)^{1-s}n^{1+s}}{k^2} \to 0.$$

2.3 ALMOST COMPLETE CONVERGENCE AND ALMOST COMPLETE CONVERGENCE RATE

Before we present the consistency result of the k-NN kernel estimate the almost complete convergence result of the kernel regression estimate $\hat{m}(x)$ of Ferraty and Vieu [30] is presented.

Theorem 2.3.1 (Ferraty and Vieu [30], p. 63) *Assume that the regression function is of continuity-type (Def. 1.3.1), furthermore assume (F), (M), (A), and (K). Additionally, suppose for the bandwidth that $h_n \to 0$ and $\frac{\log n}{nF_x(h_n)} \to 0$ as $n \to \infty$. Then we have for the Nadaraya-Watson kernel estimate for $x \in E$*

$$\lim_{n\to\infty} \hat{m}(x) = m(x) \quad \text{almost completely.}$$

The following theorem gives almost complete rates.

Theorem 2.3.2 (Ferraty and Vieu [30], p. 80) *Assume the same conditions as in Theorem 2.3.1, and a Hölder-type model (Def. 1.3.2) instead of a continuity-type model. Then we have for the Nadaraya-Watson kernel estimate for $x \in E$*

$$\hat{m}(x) - m(x) = \mathcal{O}\left(h^\beta\right) + \mathcal{O}_{a.co.}\left(\frac{\sqrt{s_n(x)\log n}}{n}\right).$$

Now we state the almost complete convergence result for the non-parametric k-NN kernel estimate, introduced in (2.1).

Theorem 2.3.3 *In the case of a continuity-type model, we suppose condition (F) for the small ball probability, (K) for the kernel function, (B) for the bandwidth k. Either assume that Condition (D1) holds with*

$$b > \max\left\{\frac{3}{2\gamma} - 1, \frac{2-\gamma}{\varepsilon_1(1-\gamma)}\right\},$$

where γ is the constant in Condition (B) and ε_1 the constant in Condition (D1). Or assume that Condition (D2) is enforced, with rate

$$b > \frac{3}{2\gamma} - 1.$$

Then we have for the k-NN kernel estimate for $x \in E$

$$\lim_{n\to\infty} \hat{m}_{k\text{-NN}}(x) = m(x) \quad \text{almost completely.}$$

Theorem 2.3.4 *In the case of a Hölder-type model, we suppose condition (F) for the small ball probability, (K) the kernel function, (B) the bandwidth* k.
If Condition (D1) holds with

$$b > \max\left\{\frac{3}{2\gamma} - 1, \frac{2-\gamma}{\varepsilon_1(1-\gamma)}\right\},$$

where γ *is the constant in Condition (B) and* ε_1 *the constant in Condition (D1). Then we have for the k-NN kernel estimate for* $x \in E$

$$\hat{m}_{k\text{-NN}}(x) - m(x) = \mathcal{O}\left(\left(F_x^{-1}\left(\frac{k}{n}\right)\right)^{\beta}\right) + \mathcal{O}_{a.co.}\left(\sqrt{\frac{\log n}{k}}\right). \tag{2.5}$$

If (D2) holds instead of (D1) with

$$b > \frac{3}{2\gamma} - 1,$$

then we have for the k-NN kernel estimate for $x \in E$

$$\begin{aligned}\hat{m}_{k\text{-NN}}(x) - m(x) = \mathcal{O}\left(\left(F_x^{-1}\left(\frac{k}{n}\right)\right)^{\beta}\right) + \mathcal{O}_{a.co.}\left(\sqrt{\frac{\log n}{k}}\right) \\ + \mathcal{O}_{a.co.}\left(\sqrt{\frac{n^{1+s}\log n}{k^2}\chi\left(x, F_x^{-1}\left(\frac{k}{n}\right)\right)^{1-s}}\right),\end{aligned} \tag{2.6}$$

where $\chi(x,h) := \max\left\{1, \frac{G_x(h)}{F_x(h)^2}\right\}$.

The covariance term $s_n(x)$ disappears in (2.5). The Condition (D1) and the condition on the rate b implies that term II in (2.4) decays faster than term I. We get

$$s_n(x) = \mathcal{O}\left(\frac{n}{F_x(h)}\right),$$

see Lemma 11.5 in [30, p. 166]. If Condition (D2) instead of (D1) is assumed we get three terms for the rate (see (2.6)). The first one in (2.6) has its origin in the regularity of the regression function, the second one stems from term I in (2.4) and the third one represents the dependence of the random variables (compare term II in (2.4)).

2.4 TECHNICAL TOOLS

Because of the randomness of the smoothing parameter $H_{n,k}$, it is not possible to use the same tools for proving the consistency as in the kernel estimation. The necessary tools are presented in this section. The following two lemmas of Burba

et al. [11] are generalisations of a result firstly presented by Collomb [14]. In our opinion, Burba et alii's [11] Lemmas 2.4.1 and 2.4.2 are valid for dependent random variables as in the original lemma from Collomb [14]. We checked the proof from Burba et al. against Collomb's proof; we did not find any reason why Burba et al. [11] assume independence. On reflection, this assumption appears unnecessary.

Let $(A_i, B_i)_{i=1}^n$ be a sequence of random variables with values in $(\Omega \times \mathbb{R}, \mathcal{A} \otimes \mathcal{B})$, not necessarily identically distributed or independent. Let $k : \mathbb{R} \times \Omega \to \mathbb{R}^+$ be a measurable function with the property

$$z \leqslant z' \Rightarrow \forall \omega \in \Omega : \; k(z, \omega) \leqslant k(z', \omega).$$

Let H be a real-valued random variable. Then define

$$\forall n \in \mathbb{N} : \; c_n(H) = \frac{\sum_{i=1}^{n} B_i k(H, A_i)}{\sum_{i=1}^{n} k(H, A_i)}. \tag{2.7}$$

Lemma 2.4.1 (Burba et al. [11]) *Let (D_n) be a sequence of real random variables and (u_n) be a decreasing sequence of positive numbers.*

- *If $l = \lim_n u_n \neq 0$ and if, for all increasing sequences $\beta_n \in (0,1)$, there exist two sequences of real random variables $(D_n^-(\beta_n))$ and $(D_n^+(\beta_n))$ (depending on the sequence (β_n)) such that*

  (L1) $\forall n \in \mathbb{N}\ D_n^- \leqslant D_n^+$ *and* $1_{[D_n^- \leqslant D_n \leqslant D_n^+]} \to 1$ *almost completely,*

  (L2) $\frac{\sum_{i=1}^{n} k(D_n^-, A_i)}{\sum_{i=1}^{n} k(D_n^+, A_i)} - \beta_n = \mathcal{O}_{a.co.}(u_n)$,

  (L3) Assume there exists a real positive number c such that $c_n(D_n^-) - c = \mathcal{O}_{a.co.}(u_n)$ *and* $c_n(D_n^+) - c = \mathcal{O}_{a.co.}(u_n)$.

  Then $c_n(D_n) - c = \mathcal{O}_{a.co.}(u_n)$.

- *If $l = 0$ and if (L1), (L2), and (L3) hold for any increasing sequence $\beta_n \in (0,1)$ with limit 1, the same conclusion holds.*

Lemma 2.4.2 (Burba et al. [11]) *Let (D_n) be a sequence of real random variables and $(v_n)_n$ a decreasing positive sequence.*

- *If $l' = \lim_n v_n \neq 0$ and if, for all increasing sequences $\beta_n \in (0,1)$, there exist two sequences of real random variables $(D_n^-(\beta_n))$ and $(D_n^+(\beta_n))$ such that*

  (L1') $D_n^- \leqslant D_n^+\ \forall n \in \mathbb{N}$ *and* $1_{[D_n^- \leqslant D_n \leqslant D_n^+]} \to 1$ *almost completely,*

  (L2') $\frac{\sum_{i=1}^{n} k(D_n^-, A_i)}{\sum_{i=1}^{n} k(D_n^+, A_i)} - \beta_n = o_{a.co.}(v_n)$,

  (L3') Assume there exists a real positive number c such that $c_n(D_n^-) - c = o_{a.co.}(v_n)$ *and* $c_n(D_n^+) - c = o_{a.co.}(v_n)$.

Then $c_n(D_n) - c = o_{a.co.}(v_n)$,

- *If* $l' = 0$ *and if (L1'), (L2'), and (L3') are checked for any increasing sequence* $\beta_n \in (0,1)$ *with limit* 1, *the same result holds.*

Burba et al. [11] use in their consistency proof of the k-NN kernel estimate for independent data a Chernoff-type exponential inequality to ckeck Conditions (L1) or (L1'). In the case of α-mixing random variables however, we cannot use that exponential inequality. Instead we use the following lemma of Bradley [5] and Lemma 2.4.4.

Lemma 2.4.3 (Bradley [5], p. 20) *Let* (X, Y) *be a* $\mathbb{R}^r \times \mathbb{R}$ *valued random vector, such that* $Y \in L_p(P)$ *for some* $p \in [1, \infty]$. *Let* d *be a real number such that* $\|Y + d\|_p > 0$ *and* $\varepsilon \in (0, \|Y + d\|_p]$. *Then there exists a random variable* Z *such that*

- $P_Z = P_Y$ *and* Z *is independent of* X,
- $P(|Z - Y| > \varepsilon) \leqslant 11 \left(\frac{\|Y+d\|_p}{\varepsilon}\right)^{\frac{p}{2p+1}} [\alpha(\sigma(X), \sigma(Y))]^{\frac{p}{2p+1}}$, *where* $\sigma(X)$ *is the* σ-*Algebra generated by* X.

The following lemma is needed in our proofs for technical reasons.

Lemma 2.4.4 *Let* (X_i) *be an arithmetically* α*-mixing sequence in the semi-metric space* (E, d), $\alpha(n) \leqslant cn^{-b}$, *with* $b, c > 0$. *Define* $\Delta_i(x) := 1_{B(x,h)}(X_i)$. *Then we have*

$$\sum_{i,j=1}^{n} |\mathrm{Cov}\left(\Delta_i(x), \Delta_j(x)\right)| = \mathcal{O}\left(nF_x(h)\right) + \mathcal{O}\left(\chi(x,h)^{1-s} n^{1+s}\right),$$

where $\chi(x,h) := \max\{G_x(h), F_x(h)^2\}$ *and* $s = \frac{1}{b+1}$.

Proof of Lemma 2.4.4:
The proof of this lemma is identical to that of Lemma 3.2 in [29], except for the choice of the parameter s.

□

2.5 PROOFS

Proof of Theorem 2.3.3:
To prove this theorem we apply Lemma 2.4.2. The main difference to the proof of the independent case in [11] concerns verification of (L1'). To verify (L2') and (L3') we need only small modifications.

Let $v_n = 1$, $c_n(H_{n,k}) = \hat{m}_{k\text{-NN}}(x)$ and $c = m(x)$. Choose $\beta \in (0,1)$ arbitrarily, D_n^+ and D_n^- such that

$$F_x(D_n^+) = \frac{1}{\sqrt{\beta}} \frac{k}{n}, \quad \text{and} \quad F_x(D_n^-) = \sqrt{\beta} \frac{k}{n}.$$

Define $h^+ := D_n^+ = F^{-1}\left(\sqrt{\beta}\frac{k}{n}\right)$ and $h^- := D_n^- = F^{-1}\left(\frac{1}{\sqrt{\beta}}\frac{k}{n}\right)$.

To apply Theorem 2.3.1, we have to show that the covariance term s_n fulfils following condition: there exists a $\theta > 2$ such that

$$s_n^{-(b+1)} = o\left(n^{-\theta}\right), \tag{2.8}$$

where b is the rate of the mixing coefficient. If (D1) and the condition on the rate b of the mixing coefficient holds, we have by Lemma 11.5 in [30, p. 166]

$$\begin{aligned} s_n(x) &= \mathcal{O}\left(\frac{n}{F_x(h^+)}\right) \\ &= \mathcal{O}\left(\frac{n^2}{k}\right). \end{aligned}$$

The same is true for the bandwidth h^-. It can be easily seen that there exists an $\theta > 2$ such that (2.8) holds. In the case of (D2), we have

$$s_n(x) = \mathcal{O}\left(\frac{n^2}{k}\right) + \mathcal{O}\left(\chi(x, h^+)^{1-s} n^{1+s}\right).$$

Since $\chi(x, h^+)^{1-s} n^{1+s} > 0$ for all n, it turns out that (2.8) holds under Condition (D2) as well.

Consequently, we are able to apply Theorem 2.3.1 to quarantee

$$\begin{aligned} &c_n(D_n^+) \to c \text{ almost completely, and} \\ &c_n(D_n^-) \to c \text{ almost completely.} \end{aligned}$$

Thus Condition (L3′) is verified.

In [30, p. 162] Ferraty and Vieu proved under the conditions of Theorem 2.3.1 that

$$\frac{1}{nF_x(h)} \sum_{i=1}^{n} K(h^{-1} d(x, X_i)) \to 1 \text{ almost completely.} \tag{2.9}$$

By (2.9) we have

$$\begin{aligned} &\frac{1}{nF_x(h^+)} \sum_{i=1}^{n} K(h^{+^{-1}} d(x, X_i)) \to 1 \text{ almost completely and} \\ &\frac{1}{nF_x(h^-)} \sum_{i=1}^{n} K(h^{-^{-1}} d(x, X_i)) \to 1 \text{ almost completely.} \end{aligned}$$

We get

$$\frac{\sum_{i=1}^{n} K(h^{+^{-1}} d(x, X_i))}{\sum_{i=1}^{n} K(h^{-^{-1}} d(x, X_i))} \to \beta.$$

Condition (L2′) is proved.

Finally, we check (L1′),

$$\forall \varepsilon > 0 : \sum_{n=1}^{\infty} P\left(|1_{\{D_n^- \leqslant H_{n,k} \leqslant D_n^+\}} - 1| > \varepsilon\right) < \infty.$$

Let $\varepsilon > 0$ be fixed. We know that

$$P\left(|1_{\{D_n^- \leqslant H_{n,k} \leqslant D_n^+\}} - 1| > \varepsilon\right) \leqslant P\left(H_{n,k} < D_n^-\right) + P\left(H_{n,k} > D_n^+\right). \tag{2.10}$$

For the two terms in (2.10) we obtain

$$\begin{aligned} P\left(H_{n,k} < D_n^-\right) &\leqslant P\left(\sum_{i=1}^{n} 1_{B(x,D_n^-)}(X_i) > k\right) \\ &\leqslant P\left(\sum_{i=1}^{n} \left(1_{B(x,D_n^-)}(X_i) - F_x(D_n^-)\right) > k - nF_x(D_n^-)\right) \\ &=: P_{1n} \end{aligned} \tag{2.11}$$

and

$$\begin{aligned} P\left(H_{n,k} > D_n^+\right) &\leqslant P\left(\sum_{i=1}^{n} 1_{B(x,D_n^+)}(X_i) < k\right) \\ &\leqslant P\left(\sum_{i=1}^{n} \left(1_{B(x,D_n^+)}(X_i) - F_x(D_n^+)\right) < k - nF_x(D_n^+)\right) \\ &=: P_{2n} \end{aligned} \tag{2.12}$$

In the second step of (2.11) and (2.12), we centred the random variables $1_{B(x,D_n^-)}(X_i)$ and $1_{B(x,D_n^+)}(X_i)$. It holds

$$E\left[1_{B(x,D_n^-)}(X_i)\right] = F_x(D_n^-) \text{ and } E\left[1_{B(x,D_n^+)}(X_i)\right] = F_x(D_n^+).$$

At this step, Burba et al. [11] use the independence of the random variables. The plan here is to split the data into a block scheme as is done by Modha and Masry [52], Oliveira [54], Tran [67] or Lu and Cheng [48]. Afterwards we are applying Lemma 2.4.3.

Divide the set $\{1, \ldots, n\}$ into blocks of length $2l_n$, set $m_n = [n/2l_n]$, where $[\cdot]$ is the Gaussian bracket and $f_n = n - 2l_n m_n < 2l_n$. The sequences are chosen such that $m_n \to \infty$ and $f_n \to \infty$. l_n is specified later on in the proof, see (2.16). By this choice we have $n = 2l_n m_n + f_n$.

Firstly, we examine term P_{1n}. Let

$$U_n(j) := \sum_{i=(j-1)l_n+1}^{jl_n} \left(1_{B(x,D_n^-)}(X_i) - F_x(D_n^-)\right),$$

and define

$$B_{n1} := \sum_{j=1}^{m_n} U_n(2j-1), \quad B_{n2} := \sum_{j=1}^{m_n} U_n(2j), \text{ and}$$

$$R_n := \sum_{i=2l_n m_n+1}^{n} \left(1_{B(x,D_n^-)}(X_i) - F_x(D_n^-)\right).$$

We get

$$\begin{aligned} P_{1n} &\leqslant P\left(B_{n1} > \frac{k - nF_x(D_n^-)}{3}\right) + P\left(B_{n2} > \frac{k - nF_x(D_n^-)}{3}\right) \\ &\quad + P\left(R_n > \frac{k - nF_x(D_n^-)}{3}\right) \\ &=: P_{1n}^{(1)} + P_{1n}^{(2)} + P_{1n}^{(3)} \end{aligned} \tag{2.13}$$

Let us consider $P_{1n}^{(1)}$.

Lemma 2.4.3 with $d := l_n m_n$ leads to

$$0 < l_n m_n \leqslant \|U_n(2j-1) + d_n\|_\infty \leqslant 2l_n + l_n m_n.$$

Because of $m_n l_n = \mathcal{O}(n)$ and $\frac{k}{n} \to 0$, we have

$$\varepsilon := \frac{k - nF_x(D_n^-)}{6m_n} = \frac{k(1-\sqrt{\beta})}{6m_n} \in (0, \|U_n(2j-1) + d_n\|_\infty].$$

This choice of ε is motivated by (2.15) below. By Lemma 2.4.3 we can construct $\left(\tilde{U}_n(2j-1)\right)_{j=1}^{m_n}$ such that

- the random variables $\left(\tilde{U}_n(2j-1)\right)_{j=1}^{m_n}$ are independent,
- $\tilde{U}_n(2j-1)$ has the same distribution as $U_n(2j-1)$ for $j = 1, \ldots, m_n$,
- and

$$\begin{aligned} P\left(|\tilde{U}_n(2j-1) - U_n(2j-1)| > \varepsilon\right) \leqslant 11 &\left(\frac{\|U_n(2j-1) + d\|_\infty}{\varepsilon}\right)^{\frac{1}{2}} \cdot \\ &\cdot \sup |P(AB) - P(A)P(B)|, \end{aligned}$$

  where the supremum is taken over all sets A and B with

$$A, B \in \sigma\left(U_n(1), U_n(3), \ldots, U_n(2m_n - 1)\right).$$

This leads to

$$\begin{aligned} P_{1n}^{(1)} &= P\left(\sum_{j=1}^{m_n} \left[\tilde{U}_n(2j-1) + (U_n(2j-1) - \tilde{U}_n(2j-1))\right] > \frac{k - nF_x(D_n^-)}{3}\right) \\ &\leqslant P\left(\sum_{j=1}^{m_n} \tilde{U}_n(2j-1) > \frac{k - nF_x(D_n^-)}{6}\right) \\ &\quad + P\left(\sum_{j=1}^{m_n} (U_n(2j-1) - \tilde{U}_n(2j-1)) > \frac{k - nF_x(D_n^-)}{6}\right) \\ &=: P_{1n}^{(11)} + P_{1n}^{(12)}. \end{aligned} \tag{2.14}$$

Applying Lemma 2.4.3 on $P_{1n}^{(12)}$,

$$\begin{aligned} P_{1n}^{(12)} &\leqslant \sum_{j=1}^{m_n} P\left((U_n(2j-1) - \tilde{U}_n(2j-1)) > \frac{k - nF_x(D_n^-)}{6m_n} \right) \\ &\leqslant m_n \left(\frac{6m_n l_n (m_n+1)}{k(1-\sqrt{\beta})} \right)^{\frac{1}{2}} \alpha(l_n) \\ &= \left(\frac{6m_n^3 l_n^4 (m_n+1)}{l_n^3 k(1-\sqrt{\beta})} \right)^{\frac{1}{2}} \alpha(l_n) \\ &\leqslant C \frac{n^2}{l_n^{\frac{3}{2}} k} \alpha(l_n). \end{aligned} \tag{2.15}$$

We choose the sequence l_n such that

$$l_n^a = \frac{n^2}{2^a r^a k}, \tag{2.16}$$

where r is a positive constant specified below and $a > 2/\gamma - 1$. By the condition on the mixing coefficient b and some calculations

$$\begin{aligned} \frac{n^2}{l_n^{3/2} k} \alpha(l_n) &= C \left(\frac{n^{2/a}}{k^{1/a}} \right)^{a-3/2} \left(\frac{n^{2/a}}{k^{1/a}} \right)^{-b} \\ &= C n^{(2-\gamma)(a-3/2-b)/a} \\ &\leqslant n^{-l} \end{aligned}$$

for some $l > 1$. Consequently, by the assumptions we arrive at

$$\sum_{n=1}^{\infty} P_{1n}^{(12)} < \infty. \tag{2.17}$$

Apply now Markov's inequality on term $P_{1n}^{(11)}$ for some $t > 0$,

$$\begin{aligned} &P\left(\sum_{j=1}^{m_n} \tilde{U}_n(2j-1) > \frac{k - nF_x(D_n^-)}{6} \right) \\ &\leqslant \exp\left(-t \frac{k - nF_x(D_n^-)}{6} \right) E\left[\exp\left(t \sum_{j=1}^{m_n} \tilde{U}_n(2j-1) \right) \right]. \end{aligned} \tag{2.18}$$

Due to the independence of the random variables $(\tilde{U}_n(2j-1))_{j=1}^{m_n}$, we have

$$E\left[\exp\left(t \sum_{j=1}^{m_n} \tilde{U}_n(2j-1) \right) \right] = \prod_{j=1}^{m_n} E\left[\exp\left(t\tilde{U}_n(2j-1) \right) \right]. \tag{2.19}$$

Choose $t := r \log n/k$, then we obtain together with l_n as defined in (2.16)

$$\begin{aligned} t|\tilde{U}_n(2j-1)| &\leqslant \frac{2r l_n \log n}{k} \\ &= \frac{\log(n) n^{\frac{2}{a}}}{k^{1+\frac{1}{a}}} \\ &= \log n \left(\frac{n^2}{k^{a+1}} \right)^{\frac{1}{a}}. \end{aligned}$$

In this step, we need the number of neighbours to be a power in n, i.e. $k \sim n^\gamma$. By the choice of $a > 2/\gamma - 1$, we have for large n that $t|\tilde{U}_n(2j-1)| \leqslant 1$. In the next step we use the same idea as Craig [16] in his proof. We have for large n

$$\exp\left(t\tilde{U}_n(2j-1)\right) \leqslant 1 + t\tilde{U}_n(2j-1) + t^2\tilde{U}_n(2j-1)^2.$$

The random variable $\tilde{U}_n(2j-1)$ has the same distribution as the centred random variable $U_n(2j-1)$. Hence we know that the expectation of the linear term is zero, $E\left[\tilde{U}_n(2j-1)\right] = 0$. With this and $1 + x \leqslant \exp(x)$ we get

$$\begin{aligned} E\left[\exp\left(t\tilde{U}_n(2j-1)\right)\right] &\leqslant 1 + E\left[t^2\tilde{U}_n(2j-1)^2\right] \\ &\leqslant \exp\left(t^2 E\left[\tilde{U}_n(2j-1)^2\right]\right). \end{aligned} \tag{2.20}$$

Furthermore, because $\tilde{U}_n(2j-1)$ and $U_n(2j-1)$ have the same distribution function and by some calculations, it follows that

$$\sum_{j=1}^{m_n} E\left[\tilde{U}_n(2j-1)^2\right] \leqslant \sum_{i,j=1}^{n} \left|\mathrm{Cov}(1_{B(x,D_n^-)}(X_i), 1_{B(x,D_n^-)}(X_j))\right|.$$

Since $F_x(D_n^-) = \sqrt{\beta}\frac{k}{n}$ and $k \sim n^\gamma$, we know that

$$F_x(D_n^-) = \mathcal{O}\left(n^{\gamma-1}\right).$$

We apply Lemma 2.4.4 and get in the case of (D2)

$$\begin{aligned} \sum_{j=1}^{m_n} E\left[\tilde{U}_n(2j-1)^2\right] &\leqslant C_1 n F_x(D_n^-) + C_2\chi(D_n^-)^{1-s}n^{1+s} \\ &= C_1\sqrt{\beta}k + C_2\chi(D_n^-)^{1-s}n^{1+s}, \end{aligned} \tag{2.21}$$

and in the case of (D1)

$$\begin{aligned} \sum_{j=1}^{m_n} E\left[\tilde{U}_n(2j-1)^2\right] &\leqslant C_1 n F_x(D_n^-) \\ &= C_1\sqrt{\beta}k. \end{aligned}$$

Below, we present the arguments if Condition (D2) holds, because in the case of (D1) the rationale follows the same line. By (2.19), (2.20), (2.21), and $t := r\log n/k$, we have for the second term in (2.18)

$$\begin{aligned} E\left[\exp\left(t\sum_{j=1}^{m_n}\tilde{U}_n(2j-1)\right)\right] \leqslant &\exp\left(C_1\sqrt{\beta}r^2\frac{(\log n)^2}{k}\right)\cdot \\ &\cdot\exp\left(C_2\sqrt{\beta}r^2\frac{(\log n)^2\chi(D_n^-)^{1-s}n^{1+s}}{k^2}\right). \end{aligned} \tag{2.22}$$

By $k \sim n^\gamma$, we know that the first term in (2.22) satisfies

$$\exp\left(C_1\sqrt{\beta}r^2\frac{(\log n)^2}{k}\right) \to 1 \quad \text{as} \quad n \to \infty.$$

If (D2) holds, we have for the second term in (2.22)

$$\exp\left(C_2\sqrt{\beta}\mu^2\frac{(\log n)^2\chi(D_n^-)^{1-s}n^{1+s}}{k^2}\right)\to 1 \text{ as } n\to\infty.$$

Since $F_x(D_n^-)=\sqrt{\beta}\frac{k}{n}$, $t=r\log n/k$, and by choosing $r>6/(1-\sqrt{\beta})$, we find for the first term in (2.18)

$$\begin{aligned}\exp\left(-t\frac{k-nF_x(D_n^-)}{6}\right)&=\exp\left(-\frac{r(1-\sqrt{\beta})}{6}\log(n)\right)\\&=n^{-\frac{r(1-\sqrt{\beta})}{6}}\\&\leqslant n^{-l}\end{aligned}$$

for some $l>1$. By this,

$$\sum_{n=1}^{\infty}P_{1n}^{(11)}<\infty \tag{2.23}$$

Now, combine relations (2.17) and (2.23) to obtain

$$\sum_{n=1}^{\infty}P_{1n}^{(1)}\leqslant\sum_{n=1}^{\infty}P_{1n}^{(11)}+\sum_{n=1}^{\infty}P_{1n}^{(12)}<\infty.$$

By similar arguments as for $P_{1n}^{(1)}$ we receive

$$\sum_{n=1}^{\infty}P_{1n}^{(2)}<\infty.$$

Finally, we examine

$$P_{1n}^{(3)}=P\left(R_n>\frac{k-nF_x(D_n^-)}{3}\right).$$

We know that

$$\begin{aligned}|R_n|&=\left|\sum_{i=2l_nm_n+1}^{n}\left(1_{B(x,D_n^-)}(X_i)-F_x(D_n^-)\right)\right|\\&\leqslant\sum_{i=2l_nm_n+1}^{n}(1_{B(x,D_n^-)}(X_i)+F_x(D_n^-))\\&\leqslant 2\sum_{i=2l_nm_n+1}^{n}\\&\leqslant 4l_n.\end{aligned}$$

and

$$\frac{k-nF_x(D_n^-)}{3}=\mathcal{O}(k).$$

Together with the choice of l_n in (2.16) and the condition on the parameter $a > 2/\gamma - 1$ we have $k > l_n$ for large n. This implies

$$\sum_{n=1}^{\infty} P_{1n}^{(3)} < \infty.$$

Finally, we get

$$\sum_{n=1}^{\infty} P_{1n} \leqslant \sum_{n=1}^{\infty} P_{1n}^{(1)} + \sum_{n=1}^{\infty} P_{1n}^{(2)} + \sum_{n=1}^{\infty} P_{1n}^{(3)} < \infty.$$

Analysis of P_{2n} is similar to that of P_{1n}. By the definition of $nF_x(D_n^+)$

$$k - nF_x(D_n^+) = k\frac{\sqrt{\beta} - 1}{\sqrt{\beta}} < 0,$$

we find

$$P_{2n} = P\left(\sum_{i=1}^{n}\left(F_x(D_n^+) - 1_{B(x,D_n^+)}(X_i)\right) > nF_x(D_n^+) - k\right).$$

Then by similar reasoning as for P_{1n}, we get

$$\sum_{n=1}^{\infty} P_{2n} < \infty.$$

This finishes the proof of Condition (L1'), which states that

$$1_{[D_n^- \leqslant D_n \leqslant D_n^+]} \to 1 \text{ almost completely.}$$

Now, we are in the position to apply Lemma 2.4.2 to obtain the desired result,

$$\lim_{n\to\infty} \hat{m}_{k\text{-NN}}(x) = m(x) \text{ almost completely.}$$

□

Proof of Theorem 2.3.4:
To prove this theorem we use Lemma 2.4.1 from Burba et al. [11]. The conditions of Lemma 2.4.1 are proven in a similar manner as in the proof of Theorem 2.3.4. Condition (L1) is the same as (L1') of Lemma 2.4.2. So the proof can be omitted here. Conditions (L2) and (L3) are checked in a similar way as in the proof of Theorem 2.3.3. In [30, p. 162] Ferraty and Vieu prove under the conditions of Theorem 2.3.2 that

$$\frac{1}{n}\sum_{i=1}^{n} K(h^{-1}d(x, X_i)) = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s_n(x)\log n}}{n}\right). \tag{2.24}$$

Choose β_n as an increasing sequence in $(0, 1)$ with limit 1. Furthermore, choose D_n^+ and D_n^- such that

$$F_x(D_n^+) = \frac{1}{\sqrt{\beta_n}}\frac{k}{n} \quad \text{and} \quad F_x(D_n^-) = \sqrt{\beta_n}\frac{k}{n}.$$

If (D1) holds, then

$$\begin{aligned} s_n(x) &= \mathcal{O}\left(\frac{n}{F_x(h^+)}\right) \\ &= \mathcal{O}\left(\frac{n^2}{k}\right). \end{aligned} \tag{2.25}$$

Similar is true for the bandwidth h^-. In the case of (D2), we have for both bandwidth sequences h^- and h^+

$$s_n(x) = \mathcal{O}\left(\frac{n^2}{k}\right) + \mathcal{O}\left(\chi(x,h)^{1-s} n^{1+s}\right). \tag{2.26}$$

Now we are able to apply Theorem 2.3.2 with

$$h^+ = D_n^+ = F^{-1}\left(\sqrt{\beta_n}\frac{k}{n}\right) \text{ and } h^- = D_n^- = F^{-1}\left(\frac{1}{\sqrt{\beta_n}}\frac{k}{n}\right)$$

to get

$$\begin{aligned} c_n(D_n^+) &= \mathcal{O}\left(\left(F_x^{-1}\left(\frac{k}{n}\right)\right)^\beta\right) + \mathcal{O}_{a.co.}\left(\frac{\sqrt{s_n(x)\log n}}{n}\right) \text{ and} \\ c_n(D_n^-) &= \mathcal{O}\left(\left(F_x^{-1}\left(\frac{k}{n}\right)\right)^\beta\right) + \mathcal{O}_{a.co.}\left(\frac{\sqrt{s_n(x)\log n}}{n}\right). \end{aligned}$$

That verifies Condition (L3') is verified. Now, by (2.24) and the same choice of h^+ and h^- as above, we have

$$\begin{aligned} \frac{1}{nF_x(h^+)}\sum_{i=1}^n K(h^{+^{-1}} d(x,X_i)) &= \sqrt{\beta_n}\frac{k}{n} + \mathcal{O}_{a.co.}\left(\frac{\sqrt{s_n(x)\log n}}{n}\right) \text{ and} \\ \frac{1}{nF_x(h^-)}\sum_{i=1}^n K(h^{-^{-1}} d(x,X_i)) &= \sqrt{\beta_n}\frac{k}{n} + \mathcal{O}_{a.co.}\left(\frac{\sqrt{s_n(x)\log n}}{n}\right). \end{aligned}$$

By this, we obtain

$$\frac{\sum_{i=1}^n K(h^{+^{-1}} d(x,X_i))}{\sum_{i=1}^n K(h^{-^{-1}} d(x,X_i))} - \beta_n = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s_n(x)\log n}}{n}\right).$$

To check Condition (L2') we estimate $s_n(x)$ by bounds obtained either by Condition (D1) and $b > (2-\gamma)/(\varepsilon_1(1-\gamma))$ or by (D2), see (2.25) or (2.26). This completes this proof. □

2.6 APPLICATIONS AND RELATED RESULTS

Applications

In the context of functional data analysis the k-NN kernel estimate was first introduced in the monograph of Ferraty and Vieu [30]. There the authors give numerical

examples for the k-NN estimate. They tested it on different data sets, such as electrical consumption in the U.S. [30, p. 200]. In [26], Ferraty et al. examined a data set describing the El Niño phenomenon. Other interesting examples can be found in the R-package *fds (functional data sets)* or Bosq [6, pp. 247]. For both data sets the assumption of α-mixing is plausible. If we have for example a look on the electrical consumption data set, it makes sense that the electrical consumption of the year which we want to predict is more dependent on the near past than on years afterwards.

Related Results

Here we want to outline how to make a robust k-NN kernel estimate. As already mentioned in the introduction, the k-NN estimate is prone to outliers. This disadvantage can be treated by robust regression estimation. For functional data analysis Azzedine et al. [2] introduce a robust non-parametric regression estimate for independent data. Attouch et al. [1] prove the asymptotic normality for the non-parametric regression estimate for α-mixing data. Crambes et al. [17] present results dealing with the L_p error for independent and α-mixing data.

In robust estimation the non-parametric model θ_x can be defined as the root t of the following equation

$$\Psi(x,t) := E\left[\psi_x(Y,t)|X=x\right] = 0. \tag{2.27}$$

The model θ_x is called the $\psi_x(Y,t)$-regression and is a generalisation of the classical regression function. If we choose for example $\psi_x(Y,t) = Y - t$ then we have $\theta_x = m(x)$.

In the case of α-mixing data almost complete convergence and almost complete convergence rate are not yet proven for robust kernel estimate. These results can be easily obtained by arguments similar to those of Azzedine et al. [2] and those for the classical regression function estimation. By such a result and similar arguments as in this section: we get almost complete convergence and related rates for a robust k-NN non-parametric estimate.

Attouch et al. [1], Azzedine et al. [2], or Crambes et al. [17] suggest in their application the L_1-L_2 function

$$\psi(t) = t/\sqrt{(1+t^2)/2}$$

and $\psi_x(t) := \psi\left(t/M(x)\right)$, where $M(x) := \text{med}|Y - \text{med}(Y|X=x)|$ with $\text{med}(Y|X=x)$ as the conditional median of Y given $X = x$. We get the consistency for the kernel estimate of conditional distribution function directly by choosing in (2.7) for $B_i = 1_{(-\infty,y]}(Y_i)$ with Y_i as a real valued random variable distributed as Y, and by this a consistent kernel estimate of $\text{med}(Y|X=x)$.

Alternatively, if one has consistency results for a robust k-NN kernel estimate, we can choose $\psi_x(Y,t) = 1_{[Y \geqslant t]} - 1/2$, to get immediately the consistency for the kernel estimate of the conditional distribution function.

3

UNIFORM ALMOST COMPLETE CONVERGENCE RATES FOR NON-PARAMETRIC ESTIMATES FOR α-MIXING FUNCTIONAL DATA

3.1 INTRODUCTION

This chapter focuses on the uniform convergence of non-parametric estimates of various conditional quantities, such as the conditional expectation, the conditional distribution function, and the conditional density function, assuming α-mixing functional random variables. Uniform convergence with such conditional quantities has been successfully applied for independent data by Ferraty et al. [23]. For the dependent α-mixing case, we have the same applications as in the independent case, as for example bandwidth selection, Behenni et al. [3], Rachdi and Vieu [58], or Chapter 4, additive prediction and boosting, Ferraty and Vieu [32], or building confidence bands by bootstrapping [27]. More references to applications can be found in [23].

In view of non-parametric functional regression, Ferraty and Vieu examine the uniform convergence for α-mixing data in [29] and the errata [31]. The same authors, in an earlier publication [22], analyse the uniform convergence for dependent data where random variables are of the fractal-type. In the errata [31] of [29] Ferraty and Vieu state that the assumption of compactness on the set, where the uniform convergence is proven, is a necessity, namely a finite number of open balls is needed for covering the set of investigation. Since Ferraty and Vieu give no proof of almost complete uniform convergence in [29] and [31], we carry it out here in detail with some modified assumptions. The idea is based on an ably decomposition and the Fuk-Nagaev exponential inequality, see [51] or [62]. In addition to the proof of the kernel estimate of the conditional expectation, we prove almost complete uniform convergence for the kernel estimates of the other above-mentioned conditional quantities. The pointwise almost complete convergence and the rate of these kernel estimates for α-mixing random variables can be found in the monograph of Ferraty and Vieu [30]. To date, the uniform convergence for the kernel estimate of the conditional distribution function and the conditional density function has not been examined for functional α-mixing data. The independent case is examined by Ferraty et al. [23].

The second reason why we examine uniform convergence is that Ferraty and Vieu assume in [29] and [31] that the covering numbers increases polynomially. In the examples in Section 3.2.2 or Ferraty et al. [29] it can be seen that interesting function spaces exist that have covering number that grow exponentially. For inde-

pendent data, Ferraty et al. [29] expand the result of uniform convergence to such a class of function spaces. We take this step for dependent data in this chapter. The price we have to pay for this is that we have to weaken the dependence of the functional random variables. After a closer look at the proofs, it can be seen that under the condition of arithmetic mixing, the extension to a wider function class does not work. We have to assume that the data is geometrically mixing. Furthermore, to estimate the conditional expectation we have to assume that the tails of the probability distribution function of the response variable Y decays exponentially. For all estimates we split our results into two parts; the first for arithmetically mixing random variables and the second for geometrically mixing random variables.

This chapter is organised as follows: Firstly, in Section 3.2.1 we introduce the general description of the exponential inequality used in the proofs and after that two versions of that inequality corresponding to the two mixing conditions. In Section 3.2.2, we present some topological terms and definitions. Furthermore, we give some examples of covering numbers for some commonly used function spaces. In Section 3.3 we give the almost complete convergence rate of the kernel estimate of the generalised regression function. As already mentioned, we examine the two cases of mixing separately. In the sections thereafter we show in the same manner, the almost complete convergence rate of the kernel estimate of the non-parametric conditional distribution function and the kernel estimate of the non-parametric conditional density function.

3.2 PRELIMINARIES

3.2.1 *Exponential Inequalities for Mixing Random Variables*

This section begins by introducing the Fuk-Nagaev exponential inequality. It is the main tool for proving the uniform convergence of all kernel estimates that are examined in this chapter. The proof of this inequality can be found in Theorem 6.2 in the monograph of Rio [62, p. 84].

Theorem 3.2.1 (Rio [62]) *Let (X_i) be a real-valued and centered α-mixing sequence. Let $Q = \sup_i Q_i$, where*

$$Q_i(u) := \inf\{t \mid P(|X_i| > t) \leqslant u\} \text{ and } s_n := \sum_{i,j=1}^{n} |\mathrm{Cov}(X_i, X_j)|.$$

Let $R(u) := \alpha^{-1}(u)Q(u)$ and $H(u) := R^{-1}(u)$ be the generalised inverse of R. Then we have for $\lambda > 0$ and $r \geqslant 1$

$$P\left(\sup_{k\in[1,n]} |S_k| \geqslant 4\lambda\right) \leqslant 4\left(1 + \frac{\lambda^2}{rs_n}\right)^{-\frac{r}{2}} + 4\frac{n}{\lambda}\int_0^{H(\frac{\lambda}{r})} Q(u)\,du, \tag{3.1}$$

where $S_k := \sum_{i=1}^{k} X_i$.

For arithmetical or geometrical mixing random variables, we get different estimates for the integral in (3.1) Corollary 3.2.1 is for the arithmetical case and Corollary 3.2.2 for the geometrical case.

Corollary 3.2.1 (Rio [62]) *In addition to the conditions of Theorem 3.2.1, assume that*

$$\exists c \geqslant 1,\ b > 1 :\ \alpha(n) \leqslant cn^{-b} \quad \text{for all} \quad n > 0.$$

Assume for all $i \in \mathbb{N}$ *and some* $p > 2$

$$P(|X_i| > t) \leqslant t^{-p},$$

then we have

$$\frac{4}{\lambda} \int_0^{H(\frac{\lambda}{r})} Q(u)du \leqslant 4\frac{C}{r}\left(\frac{\lambda}{r}\right)^{-(b+1)p/(b+p)},$$

where $C = \frac{2p}{(2p-1)}(2^b c)^{(p-1)/(b+p)}$. *If the* X_i *are bounded,* $\|X_i\|_\infty < \infty$, *then we have*

$$\frac{4}{\lambda} \int_0^{H(\frac{\lambda}{r})} Q(u)du \leqslant 2\frac{c}{r}\left(\frac{\lambda}{r}\right)^{-(b+1)}.$$

Corollary 3.2.2 (Merlevede et al. [51], Rio [62]) *In addition to the conditions of Theorem 3.2.1, assume that*

$$\exists b, c > 0 :\ \alpha(n) \leqslant \exp\left(-cn^b\right) \quad \text{for all} \quad n > 0,$$

further there exists a constant $p \in (0, \infty]$ *and* $C > 0$ *such that*

$$\sup_i P(|X_i| > t) \leqslant C\exp\left(-t^p\right) \quad \text{for all} \quad t > 0.$$

If the random variables are bounded, $\|X_i\|_\infty < \infty$ *for all* $i \in \mathbb{N}$ *we have* $p = \infty$. *Let* $\frac{1}{a} = \frac{1}{b} + \frac{1}{p}$, *then we have for all* $\lambda > 0$ *and* $r \geqslant 1$

$$\frac{4}{\lambda} \int_0^{H(\lambda/r)} Q(u)du \leqslant 4\frac{C}{\lambda}\exp\left(-c\left(\frac{\lambda}{r}\right)^a\right).$$

The following corollary is quoted from Ferraty and Vieu [30], Corollary A.12. The formulation will be for arithmetic mixing random variables, but it can be easily seen that the conditions for the geometric mixing case, see Corollary 3.2.2, are also fulfilled.

Corollary 3.2.3 (Ferraty and Vieu [30], p. 237 et seq.) *Assume we have a sequence* $(X_{i,n})$ *of mixing random variables, depending on* n, *with arithmetic coefficient* $b > 1$. *Let* $u_n := n^{-2}s_n \log n$ *be a deterministic sequence. Furthermore, assume that one of the following two conditions are satisfied*

i) $\exists p > 2$, $\exists \theta > 2$, $\exists M_n < \infty$, *such that* $\forall t > M_n$ *we have* $P(|X_1| > t) \leqslant t^{-p}$ *and* $s_n^{-(b+1)p/(b+p)} = O(n^{-\theta})$.

ii) $\exists M_n < \infty$, $\exists \theta > 2$ *such that* $|X_{1,n}| \leqslant M_n$ *and* $s_n^{-(b+1)} = O(n^{-\theta})$.

Then we have

$$\frac{1}{n}\sum_{i=1}^{n} X_{i,n} = \mathcal{O}_{a.co.}\left(\sqrt{u_n}\right).$$

3.2.2 *Topological Aspects*

In this section, we introduce some topological terms, such as *pre-compact*, *covering number* and *entropy*. Afterwards, we present some examples of covering numbers for some commonly used spaces.

3.2.2.1 *Basic Notations*

Let (E, d) be a semi-metric space and $S_E \subset E$ be a closed and totally bounded set. As in the case of pointwise convergence, an assumption on the small ball probability of the functional variable X for the uniform convergence on S_E is needed.

Condition on the small ball probability

Assume that there exists uniformly for all $x \in S_E$ a function $\phi(h)$ such that

(F) $\exists C, C' > 0$ and $\forall x \in S_E$:

$$0 < C\phi(h) \leqslant P(X \in B(x,h)) \leqslant C'\phi(h) < \infty.$$

This is a strict assumption in view of pointwise convergence as we need such a concentration function for all $x \in S_E$. By this function ϕ, the concentration of the random variable in a ball with radius h can be uniformly controlled. Recall that in Section 1.6 we gave an example of an exponential-type process (the Ornstein-Uhlenbeck process). By choosing

$$C' = \sup_{x \in S_E} C_x \text{ and } C = \inf_{x \in S_E} C_x,$$

we have an example of the existence of such a measure on a compact set. Ferraty et al. [23] present the Onsager-Machlup function,

$$\forall x, y \in S_E : \ \mathcal{F}_X(x,y) := \log\left(\lim_{h \to 0} \frac{P(B(x,h))}{P(B(y,h))}\right)$$

for verifying that condition. If we have for the Onsager-Machlup function

$$\forall x \in S_E : \ |\mathcal{F}_X(x,0)| \leqslant C < \infty,$$

then Condition (F) is verified. For more references we refer to Ferraty et al. [23].

As Lian [47, p. 34] describes, Condition (F) automatically implies the total boundedness of the set S_E. Therefore, the total boundedness of S_E is not explicitly listed.

For some more references on that topic, we refer to Section 1.6. For completeness we introduce the definition of *pre-compact* and some related notions. Afterwards, we present some examples.

Definition 3.2.1 *A set S of a space E is called pre-compact or totally bounded, if we have for all $\varepsilon > 0$ a finite number of elements $x_1, \dots, x_k \in E$, such that*

$$S = \bigcup_{i=1}^{k} B(x_i, \varepsilon),$$

where $B(x_i, \varepsilon) := \{x \in S | d(x, x_i) < \varepsilon\}$ is an open ball in E. The covering number $N(S, d, \varepsilon)$ is then the smallest $n \in \mathbb{N}$ such that S_E is covered by n-balls.

Similar to , there exists the entropy number.

Definition 3.2.2 *Let $n \in \mathbb{N}$ be fixed, S be a pre-compact set in E, then the entropy number is defined as*

$$\begin{aligned}\varepsilon_n(S) &:= \inf\{\varepsilon > 0 | \exists \varepsilon\text{-net for } S \text{ in } E \text{ with } q \leqslant n \text{ elements}\} \\ &= \inf\{\varepsilon > 0 | N(S, d, \varepsilon) \leqslant n\}.\end{aligned}$$

For a deeper insight on entropy numbers, we refer to the monograph of Carl und Stephani [12]. In our proofs, we will use the following notion:

Definition 3.2.3 *Let $\varepsilon > 0$ be fixed, S be a pre-compact space and $N(S, d, \varepsilon)$ the smallest covering number of open balls that covers the space S. Then*

$$K_S(\varepsilon) := \log\left(N(S, d, \varepsilon)\right)$$

is known as the Kolmogorov ε–entropy.

The concept of ε–entropy is first introduced by Kolmogorov and Tihomirov [43]; the paper cited here is an English translation of the original Russian paper.

3.2.2.2 *Some Examples for the Kolmogorov ε–Entropy*

The intention of this section is to present some spaces with their corresponding Kolmogorov ε–entropy. We extract these examples out of the paper of Ferraty et al. [23], the monograph of Steinwart and Christmann [65], and the monograph of Carl and Stephani [12].

Closed Set in a Finite-dimensional Banach Space (Carl and Stephani [12, p. 9])

Let E be a m-dimensional Banach space, $m \in \mathbb{N}$. Let U_E be the closed unit ball in E; then we have for the entropy

$$n^{-\frac{1}{m}} \leqslant \varepsilon_n(U_E) \leqslant 4n^{-\frac{1}{m}},$$

and we get then for the Kolmogorov ε–entropy

$$m \log\left(\frac{1}{\varepsilon}\right) \leqslant K_{U_E}(\varepsilon) \leqslant m\left(\log(4) + \log\left(\frac{1}{\varepsilon}\right)\right).$$

Compact Set in a Hilbert Space with a Projection-based Semi-metric (Ferraty et al. [27])

Let $S \subset H$ be a compact set in a Hilbert space H. Ferraty et al. [27] prove in their Proposition 3.1.1 that in the case of a projection-based semi-metric,

$$d(x,y) = \sqrt{\sum_{i=1}^{k} < x-y, u_i >^2},$$

where $x, y \in S$, (u_i) is an orthonormal basis of H and $k \in \mathbb{N}$, the ε–entropy behaves like the entropy of a finite-dimensional Banach space, namely

$$k \log\left(\frac{1}{\varepsilon}\right) \leqslant K_S(\varepsilon) \leqslant k\left(\log(4) + \log\left(\frac{1}{\varepsilon}\right)\right).$$

Closed Ball in a Sobolev Space (Ferraty et al. [23])

Let $W_2^1(r)$ be the space of functions $f(t)$ that are defined on the interval $[0, 2\pi)$ with periodic boundary conditions, and let the following inequality be valid

$$\frac{1}{2\pi}\int_0^{2\pi} f^2(t)dt + \frac{1}{2\pi}\int_0^{2\pi} f^{(1)^2}(t)dt \leqslant r,$$

then

$$K_{W_2^1(r)}(\varepsilon) \leqslant C\left(\frac{r}{\varepsilon}\right)^{\frac{1}{m}}.$$

Open Ball in a Sobolev Space (Steinwart und Christmann [65, p. 518])

Let X be an open ball in $\mathbb{R}^d$. Then we have for all $m > \frac{d}{2}$ two positive constants $c_m(X)$ and $\tilde{c}_m(X)$, such that

$$c_m(X)n^{-\frac{m}{d}} \leqslant e_n(id : W^m(X) \to L_\infty(X)) \leqslant \tilde{c}_m(X)n^{-\frac{m}{d}}. \tag{3.2}$$

For the L_2-norm and for all $m \geqslant 0$, there exist also two constants, different from the ones above, such that

$$c_m(X)n^{-\frac{m}{d}} \leqslant e_n(id : W^m(X) \to L_2(X)) \leqslant \tilde{c}_m(X)n^{-\frac{m}{d}}, \tag{3.3}$$

where e_n is known as the dyadic entropy number. By Lemma 6.2.1 [65, p. 221] we get for these two cases, (3.2) and (3.3), following Kolmogorov ε–entropy

$$K_S(\varepsilon) \leqslant \log(4)\left(\frac{\tilde{c}_m(X)}{\varepsilon}\right)^{\frac{d}{m}}.$$

Unitball of the Cameron-Martin Space (Ferraty et al. [23])

This example originates from the publication by van der Vaart and van Zanten [68]. Let μ be a spectral measure on $\mathbb{R}$ with the following condition,

$$\int \exp(\delta|\lambda|)\mu(d\lambda) < \infty$$

for some $\delta > 0$. Let $W := (W_t : t \geqslant 0)$ be a centred, mean-square continuous Gaussian process. By Lemma 2.1 of van der Vaart und van Zanten [68] the Reproducing Kernel Hilbert Space H is expressed as

$$H = \{t \mapsto \mathrm{Re}\left((\mathcal{F}h)(t)\right) | h \in L_2(\mu)\} \text{ with } t \in [0,1].$$

$(\mathcal{F}h)(t)$ is the Fourier transformation $\mathcal{F}h : \mathbb{R} \to \mathbb{C}$ of the functions h relative to the spectral measure μ,

$$(\mathcal{F}h)(t) = \int \exp(itx)h(x)\mu(dx).$$

As a result of Lemma 2.3 of [68] we have for the Kolmogorov ε–entropy relative to the supremum norm of the unit ball U_H in H

$$K_S(\varepsilon) \leqslant C\left(\log\left(\frac{1}{\varepsilon}\right)\right)^2.$$

Link between the Kolmogorov ε–Entropy and the Small Ball Probability

A precise link between the small ball probability of a Gaussian measure on a separable Banach space and the Kolmogorov ε–entropy of the unit ball of the RKHS H generated by the Gaussian measure was discovered by Kuelbs and Li [44]. With this result it is possible to calculate the small ball probability from the Kolmogorov ε–entropy and vice versa. As already shown in the dependent case, the uniform convergence for functional data depends on the behaviour of the small ball probability and the Kolmogorov ε–entropy. The link between these two properties may be interesting, therefore an example is presented. We outline a result of the paper by Li and Linde [46], which is an extension of the paper of Kuelbs and Li [44].

Let P be a centred Gaussian measure on a real separable Banach space $(E, \|\cdot\|)$ and H_P the Hilbert space generated by P, see Li and Linde [46]. Then, as a consequence of the Theorem 1.1 and Theorem 1.2 from Li and Linde's paper [46] we have the equivalence of

$$-\log P(\|X\| \leqslant \varepsilon) \sim \varepsilon^{-\alpha}\left(\log\left(\frac{1}{\varepsilon}\right)\right)^{\beta}$$

and

$$K_{U_X}(\varepsilon) \sim \varepsilon^{-\frac{2\alpha}{2+\alpha}}\left(\log\left(\frac{1}{\varepsilon}\right)\right)^{\frac{2\beta}{2+\alpha}},$$

where $\beta \in \mathbb{R}$, $\alpha > 0$ and U_X is the unit ball in H_P.

For some more examples and a wider reference we refer to Section 1.6 and the papers cited above. In Subsection 3.3.3, after the main results and the proofs, we will make some more comments on that circumstance in view of our considerations.

3.3 THE REGRESSION FUNCTION

3.3.1 *Notations and Assumptions*

In this section, we focus on the generalised regression function

$$m_\varphi(\cdot) = E\left[\varphi(Y)|X = \cdot\right],$$

where $\varphi : \mathbb{R} \to \mathbb{R}$ is a Borel-measurable function. We deviate from the notion of the chapters before, because the results we get in this section can be directly transferred to the conditional distribution function by the choice $\varphi(Y) := 1_{[-\infty,y]}(Y)$.

The kernel estimate of the generalised regression function is given for $x \in E$ as

$$\hat{m}_\varphi(x) = \sum_{i=1}^{n} \varphi(Y_i) \frac{K\left(h_n^{-1} d(x, X_i)\right)}{\sum_{j=1}^{n} K\left(h_n^{-1} d(x, X_j)\right)}, \quad \text{if} \quad \sum_{j=1}^{n} K\left(h_n^{-1} d(x, X_i)\right) \neq 0, \tag{3.4}$$

otherwise $\hat{m}_\varphi(x) = 0$. The terms of covariance, which are a measure of dependence, are denoted by

$$s_{n,1}(x) := \sum_{i,j=1}^{n} |Cov(\Delta_i(x), \Delta_j(x))| \text{ and}$$
$$s_{n,2}(x) := \sum_{i,j=1}^{n} |Cov(Y_i\Delta_i(x), Y_j\Delta_j(x))|,$$

where

$$\Delta_i(x) := \frac{K(h^{-1} d(x, X_i))}{E\left[K(h^{-1} d(x, X_1))\right]}.$$

Furthermore, we define

$$s^*_{n,1} := \sup_{x \in S_E} s_{n,1}(x) \text{ and}$$
$$s^*_{n,2} := \sup_{x \in S_E} s_{n,2}(x).$$

We will prove the almost complete uniform convergence of the kernel estimate defined in (3.4) on a compact subset S_E of a function semi-metric space E.

In the next section, we present the assumptions.

Condition on the regularity of the generalised regression function

(R1) Assume that the generalised regression function is of Hölder-type,

$$m_\varphi \in L^\beta(E),$$

for some $\beta > 0$.

Condition on the response variable Y

(M1) Assume that the conditioned moments of $\varphi(Y)$ are uniformly bounded,

$$\forall m > 2 : \; E\left[|\varphi(Y)|^m | X = x\right] = \delta_m(x) < C < \infty.$$

Condition on the kernel function K

(K) Assume that the kernel function K is of continuous- or of discontinuous-type. Furthermore, assume for continuous-type kernel functions that there exists $C > 0$ and $\varepsilon_0 > 0$

$$\forall\, 0 < \varepsilon < \varepsilon_0 : \; \int_0^{\varepsilon} F(u)\,du > C\varepsilon F(\varepsilon).$$

Condition on the Kolmogorov ε–entropy

Initially, we consider arithmetically mixing random variables. As can be seen in the proof of Theorem 3.3.1 and the related lemmas, the covering number of the compact set S_E is of a polynomial order.

(E1) Let (E, d) be a semi-metric space, let $\varepsilon > 0$ and $C > 0$, then the condition needs to hold

$$K_{S_E}(\varepsilon) \leqslant \tilde{C}\tau \log\left(\frac{1}{\varepsilon}\right),$$

where τ on E.

If we take a closer look at the examples of Section 3.2.2, it can be seen that Condition (E1) is restrictive. Under this assumption, we do not get a uniform convergence result for some interesting function spaces. In the set of our examples, we are restricted to finite-dimensional Banach spaces. As we can see in the second example, this problem can be avoided on compact subsets of infinite-dimensional Hilbert spaces by using a projection-based semi-metric. There exist also some non finite dimensional examples, see e. g. Ferraty et al. [23] or Ferraty et al. [27]. Therefore, this assumption (E1) can be weakened so that the uniform convergence results are valid on a larger class of function spaces.

Conditions on the mixing coefficient $\alpha(n)$

(A1) Assume the data $(X_i, Y_i)_{i=1}^n$ is α-mixing with mixing coefficient,

$$\exists c > 0,\; a > 0 : \; \alpha(n) \leqslant c n^{-b},$$

with $b > 1$.

Conditions on the covariance term s_n

(D1) Let $s_n := \max\{s_{n,1}^*, s_{n,2}^*\}$, then assume that

$$s_n^{\frac{-p(b+1)}{2(b+p)}} = o(n^{-\theta})$$

for a $\theta > \tau + 2$, where τ is defined as in (E1).

Furthermore, C is in all proofs a generic positive and finite constant.

3.3.2 *Main Results*

The Arithmetically Mixing Case

We rewrite the generalised regression estimate as follows

$$\hat{m}_\varphi(x) = \frac{\hat{g}_\varphi(x)}{\hat{f}(x)},$$

where

$$\hat{g}_\varphi(x) := \frac{1}{n}\sum_{i=1}^{n} \varphi(Y_i)\Delta_i(x),\ \hat{f}(x) := \frac{1}{n}\sum_{i=1}^{n} \Delta_i(x). \tag{3.5}$$

Theorem 3.3.1 (Arithmetically Mixing) *With Conditions (F), (K), (R1), (M1), (E1), (A1), and (D1), we have*

$$\sup_{x\in S_E} |\hat{m}_\varphi(x) - m_\varphi(x)| = \mathcal{O}\left(h^\beta\right) + \mathcal{O}_{a.co.}\left(\frac{\sqrt{s_n \log n}}{n}\right).$$

Proof:
As the denominator of the kernel estimate depends on the random variables, we need to decompose the difference between the estimator and the regression function. This decomposition is as in the proof of the pointwise convergence

$$\begin{aligned}&\hat{m}_\varphi(x) - m_\varphi(x)\\ &= \frac{1}{\hat{f}(x)}\left[\hat{g}_\varphi(x) - E\left[\hat{g}_\varphi(x)\right] - \left(m_\varphi(x) - E\left[\hat{g}_\varphi(x)\right]\right)\right] - \frac{m_\varphi(x)}{\hat{f}(x)}\left[\hat{f}(x) - 1\right].\end{aligned} \tag{3.6}$$

With (3.6), we get with some calculations

$$\begin{aligned}\sup_{x\in S_E} |\hat{m}_\varphi(x) - m_\varphi(x)| \leqslant &\frac{1}{\inf\limits_{x\in S_E} |\hat{f}(x)|} \underbrace{\sup_{x\in S_E} |\hat{g}_\varphi(x) - E\left[\hat{g}_\varphi(x)\right]|}_{\text{I}} +\\ &\frac{1}{\inf\limits_{x\in S_E} |\hat{f}(x)|} \underbrace{\sup_{x\in S_E} |m_\varphi(x) - E\left[\hat{g}_\varphi(x)\right]|}_{\text{II}} +\\ &\frac{\sup\limits_{x\in S_E} |m_\varphi(x)|}{\inf\limits_{x\in S_E} |\hat{f}(x)|} \underbrace{\sup_{x\in S_E} |\hat{f}(x) - 1|}_{\text{III}}.\end{aligned} \tag{3.7}$$

For the bias term II only deterministic properties are needed, therefore there is no difference from the proof of the pointwise independent case. Term I is examined in Lemma 3.3.3 and term III in Lemma 3.3.1. For the infimum of $\hat{f}(x)$, see Lemma 3.3.2. □

First of all, we will take a closer look at term III of (3.7). We have $E\left[\hat{f}(x)\right] = 1$.

Lemma 3.3.1 *With Conditions of Theorem 3.3.1, we have*

$$\sup_{x\in S_E}\left|\hat{f}(x)-E\left[\hat{f}(x)\right]\right| = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,1}\log n}}{n}\right).$$

Proof:
The proof of this lemma is based on the proof of uniform convergence for independent random variables, see Ferraty et al. [23]. Analogously, we estimate the supremum by three terms A_1, A_2, and A_3. The difference from the independent case is that we use another exponential inequality here for α-mixing random variables. The Fuk-Nagaev exponential inequality is applied here, see in Section 3.2.1 Theorem 3.2.1, Corollary 3.2.1, and Corollary 3.2.2.

Choose for the ε–net $\varepsilon := \frac{\log n}{n}$ and denote as $x_{k(x)}$ the closest point of the ε–net x_k to $x \in S_E$. Let us have a closer look at the following decomposition

$$\sup_{x\in S_E}\left|\hat{f}(x)-E\left[\hat{f}(x)\right]\right| \leqslant A_1 + A_2 + A_3,$$

where

$$\begin{aligned}
A_1 &= \sup_{x\in S_E}\left|\hat{f}(x)-\hat{f}(x_{k(x)})\right|,\\
A_2 &= \sup_{x\in S_E}\left|\hat{f}(x_{k(x)})-E\left[\hat{f}(x_{k(x)})\right]\right|, \text{ and}\\
A_3 &= \sup_{x\in S_E}\left|E\left[\hat{f}(x_{k(x)})\right]-E\left[\hat{f}(x)\right]\right|.
\end{aligned}$$

First, we examine, as in [23], the case of continuous-type kernel functions. Because of the condition $K(1) = 0$, we have that K is Lipschitz continuous in $[0, 1]$. Define $K_h(\cdot) := K(h^{-1}(\cdot))$, then we have

$$\begin{aligned}
A_1 &= \frac{1}{n}\sup_{x\in S_E}\sum_{i=1}^{n}\left|\frac{K_h(d(x,X_i))}{E\left[K_h(d(x,X_1))\right]}-\frac{K_h(d(x_{k(x)},X_i))}{E\left[K_h(d(x_{k(x)},X_1))\right]}\right|\\
&\leqslant \sup_{x\in S_E}\frac{C}{nF(h)}\sum_{i=1}^{n}\left|K_h(d(x,X_i))-K_h(d(x_{k(x)},X_i))\right|\\
&= \sup_{x\in S_E}\frac{C}{nF(h)}\sum_{i=1}^{n}\left|K_h(d(x,X_i))-K_h(d(x_{k(x)},X_i))\right|1_{B(x_{k(x)},h)\cup B(x,h)}(X_i)\\
&\leqslant \sup_{x\in S_E}\frac{C}{nhF(h)}\sum_{i=1}^{n}\varepsilon 1_{B(x_{k(x)},h)\cup B(x,h)}(X_i).
\end{aligned}$$

We got the last step by using the Lipschitz continuity of the kernel function K. Let

$$W_i := \frac{C\varepsilon}{hF(h)}1_{B(x_{k(x)},h)\cup B(x,h)}(X_i),$$

then the random variable W_i is bounded,

$$|W_1| \leqslant \frac{C\varepsilon}{hF(h)} =: M_n < \infty. \tag{3.8}$$

Because of (3.8) and the Conditions (A1) and (D1), we are able to apply Corollary 3.2.3,

$$A_1 = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,1}\log n}}{n}\right).$$

For continuous-type kernel functions we have the problem that this type is only Lipschitz continuous on the interval $[0,1)$. Therefore, we use the term A_1 for the same decomposition as in the paper of Ferraty et al. [23] in the independent case as follows

$$\begin{aligned} A_1 &= \sup_{x\in S_E}\left|\hat{f}(x) - \hat{f}(x_{k(x)})\right| \\ &\leqslant C(A_{11} + A_{12} + A_{13}), \end{aligned} \tag{3.9}$$

where

$$\begin{aligned} A_{11} &= \sup_{x\in S_E}\frac{1}{nF(h)}|K_h(d(x,X_i)) - K_h(d(x_{k(x)},X_i))|1_{B(x_{k(x)},h)\cap B(x,h)}(X_i), \\ A_{12} &= \sup_{x\in S_E}\frac{1}{nF(h)}\sum_{i=1}^{n} K_h(d(x,X_i))1_{\overline{B(x_{k(x)},h)}\cap B(x,h)}(X_i), \text{ and} \\ A_{13} &= \sup_{x\in S_E}\frac{1}{nF(h)}\sum_{i=1}^{n} K_h(d(x_{k(x)},X_i))1_{\overline{B(x,h)}\cap B(x_{k(x)},h)}(X_i). \end{aligned}$$

On the basis of this estimate, we are now able to use in the case of term A_{11} the Lipschitz continuity of the kernel function. Furthermore, A_{11} is bounded so that we are able to apply Corollary 3.2.3 by the same arguments as in continuous-type kernel functions,

$$A_{11} = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,1}\log n}}{n}\right). \tag{3.10}$$

For the terms A_{12} and A_{13}, we use the fact that the kernel functions are bounded. By Corollary 3.2.3, we get

$$A_{12} = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,1}\log n}}{n}\right) \quad \text{and} \quad A_{13} = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,1}\log n}}{n}\right). \tag{3.11}$$

Following (3.10) and (3.11), we have

$$A_1 = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,1}\log n}}{n}\right).$$

Next, examine term A_2,

$$\begin{aligned} A_2 &= \sup_{x\in S_E}\left|\hat{f}(x_{k(x)}) - E\left[\hat{f}(x_{k(x)})\right]\right| \\ &= \max_{k=1,\ldots,N(S_E,d,\varepsilon)}\left|\hat{f}(x_{k(x)}) - E\left[\hat{f}(x_{k(x)})\right]\right|. \end{aligned} \tag{3.12}$$

We obtain

$$\begin{aligned}\hat{f}(x_{k(x)}) - E\left[\hat{f}(x_{k(x)})\right] &= \frac{1}{n}\sum_{i=1}^{n}\left(\Delta_i(x_{k(x)}) - E\left[\Delta_i(x_{k(x)})\right]\right)\\ &= \frac{1}{n}\sum_{i=1}^{n} W_{ki}.\end{aligned}$$

With the help of Conditions (K) and (F) we receive

$$|W_{k1}| \leqslant CF(h)^{-1}. \tag{3.13}$$

Now we apply the Fuk-Nagaev inequality. Let $\eta > 0$ and $r \geqslant 1$,

$$\begin{aligned}&P\left(\max_{k=1,\dots,N(S_E,d,\varepsilon)} \frac{1}{n}\left|\sum_{i=1}^{n} W_{ki}\right| \geqslant \eta\right)\\ &\leqslant \sum_{k=1}^{N(S_E,d,\varepsilon)} P\left(\left|\sum_{i=1}^{n} W_{ki}\right| \geqslant n\eta\right)\\ &\leqslant N(S_E,d,\varepsilon)\max_{k=1,\dots,N(S_E,d,\varepsilon)} P\left(\left|\sum_{i=1}^{n} W_{ki}\right| \geqslant n\eta\right)\\ &\leqslant N(S_E,d,\varepsilon)\left[\frac{nC}{r}\left(\frac{r}{n\eta}\right)^{a+1} + 4\left(1+\frac{\eta^2 n^2}{r s^*_{n,1}}\right)^{-\frac{r}{2}}\right]\\ &=: T_{1n} + T_{2n}.\end{aligned} \tag{3.14}$$

Choose $r := (\log n)^{1+\gamma}$ for some $\gamma > 0$ and $\eta := \eta_0 \frac{\sqrt{s^*_{n,1}\log n}}{n}$ for some $\eta_0 > 0$. We find for term T_{1n},

$$\begin{aligned}T_{1n} &= CN(S_E,d,\varepsilon)\left[\frac{n}{(\log n)^{1+\gamma}}\left(\frac{n(\log n)^{1+\gamma}}{n\eta_0\sqrt{s^*_{n,1}\log n}}\right)^{b+1}\right]\\ &= C\eta_0^{-(b+1)} N(S_E,d,\varepsilon)(\log n)^{\frac{2b(\gamma+1)-1}{2}} n s^{*-\frac{(b+1)}{2}}_{n,1}\\ &\leqslant C\eta_0^{-(b+1)}(\log n)^{\frac{2b(\gamma+1)-1}{2}-\tau} n^{1+\tau-\theta}.\end{aligned}$$

The last line results from the Condition $s^{*-\frac{(b+1)}{2(b+p)}}_{n,1} = o(n^{-\theta})$ and Condition (E1). For $\theta > \tau + 2$ and $b > 1$ we obtain

$$\sum_{n=1}^{\infty} T_{1n} \leqslant C\eta_0^{-(b+1)} \sum_{n=1}^{\infty} (\log n)^{\frac{2b(\gamma+1)-1}{2}-\tau} n^{1+\tau-\theta} < \infty.$$

Finally, we consider term T_{2n},

$$\begin{aligned}T_{2n} &= 4N(S_E,d,\varepsilon)\left(1+\frac{\eta^2 n^2}{r s^*_{n,1}}\right)^{-\frac{r}{2}}\\ &= 4N(S_E,d,\varepsilon)\left(1+\frac{\eta^2 n^2}{(\log n)^{1+\gamma} s^*_{n,1}}\right)^{-\frac{(\log n)^{1+\gamma}}{2}}.\end{aligned}$$

Because of the choice of r, we are able to use the Taylor expansion of $\log(1+x) = x - \frac{x^2}{2} + o(x^2)$ for $x \to 0$ and with the above choice of η we arrive at

$$\begin{aligned} T_{2n} &= 4N(S_E, d, \varepsilon) \exp\left(-\frac{(\log n)^{1+\gamma}}{2} \log\left(1 + \frac{\eta^2 n^2}{(\log n)^{1+\gamma} s^*_{n,1}}\right)\right) \\ &= 4N(S_E, d, \varepsilon) \exp\left(-\frac{(\log n)^{1+\gamma}}{2} \log\left(1 + \eta_0^2 (\log n)^{-\gamma}\right)\right) \\ &\leqslant 4N(S_E, d, \varepsilon) \exp\left(-\frac{1}{2} C \eta_0^2 \log n\right). \end{aligned}$$

Choose η_0 such that $l := \frac{1}{2} C \eta_0^2 > \tau + 1$. We receive

$$\begin{aligned} T_{2n} &= 4N(S_E, d, \varepsilon) n^{-l} \\ &\leqslant C n^{\tau - l}. \end{aligned}$$

Out of it, we obtain

$$\sum_{n=1}^{\infty} T_{2n} \leqslant \sum_{n=1}^{\infty} C n^{\tau - l} < \infty.$$

Hence it follows

$$A_2 = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,1} \log n}}{n}\right).$$

For the third term in (3.9)

$$\begin{aligned} A_3 &= \sup_{x \in S_E} \left| E\left[\hat{f}(x_{k(x)})\right] - E\left[\hat{f}(x)\right] \right| \\ &\leqslant E\left[\sup_{x \in S_E} \left| \hat{f}(x_{k(x)}) - \hat{f}(x) \right|\right] \end{aligned}$$

we obtain, by similar arguments as for A_1,

$$A_3 = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,1} \log n}}{n}\right).$$

□

In the decomposition in (3.6), we have to handle the reciprocal of the infimum of $\hat{f}(x)$.

Lemma 3.3.2 (Ferraty et al. [23]) *With Conditions (F), (K), (E1), (A1), and (D1), we have*

$$\sum_{n=1}^{\infty} P\left(\inf_{x \in S_E} \hat{f}(x) < \frac{1}{2}\right) < \infty.$$

Proof:
This proof differs from the case of independent random variables only in the way that Lemma 3.3.1 is applied here instead of Lemma 8 as in the paper by Ferraty et al. [23].

□

Lemma 3.3.3 *With Conditions (F), (K), (E1), (M1), (A1), and (D1), we have*

$$\sup_{x\in S_E} |\hat{g}_\varphi(x) - \mathrm{E}\left[\hat{g}_\varphi(x)\right]| = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,2}\log n}}{n}\right).$$

Proof:
The proof of this lemma is similar to the proof of the previous Lemma 3.3.1. In addition, we have the response variable Y here. Therefore, we have to apply the Fuk-Nagaev exponential inequality for unbounded random variables. First decompose the supremum into three terms as follows and examine them separately,

$$\sup_{x\in S_E} |\hat{g}_\varphi(x) - \mathrm{E}\left[\hat{g}_\varphi(x)\right]| \leqslant A_1 + A_2 + A_3,$$

with

$$\begin{aligned}
A_1 &= \sup_{x\in S_E} \left|\hat{g}_\varphi(x) - \hat{g}_\varphi(x_{k(x)})\right|, \\
A_2 &= \sup_{x\in S_E} \left|\hat{g}_\varphi(x_{k(x)}) - \mathrm{E}\left[\hat{g}_\varphi(x_{k(x)})\right]\right|, \text{ and} \\
A_3 &= \sup_{x\in S_E} \left|\mathrm{E}\left[\hat{g}_\varphi(x_{k(x)})\right] - \mathrm{E}\left[\hat{g}_\varphi(x)\right]\right|.
\end{aligned}$$

The proof of the rate of convergence of the terms A_1 and A_3 is analogous to the terms A_1 and A_3 in the proof of Lemma 3.3.1. Here we use the same case-by-case study of two types of kernel functions. By Condition (M1) we can then apply Corollary 3.2.3 for unbounded random variables and we obtain

$$A_1 = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,2}\log n}}{n}\right) \text{ and } A_3 = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,2}\log n}}{n}\right).$$

Term A_2 needs some more careful consideration. As for A_2 in the proof of Lemma 3.3.1, we apply the Fuk-Nagaev inequality here to receive

$$\begin{aligned}
A_2 &= \sup_{x\in S_E} \left|\hat{g}_\varphi(x_{k(x)}) - \mathrm{E}\left[\hat{g}_\varphi(x_{k(x)})\right]\right| \\
&= \max_{k=1,\ldots,N(S_E,d,\varepsilon)} \left|\hat{g}_\varphi(x_{k(x)}) - \mathrm{E}\left[\hat{g}_\varphi(x_{k(x)})\right]\right|.
\end{aligned}$$

We arrive at

$$\begin{aligned}
\hat{g}_\varphi(x_{k(x)}) - \mathrm{E}\left[\hat{g}_\varphi(x_{k(x)})\right] &= \frac{1}{n}\sum_{i=1}^{n}\left(\Delta_i(x_{k(x)})\varphi(Y_i) - \mathrm{E}\left[\Delta_i(x_{k(x)})\varphi(Y_i)\right]\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} W_{ki}.
\end{aligned}$$

Now we have to check the conditions of Corollary 3.2.1 for the random variable W_{ki}. The calculation of this verification is nearly identical to the independent case and can be found in the proof of Lemma 6.3 in [30, p. 65]. The difference is that the measurable function φ is applied to the response variable.

As the conditions of Corollary 3.2.1 are now verified, we are now able to apply the Fuk-Nagaev inequality as we did in Lemma 3.3.1,

$$\begin{aligned}
P(A_2 \geqslant \eta) &= P\left(\max_{k=1,\dots,N(S_E,d,\varepsilon)} \frac{1}{n} \left| \sum_{i=1}^{n} W_{ki} \right| \geqslant \eta \right) \\
&\leqslant \sum_{k=1}^{N(S_E,d,\varepsilon)} P\left(\left| \sum_{i=1}^{n} W_{ki} \right| \geqslant n\eta \right) \\
&\leqslant N(S_E,d,\varepsilon) \max_{k=1,\dots,N(S_E,d,\varepsilon)} P\left(\left| \sum_{i=1}^{n} W_{ki} \right| \geqslant n\eta \right) \\
&\leqslant N(S_E,d,\varepsilon) \left[\frac{nC}{r} \left(\frac{r}{n\eta} \right)^{\frac{(b+1)p}{b+p}} + 4 \left(1 + \frac{\eta^2 n^2}{r s^*_{n,2}} \right)^{-\frac{r}{2}} \right] \\
&=: T_{1n} + T_{2n}.
\end{aligned}$$

We treat the term T_{2n} as we did in the proof of Lemma 3.3.1, but because of the unbounded response variable Y, the examination of the term T_{1n} is different from that in Lemma 3.3.1,

$$\begin{aligned}
T_{1n} &= C N(S_E,d,\varepsilon) \left[\frac{n}{(\log n)^{1+\gamma}} \left(\frac{n(\log n)^{1+\gamma}}{n\eta_0 \sqrt{s^*_{n,2} \log n}} \right)^{\frac{(b+1)p}{b+p}} \right] \\
&\leqslant C\eta_0^{-\frac{(b+1)p}{b+p}} N(S_E,d,\varepsilon) \log n^{1+\gamma+\frac{(1/2+\gamma)(b+1)p}{(b+p)}} n s^{*-\frac{(b+1)p}{2(b+p)}}_{n,2} \\
&\leqslant C\eta_0^{-\frac{(b+1)p}{b+p}} N(S_E,d,\varepsilon) \log n^{1+\gamma+\frac{(1/2+)\gamma(b+1)p}{(b+p)}} n^{1-\theta} \\
&\leqslant C\eta_0^{-\frac{(b+1)p}{b+p}} \log n^{1+b+\frac{(1/2+\gamma)(b+1)p}{(b+p)}} n^{1+\tau-\theta}.
\end{aligned}$$

For $\theta > \tau + 2$, we arrive at

$$A_2 = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,2} \log n}}{n} \right).$$

This finishes the proof. □

The Geometrically Mixing Case

To get uniform convergence results for a larger class of function spaces a price has to be paid, namely the sequence of random variables has to be geometrically mixing, we have to be more restrictive regarding the response variable Y, and we need some restriction on the covariance term s_n.

Condition on the mixing coefficient

(A2) Assume the data $(X_i, Y_i)_{i=1}^n$ is α-mixing with exponentially decaying mixing coefficient, i.e.

$$\exists c,\ b > 0 :\ \alpha(n) \leqslant \exp\left(-cn^b\right).$$

Condition on the response variable Y

Furthermore, here we need the tails of the probability distribution function of the response variable Y to decline exponentially.

(M2) Assume for response variable Y

$$P\left(|Y| > t\right) \leqslant C \exp\left(-t^p\right),$$

for some $C > 0$ and $p \in (0, \infty]$, where $p = \infty$ means that the response variable Y is bounded.

Condition on the Kolmogorov ε–entropy

The extension of the result from Section 3.3.2 is that now the entropy number grows polynomially and no longer needs logarithmic growth as in Condition (E1). This generalisation increases the amount of function spaces, see Section 3.2.2. In addition, for this case we need a lower bound on the entropy number so that it does not decrease too quickly.

(E2) Assume that we have for the Kolmogorov ε–entropy

$$K_S\left(\varepsilon\right) \sim \varepsilon^{-\alpha},$$

with $\alpha \in (0, 1)$.

Condition on the covariance term s_n

(D2) Assume for the covariance term s_n that we have for some $\delta > 0$ arbitrarily small, α same as in (E2), and $a = bp/(b+p)$ large enough, with b from (A2) and p from (M2), that

$$C_1 n^{\alpha\left(1+\frac{2}{a}+\delta\right)} < s_n < C_2 n^{2-\alpha}. \tag{3.15}$$

The lower bound in 3.15 is needed for the estimate where the exponential inequality is applied, the upper bound such that the convergence rate of Theorem 3.3.2 does not degenerate. As we saw in Chapter 2, the covariance term s_n depends on the small ball probability $F(h)$ and the joint distribution function $G(h)$, so the bandwidth has to be chosen correctly in this sense. We refer for this to the following Section 3.3.3.

Theorem 3.3.2 (Geometrically Mixing) *With Conditions (F), (K), (R), (M2), (E2), (A2), and (D2), we have*

$$\sup_{x \in S_E} |\hat{m}_\varphi(x) - m_\varphi(x)| = \mathcal{O}\left(h^\beta\right) + \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,1} K_{S_E}\left(\frac{1}{n}\right)}}{n}\right).$$

Proof:
This proof is based on the same decomposition as in Theorem 3.3.1. The assertion follows then from Lemma 3.3.2 and the successive Lemma 3.3.4 and Lemma 3.3.5. □

Lemma 3.3.4 *With the conditions from Theorem 3.3.2, we have*

$$\sup_{x\in S_E} |\hat{g}_\varphi(x) - E[\hat{g}_\varphi(x)]| = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,2} K_{S_E}\left(\frac{1}{n}\right)}}{n}\right).$$

Proof:
Decompose $\sup_{x\in S_E} |\hat{g}_\varphi(x) - E[\hat{g}_\varphi(x)]|$ as in Lemma 3.3.3 into three terms,

$$\begin{aligned}
A_1 &= \sup_{x\in S_E} \left|\hat{g}_\varphi(x) - \hat{g}_\varphi(x_{k(x)})\right|, \\
A_2 &= \sup_{x\in S_E} \left|\hat{g}_\varphi(x_{k(x)}) - E\left[\hat{g}_\varphi(x_{k(x)})\right]\right|, \text{ and} \\
A_3 &= \sup_{x\in S_E} \left|E\left[\hat{g}_\varphi(x_{k(x)})\right] - E[\hat{g}_\varphi(x)]\right|.
\end{aligned}$$

First, let us have a closer look at A_2 and denote $\varepsilon := 1/n$. As in the proof of Lemma 3.3.3 we get

$$\begin{aligned}
A_2 &= \sup_{x\in S_E} \left|\hat{g}_\varphi(x_{k(x)}) - E\left[\hat{g}_\varphi(x_{k(x)})\right]\right| \\
&= \max_{k=1,\dots,N(S_E,d,\varepsilon)} \left|\hat{g}_\varphi(x_{k(x)}) - E\left[\hat{g}_\varphi(x_{k(x)})\right]\right|.
\end{aligned}$$

Let us examine

$$\begin{aligned}
\hat{g}_\varphi(x_{k(x)}) - E\left[\hat{g}_\varphi(x_{k(x)})\right] &= \frac{1}{n}\sum_{i=1}^{n} \left(\Delta_i(x_{k(x)})\varphi(Y_i) - E\left[\Delta_i(x_{k(x)})\varphi(Y_i)\right]\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} W_{ki}.
\end{aligned}$$

By applying Theorem 3.2.1 and Corollary 3.2.2, we find

$$\begin{aligned}
P(A_2 \geqslant \eta) &\leqslant N(S_E, d, \varepsilon) \max_{k=1,\dots,N(S_E,d,\varepsilon)} P\left(\left|\sum_{i=1}^{n} W_{ki}\right| \geqslant n\eta\right) \\
&\leqslant N(S_E, d, \varepsilon)\left[4n\frac{C}{n\eta}\exp\left(-c\left(\frac{n\eta}{r}\right)^a\right) + 4\left(1+\frac{\eta^2 n^2}{r s^*_{n,2}}\right)^{-\frac{r}{2}}\right] \\
&=: T_{1n} + T_{2n}, \qquad (3.16)
\end{aligned}$$

where $a = pb/(b+p)$.

To begin with, we examine term T_{2n}. Choose $\eta := \eta_0 \sqrt{s^*_{n,2} K_{S_E}(1/n)/n^2}$ with some $\eta_0 > 0$ and choose $r := n^{\alpha+\delta}$ for $\delta > 0$. We obtain

$$\begin{aligned} T_{2n} &= 4N(S_E, d, \varepsilon) \exp\left(-\frac{n^{\alpha+\delta}}{2} \log\left(1 + \frac{\eta^2 n^2}{s^*_{n,2} n^{n^{\alpha+\delta}}}\right)\right) \\ &= 4N(S_E, d, \varepsilon) \exp\left(-\frac{n^{\alpha+\delta}}{2} \log\left(1 + \frac{\eta_0^2 K_{S_E}(\varepsilon)}{n^{\alpha+\delta}}\right)\right). \end{aligned} \tag{3.17}$$

We have $\eta_0^2 K_{S_E}(\varepsilon)/n^{\alpha+\delta} \to 0$ as $n \to \infty$. Because of this we can use the Taylor-expansion of $\log(1+x) = x - \frac{x^2}{2} + o(x^2)$ as $x \to 0$,

$$T_{2n} \leqslant \exp\left(\left(1 - \frac{\eta_0^2}{2}\right) K_{S_E}(\varepsilon)\right).$$

As soon as we choose η_0 such that $\eta_0^2/2$ is larger than 1, we get for large n

$$T_{2n} \leqslant n^{-1-\kappa_1} \tag{3.18}$$

for a $\kappa_1 > 0$.

Next, we examine term T_{1n}, η and r are the same as above. We receive then by some calculation and Conditions (E2) and (D2)

$$\begin{aligned} T_{1n} &= 4C\eta_0^{-1} N(S_E, d, \varepsilon) \frac{1}{\sqrt{s^*_{n,2} K_{S_E}(\varepsilon)}} \exp\left(-c\eta_0^a n^{-a(\alpha+\delta)} (s^*_{n,2} K_{S_E}(\varepsilon))^{\frac{a}{2}}\right) \\ &= 4C\eta_0^{-1} \frac{1}{\sqrt{s^*_{n,2} K_{S_E}(\varepsilon)}} \exp\left(K_{S_E}(\varepsilon) - c\eta_0^a n^{-a(\alpha+\delta)} (s^*_{n,2} K_{S_E}(\varepsilon))^{\frac{a}{2}}\right) \\ &\leqslant 4C\eta_0^{-1} \frac{1}{\sqrt{s^*_{n,2} K_{S_E}(\varepsilon)}} \exp\left(C_1 n^\alpha - C_2 \eta_0^a n^{-a(\alpha+\delta)} (s^*_{n,2} n^\alpha)^{\frac{a}{2}}\right) \\ &= 4C\eta_0^{-1} \frac{1}{\sqrt{s^*_{n,2} K_{S_E}(\varepsilon)}} \exp\left(C_1 n^\alpha - C\eta_0^a n^{-a(\frac{\alpha}{2}+\delta)} (s^*_{n,2})^{\frac{a}{2}}\right) \\ &\leqslant 4C\eta_0^{-1} \frac{1}{\sqrt{Cn^{1+\alpha+\delta}}} \exp\left(n^\alpha (C_1 - C\eta_0^a n^{-a(\frac{\alpha}{2}+\delta)-\alpha} (s^*_{n,2})^{\frac{a}{2}}\right) \\ &< 4C\eta_0^{-1} \frac{1}{\sqrt{Cn^{1+\alpha+\delta}}} \exp\left(n^\alpha (C_1 - C\eta_0^a n^t)\right) \end{aligned} \tag{3.19}$$

for some $t > 0$. So there exists a constant $\kappa_2 > 0$ such that

$$T_{1n} \leqslant n^{-1-\kappa_2}. \tag{3.20}$$

Combine relations (3.18) and (3.20) to achieve

$$A_2 = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,2} K_{S_E}\left(\frac{1}{n}\right)}}{n}\right).$$

Finally, we have to examine the terms A_1 and A_3. As in the proof of Lemma 3.3.1, we need a case-by-case analysis for the two types of kernel functions. Here we

have that the terms of investigation are uniform in x, as can be seen in the proof of Lemma 3.3.1. Therefore the calculation of the terms A_1 and A_3 is easier than that of term A_2. By applying Corollary 3.2.3, we arrive at

$$A_1 = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,2} K_{S_E}\left(\frac{1}{n}\right)}}{n}\right) \text{ and } A_3 = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,2} K_{S_E}\left(\frac{1}{n}\right)}}{n}\right).$$

This finishes the proof of this lemma. □

Lemma 3.3.5 *With Conditions (F), (K), (E2), (A2), and (D2), we have*

$$\sup_{x \in S_E} |\hat{f}(x) - 1| = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,1} K_{S_E}\left(\frac{1}{n}\right)}}{n}\right).$$

Proof:
This proof is similar to the proof of Lemma 3.3.1 and Lemma 3.3.4. As we did there, the difference $|\hat{f}(x) - 1|$ is decomposed into three parts, denoted by A_1, A_2, and A_3. As in the proof of Lemma 3.3.4, A_2 is treated in the same way as the term A_2 in the proof of Lemma 3.3.4. Furthermore, we get the desired result for both terms A_1 and A_3 with the same arguments as in Lemma 3.3.4 for the terms A_1 and A_3. □

3.3.3 *Comments and Application*

The aim of this subsection is to give some comments on the convergence rate, especially on the choice of the upper bound of the Kolmogorov ε–entropy. To avoid getting a degenerate convergence rate, one needs to control the covariance term s_n, see Condition (D2). On this account and for simplicity we adopt a result of Ferraty and Vieu's [30] for the covariance term s_n. We assume that the joint distribution function declines fast enough so that the small ball probability would dominate the convergence rate.

On the Covariance Term s_n

For a start, we need a condition related to the joint distribution of (X_j, X_i).

(G) Assume that there exists as in Condition (F) a function $G(h)$ such that we have for all x in S_E

$$0 < C_1 G(h) \leqslant \sup_{i \neq j} P\left((X_j, X_i) \in B(x,h) \times B(x,h)\right) \leqslant C_2 G(h) < \infty.$$

for some $C_1, C_2 > 0$.

(B1) Assume that there exists a constant $\varepsilon_1 \in (0,1]$ such that

$$G(h) = \mathcal{O}\left(F(h)^{1+\varepsilon_1}\right)$$

For some comments on Condition (B1) we refer to the Condition (D1) in Chapter 2. Furthermore, we need the conditional expectation of the response variable to be bounded.

(B2) Assume for the bandwidth h that

$$\exists \varepsilon_2 \in (0,1)\ F(h) = \mathcal{O}\left(n^{-\varepsilon_2}\right).$$

(Z) Assume we have for a positive and finite constant C

$$\sup_{i \neq j} E\left[|Y_i Y_j| | (X_i, X_j)\right] \leqslant C.$$

Ferraty and Vieu consider the following lemma in the case of pointwise convergence, see their monograph [30]. The proof of the expansion to the case of uniform convergence is similar to their proof.

Lemma 3.3.6 (Ferraty and Vieu [30], p. 163 et seq.) *Assume Conditions (F), (K), (B1), (B2), and (Z). Then we have in the case of geometrically mixing random variables*

$$s_n = \mathcal{O}\left(\frac{n}{F(h)}\right).$$

For the case of arithmetically mixing random variables, the same result is obtained by an additional condition on the rate, we need

$$b > \frac{1+\varepsilon_1}{\varepsilon_1 \varepsilon_2}.$$

If Condition (B1) is omitted, we get, as already seen in Chapter 2, an additional term which presents the dependence of the random variables. We consider here now just the case when Condition (B1) is in force, as the arguments would be the same if the term of dependence were being discussed. By this we get for the arithmetical case.

Theorem 3.3.3 *With Conditions of Theorem 3.3.1 and additionally with the Conditions (G), (B1), (B2), and (Z), we have*

$$\sup_{x \in S_E} |\hat{m}_\varphi(x) - m_\varphi(x)| = \mathcal{O}\left(h^\beta\right) + \mathcal{O}_{a.co.}\left(\sqrt{\frac{\log n}{nF(h)}}\right).$$

We get a similar result for the geometrically mixing case.

Theorem 3.3.4 *With Conditions of Theorem 3.3.2 and additionally with Conditions (G), (B1), (B2), and (Z), we have*

$$\sup_{x \in S_E} |\hat{m}_\varphi(x) - m_\varphi(x)| = \mathcal{O}\left(h^\beta\right) + \mathcal{O}_{a.co.}\left(\sqrt{\frac{K_{S_E}\left(\frac{1}{n}\right)}{nF(h)}}\right).$$

A Comment on the Convergence Rate and the Kolmogorov ε–entropy

Firstly, if we do not want to get a degenerated convergence rate, we need for $n \to \infty$ that

$$\frac{K_{S_E}\left(\frac{1}{n}\right)}{nF(h)} \to 0.$$

As the small ball probability $F(h) \to 0$ for $n \to \infty$, we get by a suitably chosen bandwidth h,

$$nF(h) = \mathcal{O}\left(n^t\right)$$

for some $t \in (0,1)$. Secondly, if we have a look at the proof of Theorem 3.3.2, especially at the examination of the term T_{1n}, see (3.16) and (3.19), it is not possible to select a faster increasing denominator as for example $n^\tau F(h)$ for some $\tau > 1$. It is for this reason not possible by means of the techniques used in the proof of Theorem 3.3.2 to get a convergence result for functional spaces, whose Kolmogorov ε–entropy is of higher order than n^α for an $\alpha \in (0,1)$.

If we have a look at the example in Section 3.2.2, where we examined a direct link between the small ball probability and the Kolmogorov ε–entropy, we can see that a process that allows a useful result exists. It would be interesting if such links also existed for the other results in the example Section 3.2.2.

Application of the Uniform Convergence for Bandwidth Selection and an Open Problem

An application of the results obtained in Section 3.3 is for instance the bandwidth selection, for example by bootstrapping or by cross-validation. Here we give an outlook upon the bandwidth selection by cross-validation in the α-mixing case. The case of finite-dimensional density estimation for dependent random variables is examined by Hart and Vieu [40] or by Kim [42]. The case of non-parametric regression is examined by Härdle and Vieu [39]. The previously cited papers prove their results under the condition that the random variable X is $\mathbb{R}$–valued. One of the main tools in their proof is to apply Davidoff's inequality. In the examination of cross-validation a term of the form

$$g(X_{j(1)},\ldots,X_{j(p)}) = \prod_{\substack{r,s=1\\ s\neq r}}^{q} K\left(h^{-1}(X_{j(r)} - X_{j(s)})\right)^{\beta_{r,s}} \prod_{i=1}^{p} g_{j(i)}(X_{j(i)}) \tag{3.21}$$

is obtained. For more details see Proposition 1 in Hart and Vieu [40]. To apply Davidoff's inequality the authors used the Fourier transformation for separating the random variables in the kernel function $K\left(h^{-1}(X_{j(r)} - X_{j(s)})\right)$ such that the form

$$\prod_{r=1}^{p} h_r(X_{j(r)})$$

is obtained for the term in (3.21). In the multivariate case or in the functional case, this idea of using Fourier transformation, does not work, as we have as input of the kernel function

$$K\left(h^{-1}\|X_{j(r)} - X_{j(s)}\|\right) \quad \text{or} \quad K\left(h^{-1}d(X_{j(r)} - X_{j(s)})\right),$$

nonlinear expressions. Because of this we are not able to use the properties of the Euler function to separate the random variables $X_{j(r)}$ and $X_{j(s)}$ in $d(X_{j(r)} - X_{j(s)})$.

Possibly, instead of using Davidoff's Lemma, Bradley's Lemma, see Lemma 2.4.3 in Chapter 2, can be used to solve this problem in the multivariate and functional context.

3.4 THE CONDITIONAL DISTRIBUTION FUNCTION

3.4.1 *Notations and Assumptions*

The objective of this section is to focus on the non-parametric kernel estimate of the conditional distribution function for α-mixing data,

$$F^x(y) = P(Y \leqslant y | X = x), \tag{3.22}$$

where $x \in E$ and $y \in \mathbb{R}$. We will assume that this function exists. For a discussion of its existence and for a bibliography, we refer to the monograph by Ferraty and Vieu [30, p. 51].

The conditional distribution function can be written as the conditional expectation of an indicator function

$$F^x(y) = E\left[1_{[-\infty,y]}(Y) | X = x\right].$$

Now let us have a look at the generalised regression function in Section 3.3. If we choose there

$$\varphi(Y_i) := 1_{[-\infty,y]}(Y_i),$$

we directly obtain a non-smooth kernel estimate of the conditional distribution function

$$\hat{F}^x(y) = \sum_{i=1}^{n} 1_{[-\infty,y]}(Y_i) \frac{K\left(h_n^{-1} d(x, X_i)\right)}{\sum_{i=1}^{n} K\left(h_n^{-1} d(x, X_i)\right)}, \text{ if } \sum_{j=1}^{n} K\left(h_n^{-1} d(x, X_i)\right) \neq 0,$$

otherwise $\hat{F}^x(y) = 0$, for $x \in E$ and $y \in \mathbb{R}$. Thus, we receive an uniform convergence result on a compact set S_E by applying the theorems in Section 3.3.

Theorem 3.4.1 *With Conditions (F), (K), (R1), (M1), (E1), (A1), and (D1), we have*

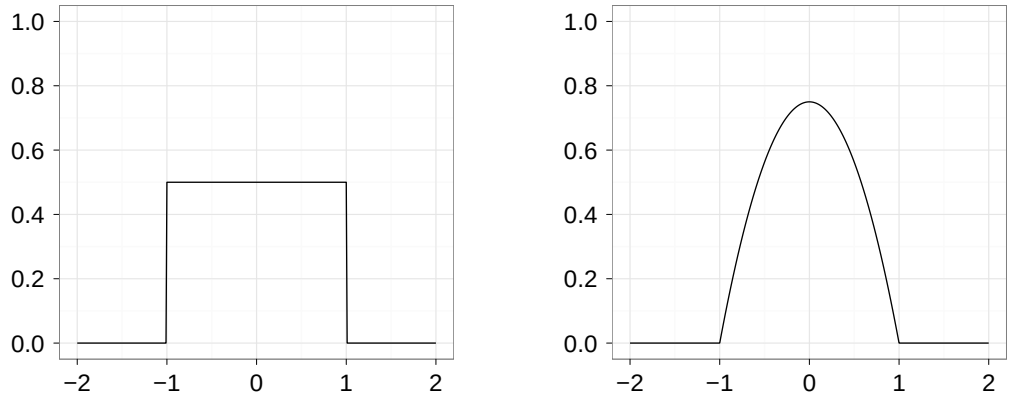
$$\sup_{x \in S_E} |\hat{F}^x(y) - F^x(y)| = \mathcal{O}\left(h^\beta\right) + \mathcal{O}_{a.co.}\left(\frac{\sqrt{s_n \log n}}{n}\right).$$

In the case of geometric mixing, we get for the same reason a convergence result on the set S_E.

Following the arguments of Ferraty and Vieu [30] and their denoted references, a smooth version of this kernel estimate can be designed. To do this a another type of kernel function can be defined.

Definition 3.4.1 *We call a kernel function K_0 of classical-type, if $\int K_0(u)du = 1$, K_0 has the support $[-1, 1]$, and $K_0(u) > 0$ for all $u \in [-1, 1]$.*

For example, the symmetric boxed or symmetric quadratic kernel function in Figure 3.1 are of classical-type.



(a) Box kernel

(b) Quadratic kernel

Figure 3.1: Two common symmetric kernel functions of classical-type.

If in addition we smooth the response variable Y, we get a smooth kernel estimate of the conditional distribution function,

$$\hat{F}^x(y) = \sum_{i=1}^{n} \Gamma_i(y) \frac{K\left(h_n^{-1} d(x, X_i)\right)}{\sum_{j=1}^{n} K\left(h_n^{-1} d(x, X_j)\right)}, \quad \text{if} \quad \sum_{j=1}^{n} K\left(h_n^{-1} d(x, X_i)\right) \neq 0,$$

otherwise $\hat{F}^x(y) = 0$, for $x \in E$ and $y \in \mathbb{R}$, with

$$\Gamma_i(y) := H(g^{-1}(y - Y_i)).$$

H is an integrated kernel

$$H(u) := \int_{-\infty}^{u} K_0(v)dv,$$

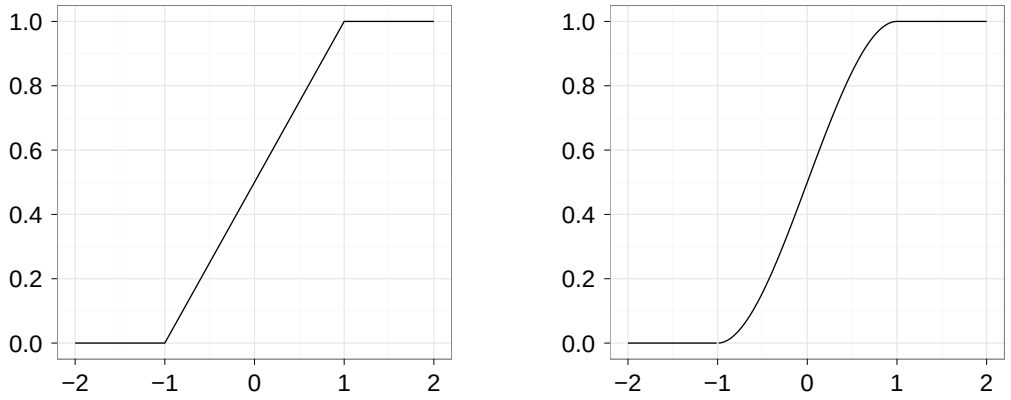
where K_0 is a kernel function of classical-type and $g := g_n$ is a smoothing parameter. The integrated kernel H acts as a local weighting in the response variable. If y is smaller than Y_i, then we have a small weight, on the other hand if y is larger than Y_i, the weight is closer to one. A price of getting a smooth estimate which a practitioner has to pay is that now two optimal smoothing parameters have to be estimated. For a more detailed reference to this type of kernel estimate we refer again to the monograph by Ferraty and Vieu [30, p. 55 et seqq.].

As in Section 3.3 we split the kernel estimate into a numerator and denominator,

$$\hat{m}_3^x(y) = \frac{1}{n} \sum_{i=1}^{n} \Gamma_i(y) \Delta_i(x), \quad \text{where} \quad \Delta_i(x) := \frac{K(h^{-1} d(x, X_i))}{E\left[K(h^{-1} d(x, X_1))\right]},$$

then we can write

$$\hat{F}^x(y) = \frac{\hat{m}_3^x(y)}{\hat{f}(x)},$$



(a) Integrated box kernel

(b) Integrated quadratic kernel

Figure 3.2: This Figure shows two common integrated kernel functions.

where $\hat{f}(x)$ is defined as in (3.5). Before we present the uniform convergence results, we introduce some conditions. Here we replace the covariance term $s_{n,2}$ of the generalised regression function defined in Section 3.3 by the following covariance term

$$s_{n,3}(x,y) := \sum_{i,j=1}^{n} |\mathrm{Cov}(\Gamma_i(y)\Delta_i(x), \Gamma_j(y)\Delta_j(x))| \quad x \in S_E,\ y \in S_{\mathbb{R}}.$$

To get a convergence result in both variables x, y, we examine the case, where $S_{\mathbb{R}} \subset \mathbb{R}$ is compact.

Condition on the regularity of the conditional distribution function

(R2) Assume that the conditional distribution function is of Hölder-type. Let $C_1, C_2 > 0$ and $\beta_1, \beta_2 > 0$, such that for all $y_1, y_2 \in S_{\mathbb{R}}$ and $x_1, x_2 \in S_E$ we have

$$|F^{x_1}(y_1) - F^{x_2}(y_2)| \leqslant C_1 d(x_1, x_2)^{\beta_1} + C_2 |y_1 - y_2|^{\beta_2}.$$

Condition on the covariance terms

(D2) Define $s^*_{n,3} := \sup_{x \in S_E} \sup_{y \in S_{\mathbb{R}}} (s_{n,3}(x,y))$ and define $s_n := \max(s^*_{n,1}, s^*_{n,3})$, then assume analogous to Condition (D1)

$$s_n^{-(a+1)} = o(n^{-\theta})$$

for some $\theta > \tau + \frac{\beta_2}{2} + 2$, where β_2 is the constant from Condition (R) and τ from Condition (E1).

As in the case of non-parametric generalised regression estimate, we divide the results here again into two parts.

3.4.2 *Main Results*

The Arithmetically Mixing Case

Theorem 3.4.2 (Arithmetically Mixing) *Assume the Conditions (F), (K), (R2), (E1), (A1), and (D2) hold and we have for the bandwidth g that $\lim_{n\to\infty} gn^{\beta_2/2} = \infty$, where β_2 is the second Hölder constant of Condition (R). Then we have*

$$\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}} |\hat{F}^x(y) - F^x(y)| = \mathcal{O}\left(h^{\beta_1}\right) + \mathcal{O}\left(g^{\beta_2}\right) + \mathcal{O}_{a.co.}\left(\frac{\sqrt{s_n \log n}}{n}\right).$$

Proof:
Ferraty and Vieu prove in their monograph [30, p. 186] the uniform convergence in y for the kernel estimate on a compact set in $\mathbb{R}$ for the α-mixing case. In the paper by Ferraty et al. [23] uniform convergence for the non-smooth kernel estimate in both variables is proven for dependent data.

Here, we combine the ideas of those two proofs and the idea we used in the proof of the generalised regression function. The following decomposition in (3.23) is based on the proofs mentioned in [30, p. 186] or [23],

$$\begin{aligned}\hat{F}^x(y) - F^x(y) =& \frac{1}{\hat{f}(x)}\left(\hat{m}_3^x(y) - E\left[\hat{m}_3^x(y)\right] - \left(F^x(y) - E\left[\hat{m}_3^x(y)\right]\right)\right] \\ &+ \frac{F^x(y)}{\hat{f}(x)}\left(E\left[\hat{f}(x)\right] - \hat{f}(x)\right). \end{aligned} \tag{3.23}$$

From (3.23) we get

$$\begin{aligned}\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}\left|\hat{F}^x(y) - F^x(y)\right| \leqslant& \frac{1}{\inf_{x\in S_E}\hat{f}(x)}\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}|\hat{m}_3^x(y) - E\left[\hat{m}_3^x(y)\right]| \\ &+ \frac{1}{\inf_{x\in S_E}\hat{f}(x)}\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}|F^x(y) - E\left[\hat{m}_3^x(y)\right)]| \\ &+ \frac{\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}} F^x(y)}{\inf_{x\in S_E}\hat{f}(x)}\sup_{x\in S_E}\left|E\left[\hat{f}(x)\right] - \hat{f}(x)\right|.\end{aligned}$$

This theorem is then a result of Lemma 3.3.1 and Lemma 3.3.2 of Section 3.3 and the successive Lemma 3.4.1 and Lemma 3.4.2. □

Lemma 3.4.1 *Under the conditions of Theorem 3.4.2, we have*

$$\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}|E\left[\hat{m}_3(x)\right] - F^x(y)| = \mathcal{O}\left(h^{\beta_1}\right) + \mathcal{O}\left(g^{\beta_2}\right).$$

Proof:
As in the independent case, just the Hölder continuity of the conditional distribution function is used. The result can be taken from Ferraty and Vieu [30, p. 84].

□

Lemma 3.4.2 *With the Conditions of Theorem 3.4.2, we have*

$$\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}|\hat{m}_3^x(y)-\mathrm{E}\left[\hat{m}_3^x(y)\right]|=\mathcal{O}_{a.co.}\left(\frac{\sqrt{s_n\log n}}{n}\right).$$

Proof:
This proof is analogous to the proof in the independent case, see [23], and the pointwise case of α-mixing random variables, see Proposition 11.20 in the monograph by Ferraty and Vieu [30, p. 185]. Because $S_{\mathbb{R}}$ is a compact set in $\mathbb{R}$, there exists a finite number z_n of open balls with radius r_n such that the union of these balls contains the set $S_{\mathbb{R}}$, namely

$$S_{\mathbb{R}}\subset\bigcup_{k=1}^{z_n}(t_k-r_n,t_k+r_n).$$

Choose the radius of the balls as $r_n=n^{-\beta_2/2}$, then the covering number grows polynomially, we have $z_n\leqslant n^{\beta_2/2}$. Denote the closest point t_k to $y\in S_{\mathbb{R}}$ as $t(y)$,

$$t(y):=\underset{\{t_1,\ldots,t_{z_n}\}}{\arg\min}|t_k-y|.$$

Now we decompose the difference as follows

$$\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}|\hat{m}_3^x(y)-\mathrm{E}\left[\hat{m}_3^x(y)\right]|\leqslant A_1+A_2+A_3+A_4+A_5,$$

where

$$\begin{aligned}
A_1&=\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}\left|\hat{m}_3^x(y)-\hat{m}_3^{x_{k(x)}}(y)\right|,\\
A_2&=\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}\left|\hat{m}_3^{x_{k(x)}}(y)-\hat{m}_3^{x_{k(x)}}(t(y))\right|,\\
A_3&=\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}\left|\hat{m}_3^{x_{k(x)}}(t(y))-\mathrm{E}\left[\hat{m}_3^{x_{k(x)}}(t(y))\right]\right|,\\
A_4&=\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}\left|\mathrm{E}\left[\hat{m}_3^{x_{k(x)}}(t(y))\right]-\mathrm{E}\left[\hat{m}_3^{x_{k(x)}}(y)\right]\right|,\text{ and}\\
A_5&=\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}\left|\mathrm{E}\left[\hat{m}_3^{x_{k(x)}}(y)\right]-\mathrm{E}\left[\hat{m}_3^x(y)\right]\right|.
\end{aligned}$$

First we examine term A_1,

$$A_1=\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}\left|\sum_{i=1}^{n}\Gamma_i(y)\left(\Delta_i(x)-\Delta_i(x_{k(x)})\right)\right|,$$

and then term A_5,

$$\begin{aligned}
A_5&=\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}\left|\mathrm{E}\left[\hat{m}_3^{x_{k(x)}}(y)\right]-\mathrm{E}\left[\hat{m}_3^x(y)\right]\right|\\
&\leqslant\mathrm{E}\left[\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}\left|\hat{m}_3^{x_{k(x)}}(y)-\hat{m}_3^x(y)\right|\right].
\end{aligned}$$

Using for A_1 and A_5 the same case-by-case study for the two types of kernel functions as in the proof of Theorem 3.3.1 for the terms A_1 and A_3, we obtain

$$A_1 = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,3}\log n}}{n}\right) \quad \text{and} \quad A_5 = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,3}\log n}}{n}\right).$$

Next, examine the term A_2. As $H^{(1)} = K_0$, we know that the integrated kernel function H is Hölder continuous with order 1, we get by some calculations

$$\begin{aligned}
|\hat{m}_3^{x_{k(x)}}(y) - \hat{m}_3^{x_{k(x)}}(t(y))| &= \frac{1}{n}\sum_{i=1}^{n} \Delta_i(x_{k(x)})\left(\Gamma_i(y) - \Gamma_i(t(y))\right) \\
&\leqslant \frac{|y - t(y)|}{g}\frac{1}{n}\sum_{i=1}^{n}\Delta_i(x_{k(x)}) \\
&\leqslant \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i(x_{k(x)})}{gn^{\frac{\beta_2}{2}}} \\
&= \frac{1}{n}\sum_{i=1}^{n} W_{ik}.
\end{aligned}$$

As a consequence of

$$|W_{1k}| \leqslant \frac{C}{gn^{\frac{\beta_2}{2}}F(h)} < \infty$$

and Condition (D2), we can apply Corollary 3.2.3. In this case we consider a slightly modified covariance term

$$\tilde{s}_{n,1}(x) = \sum_{i,j=1}^{n}\left|\mathrm{Cov}\left(\frac{\Delta_i(x)}{gn^{\beta_2/2}}, \frac{\Delta_j(x)}{gn^{\beta_2/2}}\right)\right|,$$

with $\tilde{s}_{n,1}(x) = 1/(g^2 n^{\beta_2})\ s_{n,1}(x)$. Since $gn^{\beta_2/2} \to \infty$ as $n \to \infty$, $s_{n,1}(x)$ is an upper bound of $\tilde{s}_{n,1}(x)$. Apply Corollary 3.2.3 to find

$$\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}|\hat{m}_3^{x_{k(x)}}(y) - \hat{m}_3^{x_{k(x)}}(t(y))| = \mathcal{O}_{a.co.}\left(\frac{\sqrt{\tilde{s}^*_{n,1}\log n}}{n}\right),$$

where $\tilde{s}^*_{n,1} := \sup_{x\in S_E}(\tilde{s}_{n,1}(x))$ and due to

$$P(|X| > w_2) \leqslant P(|X| > w_1) \tag{3.24}$$

for $w_1 \leqslant w_2$ and $\tilde{s}_{n,1}(x) \leqslant s_{n,1}(x)$ we arrive at

$$A_2 = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,1}\log n}}{n}\right).$$

By a similar argument, we find for A_4,

$$A_4 = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,1}\log n}}{n}\right).$$

Again, similarly to the proof of Theorem 3.3.1, we have a closer look at the term in the middle,

$$\begin{aligned} A_3 &= \sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}\left|\hat{m}_3^{x_{k(x)}}(t(y)) - E\left[\hat{m}_3^{x_{k(x)}}(t(y))\right]\right| \\ &= \max_{x_k\in\{x_1,\dots,x_{N(S_E,d,\varepsilon)}\}}\max_{t\in\{t_1,\dots,t_{z_n}\}}\left|\hat{m}_3^{x_k}(t) - E\left[\hat{m}_3^{x_k}(t)\right]\right|. \end{aligned}$$

Rewrite this difference

$$\begin{aligned} \hat{m}_3^{x_k}(t) - E\left[\hat{m}_3^{x_k}(t)\right] &= \frac{1}{n}\sum_{i=1}^{n}\left(\Delta_i(x_k)\Gamma_i(y) - E\left[\Delta_i(x_k)\Gamma_i(y)\right]\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} W_{ki}(t). \end{aligned}$$

As a result of the boundedness of the kernel function K and the integrated kernel function H and Conditions (F) and (K), the random variable W_{ki} is bounded. Apply Theorem 3.2.1 and Corollary 3.2.1 to obtain

$$\begin{aligned} P(A_3 > \eta) &= P\left(\max_{k\in\{1,\dots,N(S_E,d,\varepsilon)\}}\max_{t\in\{t_1,\dots,t_{z_n}\}}\left|\frac{1}{n}\sum_{i=1}^{n} W_{ki}(t)\right| > \eta\right) \\ &\leqslant z_n N(S_E,d,\varepsilon)\max_k\max_t P\left(\left|\sum_{i=1}^{n} W_{ki}(t)\right| > n\eta\right) \\ &\leqslant z_n N(S_E,d,\varepsilon)\left[\frac{nC}{r}\left(\frac{r}{n\eta}\right)^{a+1} + 4\left(1+\frac{\eta^2n^2}{rs^*_{n,3}}\right)^{-\frac{r}{2}}\right] \\ &=: T_{1n} + T_{2n}. \end{aligned} \tag{3.25}$$

Similarly to the proof of Lemma 3.3.1 we get for the terms T_{1n} and T_{2n} an estimation as follows, but additional to the proof of Lemma 3.3.1 we get here the extra term z_n, the covering number of the compact set $S_{\mathbb{R}}$. Let $\eta := \eta_0\sqrt{s_n\log n}/n$ for some $\eta_0 > 0$ and $r := (\log n)^{1+b}$ for $b > 0$. Then we have

$$\begin{aligned} T_{1n} &= Cz_nN(S_E,d,\varepsilon)\left[\frac{n}{(\log n)^{1+b}}\left(\frac{n(\log n)^{1+b}}{n\eta_0 s^*_{n,3}\sqrt{\log n}}\right)^{a+1}\right] \\ &= C\eta_0^{-(a+1)}z_nN(S_E,d,\varepsilon)(\log n)^{\frac{2a(b+1)-1}{2}}ns^{*-(a+1)}_{n,3} \\ &\leqslant C\eta_0^{-(a+1)}(\log n)^{\frac{2a(b+1)-1}{2}-\tau}n^{1+\tau+\frac{\beta_2}{2}-\theta}. \end{aligned}$$

For $\theta > \frac{\beta_2}{2}+\tau+2$, we get

$$\sum_{n=1}^{\infty}(\log n)^{\frac{2a(b+1)-1}{2}-\tau}n^{1+\tau+\frac{\beta_2}{2}-\theta} < \infty. \tag{3.26}$$

By a similar argument as in the proof of Lemma 3.3.1 we get for the second term in (3.25),

$$T_{2n} \leqslant 4z_n N(S_E, d, \varepsilon) \exp\left(-\frac{1}{2} C\eta_0^2 \log n\right).$$

Choose η_0 such that

$$l = \frac{1}{2} C\eta_0^2 > \frac{\beta_2}{2} + \tau + 1.$$

Then, we obtain by the properties on the covering number of S_E and $S_{\mathbb{R}}$, namely

$$N(S_E, d, \varepsilon) \leqslant Cn^{\tau} \text{ and } z_n \leqslant Cn^{\beta_2/2},$$

the following estimate

$$\begin{aligned} T_{2n} &= 4z_n N(S_E, d, \varepsilon) n^{-l} \\ &\leqslant Cn^{\frac{\beta_2}{2}+\tau-l}. \end{aligned}$$

This implies

$$\sum_{i=1}^{\infty} n^{\beta/2+\tau-\beta} < \infty. \tag{3.27}$$

Finally, combine relations (3.26) and (3.27) to find

$$A_3 = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s_{n,3}^* \log n}}{n}\right).$$

This finishes the proof. □

The Geometrically Mixing Case

Similarly as in the case of the regression function we get the uniform convergence rates for a larger class of functional spaces for geometrically mixing random variables. For the conditional distribution function we do not need Condition (M2) on the response variable. Because of this, we get here for the constant $p = \infty$ and out of this we have $a = b$.

Theorem 3.4.3 (Geometrically Mixing) *With Conditions (F), (K), (R2), (E2), (A2), and (D2), we have*

$$\sup_{x \in S_E} \sup_{y \in S_{\mathbb{R}}} |\hat{F}^x(y) - F^x(y)| = \mathcal{O}\left(h^{\beta_1}\right) + \mathcal{O}\left(g^{\beta_2}\right) + \mathcal{O}_{a.co.}\left(\frac{\sqrt{s_n K_{S_E}\left(\frac{1}{n}\right)}}{n}\right).$$

Proof:
By the same decomposition as in the case of arithmetically mixing random variables in the proof of Theorem 3.4.2 and by applying then Lemma 3.3.2, Lemma 3.3.5, Lemma 3.4.1, and the successive Lemma 3.4.3, we finish the proof. □

Lemma 3.4.3 *By the same Conditions as in Theorem 3.4.3, we have*

$$\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}|\hat{m}_3^x(y)-E[\hat{m}_3^x(y)]|=\mathcal{O}_{a.co.}\left(\frac{\sqrt{s_n K_{S_E}\left(\frac{1}{n}\right)}}{n}\right).$$

Proof:
To prove this lemma we need just minimal modifications of the proof for the generalised regression function. In addition to the proof of the regression function estimate we have here the covering number z_n of the compact space $S_{\mathbb{R}}$. If we have a closer look at the proof of Lemma 3.3.4, it can be seen that this additional factor in the estimation has just minor influence in the proof, since the covering number of S_E is of exponential order. Another difference to the proof of Lemma 3.3.4 is that we have to consider the terms D_2 and D_4 similarly as in the proof of Lemma 3.4.2, where the assumption $gn^{\beta_2/2}\to\infty$ as $n\to\infty$ is needed. By doing this, this lemma is proven. □

3.5 THE CONDITIONAL DENSITY FUNCTION

3.5.1 *Notations and Assumptions*

In this section the kernel estimate of the conditional density function for α-mixing random variables is examined. We assume here that the probability distribution of Y given X is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$. Then we denote $f^x(y)$ as the corresponding density function for the pair (x, y). If it is assumed that the conditional distribution function is differentiable in y, the density can be written as

$$\forall y \in \mathbb{R}: \; f^x(y) = \frac{\partial}{\partial y} F^x(y).$$

By this a kernel estimate $\hat{f}^x(y)$ of $f^x(y)$ follows directly as

$$\hat{f}^x(y) = \sum_{i=1}^{n} g_n^{-1} K_0(g_n^{-1}(y - Y_i)) \frac{K\left(h_n^{-1} d(x, X_i)\right)}{\sum_{j=1}^{n} K\left(h_n^{-1} d(x, X_j)\right)}$$

$$\text{if } \sum_{j=1}^{n} K\left(h_n^{-1} d(x, X_i)\right) \neq 0,$$

otherwise $\hat{f}^x(y) = 0$, for $x \in E$ and $y \in \mathbb{R}$. K_0 is a classical-type kernel function and $h := h_n$ and $g := g_n$ are bandwidths. Define as in the sections before the corresponding covariance term

$$s_{n,4}(x, y) := \sum_{i,j=1}^{n} |\mathrm{Cov}(\Omega_i(y)\Delta_i(x), \Omega_j(y)\Delta_j(x))|,$$

where

$$\Omega_i(y) := g^{-1} K_0(g^{-1}(y - Y_i)) \;\text{ and }\; \Delta_i(x) := \frac{K(h^{-1} d(x, X_i))}{E\left[K(h^{-1} d(x, X_1))\right]}.$$

Let $S_{\mathbb{R}} \subset \mathbb{R}$ be compact.

Condition on the classical-type kernel function

(H1) Assume that the classical-type kernel function K_0 is Lipschitz continuous and bounded, namely let $C > 0$ and for all $u_1, u_2 \in \mathbb{R}$ we have

$$|K_0(u_1) - K_0(u_2)| \leqslant C|u_1 - u_2|.$$

(H2) Furthermore, assume for the classical-type kernel function K_0

$$\int |t|^{\beta_2} K_0(t)\,dt < \infty \text{ and}$$
$$\int K_0^2(t)\,dt < \infty.$$

Condition on the regularity of the conditional density function

(R3) Assume that the conditional density function $f^x(y)$ is of Hölder-type. There exist constants $C_1, C_2 > 0$ such that we have for all $x_1, x_2 \in S_E$ and $y_1, y_2 \in \mathbb{R}$

$$|f^{x_1}(y_1) - f^{x_2}(y_2)| \leqslant C_1(d(x_1, x_2))^{\beta_1} + C_2|y_1 - y_2|^{\beta_2}.$$

Condition on the covariance terms

(D3) Define $s^*_{n,4} := \sup_{x \in S_{\mathbb{R}}} \sup_{y \in S_E} (s_{n,4}(x, y))$ and $s_n := \max\left\{s^*_{n,1}, s^*_{n,4}\right\}$, then assume

$$s_n^{-(a+1)} = o(n^{-\theta}),$$

where $\theta > \tau + \beta_2 + 2$ and β_2 is the constant of Condition (R) and τ the constant of Condition (E1).

3.5.2 *Main Results*

The Arithmetically Mixing Case

Theorem 3.5.1 (Arithmetically Mixing) *Assume Conditions (F), (K), (R3), (E1), (H1), (H2), (R), (A1), and (D3) hold, and we have for the bandwidth g that $\lim_{n\to\infty} gn^{\beta_2/2} = \infty$, where β_2 is the second Hölder constant of Condition (R3), then we have*

$$\sup_{x \in S_E} \sup_{y \in S_{\mathbb{R}}} |\hat{f}^x(y) - f^x(y)| = \mathcal{O}\left(h^{\beta_1}\right) + \mathcal{O}\left(g^{\beta_2}\right) + \mathcal{O}_{a.co.}\left(\frac{\sqrt{s_n \log n}}{n}\right).$$

Proof:
Analogously to the proofs of Theorem 3.3.1 or Theorem 3.4.2 we fall back to the ideas of proving uniform consistency in the independent case, Theorem 5 of Ferraty et al. [23] or Ferraty and Vieu [30]. First, the difference of the kernel estimate and the conditional density function is divided as in the independent case as follows

$$\begin{aligned}\hat{f}^x(y) - f^x(y) =& \frac{1}{\hat{f}(x)}\left(\hat{m}_4^x(y) - E\left[\hat{m}_4^x(y)\right] - \left(f^x(y) - E\left[\hat{m}_4^x(y)\right]\right)\right) \\ &+ \frac{f^x(y)}{\hat{f}(x)}\left(E\left[\hat{f}(x)\right] - \hat{f}(x)\right), \end{aligned} \tag{3.28}$$

where

$$\hat{m}_4^x(y) = \frac{1}{n}\sum_{i=1}^{n} \Omega_{in}(y)\Delta_i(x).$$

Apply the supremum on these terms as in the proof of Theorem 3.3.1 or Theorem 3.4.2. Afterwards, the proof is a conclusion of Lemma 3.3.1 and Lemma 3.3.2 of Section 3.3 and the successive Lemma 3.5.1 and Lemma 3.5.2. □

Lemma 3.5.1 *By Conditions (F), (K), (H1), (H2), and (R), we have*

$$\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}|f^x(y)-E\left[\hat{m}_4^x(y)\right]|=O(h^{\beta_1})+O(g^{\beta_2}).$$

Proof:
This proof is the same as in the independent case. The result is obtained by using the Hölder continuity (H1) and the Condition (H2). For the proof see Lemma 14 in the paper by Ferraty et al. [23]. □

Lemma 3.5.2 *With Conditions of Theorem 3.5.1, we have*

$$\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}|\hat{m}_4^x(y)-E\left[\hat{m}_4^x(y)\right]|=\mathcal{O}_{a.co.}\left(\frac{\sqrt{s_n\log n}}{n}\right).$$

Proof:
Analogously to the proof of the consistency of the conditional distribution function in Theorem 3.4.2, the compact set $S_{\mathbb{R}}$ can be covered by a finite number of open balls, namely

$$S_{\mathbb{R}}\subset\bigcup_{k=1}^{z_n}(t_k-r_n,t_k+r_n).$$

By choosing the radius of the balls as $r_n=Cn^{-\beta_2}$, we get for the covering number $z_n\leqslant n^{\beta_2}$. Let $t(y)$ be the closest point of the centres of the balls to a point $y\in S_{\mathbb{R}}$,

$$t(y):=\underset{\{t_1,\ldots,t_{z_n}\}}{\arg\min}\,|t_k-y|.$$

Decompose the difference as follows

$$\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}\left|\hat{m}_4^x(y)-E\left[\hat{m}_4^x(y)\right]\right|\leqslant A_1+A_2+A_3+A_4+A_5,$$

where

$$\begin{aligned}
A_1&=\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}\left|\hat{m}_4^x(y)-\hat{m}_4^{x_{k(x)}}(y)\right|,\\
A_2&=\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}\left|\hat{m}_4^{x_{k(x)}}(y)-\hat{m}_4^{x_{k(x)}}(t(y))\right|,\\
A_3&=\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}\left|\hat{m}_4^{x_{k(x)}}(t(y))-E\left[\hat{m}_4^{x_{k(x)}}(t(y))\right]\right|,\\
A_4&=\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}\left|E\left[\hat{m}_4^{x_{k(x)}}(t(y))\right]-E\left[\hat{m}_4^{x_{k(x)}}(y)\right]\right|,\text{ and}\\
A_5&=\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}\left|E\left[\hat{m}_4^{x_{k(x)}}(y)\right]-E\left[\hat{m}_4^x(y)\right]\right|.
\end{aligned}$$

The proof for the terms A_1 and A_5 is analogous to the terms A_1 and A_3 in the proof of Lemma 3.3.1. Similarly to there, a case-by-case analysis for the two type of kernel functions has to be done here. We obtain

$$A_1 = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,4}\log n}}{n}\right) \quad \text{and} \quad A_5 = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,4}\log n}}{n}\right).$$

Term A_2 and A_4 are similar to D_2 and D_4 in the proof of the conditional distribution function in Lemma 3.4.2. Rewrite the sum, apply Condition (H1) and the Lipschitz continuity of the kernel function K_0,

$$\begin{aligned}
|\hat{m}_4^{x_{k(x)}}(y) - \hat{m}_4^{x_{k(x)}}(t(y))| &= \left|\frac{1}{n}\sum_{i=1}^{n}\left(\Omega_i(y) - \Omega_i(t(y))\right)\Delta_i(x_{k(x)})\right| \\
&\leqslant \left|\frac{|y - t(y)|}{g^2 n}\sum_{i=1}^{n}\Delta_i(x_{k(x)})\right| \\
&\leqslant \left|\frac{r_n}{g^2 n}\sum_{i=1}^{n}\Delta_i(x_{k(x)})\right| \\
&= \left|\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i(x_{k(x)})}{g^2 n^{\beta_2}}\right| \\
&= \left|\frac{1}{n}\sum_{i=1}^{n} W_{ki}\right|.
\end{aligned}$$

Due to

$$|W_{k1}| \leqslant \frac{C}{g^2 n^{\beta_2} F(h)},$$

and by Condition (D3), we can now apply Corollary 3.2.3. Analogously to the proof of Lemma 3.4.2 we have here a slightly modified covariance term

$$\tilde{s}_{n,1}(x) = \sum_{i,j=1}^{n}\left|\mathrm{Cov}\left(\frac{\Delta_i(x)}{g^2 n^{\beta_2}}, \frac{\Delta_j(x)}{g^2 n^{\beta_2}}\right)\right|.$$

Similarly, we have that $\tilde{s}_{n,1}(x) = 1/(g^4 n^{2\beta_2})\, s_{n,1}(x)$. Since $g n^{\beta_2/2} \to \infty$ as $n \to \infty$, $s_{n,1}(x)$ is an upper bound of $\tilde{s}_{n,1}(x)$. By applying now Corollary 3.2.3, we get

$$\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}|\hat{m}_4^{x_{k(x)}}(y) - \hat{m}_4^{x_{k(x)}}(t(y))| = \mathcal{O}_{a.co.}\left(\frac{\sqrt{\tilde{s}^*_{n,1}\log n}}{n}\right). \tag{3.29}$$

Finally, out of (3.29) and the fact of (3.24), we arrive at

$$\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}|\hat{m}_4^{x_{k(x)}}(y) - \hat{m}_4^{x_{k(x)}}(t(y))| = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s_n\log n}}{n}\right).$$

The rate for A_4 we get by

$$\begin{aligned} A_4 &= \sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}\left|E\left[\hat{m}_4^{x_{k(x)}}(t(y))\right]-E\left[\hat{m}_4^{x_{k(x)}}(y)\right]\right| \\ &\leqslant E\left[\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}\left|\hat{m}_4^{x_{k(x)}}(t(y))-\hat{m}_4^{x_{k(x)}}(y)\right|\right] \end{aligned}$$

and a similar calculation as for A_2. Finally, term A_3 has to be examined. This calculation is analogous to that of term D_3 in the proof of Lemma 3.4.2. We just replace the definition of the covering number from $z_n \leqslant Cn^{\beta_2/2}$ by $z_n \leqslant Cn^{\beta_2}$ and use instead of (D2) the Condition (D3). We arrive at

$$A_3 = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s^*_{n,4}\log n}}{n}\right).$$

□

The Geometrically Mixing Case

Similarly to the conditional distribution function we just need to examine the bias term $\hat{m}_4^x(y) - E\left[\hat{m}_4^x(y)\right]$ of the decomposition in (3.28) to get the uniform almost complete convergence result.

Theorem 3.5.2 (Geometrically Mixing) *Assume Conditions (F), (K), (R3), (E2), (H1), (H2), (A2), and (D2) hold, where in (D2) the covariance term s_n is defined as in (D3), and assume for the bandwidth g that $\lim_{n\to\infty} gn^{\beta_2/2} = \infty$, where β_2 is the second Hölder constant of Condition (R3), then we have*

$$\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}|\hat{f}^x(y) - f^x(y)| = \mathcal{O}\left(h^{\beta_1}\right) + \mathcal{O}\left(g^{\beta_2}\right) + \mathcal{O}_{a.co.}\left(\frac{\sqrt{s_n K_{S_E}\left(\frac{1}{n}\right)}}{n}\right).$$

Proof:
By the same decomposition as in the case of arithmetically mixing random variables in the proof of Theorem 3.5.1 and by applying then Lemma 3.3.2, the reciprocal of $\hat{f}(x)$, Lemma 3.3.5, the rate of $\hat{f}(x) - 1$, Lemma 3.4.1, the bias and the successive Lemma 3.5.3, we get the result. □

Lemma 3.5.3 *With the Conditions of Theorem 3.5.2, we have*

$$\sup_{x\in S_E}\sup_{y\in S_{\mathbb{R}}}|\hat{m}_4^x(y) - E\left[\hat{m}_4^x(y)\right]| = \mathcal{O}_{a.co.}\left(\frac{\sqrt{s_n K_{S_E}\left(\frac{1}{n}\right)}}{n}\right).$$

Proof:
The argumentation of the proof of this lemma is the same as that for Lemma 3.4.3.
□

4

BOOTSTRAPPING IN NON-PARAMETRIC REGRESSION FOR BANDWIDTH SELECTION

4.1 INTRODUCTION

This chapter focuses on the issue of bandwidth choice in non-parametric functional regression. The primary emphasis is on a data-driven local selection procedure. Traditionally, the mean squared error is used as a measure of accuracy for choosing the smoothing parameter. As we want a local method, the pointwise mean squared error is considered. We present a bootstrap method for choosing the bandwidth by using this accuracy measure and we prove the asymptotic optimality of this selection procedure. As an open problem, such a method is firstly mentioned in the functional context by Ferraty et al. [25]. In Ferraty et al. [27] the validity of the bootstrap in non-parametric regression is treated. There they proved that the distribution of $m(x) - \hat{m}(x)$ can be approximated by a bootstrap kernel estimate. Our proof is strongly connected to the lemmas and proofs of that work.

This bootstrap procedure enlarges the numbers of methods for choosing the smoothing parameter in the non-parametric functional regression analysis. Rachdi and Vieu [58] prove the asymptotic optimality of global cross-validation. Behenni et al. [3] adapt the global cross-validation method to a local one. The optimal bandwidth, in the sense of minimising the mean squared error, can also be estimated by a plugin approach, for example by estimation of the constants in the expansion of the mean squared error, see Theorem 4.3.1.

The ideas in this section my be extended to the k-NN kernel estimate, where the bandwidth depends on how concentrated the data is around the point of interest x. Burba et al. [11] prove the almost complete convergence, a validation of a cross-validation or bootstrap procedure for the k-NN kernel estimate is outstanding. Ouyang et al. [55] treated cross-validation for the k-NN kernel estimate, maybe it can be transferred.

4.2 PRELIMINARIES

4.2.1 *Description of the Kernel Estimate*

On the basis of i. i. d. observations $D_n := ((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))$ distributed as (X, Y) with unknown distribution P, we focus on the functional non-parametric regression model

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

The errors ε_i are distributed as a random variable ε that satisfies

$$E[\varepsilon|X] = 0$$

and we denote by

$$\sigma_\varepsilon^2(X) := E\left[\varepsilon^2 \middle| X\right]$$

the second conditional moment. If $\sigma_\varepsilon^2(X) \equiv const$, we speak of homoscedastic residuals else we speak of heteroscedastic residuals. The kernel estimate of the regression function $m(x)$ is expressed by

$$\hat{m}_{h_n}(x) = \sum_{i=1}^{n} Y_i \frac{K\left(h_n^{-1} d(x, X_i)\right)}{\sum_{i=1}^{n} K\left(h_n^{-1} d(x, X_i)\right)}, \quad \text{if} \quad \sum_{j=1}^{n} K\left(h_n^{-1} d(x, X_i)\right) \neq 0, \tag{4.1}$$

otherwise $\hat{m}_{h_n}(x) = 0$, for $x \in E$ fixed. K is a kernel function and $h := h_n$ is a positive sequence. We differ in this chapter from the notation of the kernel estimate of the previous chapters as we examine the choice of the bandwidth here.

4.2.2 *Motivation of this Bandwidth Selection Procedure*

It is natural to measure the performance of a regression function estimate by a L_2–error criterion. Recall Section 1.1, where we are looking for a measurable function $f : E \to \mathbb{R}$ such that

$$E\left[(Y - f(X))^2\right]$$

is minimal. The solution of this minimisation is the regression function $m(X)$. It can be proven that the regression function is also the solution of minimising the following term

$$E\left[(Y - f(X))^2 | X = x\right].$$

This regression function is then approximated by the kernel estimate given in (4.1). Thus, we want to choose a bandwidth h depending on $x \in E$ such that

$$E\left[(Y - \hat{m}_h(X))^2 | X = x\right] \text{ is close to } E\left[(Y - m(X))^2 | X = x\right].$$

By adding a null and a simple calculation, we find

$$\begin{aligned} E\left[(Y-\hat{m}_h(X))^2|X=x\right] &= E\left[(Y-m(X)+m(X)-\hat{m}_h(X))^2|X=x\right] \\ &= E\left[(m(X)-\hat{m}_h(X))^2|X=x\right] \\ &\quad + E\left[(Y-m(X))^2|X=x\right]. \end{aligned} \tag{4.2}$$

The left side in (4.2) decreases as soon as $\hat{m}_h(x)$ is close to $m(x)$. As the distribution of (X,Y) is unknown and therefore $m(x)$, it is not feasible to calculate $(m(x)-\hat{m}_h(x))^2$. One possible way out is bootstrapping. In functional non-parametric regression Ferraty et al. [27] show that the distribution of $m(x)-\hat{m}_h(x)$ can be approximated by a bootstrap approximation $m^*_{gh}(x)-\hat{m}_g(x)$, with $m^*_{gh}(x)$ as the bootstrap kernel estimate, which will be defined later on, and $\hat{m}_g(x)$ as the pilot kernel estimate with bandwidth g. We will show in the following sections how to construct that bootstrap kernel estimate $m^*_{gh}(x)$ and prove the asymptotic optimality of the bootstrap method for bandwidth selection. To be more precise, we will show

$$nF_x(h)\left(E\left[(m^*_{gh}(x)-\hat{m}_g(x))^2|D_n\right]-E\left[(m(x)-\hat{m}_h(x))^2\right]\right)\to 0 \text{ a.s. } n\to\infty.$$

4.3 BOOTSTRAP IN FUNCTIONAL NON-PARAMETRIC REGRESSION

Ferraty et al. [27] prove the validity of the residual and wild bootstrap in non-parametric functional regression. As already mentioned above, they use the bootstrap procedure for approximating the distribution function of $m(\cdot)-\hat{m}_h(\cdot)$ by the conditioned distribution function of $m^*_{gh}(x)-\hat{m}_g(x)$ on the data, more precisely

$$\begin{aligned} \sup_{y\in\mathbb{R}} \Big| & P\left(\sqrt{nF_x(h)}(m(x)-\hat{m}_h(x))\leqslant y\right) \\ & - P\left(\sqrt{nF_x(h)}(m^*_{gh}(x)-\hat{m}_g(x))\leqslant y|D_n\right)\Big| \to 0 \text{ a.s.} \end{aligned}$$

There are a few more techniques for proving the validity of a bootstrap procedure, see e.g. the monograph of Shao and Tu [64].

The idea that we are using for proving the asymptotic validity of this adaptive bandwidth selection procedure is based on the work of Manteiga et al. [50] or Hall et al. [35]. In both papers the authors prove in the finite-dimensional case that the bootstrap procedure for choosing the optimal bandwidth is valid.

4.3.1 *Bootstrap Procedure*

Herein the procedure for residual and wild bootstrapping is presented. As the response variable Y is a real-valued random variable, the error ε, which is bootstrapped in this functional context, as for instance in [37], is also a real-valued random variable. In the case of residual bootstrap we assume that the error is homoscedastic, i. e. $\sigma^2_\varepsilon(X)\equiv const$. On the next lines we present the bootstrap procedure used in non-parametric regression, see e.g. [27].

S1 : Use data D_n to get a kernel estimate $\hat{m}_g(x)$ with a bandwidth g, chosen by a consistent bandwidth selection procedure, for example cross-validation. The

bandwidth g is called the *pilot bandwidth* and $\hat{m}_g(x)$ the *pilot kernel estimate*. Calculate with this pilot kernel estimate the residuals

$$\hat{\varepsilon}_{i,g} = Y_i - \hat{m}_g(X_i) \text{ for } i = 1, \dots, n.$$

S2 : This step is divided into the homoscedastic and the heteroscedastic case.

a) *Residual Bootstrapping*, $\sigma_\varepsilon^2(X) \equiv \text{const}$
Draw n i. i. d. random variables $\varepsilon_1^*, \dots, \varepsilon_n^*$ from the cumulative distribution of the centred residuals

$$(\tilde{\varepsilon}_{1,g}, \dots, \tilde{\varepsilon}_{n,g}) = (\hat{\varepsilon}_{1,g} - \overline{\hat{\varepsilon}}_g, \dots, \hat{\varepsilon}_{n,g} - \overline{\hat{\varepsilon}}_g),$$

where $\overline{\hat{\varepsilon}}_g$ is the empirical mean of the residuals $\hat{\varepsilon}_{1,g}, \dots, \hat{\varepsilon}_{n,g}$.

b) *Wild Bootstrapping*
The bootstrap residuals are defined as

$$\varepsilon_i^* = \hat{\varepsilon}_{i,g} V_i \text{ for } i = 1, \dots, n,$$

where $V_1, \dots, V_n$ are i. i. d. real-valued random variables with the properties

$$E[V_1] = 0 \text{ and } E\left[V_1^2\right] = 1.$$

S3 : Define the bootstrap response variables as follows,

$$Y_i^* = \hat{m}_g(X_i) + \varepsilon_i^*.$$

S4 : Calculate, based on this bootstrapped data $(X_1, Y_1^*), \dots, (X_n, Y_n^*)$, the kernel estimate $\hat{m}_{hg}^*(x)$ with bandwidth h. The index indicates that this estimate depends on the pilot kernel estimate $\hat{m}_g$ and on bandwidth h.

Repeat this bootstrap procedure B times and use the empirical distribution of $\hat{m}_{hg}^*(x) - \hat{m}_g(x)$ for selecting the bandwidth. More precisely, let H be a fixed set of bandwidths, for example the distances to the k nearest neighbours of the element x, then

$$h^*(x) := \arg\min_{h \in H} \frac{1}{B} \sum_{b=1}^{B} (\hat{m}_{hg}^*(x) - \hat{m}_g(x))^2.$$

In wild bootstrapping a random variable V_i is used for getting the residuals. In Step 2 of the above procedure Mammen [49] suggests for the V_i the following continuous auxiliary distribution

$$V_i := (1 - 2 \cdot 20^{-\frac{2}{3}})^{\frac{1}{2}} W_i + 20^{-\frac{1}{3}}(W_i^2 - 1),$$

where the W_i are i. i. d. as $N(0, 1)$. In practice the choice of this continuous auxiliary distribution provides good results. Alternatively, a two-point distribution can be chosen for the auxiliary distribution V_i. For a discussion on such alternative V_i's we refer to Davidson et al. [19].

The theoretical validation of this adaptive bandwidth selection method is given in Subsection 4.3.3.

4.3.2 *Assumptions, Notations, and Asymptotic Expansion*

All of the conditions and notations are introduced and commented on in the two papers of Ferraty, [25] and [27], as long as the consistency result here is based on the lemmas of the cited papers.

Let $B_h(x) := \{x_1 \in E : d(x_1, x) \leqslant h\}$ be a ball centred in $x \in E$ with radius h and $F_x(h) = P(d(X, x) \leqslant h)$. Furthermore, denote

$$M_0(x) = K(1) - \int_0^1 (sK(s))' \tau_{0x}(s) ds,$$

$$M_1(x) = K(1) - \int_0^1 K'(s) \tau_{0x}(s) ds, \text{ and}$$

$$M_2(x) = K^2(1) - \int_0^1 (K^2)'(s) \tau_{0x}(s) ds,$$

where $\tau_{0x}(s) = \lim_{h \to 0+} \tau_{hx}(s)$ with

$$\tau_{hx}(s) = F_x(sh)/F_x(h) = P(d(x, X) \leqslant sh \mid d(x, X) \leqslant h).$$

For example, if the small ball probability is of fractal-type, namely $F_x(h) \sim Ch^\tau$ for some $\tau, C > 0$, then $\tau_{0x}(s) = s^\tau$. The proof and some more examples can be found in Proposition 1 of Ferraty et al. [30] and for some deeper discussion on that topic see Section 4 therein.

Conditions on the regularity

(M) Assume that the regression function $m(x)$, the second conditional moment $\sigma_\varepsilon^2(x)$, and the conditional expectation $E\left[|Y| \middle| X = x\right]$ are continuous in a neighbourhood of x. Furthermore, assume that

$$\exists \varepsilon > 0 \; \forall m \in \mathbb{N} : \sup_{x_1 \in E:\, d(x_1, x) < \varepsilon} E\left[\, |Y|^m | X = x_1\right] < \infty.$$

Conditions on the conditional expectation

(D) Define

$$\phi_x(s) := E\left[m(X) - m(x) | d(X, x) = s\right],$$

then assume for all elements (x_1, s) in a neighbourhood of $(x, 0)$ that $\phi_{x_1}(0) = 0$, $\phi'_{x_1}(s)$ exists, $\phi'_{x_1}(0) \neq 0$, and $\phi'_{x_1}(s)$ is uniformly Hölder continuous of order $0 < \alpha \leqslant 1$ in (x_1, s).

This condition is a workaround in non-parametric functional regression analysis for calculating the derivative of the regression function $m(x)$. For a link between the function $\phi_x(s)$ and $m'(x)$ and for some more comments we refer to Ferraty et al. [30].

Conditions on the small ball probability

(F) Assume for all $x_1 \in E$ that $F_{x_1}(0) = 0$ and further on assume $F_{x_1}(t)/F_x(t)$ is Hölder continuous of order α in x_1, uniformly in t in a neighbourhood of 0, with the same α as in (D).

Conditions on $\tau_{hx}(s)$

(T) Assume for all $x_1 \in E$ and $0 \leqslant s \leqslant 1$ that $\tau_{0x_1}(s)$ exists and

$$\sup_{\substack{x_1 \in E \\ 0 \leqslant s \leqslant 1}} |\tau_{hx_1}(s) - \tau_{0x_1}(s)| = o(1).$$

Furthermore, assume

$$M_0(x) > 0,\ M_1(x) > 0,\ \text{and} \inf_{d(x_1,x)<\varepsilon} M_2(x) > 0$$

for some $\varepsilon > 0$ and suppose for $k = 1, 2, 3$, that $M_k(x_1)$ is Hölder continuous of order α, with α defined in (D).

Conditions on the kernel function

(K) The support of the kernel function K is $[0, 1)$, $K \in C^1[0, 1)$, $K'(s) \leqslant 0$ for $0 \leqslant s < 1$ and $K(1) > 0$.

In this chapter a slightly different definition of the kernel functions is used, as different properties are used here. Most kernel functions introduced in Chapter 1 can be modified such that they fulfil these assumptions.

Conditions on the bandwidths

(B) Let $h := h_n$ and $g := g_n$. We assume that the bandwidth h is of the same order as the optimal bandwidth, namely

$$h = O\left((nF_x(h))^{-1/2}\right).$$

The pilot bandwidth g satisfies $g \to 0$ and $h/g \to 0$ for $n \to \infty$.

The following conditions are of a technical nature. Assume $nF_x(h) \to \infty$, $h\sqrt{nF_x(h)} = O(1)$, $g^{1+\alpha}\sqrt{nF_x(h)} = o(1)$, $F_x(h+g)/F_x(g) \to 1$, $\log n(F_x(h)/F_x(g)) = o(1)$ and $gh^{\alpha-1} = O(1)$, where α is the Hölder constant defined in (D).

To make the bootstrap procedure work, the pilot bandwidth g has to be asymptotically of larger order than h. This condition is similar to the finite-dimensional case of non-parametric regression. Härdle and Marron [38] illustrate this by an asymptotic analysis of the bias term and the asymptotic analysis of the bias of the bootstrap approximation. It turns out in the proof of Lemma 4 in Ferraty et al. [27] that in the infinite-dimensional non-parametric regression this over-smoothing of the pilot kernel estimate is also needed, for the same reasons as in the finite-dimensional considerations.

For finite-dimensional non-parametric regression some alternative techniques to over-smoothing the pilot kernel estimate are developed for handling the bias term. For example, Härdle and Bowman [38] mirrore the bias-variance trade-off to the bootstrap kernel estimate. To do that they view an approximation of a decomposition of $m(x) - \hat{m}(x)$, and use the terms of this decomposition to construct an unbiased bootstrap kernel estimate. Another method is introduced by Hall [34] who estimates the bias term by subsampling. Hall samples from a smaller data set D_{m_n}, $m_n < n$, and he defines in this data set a kernel estimate. Due to the dependence of the bandwidth on the sample size, the optimal bandwidth h_{m_n} on the data set D_{m_n} is different to the optimal bandwidth h_n on the full data set D_n. Then the bias term can be estimated by the difference of the kernel estimates on D_{m_n} and D_n.

Technical condition

(E) For each n, there exists $r_n \geqslant 1$, $l_n > 0$ and curves $t_{1n}, \ldots, t_{r_n n} \in E$ such that we have a finite covering around x,

$$B(x, h) \subset \bigcup_{i=1}^{r_n} B(t_{in}, l_n),$$

with $r_n = O(n^{g/h})$ and $l_n = o(g(\sqrt{nF_x(h)})^{-1})$.

It is assumed that the covering number is of a polynomial order. In view of the paper by Ferraty et al. [23] Condition (E) can be generalised to function spaces, where the covering number is of exponential order. To get such a result, the covering number has to fulfil Condition (H5b) in Ferraty et al. [23]. Then by a modification of the proof of Lemma 6 in [27], one obtains the validity of bootstrapping in non-parametric functional regression on a larger class of function spaces.

4.3.2.1 *Mean Squared Error and the Choice of the Bandwidth*

This subsection focuses on the choice of the bandwidth in view of the mean squared error. Ferraty et al. [25] give an asymptotic evaluation of the mean squared error of the kernel estimate. This asymptotic evaluation is generalised by Delsol [20] to higher moments and α-mixing functional random variables. Before we give some comments on the choice of the bandwidth, we present the results from Ferraty et al. [25].

Theorem 4.3.1 (Ferraty et al. [25]) *Under Conditions (M), (D), (F), (T), (K), and (B), we have for the kernel estimate the following asymptotic development,*

$$E[\hat{m}_h(x)] - m(x) = \phi'(0)\frac{M_0(x)}{M_1(x)}h + O\left((nF_x(h))^{-1}\right) + o(h) \text{ and} \tag{4.3}$$

$$Var[\hat{m}_h(x)] = \frac{1}{nF_x(h)}\frac{M_2(x)}{M_1^2(x)} + o\left((nF_x(h))^{-1}\right). \tag{4.4}$$

It turns out that we have similar characteristics in the functional non-parametric regression as to finite-dimensional case, see for example Györfi et al. [33]. Analogously to the almost complete convergence rate, see Chapter 1, we have to control the bandwidth h for $n \to \infty$. Firstly, as a result of examining the bias term (4.3), one would choose a bandwidth h that converges fast to zero. But, if we consider the variance term (4.4), we need $nF_x(h) \to \infty$. Thus, a bandwidth selection method has to control this bias-variance trade-off.

4.3.3 *Main Result*

In the following theorem we show that the bootstrap method can be used for adaptive bandwidth selection in non-parametric functional regression.

Theorem 4.3.2 *Under Conditions (M), (D), (F), (T), (K), (B), and (E), we have*

$$nF_x(h)\left(E\left[(\hat{m}^*_{hg}(x) - \hat{m}_g(x))^2 \middle| D_n\right] - E\left[(\hat{m}_h(x) - m(x))^2\right]\right) \to 0 \text{ a.s.}$$

as $n \to \infty$. This theorem holds for residual as well as for wild bootstrapping.

We want to remark that the statement of our result in Theorem 4.3.2 is equivalent to the following

$$\frac{E\left[(\hat{m}_{\hat{h}}(x) - m(x))^2\right]}{\inf\limits_{h \in H_n} E\left[(\hat{m}_h(x) - m(x))^2\right]} \to 1 \text{ a.s.}$$

as $n \to \infty$, where $\hat{h}(x) := \arg\min\limits_{h \in H} E\left[(\hat{m}^*_{hg}(x) - \hat{m}_g(x))^2 \middle| D_n\right]$ and H_n is a set of bandwidths that is of the same order as the optimal bandwidth, namely $\mathcal{O}\left((nF_x(h))^{-1/2}\right)$. The idea of the proof of equivalence can be found in [36, p. 199].

Proof
First, note that the mean squared error can be described by the following bias-variance decomposition,

$$\begin{aligned} E\left[(\hat{m}_h(x) - m(x))^2\right] &= (E\left[\hat{m}_h(x)\right] - m(x))^2 + Var\left[\hat{m}_h(x)\right] \text{ and} \\ E\left[(\hat{m}^*_{hg}(x) - \hat{m}_g(x))^2 \middle| D_n\right] &= \left(E\left[\hat{m}^*_{hg}(x) \middle| D_n\right] - \hat{m}_h(x)\right)^2 \\ &\quad + Var\left[\hat{m}^*_{hg}(x) \middle| D_n\right]. \end{aligned}$$

With the help of Lemma 3 and Remark 2 of Ferraty et al. [27] we have for residual and wild bootstrapping

$$Var\left[\hat{m}^*_{hg}(x) \middle| D_n\right] = Var\left[\hat{m}_h(x)\right] + o\left((nF_x(h))^{-1}\right) \text{ a.s.}$$

Out of it we obtain

$$Var\left[\hat{m}^*_{hg}(x) \middle| D_n\right] - Var\left[\hat{m}_h(x)\right] = o\left((nF_x(h))^{-1}\right) \text{ a.s.}$$

Now, consider the bias terms,

$$\left(\mathrm{E}\left[\hat{m}_h(x)\right] - m(x)\right)^2 - \left(\mathrm{E}\left[\hat{m}^*_{hg}(x)\middle|D_n\right] - \hat{m}_h(x)\right)^2 =: B_{1n} \cdot B_{2n},$$

where

$$\begin{aligned} B_{1n} &= \mathrm{E}\left[\hat{m}_h(x)\right] - m(x) - \left(\mathrm{E}\left[\hat{m}^*_{hg}(x)\middle|D_n\right] - \hat{m}_h(x)\right) \text{ and} \\ B_{2n} &= \mathrm{E}\left[\hat{m}_h(x)\right] - m(x) + \mathrm{E}\left[\hat{m}^*_{hg}(x)\middle|D_n\right] - \hat{m}_h(x). \end{aligned}$$

By Lemma 4 of Ferraty et al. [27] we get

$$\begin{aligned} B_{1n} &= \mathrm{E}\left[\hat{m}^*_{hg}(x)\middle|D_n\right] - \hat{m}_h(x) - \mathrm{E}\left[\hat{m}_h(x)\right] + m(x) \\ &= o\left((nF_x(h))^{-1/2}\right) \text{ a.s.} \end{aligned} \tag{4.5}$$

For the second term we have

$$\begin{aligned} B_{2n} &= \mathrm{E}\left[\hat{m}_h(x)\right] - m(x) + \mathrm{E}\left[\hat{m}^*_{hg}(x)\middle|D_n\right] - \hat{m}_h(x) \\ &= 2\left(\mathrm{E}\left[\hat{m}_h(x)\right] - m(x)\right) + B_{1n}. \end{aligned} \tag{4.6}$$

The choice of the bandwidth h in Condition (B) and (4.3) lead to

$$\mathrm{E}\left[\hat{m}_h(x)\right] - m(x) = o\left((nF_x(h))^{-1/2}\right). \tag{4.7}$$

Finally, combine relation (4.5), (4.6), and (4.7) to find

$$\left(\mathrm{E}\left[\hat{m}_h(x)\right] - m(x)\right)^2 - \left(\mathrm{E}\left[\hat{m}^*_{hg}(x)\middle|D_n\right] - \hat{m}_h(x)\right)^2 = o\left((nF_x(h))^{-1}\right) \text{ a.s.}$$

This finishes the proof. □

4.4 APPLICATION

The intention of this section is to analyse the practical aspects of the proposal bandwidth selection. For all experiments we choose $B = 100$ bootstrap replications, as larger numbers of bootstrap replication do not improve the results. The set of bandwidths is defined as a set of the k nearest neighbours of the function of interest $x \in E$, $h = h(x) \in H = \{h_1, \ldots, h_k\}$, where h_i is the distance to the i^{th} neighbour of x with respect to the semi-metric d and k is chosen depending on the size of the data set. For kernel function we use the asymmetrical quadratic kernel function, namely

$$K(u) = \frac{3}{2}(1-u^2)1_{[0,1]}(u).$$

As semi-metric we choose the L_2–norm of the q^{th} derivatives of the curves,

$$\left(d_q^{deriv}(x_1, x_2)\right)^2 = \int_a^b (x_1^{(q)}(t) - x_2^{(q)}(t))^2 dt, \tag{4.8}$$

where $q \in \mathbb{N}$ is chosen depending on the data set. Procedure 1 provides the structure of our implementation. For calculating the residuals we used in all test sets

wild bootstrapping following continuous auxiliary distribution, as suggested in [49],

$$V_i := (1 - 2 \cdot 20^{-\frac{2}{3}})^{\frac{1}{2}} W_i + 20^{-\frac{1}{3}}(W_i^2 - 1),$$

where the W_i are standard normally distributed random variables.

Procedure 1 Calculate $h^*(x) = \arg\min \frac{1}{B} \sum_{b=1}^{B} (\hat{m}^*_{hg}(x) - \hat{m}_g(x))^2$

```
Calculate pilot kernel estimate m̂_g(x)
MSEmin ← 0; l* ← 1
for l = 1 to |H| do
    MSE ← 0
    for i = 1 to B do
        Calculate residuals (ε*_k)_{k=1}^n
        Calculate new response variable Y*_k = m̂_g(X_k) + ε*_k for k = 1,...,n
        Calculate new kernel estimate m̂*_{h_l,g}(x) based on this set and bandwidth h_l
        MSE ← MSE +(m̂_g(x) − m̂*_{h_l,g}(x))^2
    end for
    if l == 1 then
        MSEmin ← MSE
    else
        if MSE < MSEmin then
            l* ← l
        end if
    end if
end for
h*(x) ← h_{l*}
```

Simulated Data

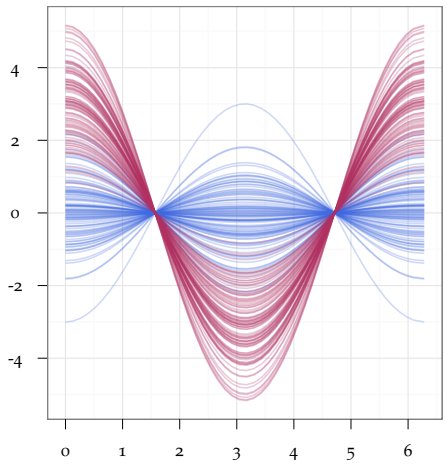
In this subsection the bootstrap method is compared to global cross-validation that is introduced for non-parametric functional regression by Rachdi and Vieu [58]. In the following we speak of homogenous functional data, if the small ball probability $P((d(x, X) \leqslant \varepsilon)$ behaves similar in ε for all $x \in E$. We speak of heterogenous data, if the small ball probability changes its behaviour, e.g. from exponential to fractal-type on the function space E.

The Data Set of Burba et al. [11]

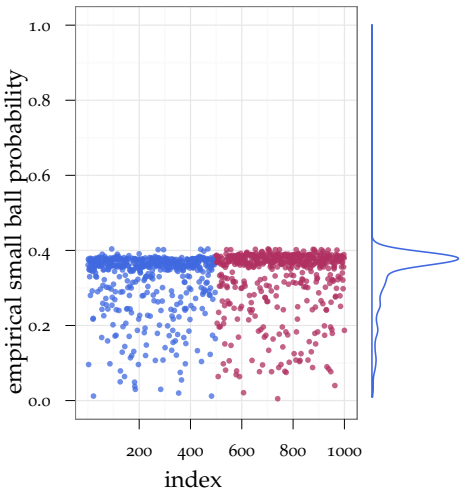
We simulated $n = 1000$ pairs of data (X_i, Y_i) with

$$X_i(t) = a_i \cos(t), \tag{4.9}$$

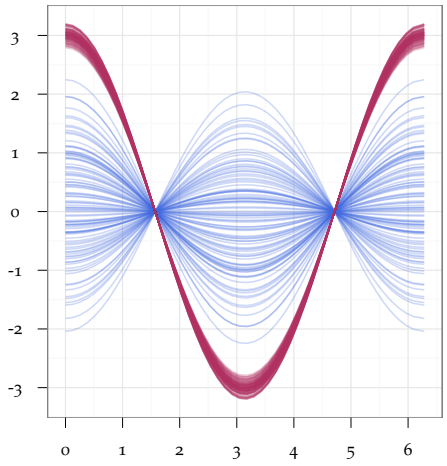
where t is a vector of equidistant points of the interval $[0, 2\pi]$. For $i = 1, \ldots, 500$ a_i is distributed as $N(0, 1)$ and for $i = 501, \ldots, 1000$ a_i is distributed as $N(3, \sigma^2)$, where $\sigma \in \{0.1, 1\}$.



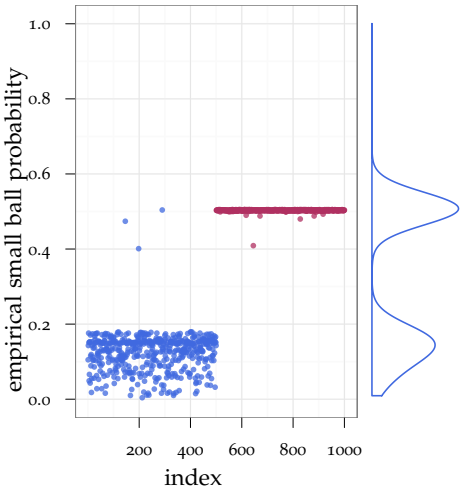
(a) A sample of 100 curves of the homogenous data set, where we have $\sigma = 1$.



(b) $\hat{F}_{X_i}(h_{opt})$ for the homogenous data set, where we have $\sigma = 1$.



(c) A sample of 100 curves of the heterogenous data set, where we have $\sigma = 0.1$.



(d) $\hat{F}_{X_i}(h_{opt})$ for the heterogenous data set, where we have $\sigma = 0.1$.

Figure 4.1: A sample of curves generated by (4.9) and values of the small ball probability $\hat{F}_{X_i}(h_{opt})$. The blue curves / dots belong to the set, where $a_i \sim N(0,1)$ and the red ones to $a_i \sim N(3,\sigma)$

We examine the regression model

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

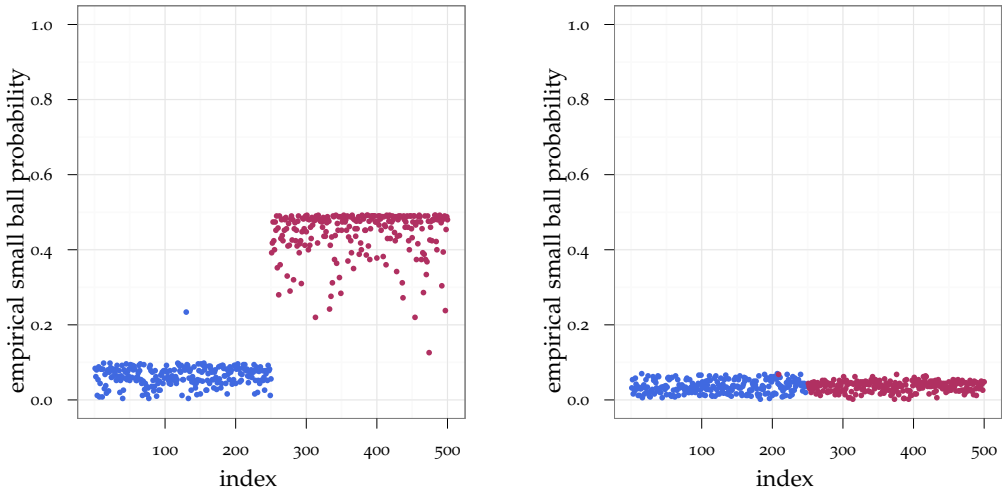
where ε_i are distributed as $N(0, 0.05)$ and the regression function is chosen as

$$m(X_i) = a_i^2 \text{ for } i = 1, \ldots, n.$$

As Burba et al. introduced in [11], the empirical small ball probability

$$\hat{F}_x(h) = \frac{1}{n}\sum_{k=1}^{n} 1_{B(x,h)}(X_k) \tag{4.10}$$

can be used to illustrate the difference between homogenous and heterogenous data. As in the plot of the empirical small ball probability $\hat{F}_{X_i}(h_{opt})$, the heterogeneity and the homogeneity of the data can be visualised, whereas these characteristica may be difficult to see in the plot of the curves. As for example in the above introduced data set we have in the case of $\sigma = 1$ homogenous data and in the case of $\sigma = 0.1$ heterogenous data. In Fig. 4.1 we plotted a sample of the curves and the associated small ball probability with a density estimate of the image of distribution of the empirical small ball probability.



(a) $\hat{F}_{X_i}(h_{opt})$, where h_{opt} is chosen by global cross-validation.

(b) $\hat{F}_{X_i}(h_{opt}(X_i))$, where $h_{opt}(X_k)$ is chosen for each X_k in the test set by bootstrapping.

Figure 4.2: The empirical small ball probability resulting from global and local bandwidth selection methods, respectively, for the same test set.

To illustrate heterogenous data sets, first an optimal bandwidth h_{opt} is chosen by a global method, here global cross-validation, then the empirical small ball probability is calculated for each data element X_i with h_{opt} and we get $\left(\hat{F}_{X_i}(h_{opt})\right)_{i=1}^{n}$. If the data is homogenous, then we have for each ball $B(X_i, h_{opt})$ a similar amount of data points therein. Firstly, we splitted the data in a learning and testing set. The learning set is used for calculating the optimal global bandwidth by cross-validation. In Fig. 4.2 we plotted then the empirical small ball probability for the test set. In Fig. 4.2 (a) we used the global bandwidth received by cross-validation. In Fig. 4.2 (b) we calculated for each element in the test set a local optimal bandwidth by bootstrapping. As can be seen in the figure, we have in the case of local bandwidth selection for all elements in the test set a similar number of neighbours. If one looks at Fig. 4.2 one may suspect that for homogenous data X_i the density estimate is unimodal and multi-modal for heterogenous data. Based on this pre-

sumption it would be probably possible to decide automatically between global and local bandwidth selection procedure.

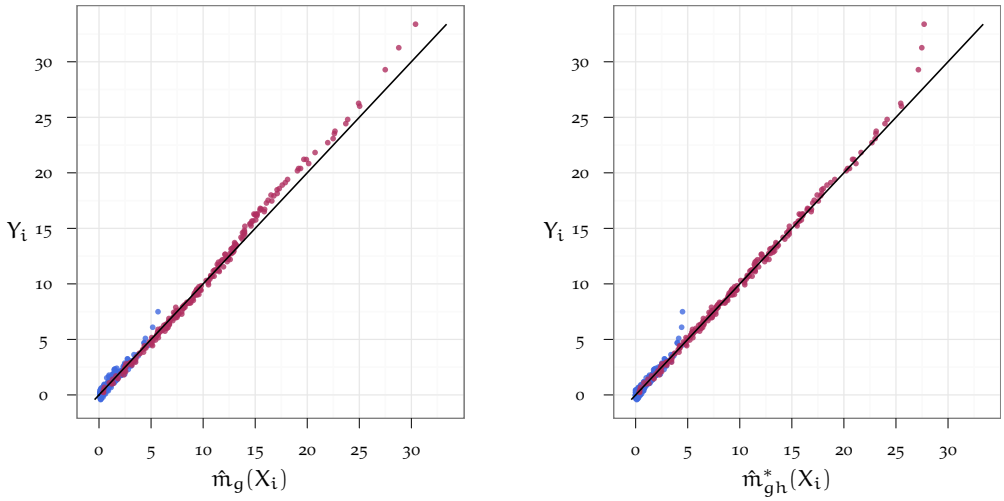
Comparing the Performance

In this section we compare our local bootstrap bandwidth selection method to the global cross-validation procedure. For pilot bandwidth estimation for our procedure we use the global cross-validation method, which is introduced for functional data analysis by Rachdi and Vieu [58]. We denote the pilot kernel estimate by $\hat{m}_g(x)$ and the bootstrap kernel estimate by $\hat{m}^*_{gh}(x)$.

For comparing the performance, we used 500 randomly chosen data pairs (X_i, Y_i) for learning and the remaining 500 pairs for testing. Afterwards, the performance of the prediction is evaluated by the empirical mean squared error (EMSE) on the test sample,

$$\frac{1}{500}\sum_{k}(Y_k - \hat{m}(X_k))^2\,,$$

where $\hat{m}(X_k)$ is one of the kernel estimates introduced above. We repeated that procedure 200 times, each time with randomly arranged learning and testing sets. The EMSE are plotted for each procedure into a box plot. Furthermore, we apply a two-sided Wilcoxon rank sum test on the EMSE values.



(a) Global cross-validation (b) Bootstrapping

Figure 4.3: The homogenous case $\sigma = 1$.

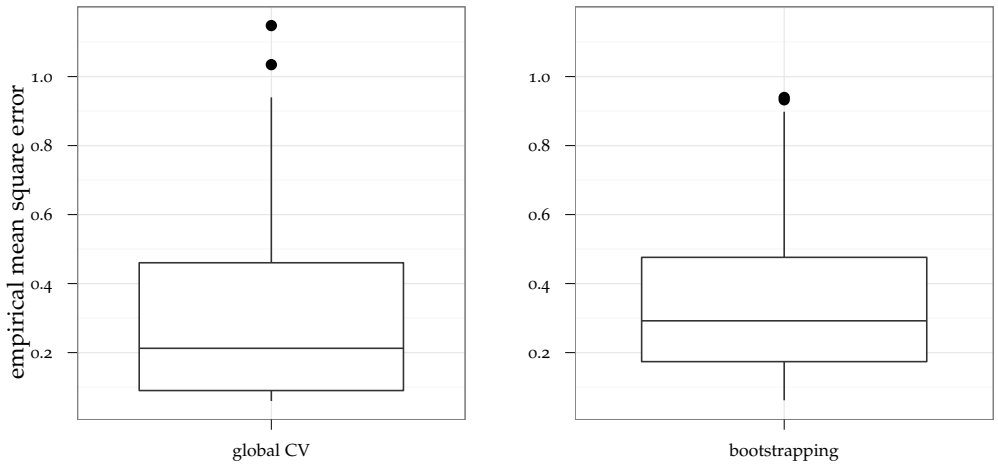
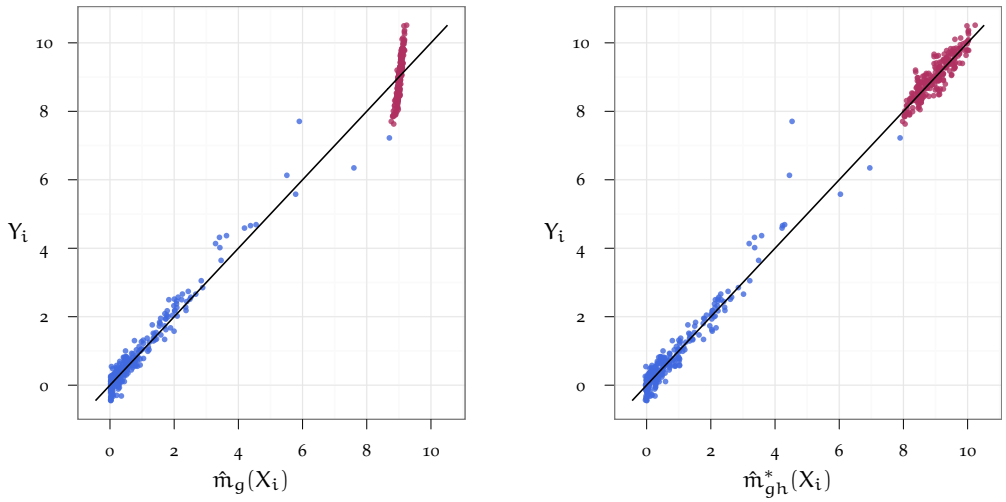


Figure 4.4: The homogenous case $\sigma = 1$. The p-value of the two-sided Wilcoxon rank sum test is $p = 0.006447$.

The result of the simulation for the homogenous case $\sigma = 1$ can be seen in Fig. 4.3. Fig. 4.3 (a) and Fig. 4.3 (b) show the prediction versus the true response. In Fig. 4.3 (a) we used global cross-validation and in Fig. 4.3 (b) the bootstrap procedure for the bandwidth selection for one sample. In this homogenous case the global cross-validation method delivers significant better predictions than the bootstrap procedure, see the p-value in the caption of Fig.4.4. We assume that the bootstrap procedure performs worse because of the outliers we got by the high and sparse values of the second group of the a_i.

Next, we examine the heterogenous case $\sigma = 0.1$.



(a) Global cross-validation

(b) Bootstrapping

Figure 4.5: The heterogenous case $\sigma = 0.1$.

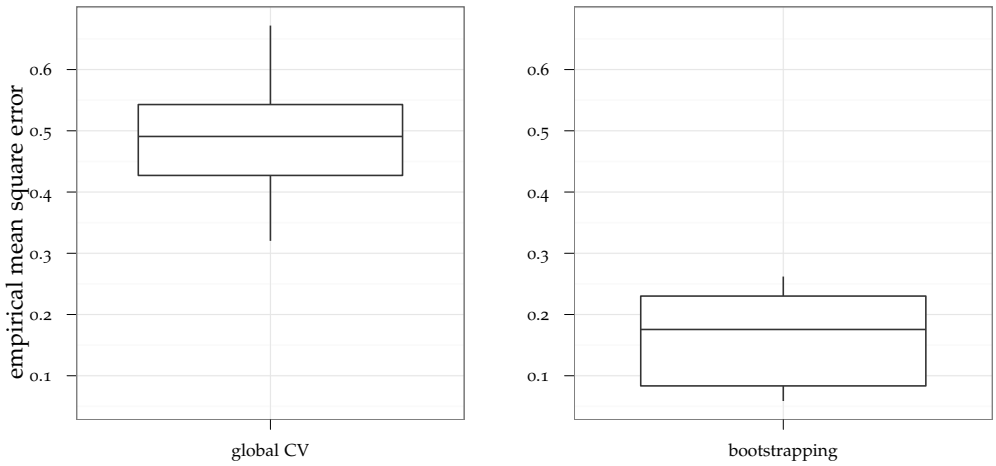


Figure 4.6: Boxplot of the mean squared error over 200 runs. The p-value of the two-sided Wilcoxon rank sum test is $p < 1e-4$.

As presumed Fig. 4.6 shows that in the case of heterogenous data we have a significant improvement of the prediction by the bootstrap procedure.

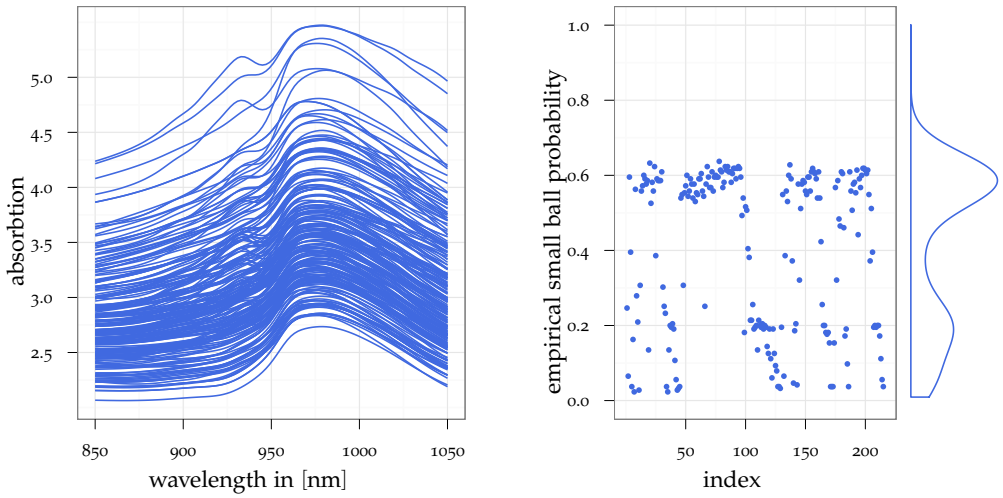
Real World Data

In this section we examine the bootstrap procedure for choosing the bandwidth on three real world data sets. All data sets are spectrometric data sets. The task here is to predict some value of interest based on the near-infrared absorbance spectrum.

For all three data sets we used 50% of randomly chosen data for learning and the remaining 50% for testing. This was repeated 200 times. For the fat and the moisture data set we used the L_2–norm of the second derivatives as semi-metric and for the octane data set the L_2–norm of the first derivative, see (4.8).

In the data sets the curves display the near-infrared absorbance spectrum on pieces of meat, wheat, or oil. For each of these pieces we have some real-valued response variable Y that represents in the case of meat the percentage of fat, for the moisture data set the percentage of the moisture of the piece of wheat and for the oil data set it represents the octane number. The first data set is widely studied in functional data analysis, see [25], [27] or [58]. In the context of functional data analysis the moisture and the octane data set were examined by Reiss and Ogden [61]. All three data sets can be found in the R package *functional data sets* downloadable on CRAN, [57].

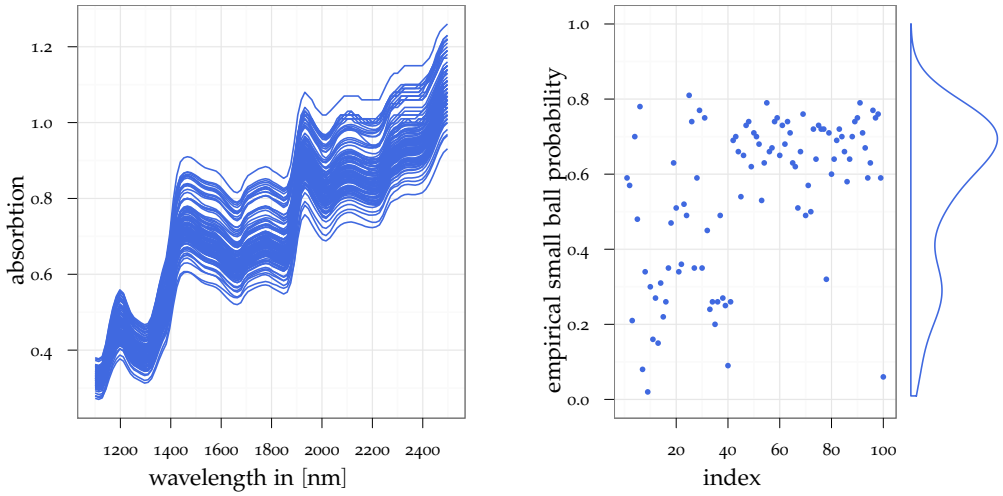
The spectral curves and the corresponding empirical small ball probability calculated by global cross-validation bandwidth can be seen for the fat data set in Fig. 4.7, for the moisture data set in Fig. 4.8, and for the octane data set in Fig. 4.9.



(a) The spectrum of the fat data set.

(b) The empirical small ball probability with bandwidth from global cross-validation.

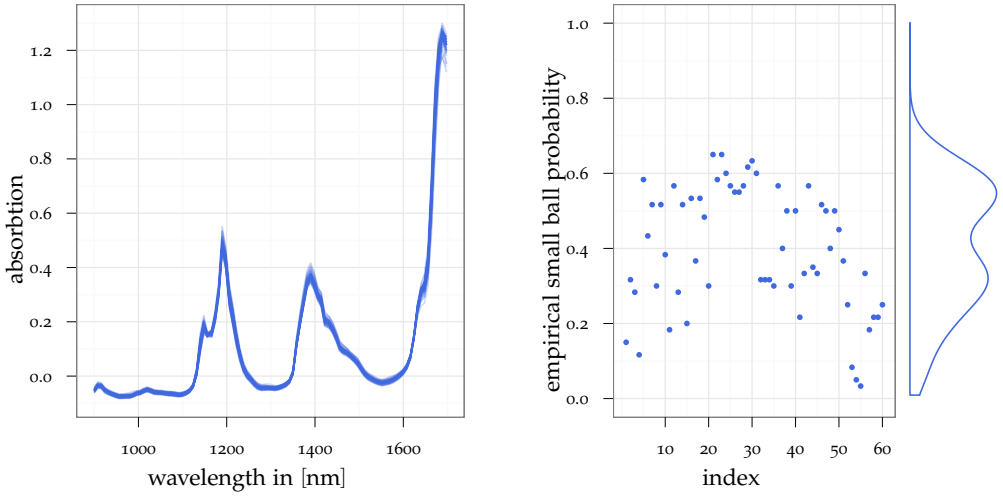
Figure 4.7: The near-infrared absorbance spectrum of the fat data set and the corresponding empirical small ball probability.



(a) The spectrum of the moisture data set.

(b) The empirical small ball probability with bandwidth from global cross-validation.

Figure 4.8: The near-infrared absorbance spectrum of the moisture data set and the corresponding empirical small ball probability.

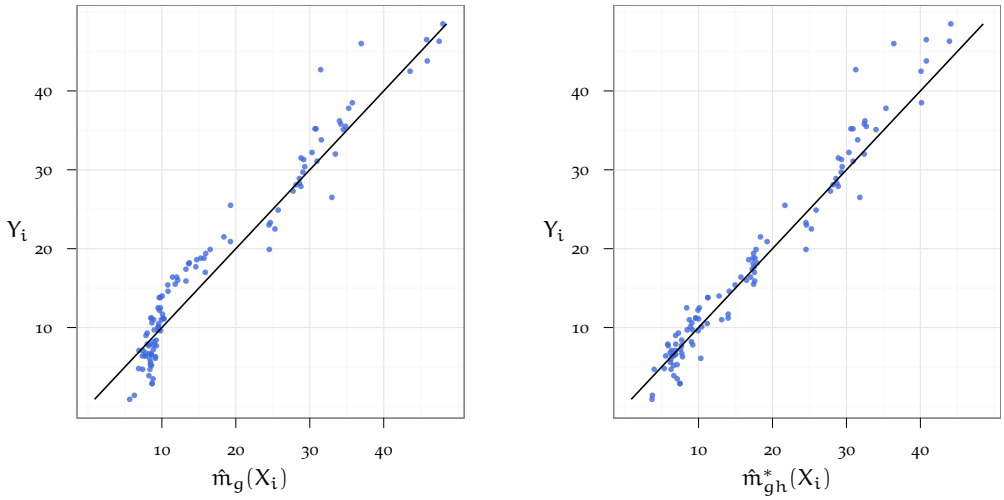


(a) The spectrum of the octane data set.

(b) The empirical small ball probability with bandwidth from global cross-validation.

Figure 4.9: The near-infrared absorbance spectrum of the octane data set and the corresponding empirical small ball probability.

By the distribution of the empirical small ball probability in subplot (b) of Fig. 4.7, Fig. 4.8, and Fig. 4.9 it can be expected that a kernel estimate with local bandwidth selection outperforms a kernel estimate with global bandwidth selection. It seems that the fat data set is more heterogenous than the moisture data set that is more heterogenous than the octane data set. But for the last two sets we have just a small sample size.



(a) Global cross-validation

(b) Bootstrapping

Figure 4.10: Prediction of the percentage of fat versus the true values.

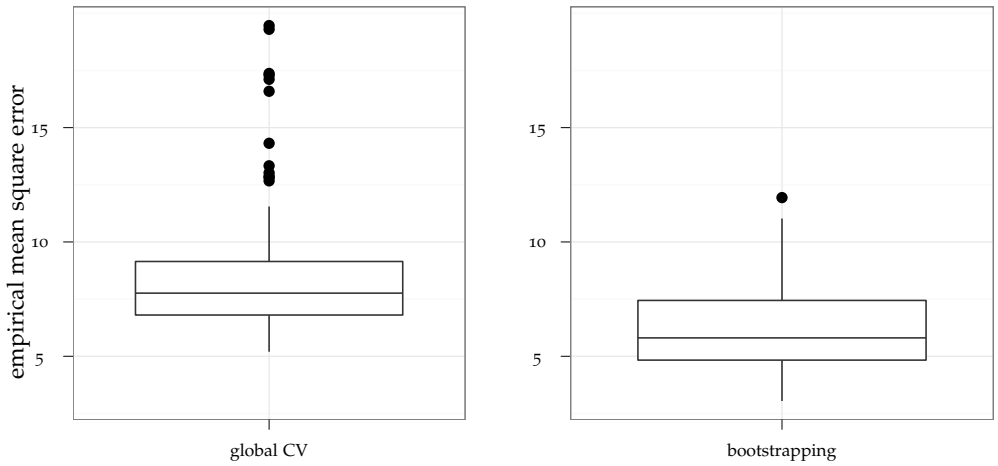
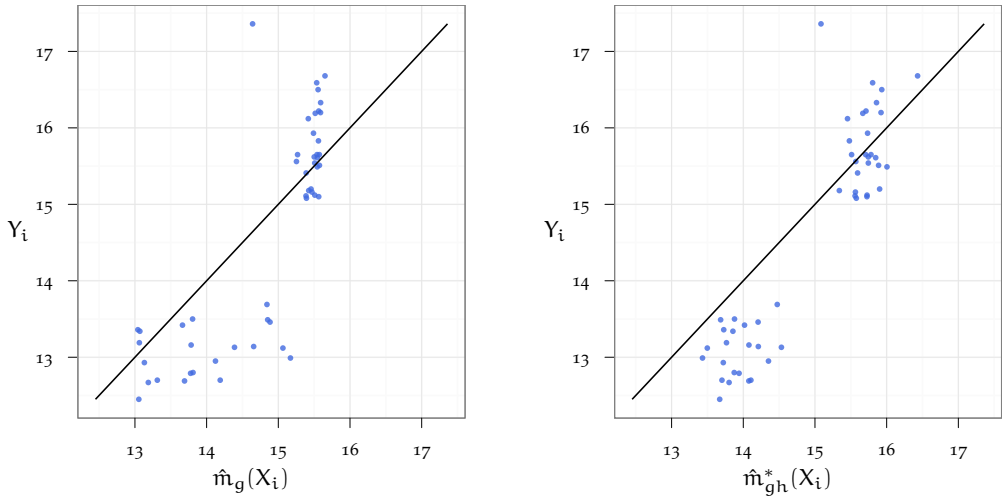


Figure 4.11: Comparison of the performance of global and local bandwidth selection procedure for the fat data set. The p-value of the two-sided Wilcoxon rank sum test is $p < 1e-4$.



(a) Global cross-validation

(b) Bootstrapping

Figure 4.12: Prediction of the percentage of moisture versus the true values.

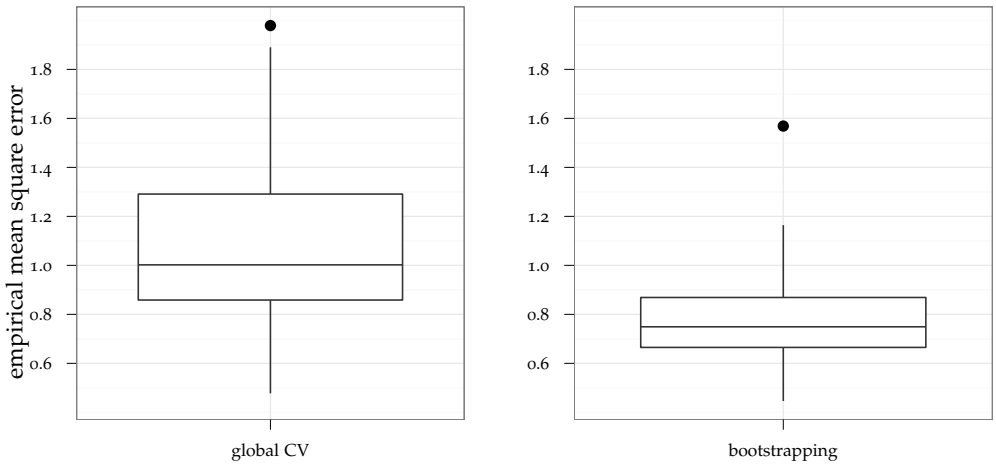
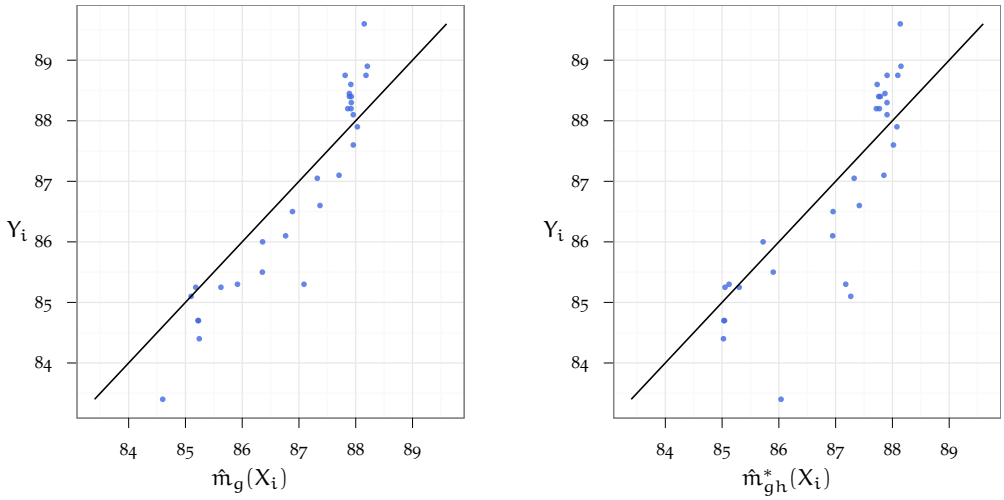


Figure 4.13: Comparison of the performance of global and local bandwidth selection procedure for the moisture data set. The p-value of the two-sided Wilcoxon rank sum test is: $p \leqslant 1e-4$.



(a) Global cross-validation

(b) Bootstrapping

Figure 4.14: Prediction of the octane number versus the true values.

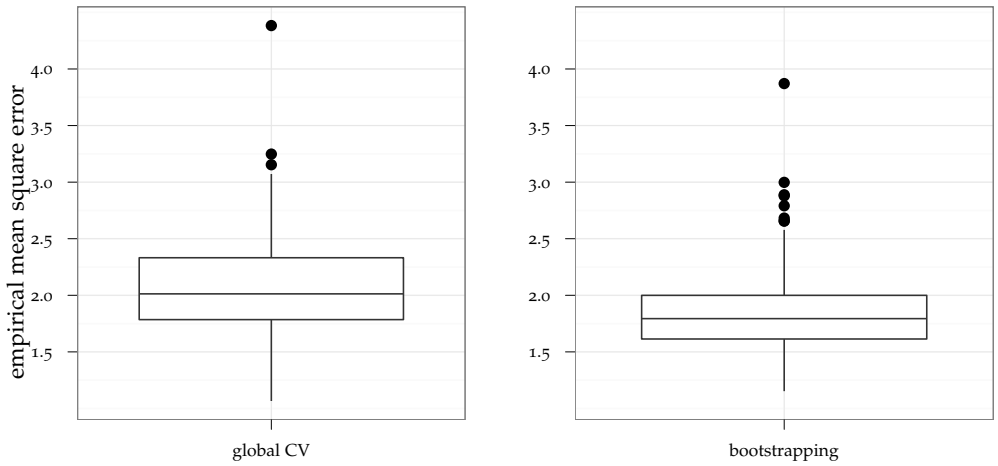


Figure 4.15: Comparison of the performance of global and local bandwidth selection procedure for the octane data set. The p-value of the two-sided Wilcoxon rank sum test is $p < 1e-04$.

In all three data sets the bootstrap procedure for bandwidth selection performs significantly better than the global cross-validation method for bandwidth selection, see the box plots in Fig. 4.11, Fig. 4.13, and Fig. 4.15.

Both, the simulation data sets and the real world data sets suggest that the distribution of the empirical small ball probability with bandwidth obtained by global cross-validation can be used for the decision of using a local or a global bandwidth selection method. For measuring the homogeneity of the data X_i one can use the measure of homogeneity introduced by Ferraty and Vieu [30]. Alternatively, this can be decided by a Hartigans' Dip test on unimodality [41] of the density estimate of the empirical small ball probability. However, we do not pursue this idea in this work.

An apparent disadvantage of our method is that it needs more computation time than the cross-validation method, as we need it for calculating the pilot kernel estimate $\hat{m}_g(x)$. However, with just little additional effort it is possible to build confidence intervals based on the bootstrapped data (X_i, Y_i^*). Furthermore, we compared in our simulations the bootstrap procedure with the k-nearest neighbour kernel estimate using global cross-validation as bandwidth selection method. We got for the k-NN kernel estimate a similar precision of the prediction as for our bandwidth selection method, so the advantage of better prediction disappears here.

LIST OF FIGURES

NOTATION AND SYMBOLS

MISCELLANEOUS

$a := b$	a is defined by b
$C, C_1, \cdots, C_6$	unspecified generic constants
n	sample size
(X_i)	short form for the sequence of f. r. v. $(X_i)_{i \geqslant 1}$
(u_i)	short form for the sequence of real numbers $(u_i)_{i \geqslant 1}$
h, h_n, g, g_n	generic notation for bandwidths
k, k_n	generic notation of the number of neighbours

SETS

$\mathbb{R}$, $\mathbb{R}^+$	set of real numbers, set of positive numbers including 0
(a, b), $[a, b]$	open or closed intervals in $\mathbb{R}$
$\mathbb{N}$	set of positive integers
$\mathbb{Z}$	set of positive and negative integers including null
$S_{\mathbb{R}}$	generic compact set of $\mathbb{R}$

FUNCTIONS

$1_A(x)$	indicator function, $1_A(x) = 1$, if $x \in A$, else $1_A(x) = 0$
$[\cdot]$	Gaussian bracket, $[x] = \max\{y \in \mathbb{Z} \mid z \leqslant x\}$, $x \in \mathbb{R}$
$f^{-1}(y)$	generalised inverse, $f^{-1}(y) = \inf\{x \mid f(x) \geqslant y\}$
max, min	maximum, minimum
sup, inf	supremum, infimum
log	natural logarithm

SPACES

E	generic functional space
(E, d)	generic semi-metric space
S_E	generic compact set of E
$C(E)$	space of continuous functions $f : E \to \mathbb{R}$
$L^\beta(E)$	space of Hölder continuous functions $f : E \to \mathbb{R}$ with parameter $\beta > 0$

NORMS AND SEMI-METRIC

$d(\cdot,\cdot)$	semi-metric
$\|\cdot\|_\infty$	supremum norm

OTHER SYMBOLS RELATED TO SEMI-METRIC SPACES

$B(x,h)$	open ball with centre x and radius h in semi-metric in E
$N(S,d,\epsilon)$	covering number
$K_S(\varepsilon)$	logarithm of $N(S,d,\epsilon)$, Kolmogorov's ε-entropy
$\varepsilon_n(S)$	entropy number
e_n	dyadic entropy number

MEASURES, PROBABILITY DISTRIBUTIONS, AND DISTRIBUTION FUNCTIONS

$(\Omega,\mathcal{A})$	generic measurable space with σ-algebra $\mathcal{A}$
$(\Omega,\mathcal{A},P)$	probability space with distribution P
P	probability distribution
$P(\cdot\|X=x)$	conditional distribution, shortly $P(\cdot\|x)$
$\mathcal{B}(\mathbb{R})$	Borel σ-algebra on $\mathbb{R}$
$\mathcal{E}_d$	σ-algebra generated by the topology of E w. r. t. d
μ, ν	unspecified measures
$\mu \ll \nu$	ν is absolutely continuous w. r. t. μ
$supp(\mu)$	support of the measure μ
$F_x(h)$	measure of $B(x,h)$ w. r. t. P

RANDOM VARIABLES AND RELATED QUANTITIES

X, X_i	generic f. r. v., explanatory variable
Y, Y_i	generic r. r. v., response variable
$H_{n,k}$	positive r. r. v., k-NN bandwidth
$E[Y\|X=x]$	conditional expectation of a r.r.v. Y given the f.r.v. X
$E[Y]$	expectation of a r. r. v. Y w.r.t. P
$Var[Y]$	variance of a r. r. v. Y
$Cov(Y_i,Y_j)$	covariance of two r. r. v. Y_i and Y_j

ESTIMATORS

$m(\cdot)$	nonlinear regression function
$\hat{m}(\cdot)$, $\hat{m}_h(\cdot)$	estimate of $m(\cdot)$
$\hat{m}_{k-NN}(\cdot)$	k-NN kernel estimate of $m(\cdot)$
$F^x(y)$	conditional distribution of some r. r. v. Y given the f.r.v. X
$\hat{F}^x(y)$	estimate of $F^x(y)$
$f^x(y)$	conditional density of some r. r. v. Y given the f.r.v. X
$\hat{f}^x(y)$	estimate of $f^x(y)$

KERNEL FUNCTIONS

K	generic notation of a asymmetrical kernel function
K_0	generic notation of a symmetrical kernel function
H	generic notation of a integrated kernel function

OTHER SYMBOLS RELATED TO ESTIMATION AND CONVERGENCE

$\mathcal{O}(\cdot)$	Landau symbol
$o(\cdot)$	Landau symbol
$\mathcal{O}_{a.co.}(\cdot)$	rate of almost complete convergence

LIST OF ABBREVIATIONS

i. i. d.	independent and identically distributed
w. r. t.	with respect to
r. r. v.	real random variable ($\mathbb{R}$-valued)
f. r. v.	functional random variable (E-valued)
a. s.	almost surely
k-NN	k nearest neighbour
et seq.	and the following
et seqq.	
RKHS	Reproducing Kernel Hilbert Space

BIBLIOGRAPHY

[1] Attouch, M., Laksaci, A., and Saïd, E. O. Asymptotic normality of a robust estimator of the regression function for functional time series data. *Journal of the Korean Statistical Society 39*, 4 (2010), 489–500.

[2] Azzedine, N., Laksaci, A., and Ould-Saïd, E. On robust nonparametric regression estimation for a functional regressor. *Statistics & Probability Letters 78*, 18 (2008), 3216–3221.

[3] Benhenni, K., Ferraty, F., Rachdi, M., and Vieu, P. Local smoothing regression with functional data. *Computational Statistics 22*, 3 (2007), 353–369.

[4] Bogachev, V. I. *Gaussian measures*. Math surveys and monographs, 62, American Mathematical Society, 1999.

[5] Bosq, D. *Nonparametric statistics for stochastic processes: estimation and prediction*. Lecture notes in statistics. Springer, 1998.

[6] Bosq, D. *Linear Processes in Function Spaces*. Springer, Berlin, 2000.

[7] Bradley, R. C. Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions. *Probability Surveys 2* (2005), 107–144.

[8] Bradley, R. C. *Introduction to Strong Mixing Conditions, Volume 1*. Kendrick Press, 2007.

[9] Bradley, R. C. *Introduction to Strong Mixing Conditions, Volume 2*. Kendrick Press, 2007.

[10] Bradley, R. C. *Introduction to Strong Mixing Conditions, Volume 3*. Kendrick Press, 2007.

[11] Burba, F., Ferraty, F., and Vieu, P. k-Nearest Neighbour method in functional nonparametric regression. *Journal of Nonparametric Statistics 21*, 4 (2009), 453–469.

[12] Carl, B., and Stephani, I. *Entropy, compactness and the approximation of operators*. Cambridge University Press, 1990.

[13] Cérou, F., and Guyader, A. Nearest neighbor classification in infinite dimension. *ESAIM: P&S 10* (2005), 340–255.

[14] Collomb, G. Estimation de la regression par la methode des k points les plus proches avec noyau : quelques propriétés de convergence ponctuelle. In *Statistique non Paramétrique Asymptotique*, J.-P. Raoult, Ed., vol. 821 of *Lecture Notes in Mathematics*. Springer Berlin / Heidelberg, 1980, pp. 159–175.

[15] Cover, T., and Hart, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory 13*, 1 (1967), 21–27.

[16] Craig, C. C. On the Tchebychef inequality of Bernstein. *Annals of Mathematical Statistics 4* (1933), 94–102.

[17] Crambes, C., Delsol, L., and Laksaci, A. Robust nonparametric estimation for functional data. *Journal of Nonparametric Statistics 20*, 7 (Oct. 2008), 573–598.

[18] Dabo-Niang, S., and Rhomari, N. Estimation non paramétrique de la régression avec variable explicative dans un espace métrique kernel regression estimation when the regressor takes values in metric space. *Comptes Rendus Mathematique 336*, 1 (2003), 75–80.

[19] Davidson, J., Monticini, A., and Peel, D. Implementing the wild bootstrap using a two-point distribution. *Economics Letters 96*, 3 (2007), 309–315.

[20] Delsol, L. *Regression sur variable fonctionelle: Estimation, Tests de structure et Application*. PhD thesis, Universite Toulouse III - Paul Sabatier, 2008.

[21] Devroye, L., Györfi, L., and Lugosi, G. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

[22] Ferraty, F., Goia, A., and Vieu, P. Functional nonparametric model for time series: a fractal approach for dimension reduction. *TEST 11*, 2 (2002), 317–344.

[23] Ferraty, F., Laksaci, A., Tadj, A., and Vieu, P. Rate of uniform consistency for nonparametric estimates with functional variables. *Journal of Statistical Planning and Inference 140*, 2 (2010), 335–352.

[24] Ferraty, F., Laksaci, A., and Vieu, P. Estimating Some Characteristics of the Conditional Distribution in Nonparametric Functional Models. *Statistical Inference for Stochastic Processes 9*, 1 (May 2006), 47–76.

[25] Ferraty, F., Mas, A., and Vieu, P. Nonparametric Regression on Functional Data: inference and Practical Aspects. *Australian & New Zealand Journal of Statistics 49*, 3 (2007).

[26] Ferraty, F., Rabhi, A., and Vieu, P. Conditional Quantiles for Dependent Functional Data with Application to the Climatic El Niño Phenomenon. *Sankhya: The Indian Journal of Statistics 67* (2005), 378–398.

[27] Ferraty, F., Van Keilegom, I., and Vieu, P. On the Validity of the Bootstrap in Non-Parametric Functional Regression. *Scandinavian Journal of Statistics 37* (2010), 286–306.

[28] Ferraty, F., and Vieu, P. Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés. *Comptes rendus de l'Académie des sciences. Série 1, Mathématique 330* (2000), 139–142.

[29] Ferraty, F., and Vieu, P. Non-parametric models for functional data, with application in regression, time-series prediction and curve discrimination. *Journal of Nonparametric Statistics 16*, 1 & 2 (2004), 111–125.

[30] Ferraty, F., and Vieu, P. *Nonparametric Functional Data Analysis*. Springer, New York, 2006.

[31] Ferraty, F., and Vieu, P. Erratum of: 'Non-parametric models for functional data, with application in regression, time-series prediction and curve discrimination'. *Journal of Nonparametric Statistics 20*, 2 (2008), 187–189.

[32] Ferraty, F., and Vieu, P. Additive prediction and boosting for functional data. *Computational Statistics & Data Analysis 53*, 4 (2009), 1400–1413.

[33] Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. *A Distribution-Free Theory of Nonparametric Regression.* Springer, New-york, 2002.

[34] Hall, P. Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *Journal of Multivariate Analysis 32*, 2 (1990), 177–203.

[35] Hall, P., Lahiri, S. N., and Polzehl, J. On Bandwidth Choice in Nonparametric Regression with Both Short- and Long-Range Dependent Errors. *The Annals of Statistics 23*, 6 (1995).

[36] Härdle, W. *Applied nonparametric regression*. Econometric Society monographs. Cambridge University Press, 1992.

[37] Härdle, W., and Bowman, A. Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands. *American Statistical Association 83*, 401 (1988).

[38] Härdle, W., and Marron, J. S. Bootstrap simultaneous error bars for nonparametric regression. *Annals of Statistics 19*, 2 (1991), 778–796.

[39] Härdle, W., and Vieu, P. Kernel regression smoothing of time series. *Journal of Time Series Analysis 13*, 3 (1992), 209–232.

[40] Hart, J., and Vieu, P. Data-driven bandwidth choice for density estimation based on dependent data. *Annals of Statistics 18*, 2 (1990), 873–890.

[41] Hartigan, J. A., and Hartigan, P. M. The Dip Test of Unimodality. *The Annals of Statistics 13*, 1 (1985), 70–84.

[42] Kim, T. Y. Asymptotically optimal bandwidth selection rules for the kernel density estimator with dependent observations. *Journal of Statistical Planning and Inference 59*, 2 (1997), 321–336.

[43] Kolmogorov, A. N., and Tihomirov, V. M. Epsilon-entropy and epsilon-capacity of sets in function spaces. *Uspehi Matematicheskikh Nauk 14*, 2 (1959), 3–86.

[44] Kuelbs, J., and Li, W. V. Metric Entropy and the Small Ball Problem for Gaussian Measures. *Journal of Functional Analysis 116*, 1 (1993), 133–157.

[45] Laloë, T. A k-nearest neighbor approach for functional regression. *Statistics & Probability Letters 78*, 10 (2008), 1189–1193.

[46] Li, W. V. Approximation, Metric Entropy and Small Ball Estimates for Gaussian Measures.

[47] Lian, H. Convergence of functional k-nearest neighbor regression estimate with functional responses. *Electronic Journal of Statistics 5* (2011), 31–40.

[48] Lu, Z., and Cheng, P. Strong consistency of nearest neighbor kernel regression estimation for stationary dependent samples. *Science in China Series A: Mathematics 41*, 9 (1998), 918–926.

[49] Mammen, E. Resampling Methods for Nonparametric Regression. In *Smoothing and Regression: Approaches, Computation, and Application*, M. G. Schimek, Ed. John Wiley, 2000.

[50] Manteiga, W. G., Miranda, M. D. M., and González, A. P. The choice of smoothing parameter in nonparametric regression through Wild Bootstrap. *Computational Statistics & Data Analysis 47*, 3 (2004), 487–515.

[51] Merlevède, F., Peligrad, M., and Rio, E. A Bernstein type inequality and moderate deviations for weakly dependent sequences. In *High Dimensional Probability V: The Luminy Volume*, C. Houdré, V. Koltchinskii, D. M. Mason, and M. Peligrad, Eds. Beachwood, Ohio, USA: Institute of Mathematical Statistics, 2009, pp. 273–292.

[52] Modha, D. S., and Masry, E. Minimum Complexity Regression Estimation with Weakly Dependent Observations. *IEEE Transactions on Information Theory 42*, 6 (1996), 2133–2145.

[53] Nadaraya, E. A. On estimating regression. *Theory of Probability and its Applications 9*, 1 (1964), 141–142.

[54] Oliveira, P. E. Nonparametric density and regression estimation functional data. Tech. rep., Departamento de Matemática, Universidade de Coimbra, 2005.

[55] Ouyang, D., Li, D., and Li, Q. Cross-validation and non-parametric k nearest-neighbour estimation. *Econometrics Journal 9*, 3 (2006), 448–471.

[56] Pesin, B. On Rigorous Mathematical Definitions of Correlation Dimension and Generalized Spectrum for Dimensions. *Journal Statistical Physics 71*, 3-4 (1993), 529–547.

[57] R-Project. http://www.r-project.org/.

[58] Rachdi, M., and Vieu, P. Nonparametric regression for functional data: automatic smoothing parameter selection. *Journal of Statistical Planning and Inference 137*, 9 (2007), 2784–2801.

[59] Ramsay, J. O., and Silverman, B. W. *Functional Data Analysis*. Springer, New York, 1997.

[60] Ramsay, J. O., and Silverman, B. W. *Applied Functional Data Analysis: Methods and Case Studies*. Springer, New York, 2002.

[61] Reiss, Philip, T., Ogden, and Todd, R. Functional Principal Component Regression and Functional Partial Least Squares. *Journal of the American Statistical Association 102*, 479 (Sept. 2007), 984–996.

[62] Rio, E. *Théorie asymptotique des processus aléatoires faiblement dépendants*, vol. 31 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer, Berlin, 2000.

[63] Rosenblatt, M. A central Limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences of the United States of America 42*, 1 (1956), 43–47.

[64] Shao, T., and Tu, D. *The Jackknife and Bootstrap*. Springer, 1995.

[65] Steinwart, I., and Christmann, A. *Support Vector Machines*. Springer, 2008.

[66] Stone, C. J. Optimal global rates of convergence for nonparametric regression. *Annals of Statistics 10*, 4 (1982), 1040–1053.

[67] Tran, L. T. Nonparametric Estimation for Time Series by Local Average Estimators. *The Annals of Statistics 42*, 2 (1993), 1040–1057.

[68] van der Vaart, A., and van Zanten, H. Bayesian inference with rescaled Gaussian process priors. *Electronic Journal of Statistics 1* (2007), 433–448.

[69] van der Vaart, A. W., and van Zanten, J. H. Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics 36*, 3 (June 2008), 1435–1463.

[70] Watson, G. S. Smooth regression analysis. *Sankhya: The Indian Journal of Statistics, Series A 26*, 4 (1964), 359–372.