Oxford Handbooks Online

Kernel Regression Estimation for Functional Data

Frédéric Ferraty and Philippe Vieu

The Oxford Handbook of Functional Data Analysis
Edited by Frédéric Ferraty and Yves Romain

Print Publication Date: Nov 2010 Subject: Physical Sciences, Mathematics
Online Publication Date: Aug 2018 DOI: 10.1093/oxfordhb/9780199568444.013.4

Abstract and Keywords

This article provides an overview of recent nonparametric and semiparametric advances in kernel regression estimation for functional data. In particular, it considers the various statistical techniques based on kernel smoothing ideas that have recently been developed for functional regression estimation problems. The article first examines nonparametric functional regression modelling before discussing three popular functional regression estimates constructed by means of kernel ideas, namely: the Nadaraya-Watson convolution kernel estimate, the kNN functional estimate, and the local linear functional estimate. Uniform asymptotic results are then presented. The article proceeds by reviewing kernel methods in semiparametric functional regression such as single functional index regression and partial linear functional regression. It also looks at the use of kernels for additive functional regression and concludes by assessing the impact of kernel methods on practical real-data analysis involving functional (curves) datasets.

Keywords: functional regression estimation, functional data, kernel smoothing, nonparametric functional regression, Nadaraya-Watson convolution kernel estimate, kNN functional estimate, local linear functional estimate, semiparametric functional regression, single functional index regression, additive functional regression

4.1 Introduction

OVER the last decade there has been great interest in developing new models for regression problems involving functional data. The main aim of these models is to overcome the lack of flexibility of the standard linear functional modeling ideas, and recent research has naturally been oriented towards nonparametric (and more recently semiparametric) models. The general presentation done in Chapter 1, provides a detailed discussion of all these various ways of modeling functional regression problems. While

Chapter 2 focuses on linear modeling, this chapter is concerned with recent nonparametric and semiparametric advances.

In the standard literature on nonparametric regression, convolution kernels play a major role (see Collomb (1981, 1985); Härdle (1990); Sarda and Vieu (2000); Schimek (2000); and Györfi *et al*. (2002) for bibliographical surveys and general discussions of standard multivariate nonparameric regression). They are many reasons for this. From a practical point of view, these estimates are quite appealing (p. 73) (and now widely used) because they are very easy to implement. This feature combines with various attractive theoretical properties, such as for instance the fact that they can reach Stone's optimal rates of convergence for multivariate nonparametric regression problems (see Stone (1982)). Finally, another nice feature of kernel ideas is the fact that they provide several different classes of estimators. In the nonparametric setting, this includes the standard Nadaraya–Watson estimator, but also the k-nearest neighbour (kNN) and local polynomial estimators. But this holds not only for nonparametric modeling, since kernel ideas can be used in most semiparameric models or in dimensionality-reduction models.

Therefore amongst the first developments linking nonparametric regression modeling and functional data analysis, kernel ideas have taken a determining role. The main goal of this chapter is to present the various statistical techniques based on kernel smoothing ideas that have recently been developed for functional regression estimation problems. It is worth noting that even if this field of statistics is rather young, since the first paper in this area is only ten years old (see Ferraty and Vieu (2000)), the literature is now sufficiently developed to make an exhaustive presentation of the results impossible. So our presentation of these methods is carried out using a necessary selective set of results. This selection has been motivated by the wish to make the presentation as simple as possible: that is why strong consistency type results have been preferentially chosen. Other kinds of results, such as L_2 consistency or asymptotic normality, are presented in a more synthetic way. The second guideline followed during the selection of material has been to emphasize the local specificities of functional data modeling: this is why special attention has been given to pointwise consistency results.

The first part of this contribution (Section 4.2) concerns nonparametric regression modelling and is organized as follows. A few asymptotic pointwise results are presented in Sections 4.2.2 and 4.2.3 for three popular functional regression estimates constructed by means of kernel ideas, namely the functional version of the Nadaraya–Watson estimate, a functional adaptation of the local linear method, and the kNN functional estimate. In each case the rates of convergence are specified and the influence of the local structure of the functional space is highlighted. In a second approach, Section 4.2.4 discusses uniform consistency type results. In this setting the existing literature is much less developed, and to our knowledge the only available result is for Nadaraya–Watson type estimates. The emphasis here is on how such a uniform theoretical result turns out to be a key tool for solving further problems of great applied impact, such as automatic bandwidth selection or bootstrapping. Finally, all the results presented so far are

commented on in Section 4.2.5, which also presents a short survey of the state-of-the-art in nonparametric functional regression.

The second part of this chapter concerns recent advances in the area of alternative models which fall between parametric (linear) models and purely nonparametric ones. Special attention is paid to additive modeling (see Section 4.4), while (p. 74) semiparametric models are discussed in Section 4.3; here special attention is given to single functional index models (see Section 4.3.2) and to partial linear functional models (see Section 4.3.3). The main point all these models have in common is the need for a preliminary purely nonparametric functional stage, and it will be emphasized how the uniform consistency developments presented earlier in Section 4.2.4 are indispensable tools for their study. As a consequence, and because uniform consistency results are only known for Nadaraya–Watson type functional estimates, all the methods developed in Sections 4.4 and 4.3 will concern such kinds of estimates.

The large range of possible models for functional regression problems opens the door to a new kind of question: which model is to be used in practice? One way of looking at this problem is to construct goodness-of-fit testing procedures, and once again kernel methods turn out to be of great interest. This field has been very underdeveloped in the literature up to now; the few existing studies will be presented in Section 4.5. Once again, the uniform consistency developments described earlier in Section 4.2.4 are necessary tools, and consequently only Nadaraya–Watson type functional estimates will be considered.

It is worth noting that Sections 4.2–4.5 will offer the opportunity for presenting an exhaustive survey of the literature on kernel methods for functional regression. However, kernel ideas can be used in other settings than regression, and Section 4.6.1 will be devoted to a brief presentation of this literature. The main focus of this chapter is on theoretical considerations, but the final section will briefly discuss the impact of kernel methods on practical real-data analysis involving functional (curves) datasets.

4.2 Kernels in nonparametric functional regression

4.2.1 Models and estimates

Consider the following regression problem:

(4.1)

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

where Y and ε are real random variables with $E(\varepsilon|X) = 0$, and where X is a covariate which is allowed to belong to some abstract topological space E which can be of infinite dimension. In what follows, P_X denotes the probability measure of the random variable X. This is called a functional regression model with scalar response. For the sake of generality, we do not restrict ourselves to Hilbert or Banach spaces (p. 75) and the topology endowing the space E is only assumed to be associated with a semi-metric d. The only restriction on the operator r is a Lipschitz-type regularity hypothesis:

(4.2)

$$\exists \beta > 0, \exists C < +\infty, \forall (x, x') \in E \times E, |r(x) - (x')| \leq Cd(x, x')^{\beta}.$$

According to the general classification of regression models presented in Chapter 1 (see Definition 1.3 therein), this model is nonparametric in the sense that the general assumption (4.2) does not allow us to characterize the operator r by a finite number of elements of E. In other words, r is a nonlinear real-valued operator defined on the functional space E.

Given a sample of independent pairs (X_i, Y_i) distributed as (X, Y), the statistical problem consists in estimating the functional nonlinear operator r. In what follows, three different estimates of r will be considered: the Nadaraya–Watson functional estimate, the kNN functional estimate, and the local linear functional estimate, each of them being constructed as a locally-weighted average. More precisely, if x is an element of E , these estimates have the following form:

(4.3)

$$\hat{r}(x) = \frac{\sum_{i=1}^{n} W_n(x, X_i) Y_i}{\sum_{i=1}^{n} W_n(x, X_i)}.$$

4.2.2 Pointwise almost-sure consistency

Throughout this section, x is a fixed point of the functional space E; and strong pointwise consistency results (with rates) are presented for three different estimates constructed in the manner of (4.3). Proofs are presented in a very synthetic way, the main objective being to emphasize their main ideas. Precise references will be given for the reader interested in having all the technical details of these proofs. Comments on these results, as well as a complementary bibliography, are referred to Section 4.2.5.

General conditions

The weights $W_n(x, X_i)$ are constructed differently according to the type of estimate to be considered, but in each case they are based on some real-valued kernel function K satisfying the following set of standard conditions:

(4.4)

$$K \geq 0, \int_R K = 1, K \text{ is Lipschitz on} [0, 1), \text{ and support}(K) \subset [0, 1).$$

Basically, the weight $W_n(x, X_i)$ vanishes as soon as X_i is "too far" from x. This is controlled by means of a smoothing parameter $h = h_n$, called bandwidth, which determines the size of the neighbourhood around x (in the sense of the topology associated with d) outside of which the weights are zero. This means that the (p. 76) estimate is not (for a fixed point x) using the whole statistical sample, but only the pairs (X_i, Y_i) for which X_i falls inside of the ball

$$B(x, h) = \{x' \in E, d(x, x') \leq h\},$$

and naturally its behavior will depend strongly on the number of such pairs. From a mathematical point of view, these local concentration effects can be controlled by means of the following small-ball probability function:

$$\phi_x(\in) = p_X(B(x, \in)),$$

which is assumed to be such that $\phi_x(\epsilon) > 0$ for any $\epsilon > 0$. For each $x \in E$, this concentration function will be linked with the bandwidth h (depending on n) through the following set of standard assumptions:

(4.5)

$$\lim_{n \to \infty} h = 0 \text{ and } \lim_{n \to \infty} \frac{n\phi_x(h)}{\log n} = \infty.$$

There is also a need for a strong link between the kernel function K and the small-ball probability function ϕ_x. Here, this is stated through the following technical hypothesis. Assume that there exist three constants $(C_1, C_2, \in_0) \in]0, +\infty[^3$ such that:

(4.6)

$$\forall \in \leq \in_0, C_1\phi_x(\in) \leq EK\left(\frac{d(x, X)}{\in}\right) \leq C_2\phi_x(\in).$$

As stated in Ferraty and Vieu (2006, Chapter 4), condition (4.6) is not too restrictive in the sense that it is satisfied in most situations.

Finally, in a standard way, we need some condition to control the tails of the distribution of Y. This is achieved by using the following moment assumption:

(4.7)

$$\forall\, m \geq 2, E(|Y|^m | X = \cdot) \text{exists and is continuous.}$$

Kernel estimation

The first estimate to be considered is a functional version of the Nadaraya–Watson convolution kernel estimate. It can be defined as follows:

Definition 4.1 *The kernel functional regression estimate, denoted from now on by* $\hat{r}_{\text{ker}}$, *is defined by the local weighted average equation* (4.3) *with the following weights*:

$$W_n(x, X_i) = W_{n,\text{ker}}(x, X_i) = K\left(\frac{d(x, X_i)}{h}\right).$$

This estimate was introduced for nonparametric functional regression problems (see Ferraty and Vieu (2000)), and various asymptotic results are now available in the literature. In the following theorem we present a simple result giving the rate of almost sure (a.s.) convergence, while more insights into the existing literature are referred to Section 4.2.5.

(p. 77) Theorem 4.1 *Under conditions* (4.2), (4.4), (4.5), (4.6), and (4.7), *we have*

(4.8)

$$\hat{r}_{\text{ker}}(x) - r(x) = O\left(h^{\beta}\right) + O\left(\sqrt{\frac{\log n}{n\phi_x(h)}}\right), \text{ a.s.}$$

Proof of Theorem 4.1 A complete proof of this result can be found in Ferraty and Vieu (2006*a*, Chapter 6). Here, we simply present its main lines of argument. The key step in the proof consists in the following decomposition:

(4.9)

$$\hat{r}_{\text{ker}}(x) = \frac{\hat{r}_{\text{ker},0}(x)}{\hat{r}_{\text{ker},1}(x)},$$

where for $j = 0$ or 1:

$$\hat{r}_{\text{ker}}(x) = \frac{1}{nEW_{n,\text{ker}}(x, X_i)} \sum_{i=1}^{n} Y_i^j W_{n,\text{ker}}(x, X_i).$$

This decomposition has been adapted to take into account the fact that no assumption about the existence of density for the functional variable X is made, and it is therefore slightly different from the classical decomposition performed in standard multivariate kernel estimation (see, for instance, Collomb (1976, 1984) for the earliest results in this area, and Sarda and Vieu (2000) for a bibliographical survey). The bias terms can be computed directly by using condition (4.2) defining nonparametric modeling, and we get

(4.10)

$$E\hat{r}_{\mathrm{ker},0}(x) = 1 \text{ and } E\hat{r}_{\mathrm{ker},1}(x) = r(x) + O(h^{\beta}).$$

Both the other terms can be expressed as sums of independent and centered real random variables:

$$\Delta_{i,j} = Y_i^j \frac{W_{n,\mathrm{ker}}(x, X_i)}{EW_{n,\mathrm{ker}}(x, X_i)} - E\left(Y_i^j \frac{W_{n,\mathrm{ker}}(x, X_i)}{EW_{n,\mathrm{ker}}(x, X_i)}\right),\ i = 1, \ldots, n,\ j = 0, 1,$$

for which a Bernstein-type exponential inequality can be used. Conditions (4.6) and (4.7) allow us to bound the moments of these variables in such a way that:

(4.11)

$$\exists\, C > 0, \forall\, m \geq 2, \quad E\Delta_{i,j}^m \leq C\phi_x(h)^{1-m}.$$

Once these bounds are obtained, one can use any kind of exponential-type inequality (for instance, Ferraty and Vieu (2006*a*), Corollary A.8-ii) to obtain, for some $\eta_0 > 0$:

(4.12)

$$\sum_{n=1}^{\infty} p\left(\frac{1}{n}\left|\sum_{i=1}^{n}\Delta_{i,j}\right| \geq \eta_0\sqrt{\frac{\log n}{n\phi_x(h)}}\right) < +\infty,$$

(p. 78) Because of the Borel–Cantelli Lemma, this is sufficient to get that:

(4.13)

$$\hat{r}_{\mathrm{ker},j}(x) - E\hat{r}_{\mathrm{ker},j}(x) = O\left(\sqrt{\frac{\log n}{n\phi_x(h)}}\right), a.s.,\ j = 0, 1.$$

Finally, Theorem 4.1 follows directly from (4.9), (4.10), and (4.13). □

Local linear smoothing

A natural extension of the standard Nadaraya–Watson estimate consists in using local linear ideas. This leads to the following local linear regression estimate:

> Definition 4.2 *The local linear functional regression estimate, denoted from now on by* $\hat{r}_{LL}$, *is defined by the local weighted average equation* (4.3) *when the weights are constructed from some known operator* β *from* E^2 *into* $\mathbb{R}$ *by*:

$$\begin{aligned} W_n(x, X_i) &= W_{n,LL}(x, X_i) \\ &= \sum_{j\neq i}^{n} \beta(X_i, x)\left(\beta(X_i, x) - \beta(X_j, x)\right) K\left(\frac{d(x, X_i)}{h}\right) K\left(\frac{d(x, X_j)}{h}\right). \end{aligned}$$

This estimate was introduced for nonparametric functional regression problems in Barrientos-Marin (2007). Note that the interest of this estimate is that it can be seen as the explicit solution in a of the minimization problem:

$$\min_{(a,b)\in R^2}\sum_{i=1}^{n}\left(Y_i-a-b\beta\left(X_i,x\right)\right)^2K\left(\frac{d\left(x,X_i\right)}{h}\right).$$

In the literature there are fewer asymptotic results than for the estimate $\hat{r}_{\ker}$. In the following theorem we present its rate of pointwise almost sure convergence, while more insights into the existing literature are referred to Section 4.2.5. The study of the estimate $\hat{r}_{LL}$ needs additional technical assumptions linking the operator β, the semi-metric d, and the kernel function K. To improve the readability of this chapter, we summarize these conditions by assuming that there exist two constants $(M_1, M_2) \in]0, +\infty[^2$ such that:

(4.14)

$$\forall\left(k,l\right)\in\mathcal{N}^*\times\mathcal{N},\ E\left(\left|\beta\left(X,x\right)\right|^lK\left(\frac{d\left(x,X\right)}{h}\right)^k\right)\leq M_1h^l\phi_x\left(h\right)$$

and

(4.15)

$$E\left(\beta\left(X,x\right)^2K\left(\frac{d\left(x,X\right)}{h}\right)\right)\geq M_2h^2\phi_x\left(h\right).$$

It is not our purpose here to discuss these conditions, but their high degree of generality appears in Lemma A.1 of Barrientos-Marin *et al*. (2010).

(p. 79) Theorem 4.2 *Under conditions* (4.2), (4.4), (4.5), (4.6), (4.7), (4.14), and (4.15) *we have*:

(4.16)

$$\hat{r}_{LL}\left(x\right)-r\left(x\right)=O\left(h^\beta\right)+O\left(\sqrt{\frac{\log n}{n\phi_x\left(h\right)}}\right),\ a.s.$$

Proof of Theorem 4.2 A complete proof of this result can be found in Barrientos-Marin *et al*. (2009). While the technical details are rather heavy, the main ideas of the proof are quite simple and will be presented now. The proof follows the same general framework as for the kernel estimate $\hat{r}_{\ker}$ in Theorem 4.1, and is based on the following decomposition

(4.17)

$$\hat{r}_{LL}\left(x\right)=\frac{\hat{r}_{LL,0}\left(x\right)}{\hat{r}_{LL,1}\left(x\right)},$$

where for $j = 0$ or 1:

$$\hat{r}_{LL}(x) = \frac{1}{nEW_{n,LL}(x, X_i)} \sum_{i=1}^{n} Y_i^j W_{n,LL}(x, X_i).$$

The bias terms can be computed by using condition (4.2), without any more difficulty than for the kernel estimate $\hat{r}_{\text{ker}}$ in Theorem 4.1, and one arrives at:

(4.18)

$$E\hat{r}_{LL,1}(x) = r(x) E\hat{r}_{LL,0}(x) + O(h^{\beta}).$$

The treatment of the dispersion terms is more difficult since they cannot be written as simple sums of i.i.d. centered random variables. The idea here is to use the following decomposition, for $j = 0, 1$:

$$\hat{r}_{LL,j}(x) = A[B_1 B_2 - C_1 C_2],$$

where

$$A = \frac{n^2 h^2 \phi_x(h)^2}{EW_{n,LL}(x, X_1)},$$

$$B_1 \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i^j K\left(\frac{d(x,X_i)}{h}\right)}{\phi_x(h)}, B_2 = \frac{1}{n} \sum_{i=1}^{n} \frac{\beta(X_i, x)^2 K\left(\frac{d(x,X_i)}{h}\right)}{h^2 \phi_x(h)},$$

and

$$C_1 \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i^j \beta(X_i, x) K\left(\frac{d(x,X_i)}{h}\right)}{h \phi_x(h)}, C_2 = \frac{1}{n} \sum_{i=1}^{n} \frac{\beta(X_i, x) K\left(\frac{d(x,X_i)}{h}\right)}{h \phi_x(h)}.$$

In other words, and because A is bounded, we have

$$\hat{r}_{LL,j}(x) - E\hat{r}_{LL,j}(x) = O((B_1 B_2 - EB_1 B_2) - (C_1 C_2 - EC_1 C_2)).$$

(p. 80) Each of the terms B_1, B_2, C_1, and C_2 can be written as a sum of i.i.d. real random variables and can therefore be treated by standard tools. To fix our ideas, if we look at the term B_1 we can write

$$B_1 = \frac{1}{n} \sum_{i=1}^{n} W_i,$$

and we can bound the moments of the variables W_i, exactly as was done to obtain (4.11), simply by using conditions (4.7), (4.14), and (4.15). In this way, we arrive at

$$E|W_i^m| = O(\phi_x(h)^{1-m}).$$

Then, exactly as we obtained (4.13) for the estimate $\hat{r}_{\text{ker}}$, the use of a standard exponential inequality leads to:

$$B_1 - EB_1 = O\left(\sqrt{\frac{\log n}{n\phi_x(h)}}\right), a.s.$$

The same kind of expression can similarly be obtained for the terms B_2, C_1, and C_2, and finally

(4.19)

$$\hat{r}_{LL,j}(x) - E\hat{r}_{LL,j}(x) = O\left(\sqrt{\frac{\log n}{n\phi_x(h)}}\right), a.s.$$

Theorem 4.2 follows directly from (4.17)–(4.19).

kNN estimation

Another extension of the standard Nadaraya–Watson estimate consists in using locally adaptive bandwidth constructed by means of k-nearest neighbour (kNN) ideas. Such a kNN functional regression estimate can be defined as follows:

> Definition 4.3 *The k-nearest neighbour functional estimate, denoted from now on by* $\hat{r}_{\text{kNN}}$, *is defined by the local weighted average equation* (4.3) *with the following weights*:
>
> $$W_n(x, X_i) = W_{n,\text{kNN}}(x, X_i) = K\left(\frac{d(x, X_i)}{H_k(x)}\right),$$
>
> *the bandwidth* $H_k(x)$ *being defined from a sequence* $k = k_n$ *of integers by*:
>
> $$H_k(x) = \inf\{h > 0,\ \#\{i, X_i \in B(x, h)\} = k\}.$$

This estimate was introduced for nonparametric functional regression problems by Burba *et al*. (2008). Its main appealing feature when compared with the estimate $\hat{r}_{\text{ker}}$ is that it uses a local adaptive bandwidth (that is, a bandwidth depending on x) which is controlled by a discrete parameter $k \in \mathbb{N}$ (rather than by a continous one h). On the other hand, its main drawback will be the additional (p. 81) high level of technicality needed to study its theoretical behavior, because of the randomness of the new parameter $H_k(x)$ which depends on the whole functional sample X_i, $i = 1, \ldots, n$. In the literature there are not as many asymptotic results as for the estimate $\hat{r}_{\text{ker}}$. In the following theorem we present its rate of pointwise almost sure convergence; more details of the existing literature are referred to Section 4.2.5. It should be natural to express the rates of convergence as functions of the number k of neighbours. However, we have decided to follow another approach in order to unify the presentation of the results for all the various estimates

studied in this chapter. This is the reason why we make the following additional restriction:

(4.20)

$$\phi_x(\cdot) \text{ is continuous and strictly increasing around} 0,$$

which allows us to define precisely the value h such that:

(4.21)

$$\phi_x(h) = \frac{k}{n}.$$

Theorem 4.3 *Assume that conditions* (4.2), (4.4), (4.6), (4.7), and (4.20) *hold. Assume also that k is such that the bandwidth defined by* (4.21) *satisfies condition* (4.5). *Then we have*

(4.22)

$$\hat{r}_{\mathrm{kNN}}(x) - r(x) = O\left(h^{\beta}\right) + O\left(\sqrt{\frac{\log n}{n\phi_x(h)}}\right), \ a.s.$$

Proof of Theorem 4.3 A complete proof of this result can be found in Burba *et al*. (2009). While technical details are rather heavy, the main ideas of the proof are quite simple and will be presented now. The key point consists in using condition (4.20) to construct, for any nonnegative increasing sequence such that $\lim_{n\to\infty}\beta_n = 1$, two deterministic bandwidths h^- and h^+ in the following way:

$$\phi_x(h^-) = \sqrt{\beta_n}\frac{k}{n} \text{ and } \phi_x(h^+) = \sqrt{\beta_n^{-1}}\frac{k}{n}.$$

Then, by a Chernoff-type inequality, one has that the random bandwidth $H_k(x)$ satisfies

(4.23)

$$1_{\{h^- \le H_k(x) \le h^+\}} \to 1, \ a.s.$$

We note that both bandwidths h^- and h^+ satisfy condition (4.5). Therefore both kernel estimates $\hat{r}_{\mathrm{kNN}}$ constructed using these bandwidths are such that Theorem 4.1 holds. Finally, by a suitable functional adaptation of standard ideas previously developed in Collomb (1980) for real variables X, one can show that the result (4.23) is enough to insure that the kernel estimate using the random bandwidth $H_k(x)$ (that is, precisely, the kNN estimate $\hat{r}_{\mathrm{kNN}}$) satisfies the same asymptotic property as both the kernel estimates $\hat{r}_{\mathrm{kNN}}$ constructed using the bandwidths h^- and h^+. □

(p. 82) ## 4.2.3 Other pointwise asymptotic results

Presentation

In addition to the pointwise almost sure consistency results stated below, it is possible to obtain asymptotic results for functional estimates with respect to other modes of convergence. However, this can be more difficult, and the existing literature (as far as we know) is restricted only to the Nadaraya–Watson kernel estimate $\hat{r}_{\text{ker}}$. In what follows we will present a few additional pointwise results for this estimate, including an L_2 asymptotic expansion, general L_p consistency results, and asymptotic normality. Because the statements of these results need tedious conditions and notation, we have decided to present them in a very synthetic way just to give the main ideas, and we refer readers to the relevant literature for more insights. Recall that x is a fixed point of the functional space E.

L_2 expansion for the kernel estimate

When stating such results, one is not only interested in finding the rates of convergence but also in obtaining precise expressions for the constant terms involved in the leading terms of the asymptotic expansion. To make this possible, it is necessary to strengthen the nonparametric model. This is done by introducing the following real-valued function:

$$\forall s \in R,\ \xi(s) = E\left[\left(r(X) - r(x)\right) \middle| d(X, x) = s\right],$$

and by changing condition (4.2) into the following:

(4.24)

$$\xi(0) = 0 \text{ and } \xi'(0) \text{ exists.}$$

The conditions on the parameters of the estimate $\hat{r}_{\text{ker}}$ also have to be slightly modified into the following ones:

(4.25)

$$\forall s \in [0, 1),\ K'(s) \leq 0 \text{ and } K(1) > 0,$$

(4.26)

$$\lim_{n\to\infty} h = 0 \text{ and } \lim_{n\to\infty} n\phi_x(h) = \infty,$$

while the concentration function also has to be such that

(4.27)

$$\phi_x(0) = 0.$$

The leading terms of the asymptotic expansion of the pointwise L_2 error of the estimate $\hat{r}_{\text{ker}}$ can be expressed by means of the following function (which is assumed to exist):

$$\tau(s) = \lim_{\epsilon \to 0} \frac{\phi_x(\epsilon s)}{\phi_x(\epsilon)}.$$

(p. 83) We refer readers to Ferraty *et al*. (2007) for a discussion on the low restrictiveness of these assumptions (and, more specifically, for the links between (4.24) and the differentiability properties of the nonlinear operator *r*, as well as for a precise expression of the function τ in various specific situations).

Theorem 4.4 *Assume that the conditions of Theorem* 4.1 *are satisfied as well as* (4.25), (4.26), and (4.27). *Then we have*:

(4.28)

$$E\left[\hat{r}_{\mathrm{ker}}(x) - r(x)\right]^2 = B^2 h^2 + V \frac{1}{n\phi_x(h)} + o\left(h^2 + \frac{1}{n\phi_x(h)}\right),$$

where

$$B = \xi(0) \frac{K(1) - \int_0^1 (sK(s))' \tau(s) ds}{K(1) - \int_0^1 K'(s) \tau(s) ds}$$

and

$$V = E\left[\varepsilon^2 | X = x\right] \frac{K^2(1) - \int_0^1 2K(s)K'(s)\tau(s) ds}{\left(K(1) - \int_0^1 (s)\tau(s) ds\right)^2}.$$

Proof of Theorem 4.4 This proof is omitted; it can be found in Ferraty *et al*. (2007), under conditions slightly weaker than those presented here. Let us just note that it is based on the decomposition (4.9) and on precise asymptotic expansions of the bias and the variance of both terms $\hat{r}_{\mathrm{ker},0}(x)$ and $\hat{r}_{\mathrm{ker},1}(x)$. □

Pointwise asymptotic normality for the kernel estimate

In the same spirit one can also specify the asymptotic distribution of the estimation error with the kernel estimate $\hat{r}_{\mathrm{ker}}$. This is achieved in the next result, where the following notation is used:

$$U_{n,x} = \sqrt{\frac{n\phi_x(h)}{V}} \left[\left(\hat{r}_{\mathrm{ker}}(x) - r(x)\right) - Bh\right].$$

Theorem 4.5 *Assume that the conditions of Theorem* 4.4 *are satisfied and that* $B \neq 0$. *Assume also that*

(4.29)

$$\lim_{n\to\infty} h\sqrt{n\phi_x(h)} < \infty.$$

Then we have:

(4.30)

$$U_{n,x} \xrightarrow{L} \mathrm{N}(0, 1).$$

Proof of Theorem 4.5 This result was obtained by Ferraty *et al*. (2007), under conditions slightly weaker than those presented here. Its proof is rather short and it combines the asymptotic expansions of the bias and variance obtained (p. 84) before (see Theorem 4.4), together with a standard central limit theorem for an array of i.i.d. centered real random variables. Note that, in the above-mentioned paper, condition (4.29) was omitted but it is obviously necessary in order that the term $o(h)$ appearing in the bias part of (4.28) does not disrupt the asymptotic normality. □

Note that, because of Slutsky's Theorem, (4.30) remains true if we change the quantities $\phi(h)$, B, and V by some consistent estimators $\hat{\phi}(h)$, $\hat{B}$ and $\hat{V}$.

$\mathbf{L_p}$ expansion for the kernel estimate

The final result on pointwise convergence that will be presented concerns the L_p errors of the estimate $\hat{r}_{\mathrm{ker}}$.

Theorem 4.6 *Under the conditions of Theorem* 4.5*, we have for any* $p \in \mathbb{N}$:

(4.31)

$$E U_{n,x}^p \to E W^p,$$

where W *is a standard* N(0, 1) *random variable.*

Proof of Theorem 4.6 This result was obtained by Delsol (2007), under conditions slightly weaker than those presented here (including non-integer values for p). The main difficulty of the proof consists in checking that the random variable $U_{n,x}^p$ is uniformly integrable. Once this is done, the convergence in distribution of $U_{n,x}$ towards W (stated before in Theorem 4.5) directly implies that the result (4.31) holds. □

Note that, subject to suitable additional condition allowing us to neglect the bias term, one directly obtains asymptotic expansions of L_p errors. This result is presented in the following corollary, which was also stated by Delsol (2007) under slightly weaker assumptions.

Corollary 4.1 *Under the conditions of Theorem* 4.5*, and if the bandwidth satisfies*

$$\lim_{n\to\infty} h\sqrt{n\phi_x(h)} = 0,$$

then we have, for any $p \in \mathbb{N}$,

(4.32)

$$E\left[\left(\hat{r}_{\ker}(x) - E\hat{r}_{\ker}(x)\right)^p\right] = \left(\frac{V}{n\phi_x(h)}\right)^{\frac{p}{2}} EW^p + o\left(\left(\frac{1}{n\phi_x(h)}\right)^{\frac{p}{2}}\right),$$

where W *is a* $\mathrm{N}(0, 1)$ *random variable and* V *is defined as in Theorem* 4.4.

Note that the same kinds of results were also obtained in Delsol (2007) for the uncentered L_p errors

$$E\left[\left(\hat{r}_{\ker}(x) - r(x)\right)^p\right],$$

(p. 85) but we have decided not to present them here because the leading terms of their asymptotic expansions are rather complicated and would require the introduction of long and tedious notation.

4.2.4 Uniform asymptotic results

Presentation

The aim of this section is to obtain uniform almost sure consistency (with rates) over some subset S of the functional space E and then to discuss how these theoretical results have great practical impact because of their possible applications to further advances in automatic bandwidth choice, bootstrapping, as well as on various other topics.

It turns out that the gap between pointwise and uniform results is not so easy to bridge in a functional setting as it is for standard multivariate nonparametric statistics. The first consequence of this difficulty is the lack of a wide variety of results, and it turns out that uniform results are only known (as far as we are aware) in a functional setting for the Nadaraya–Watson estimate $\hat{r}_{\ker}$ (while extensions to other estimates such as $\hat{r}_{\mathrm{LL}}$ or $\hat{r}_{\mathrm{kNN}}$ should be possible, but are still open questions). This is the reason why, from now on, we concentrate entirely on the kernel estimate $\hat{r}_{\ker}$. Another direct consequence of the difficulty of passing from pointwise to uniform results is the deteroriation of the rates of convergence, which will be controlled by means of Kolmogorov entropy considerations.

Entropy: a tool for controlling uniform rates of convergence

While it is well known that in multivariate nonparametric statistics the rates of convergence remain the same for pointwise as for uniform consistency (at least when the set S is compact), this is not always the case when functional variables are involved (see the final section below for more comments on this subject). Once again, the topological structure on the semi-metric space (E, d) will play a prominent role, as well of course as the complexity of the specific subset S on which uniform results are stated. A natural tool

that enables this to be clearly seen consists in linking the rates of convergence of the estimate $\hat{r}_{\text{ker}}$ with the Kolmogorov entropy (see Kolmogorov and Tikhomirov (1959)) of the set S, whose definition is recalled now.

Definition 4.4 *Let* $\epsilon > 0$ *be fixed. Let* $N_\epsilon(S)$ *be the minimal number of open balls in* E *of radius* ϵ *needed to cover* S*. Then the* ϵ-entropy *of the set* S *is defined by*

(4.33)

$$\psi_S(\epsilon) = \log(N_\epsilon(S)).$$

In the following section, the rates of uniform consistency of the kernel estimate $\hat{r}_{\text{ker}}$ over the set S will be presented. For that, we need a uniform version of the small-ball (p. 86) probability function. We recall that $\phi_x = P_X(\mathcal{B}(x, \epsilon))$ and assume that there exists a function ϕ_S and three constants α_1, α_2, and α_3 such that, for any $\epsilon > 0$:

(4.34)

$$0 < \alpha_1 \phi_S(\epsilon) \leq \inf_{x \in S} \phi_x(\epsilon) \leq \sup_{x \in S} \phi_x(\epsilon) \leq \alpha_2 \phi_S(\epsilon) < \infty \text{ and } \phi'_S(\epsilon) \leq \alpha_3.$$

One also needs the following technical conditions on the entropy ψ_S and on the concentration function ϕ_S:

(4.35)

$$\exists\, n_0, \forall\, n > n_0, \frac{(\log n)^2}{n\phi_S(h)} < \psi_S\left(\frac{\log n}{n}\right) < \frac{n\phi_S(h)}{\log n},$$

and

(4.36)

$$\exists\, \delta > 0, \sum_{i=1}^{n} e^{-\delta \psi_S\left(\frac{\log n}{n}\right)} < \infty.$$

It is not our purpose here to discuss these assumptions. Let us just note that they are not that restrictive, in the sense that they are satisfied in most cases of statistical interest (detailed comments can be found in Ferraty *et al*. (2009)).

Rates of uniform consistency of the kernel regression estimate

The following theorem states the rate of uniform consistency of the Nadaraya–Watson functional kernel estimate $\hat{r}_{\text{ker}}$ defined before. This result allows us to estimate the regression operator r at a random point (see Corollary 4.2), and this will be a key tool for further developments (see the following three sections).

Theorem 4.7 *Assume that conditions* (4.2), (4.4), (4.7), (4.34), (4.35), and (4.36) *hold. Also assume that* (4.5) and (4.6) *hold for any* $x \in S$*. Then we we have*

(4.37)

$$\sup_{x\in S}|\hat{r}_{\ker}(x)-r(x)|=O\left(h^{\beta}\right)+O\left(\sqrt{\frac{\psi_S\left(\frac{\log n}{n}\right)}{n\phi_S(h)}}\right),\ a.s.$$

Proof of Theorem 4.7 A complete proof of this result can be found in Ferraty *et al*. (2009). Here we just present its main lines of argument. As for the proof of the pointwise result (see Theorem 4.1 above), it is based on the decomposition (4.9). For the treatment of the bias terms there is no additional problem linked with uniformity, and from condition (4.2) we obtain that

(4.38)

$$\sup_{x\in S}|E\hat{r}_{\ker,0}(x)-1|=0 \text{ and } \sup_{x\in S}|E\hat{r}_{\ker,1}(x)-r(x)|=O\left(h^{\beta}\right).$$

The techninal difficulty linked with uniformity appears when computing the dispersion terms. To make the presentation simpler, let us introduce the notation $\epsilon_n = \log n/n$, $N = N_{\epsilon n}(S)$, and for $k = 1, \ldots, N$ let us denote by x_k the centers of the N balls of radius ϵ_n which are needed to cover the set S (see Definition 4.4). (p. 87) Moreover, for each $x \in S$, we denote by $x_{k(x)}$ the center of the ball which contains

x. The idea is to use, for $j = 0, 1$, the following decomposition:

(4.39)

$$\sup_{x\in S}|\hat{r}_{\ker,j}(x)-E\hat{r}_{\ker,j}(x)|\le A+B+C,$$

with

$$\begin{aligned}A &= \sup_{x\in S}|\hat{r}_{\ker,j}(x)-\hat{r}_{\ker,j}(x_{k(x)})|,\\ B &= \sup_{x\in S}|\hat{r}_{\ker,j}(x_{k(x)})-E\hat{r}_{\ker,j}(x_{k(x)})|,\\ C &= \sup_{x\in S}|E\hat{r}_{\ker,j}(x_{k(x)})-E\hat{r}_{\ker,j}(x)|.\end{aligned}$$

The term A is dealt with by using the Lipschitz property of the kernel function K, which leads to

$$A\le\frac{C}{n}\sum_{i=1}^{n}\frac{\epsilon_n Y_i^j}{h\phi_S(h)}1_{X_i\in B(x,h)\cup B(x_{k(x)},h)},$$

and one can use any kind of exponential-type inequality (for instance Corollary A. 8-ii in Ferraty and Vieu (2006)) to get:

$$A = O\left(\sqrt{\frac{\epsilon_n \log n}{nh\phi_S(h)}}\right), \text{ a.s.}$$

The same thing can be done for the term C, and finally the definition of ϵ_n and condition (4.35) allow us to get

(4.40)

$$A + C = O\left(\sqrt{\frac{\psi_S(\log n)}{n\phi_S(h)}}\right), \text{ a.s.}$$

Regarding the term B, we can write, for some $\eta > 0$:

$$P\left(B > \eta\sqrt{\frac{\psi_S(\epsilon_n)}{n\phi_S(h)}}\right) \leq N\left(\max_k P\left(\left|\hat{r}_{\ker,j}(x_k) - E\hat{r}_{\ker,j}(x_k)\right| > \eta\sqrt{\frac{\psi_S(\epsilon_n)}{n\phi_S(h)}}\right)\right).$$

By proceeding in the same way as for (4.12) and by using the Borel–Cantelli Lemma, we obtain that:

(4.41)

$$B = O\left(\sqrt{\frac{\psi_S\left(\frac{\log n}{n}\right)}{n\phi_S(h)}}\right), \text{ a.s.}$$

Finally, Theorem 4.7 follows directly from (4.9), (4.38), (4.39), (4.40), and (4.41). □

The next result is an obvious consequence of Theorem 4.7 that will be used repeatedly later.

(p. 88) Corollary 4.2 *Assume that the conditions of Theorem* 4.7 *are satisfied. Then, for any random variable Z taking values on the functional space E, we have*:

(4.42)

$$\left|\hat{r}_{\ker}(Z) - r(Z)\right| 1_{Z \in S} = O\left(h^{\beta}\right) + O\left(\sqrt{\frac{\psi_S\left(\frac{\log n}{n}\right)}{n\phi_S(h)}}\right), \text{ a.s.}$$

Application to bandwidth choice

In this section we will present some automatic data-driven bandwidth selection procedures based on cross-validation ideas, for which an asymptotic optimality result can be stated. From a technical point of view, uniform consistency results play a major role in the statement of such optimality results and that is why the literature on bandwidth

selection (at least in the infinite-dimensional setting) is restricted to only the kernel estimate $\hat{r}_{\text{ker}}$.

Again let x be a fixed functional element in E. The bandwidth-choice problem consists in finding some data-driven bandwidth $\hat{h}_x = \hat{h}_x(X_1, \ldots, X_n)$ that minimizes some error of estimation, such as for instance the L_2 error:

$$L_{2,x}(h) = E\left(\hat{r}_{\text{ker}}(x) - r(x)\right)^2.$$

Taking their inspiration from ideas developed in standard multivariate problems (see Härdle and Marron (1985) or Vieu (1991)), the following functional version of a local cross-validation criterion has recently been proposed by Benhenni *et al*. (2007):

(4.43)

$$CV_x(h) = \frac{1}{n}\sum_{j=1}^{n}\left(Y_j - \hat{r}_{\text{ker}}^{-j}(X_j)\right)^2 f_{n,x}(X_j),$$

and the bandwidth is defined by minimizing this data-driven criterion over some set $H = H_n$ of possible ones:

(4.44)

$$\hat{h}_{x,CV} = \arg\min_{h \in H} CV_x(h).$$

In this definition, $\hat{r}_{\text{ker}}^{-j}$ is the leave-one-out kernel estimate constructed by deleting the jth pair (X_j, Y_j):

(4.45)

$$\forall\, y \in E,\ r_{\text{ker}}^{-j}(y) = \frac{\sum_{i \neq j}^{n} W_{n,\text{ker}}(y, X_i) Y_i}{\sum_{i \neq j}^{n} W_{n,\text{ker}}(y, X_i)},$$

and $f_{n,x}$ is a real-valued positive weight function concentrated around the functional element x of interest. To fix our ideas, in what follows we will simply consider the following natural choice for $f_{n,x}$:

(4.46)

$$f_{n,x}(y) = 1_{y \in B(x,g)},$$

(p. 89) the parameter g being a sequence of positive real numbers that tend to zero more slowly than the smoothing parameter h to be selected. Of course, more general forms of weighting are possible (including, for instance, smooth versions of $f_{n,x}$).

The following conditons are necessary in order to obtain the optimality property of the cross-validated bandwidth $\hat{h}_{x,CV}$. Assume that there is some subset S such that $\mathscr{B}(x, g) \subset S \subset E$ for which the following conditions are satisfied:

(4.47)

$$\exists \gamma_1 > 0, \ \# H = O(n^{\gamma_1}),$$

(4.48)

$$\exists \gamma_2 \in (0, 1), \forall h \in H, g = Ch^{\gamma_2},$$

(4.49)

$$\exists \gamma_3 > 0, \forall h \in H, \phi_S(h) = O(n^{-\gamma_3}).$$

Theorem 4.8 *Assume that the conditions of Theorems* 4.4 *and* 4.7 *are satisfied for any* $h \in H$. *Assume in addition that* (4.46), (4.47), (4.48), and (4.49) *hold. Then we have*:

(4.50)

$$\lim_{n\to\infty} \frac{L_{2,x}(\hat{h}_{x,CV})}{\inf_{h\in H} L_{2,x}(h)} = 1, \ a.s.$$

Proof of Theorem 4.8 This result can be found in Benhenni *et al*. (2007) under slightly weaker assumptions. Here we just present the main ideas of the proof. The proof is based on the following decomposition:

(4.51)

$$CV_x(h) = A_1(h) + A_2(h) + A_3,$$

where

$$A_1(h) = \frac{1}{n}\sum_{j=1}^{n}\left(r_{\text{ker}}^{-j}(X_j) - r(X_j)\right)^2 f_{n,x}(X_j)$$

and

$$A_2(h) = -\frac{2}{n}\sum_{j=1}^{n}\left(\hat{r}_{\text{ker}}^{-j}(X_j) - r(X_j)\right)\left(\hat{r}_{\text{ker}}^{-j}(X_j) - Y_j\right) f_{n,x}(X_j).$$

The proof is broken down into two steps. First of all, taking inspiration from the general results stated by Marron and Härdle (1986), one can show the asymptotic equivalence between $A_1(h)$ and the quadratic error $L_{2,x}(h)$. Then, because the term A_3 does not depend on h, it suffices to show that $A_2(h)$ is of lower order than $L_{2,x}(h)$. This last point requires technical calculations, but the main ideas are to

use the asymptotic expansion for $L_{2,x}(h)$ given in Theorem 4.4 and then to show that $A_2(h)$ is of lower order than the leading terms in this expansion. This last point requires repeated use of Theorem 4.7 in order to control the terms $\left(\hat{r}_{\text{ker}}^{-j}(X_j) - r(X_j)\right)$ appearing in $A_2(h)$, and this is why the conditions assumed for Theorem 4.8 are basically those needed to obtain both Theorems 4.4 and 4.7. □

(p. 90) It is worth noting that the same kind of result can be obtained if the pointwise L_2 error is changed into an integrated one, subject to a suitable modification of the cross-validation criterion (based on the use of a weight function which no longer depends on n). This extension is not presented here (see Rachdi and Vieu (2007)) in order to avoid introducing additional long and tedious notation. Note, finally, that these results on cross-validation are of even more interest than in the case of finite-dimensional nonparametric statistics. This is because in the functional situation it is hard to figure out how alternative techniques could be developed. For instance, while plug-in techniques are a competitive alternative to bandwidth selection rules in multivariate settings, the high complexity of the constants appearing in the L_2 asymptotic expansions (see, for instance, Theorem 4.4) does not lead us to expect too much development of these ideas in the functional setting. Perhaps the bootstrap ideas to be developed in the next section could lead to an alternative bandwidth selection method, but this has still to be developed.

Applications to bootstrapping

Another important field for which the result of Theorem 4.5 has been of major interest is bootstrapping. Keeping in mind the results that exist in standard multivariate nonparametric regression (see Mammen (2000) for a discussion and extensive references), there are many ways of thinking about bootstrapping in regression problems. Extensions to functional variables are not very easy and have not been much developed; the literature is mainly devoted to residual-type bootstrapping. This is quite natural since the residuals that appear in the functional regression model (4.1) are real random variables, and therefore the gap between the multivariate and functional settings is narrower than it would be if one was resampling from the data (X_i, Y_i) themselves.

Because of the scarcity of literature in the relatively new field of functional bootstrapping, our goal is restricted to the kernel estimate $\hat{r}_{\text{ker}}$ and to a fixed functional element $x \in E$. The main goal of bootstrapping is to provide, by suitable random generation of new samples, an approximation of the distribution of the estimation error $\left(\hat{r}_{\text{ker}}(x) - r(x)\right)$ in order to overcome the problems linked with the use of the standard normal asymptotic approximation, which requires preliminary estimates of bias and variance. Given the highly complex structure one observes in the results (4.28) and (4.30), it goes without saying that bootstrapping can be expected to be even more interesting with functional data than it is in standard nonparametric statistics. It is not our purpose here to present a precise statement of the theoretical behavior of bootstrapping in functional problems, because it would require a long list of specific notation and conditions. We will

simply explain how the procedure is constructed; we refer to Ferraty *et al*. (2010) for a precise statement of its asymptotic validity.

(p. 91) Residual bootstraping uses a second bandwidth which is denoted by b; the corresponding kernel estimate is denoted by $\tilde{r}_{\ker}$. The procedure consists in repeating the following steps N_B times:

(i) Compute $\hat{\epsilon}_i = Y_i - \tilde{r}_{\ker}(X_i)$ for $i = 1, \ldots, n$.
(ii) Draw, for $i = 1, \ldots, n$, new residuals $\hat{\epsilon}_i$ and then draw new responses $\tilde{Y}_i = \tilde{r}_{\ker}(X_i) + \tilde{\epsilon}_i$ (this step will be explained in detail later at the end of this section).
(iii) Compute the kernel estimate as in Definition 4.1 but using the new sample $(X_i, \tilde{Y}_i), i = 1, \ldots, n$, (and denote by $\hat{r}^b_{\ker}$ this new estimate).

Then we approximate the distribution of the theoretical error $(\hat{r}_{\ker}(x) - r(x))$ using the empirical distribution (over the N_B replications) of the bootstrapped errors $(\hat{r}^b_{\ker}(x) - \tilde{r}_{\ker}(x))$. It is shown in Ferraty *et al*. (2010) that both distributions are asymptotically equivalent, subject to conditions controlling the respective asymptotic behaviors of h and b and subject to the way the bootstrapped residuals appearing in step (ii) are constructed. Without entering into the technical details of this proof, it is worth mentioning that step (i) of the procedure requires that the asymptotic behavior of the kernel estimate taken at each random element X_i is controlled and once again the uniformity result (and, more precisely, Corollary 4.2) plays a major role.

Let us mention that there are many ways to generate bootstrapped residuals $\tilde{\epsilon}_i$. The most popular ways for performing step (ii) are the following:

(ii*a*) The naive bootstrap, consisting in simply performing a random permutation of the original residuals $\hat{\epsilon}_i$ this approach is convenient for homoscedastic situations;
(ii*b*) The wild bootstrap, consisting in generating new variables U_i and defining $\tilde{\epsilon}_i = \hat{\epsilon}_i U_i$; this approach is convenient even in heteroscedastic situations as long as the U_i are chosen in such a way that $\tilde{\epsilon}_i$ and $\hat{\epsilon}_i$ have the same first moments (see Härdle and Marron (1991), for an example of a probability distribution available for the variables U_i).

In conclusion, let us mention that such a boostrapped approximation of the distribution error can be useful in various practical problems. Its use for confidence-interval construction has been developed in Ferraty *et al*. (2010). Its use for data-driven bandwidth selection is still an open problem, but the multivariate ideas discussed by Härdle and Bowman (1988) could certainly be adapted in some way to functional situations.

Other applications of uniform consistency results

It is not possible to list all the possible applications of the uniform consistency results stated above. The main reason for this is that such kinds of results (p. 92) (specifically, results like Corollary 4.2) are needed as long as asymptotic properties are expected for multistage estimation procedures. This can clearly be seen above in the sections on

bandwidth selection using the leave-one-out estimate (4.45) and also the section on bootstrapping using the preliminary step of the estimation of residuals (step (i) of the procedure). Therefore an obvious field of application of this result is (or will be) the study of models involving more than one operator, such as the additive model that will be studied in Section 4.4 or the various semiparametric models that will be studied in Section 4.3. Since the study of semiparametric models/estimates for functional data is a very young field, future statistical research will certainly make extensive use of these uniform results.

4.2.5 Comments and brief bibliographical survey

The aim of this section is twofold. Firstly, we comment on the hypothesis introduced earlier in order to get asymptotic results. Note that, because the general structure of the results presented above is not specific to the nonparametric setting but is common to the next sections of this chapter, the comments on the general results themselves (and, more specifically, on rates of convergence) are referred to the concluding Section 4.6. Our second aim here is to provide a short bibliographical survey for nonparametric functional regression. At the end we pay specific attention to functional discrimination, which can be seen as a special case of the general functional regression framework.

Comments on the hypothesis

As we have said in the previous sections, we have not attempted to present the results under the most general hypotheses, but have rather preferred to construct general sets of possible hypotheses that could apply to more than one theorem. The bibliographical references mentioned earlier will help readers to obtain most of the technical details. However, even if some of our conditions may appear rather technical and complicated, it should be emphasized that they are sufficiently unrestrictive to open the way for many applications.

The most important conditions that allow us to deal with the functional setting are those linked with the small-ball probability function ϕ_x. It is easy to check that all the conditions are trivially satisfied if we restrict our results to the standard multivariate situation $E = \mathbb{R}^k$ and, furthermore, they allow us to extend many earlier papers in these fields to situations in which the explanatory variable does not necessarily admit a density with respect to the Lebesgue measure on $\mathbb{R}^k$. More generally, they are trivially satisfied as long as the functional process X is of fractal type (that is, when for some $a > 0, \phi_x(\epsilon) \sim C\epsilon^a$). Even if we know that we can (p. 93) always construct a topology associated with a suitable semi-metric d such that any process X is fractal, it is worth noting that our assumptions on ϕ_x are satisfied in much more general situations than fractal ones. All the details on these questions can be found in the general discussion provided in Chapter 13 of the monography by Ferraty and Vieu (2006*a*). In the same spirit and as discussed in Ferraty *et al*. (2009), the topological conditions imposed on the entropy function $\psi_S(\epsilon)$ are satisfied in a wide range of situations (including the multivariate and fractal ones, but also many

others). This is also true of the conditions needed on the functions τ and ξ in Sections 4.4, 4.5, and 4.6 for specifying the exact leading terms of L_p errors (see Ferraty *et al*. (2007)).

Finally, the conditions imposed on the parameters of the estimates (namely, kernels and bandwidths) are not restrictive in the sense that they are similar to those widely used in multivariate nonparametric statistics.

Further literature on functional regression

Further literature on kernel functional regression

The kernel estimate $\hat{r}_{\text{ker}}$ has been studied quite a lot in the literature since its introduction by Ferraty and Vieu (2000, 2002). The study of its pointwise almost-sure consistency properties has been described in detail in Ferraty and Vieu (2006*a*), and these results have been extended to not-necessarily-independent pairs (X_i, Y_i) (with direct applications to functional time-series analysis) by Ferraty *et al*. (2002), Ferraty and Vieu (2004, 2008), and Benhenni *et al*. (2008). Chapter 5 of this book is especially devoted to this functional time-series setting. Regarding the extension of the L_p pointwise convergence of the kernel estimate under dependence conditions, the reader is referred to Delsol (2007), while asymptotic normality under dependence conditions has been obtained by Masry (2005) and Delsol (2008*a*). With regard to uniform consistency properties, the literature is restricted to Ferraty and Vieu (2004, 2008) and Ferraty *et al*. (2009). To conclude this short survey, let us note that a robust version of $\hat{r}_{\text{ker}}$ has been studied in Crambes *et al*. (2008). Let us also indicate that this estimate has been revisited in the special case when E is a finite-dimensional manifold by Pelletier (2006), and in the special discrimination problem (that is, when Y only takes a finite number of values) by Ferraty and Vieu (2003) and Abraham *et al*. (2006).

Further literature on local linear functional regression

The literature on local linear regression for functional variables is much less extensive. The first contribution to this topic was that of Barrientos-Marin (2007), and up to now only pointwise consistency results have been available; the result we presented earlier was taken from these. The reader may also look at the recent work by Baíllo and Grané (2008), Boj *et al*. (2008), and Barrientos-Marin *et al*. (2010) for (p. 94) similar results. See also Aneiros *et al*. (2008) for similar methods in a time-series framework.

Further literature on kNN regression

Regarding the asymptotic properties of the kNN functional estimate $\hat{r}_{\text{kNN}}$, the literature is in general restricted to the rate of pointwise almost-sure consistency, which we presented above and which was taken from Burba *et al*. (2009) (see also Burba *et al*. (2008) for a consistency result without rate). However, it is worth noting that kNN ideas are very popular in the classification community and a few papers have been devoted to the special case of the regression model (4.1) when Y takes only a finite number of values. This specific regression model has direct applications in functional data discrimination; references in this direction include Ferraty and Vieu (2003), Biau *et al*. (2005), and Cerou

and Guyader (2006). The next section is especially devoted to such discrimination problems.

On functional discrimination

It is worth noting that all the results presented earlier in this chapter apply directly to situations in which the response variable only takes a finite number of values, let us say $Y \in \{1, \ldots, G\}$. This is typically the case when one has to deal with supervised clasification. In this situation, one has a sample of functional objects $X_1, \ldots, X_n$, each of them being connected to one of the specific groups $1, \ldots, G$, and one wishes to estimate, for each $g = 1, \ldots, G$, the function:

$$R_g(\cdot) = P(Y = g | X = \cdot).$$

Then, once each operator R_g has been estimated (let us say by some estimator $\hat{R}_g$), one can use this to assign a group to a new functional object x simply by taking the value $\hat{g}_x$ such that

$$\hat{g}_x = \max_{\{g=1,\ldots,G\}} \hat{R}_g(x).$$

Once we have noted that each operator R_g can be seen as a specific regression operator, such as those defined in (4.1), with response variable

$$Z_g = 1 \text{ if } Y = g$$
$$Z_g = 0 \text{ if } Y \neq g,$$

it is easy to see that all the literature devoted to functional regression (including all of what has been presented above in this chapter) can be directly used in functional discrimination. An extensive discussion on nonparametric functional discrimination can be found in Chapter 8 of Ferraty and Vieu (2006*a*).

(p. 95) 4.3 Kernel methods in semiparametric functional regression

4.3.1 Presentation

Semiparametric ideas have been widely developed in standard multivariate statistics (see, for instance, Bickel *et al*. (1993), Horowitz (1998), and Härdle *et al*. (2004) for general monographs and Sperlich *et al*. (2006) for a selection of the most recent developments). The main goal of multivariate semiparametric statistics is to achieve a trade-off between models that are too flexible (as nonparametric ones can be when the dimension of the covariables is high) and too restrictive (as linear models, or more generally parametric models, can be). In other words, semiparametric modeling is a nice way to reduce

dimensionality effects in multidimensional statistics. Therefore it is quite natural to expect a major development of semiparametric ideas in the functional setting which, roughly speaking, corresponds to some infinite-dimensional situation. The newness of this topic means that there are not many developments available in the literature, and as far as we know developments only concern two specific models: the single functional index model and the partial linear functional model. The aim of this section is to present the few theoretical advances that exist concerning these models. In Chapter 1 of this volume the reader can find a more general discussion on semiparametric functionalmodels, including a precise definition of what is called a “semiparametric functional,” as well as various other models whose investigation is still an open problem.

A major common feature of semiparametric estimation (see Definition 1.5 of Chapter 1) is that we involve one, or more than one, functional nonlinear operators which have to be estimated nonparametrically. This is why the general nonparametric functional estimation techniques presented above in Section 4.2 will play a major role in semiparametric statistics. Once again because of the youth of semiparametric functional statistics as a research area, all the advances in this field concern only kernel-type estimates (such as those defined in Definition 4.1). Thus all that follows in Section 4.3 concerns such kinds of estimates, while extensions to other types of estimates (such as kNN or local linear estimates) are still open and challenging problems.

4.3.2 Single functional index regression

Presentation of the model

The single functional index model consists in assuming that the functional variable X acts on the response Y only through its projection on some (unknown) fixed (p. 96) functional element θ_0. The formulation of this idea requires us to reinforce the conditions on the space E on which X takes its values. More precisely, to insure the existence of the projection it is necessary to assume that E is a Hilbert space endowed with an inner product $\langle \cdot, \cdot \rangle_E$ (and an associated norm $|| \cdot ||_E$). The following model was first introduced in Ferraty *et al*. (2003) where its identifiability was studied.

Definition 4.5 *The single functional index regression model is defined by assuming that the regression model defined in* (4.1) *can be rewritten as*

(4.52)

$$Y = r(X) + \in = g\left(\langle X, \theta_0 \rangle_E \right) + \in,$$

where g is an unknown real function and θ_0 *is an unknown parameter in E, and where* $E(\epsilon|X) = 0$.

From now on we will assume the identifiability of this model. General conditions for insuring this identifiability are given in Ferraty *et al*. (2003). This is clearly a semiparametric model (see Definition 1.5 of Chapter 1) with a nonparametric component g and a parametric component $\theta \in E$. The statistical problem consists in estimating the operator g and the parameter θ_0 together from a sequence of n independent pairs (X_i, Y_i) each with the same distribution as (X, Y).

Kernel estimation

There are many ways to attack this problem in standard multivariate situations (that is, when $E = \mathbb{R}^p$), a selection of references includes Härdle *et al*. (1993), Hristache *et al*. (2001), and Delecroix *et al*. (2006). The literature for functional settings is much less extensive and, as far as we know, only the following two-stage estimation procedure has been studied.

First of all one considers, for each functional element $\theta \in E$, the following real-valued operator:

$$\forall\, x \in E,\ r_\theta(x) = E\left(Y \middle| \langle X, \theta \rangle = x\right).$$

Note that, under model (4.52), we have:

$$r(\cdot) = g\left(\langle \cdot, \theta_0 \rangle_E \right) = r_{\theta_0}(\cdot).$$

It is natural to estimate r_θ using kernel smoothing ideas, and this leads to:

(4.53)

$$\forall\, x \in E,\ \hat{r}_\theta(x) = \sum_{i=1}^{n} \frac{Y_i K\left(\frac{|\langle x - X_i, \theta\rangle_E|}{h}\right)}{K\left(\frac{|\langle x - X_i, \theta\rangle_E|}{h}\right)}.$$

It is worth noting that, for a fixed value of θ, the study of the estimate $\hat{r}_\theta$ is not that complicated since it can be viewed as a special case of the kernel functional estimate (p. 97) (see Definition 4.1) in which the semi-metric is taken to be

(4.54)

$$\forall\, (u, v) \in E^2,\ d_\theta(u, v) = |\langle u - v, \theta \rangle_E|.$$

In a second step, we look at the value of θ for which the estimate $\hat{r}_\theta$ is closest to the true unknown operator r_{θ_0} with respect to some measure of accuracy such as, for instance, the following mean integrated squared error:

(4.55)

$$MISE(\hat{r}_\theta) = \int_E E\left(\hat{r}_\theta(x) - r_{\theta_0}(x)\right)^2 dP_X(x).$$

Of course, because r_{θ_0} is unknown, this measure of error is uncomputable in practice, and a standard data-driven approximation of it consists in introducing the following cross-validation criterion:

(4.56)

$$CV(\theta) = \frac{1}{n}\sum_{j=1}^{n}\left(Y_j - \hat{r}_\theta^{-j}(X_j)\right)^2,$$

where

(4.57)

$$\forall\, y \in E,\ \hat{r}_\theta^{-j}(y) = \frac{\sum_{i\neq j}^{n} Y_i K\left(\frac{|\langle x - X_i, \theta\rangle_E|}{h}\right)}{\sum_{i\neq j}^{n} K\left(\frac{|\langle x - X_i, \theta\rangle_E|}{h}\right)}.$$

Now, the functional index θ is estimated by minimizing this criterion over some set of possible values Θ (which is allowed to depend on n), to be discussed later:

(4.58)

$$\hat{\theta} = \arg\min_{\theta\in\Theta} CV(\theta).$$

Finally, under model (4.52) the regression operator r is estimated by means of:

(4.59)

$$\forall\, x \in E,\ \hat{r}_{\mathrm{SFIM}}(x) = \hat{r}_{\hat{\theta}}(x).$$

The section on asymptotics below will be devoted to the statement of theoretical results that prove the good asymptotic performance of this two-stage procedure; first let us state and comment on the general conditions needed to insure the validity of the procedure.

General assumptions

As discussed before, the first step of the procedure consists, for each θ, in estimating the real-valued operator r_θ. The estimator $\hat{r}_\theta$ defined by (4.53) can be seen as a special case of the kernel estimate $\hat{r}_{\mathrm{ker}}$ with the particular semi-metric d_θ defined in (4.54). So it is not surprising that we need the same kinds of assumptions as those introduced in Section 4.2, subject of course to suitable adaptation to this new semi-metric d_θ.

General conditions on the model

The general modeling conditions on the distribution of (X, Y) are summarized, on the one hand (and in a similar way to what was done in (4.2)), by assuming that the unknown operators r_θ satisfy the following Lipschitz regularity assumption:

(4.60)

$$\exists\, \beta > 0, \exists\, C < +\infty, \forall\, \theta \in \Theta, \forall\, (x, x') \in E^2,$$
$$|r_\theta(x) - r_\theta(x')| \leq C d_\theta(x, x')^\beta,$$

and on the other hand by assuming the same moment conditions on the variable Y as in (4.7). It is now worth introducing some conditions on the small-ball probability function. These conditions will be much simpler than those used in the general case because the semi-metric d_θ acts only on the one-dimensional variable $\langle X, \theta \rangle$, and not really on the infinite-dimensional variable X as was the case in the general framework of Section 4.2. The conditions can be summarized by assuming that, when $\epsilon \to 0$, the following conditions are almost surely satisfied:

(4.61)

$$\forall\, \theta \in \Theta,\ P(d_\theta(X, X_1) \leq \in | X) \sim \in c_{X,\theta},\ \text{with } 0 < c_1 \leq c_{X,\theta} \leq c_2 < \infty.$$

It is easy to see that these conditions are not very restrictive since they are satisfied whenever the real random variables $\langle X, \theta \rangle$ have densities (having upper and lower bounds) with respect to the Lebesgue measure on $\mathbb{R}$.

Conditions needed for kernel estimation at a fixed value of θ

It now remains for us to state the conditions on the parameters of the estimate (that is, on the kernel function K and on the smoothing parameter h). The general assumptions (4.4), (4.5), and (4.6) have to be reinforced by assuming that the bandwidth is such that

(4.62)

$$\exists (d_1, d_2), \exists\, n_0,\ 0 < d_1 < d_2 < 1 \text{ and } \forall\, n > n_0,\ n^{-d_2} \le h \le n^{-d_1},$$

and by assuming that the kernel function K is strictly decreasing and such that

(4.63)

$$\exists (k_1, k_2),\ 0 < k_1 \le k_2 < \infty,\ k_1 1_{[0,1]}(t) \le K(t) \le k_2 1_{[0,1]}(t).$$

Conditions needed for the estimation of θ

To insure the good behavior of the cross-validation criterion CV as a data-driven approximation of the true theoretical error $MISE$, we must assume that the set Θ of possible values of θ is not too large. More precisely, the following cardinality condition is needed:

(4.64)

$$\exists_\alpha > 0, \exists\, C < \infty,\ \#\Theta = Cn^\alpha.$$

Finally, for simplicity of presentation, we assume that there exists some compact subset $S \subset E$ such that:

(4.65)

$$X \in S,\ a.s.$$

(p. 99) While this last condition looks rather restrictive, it can be easily passed over simply by introducing into the various criteria (that is, on CV and $MISE$) some weight function having support on S, exactly as we did before for bandwidth selection.

Some asymptotics

The main result of this section is the next theorem which states the L_2 rates of convergence of the two-stage estimate $\hat{r}_{\mathrm{SFIM}}$. This result is taken from the recent paper by Ait-Saïdi *et al*. (2008). It will not be proved in detail here, but the main lines of argument will be described. The complete proof can be found in the above-mentioned paper.

Theorem 4.9 *Assume that the model* (4.52) *holds with* $\theta_0 \in \Theta$. *Under conditions* (4.7), (4.60), (4.61), (4.62), (4.63), (4.64), and (4.65), *we have*:

(4.66)

$$\int_E E\left(\hat{r}_{\mathrm{SFIM}}(x) - r_{\theta_0}(x)\right)^2 dP_X(x) = O\left(h^{2\beta}\right) + O\left(\frac{1}{nh}\right),\ a.s.$$

Proof of Theorem 4.9 The complete proof can be found in Ait-Saïdi (2008). It is broken down into two main intermediary results.

(1) Firstly, we look for results concerning the kernel estimate $\hat{r}_\theta$, for each fixed functional element θ. One can apply the general uniform consistency results obtained earlier in Theorem 4.7 and Corollary 4.2, noting that condition (4.61) implies that the small-ball probability defined in (4.34) is now such that

$$\phi_S(\epsilon) \sim C\epsilon, \quad \text{as } \epsilon \to 0,$$

and that the entropy function ψ_S defined in (4.33) is such that

$$\psi_S(\epsilon) \sim -\log(\epsilon), \text{ as } \epsilon \to 0.$$

In summary, result (4.37) becomes:
(4.67)

$$\sup_{x \in S} |\hat{r}_\theta(x) - r_\theta(x)| = O(h^\beta) + O\left(\sqrt{\frac{\log n}{nh}}\right), \ a.s.$$

A similar route can be followed for the L_2 errors of the estimates $\hat{r}_\theta$. Since they are special cases of the estimate $\hat{r}_{\text{ker}}$, one can use the bounds obtained in (4.11), by taking $\phi_x(h) = h$, to show that:
(4.68)

$$\text{var}(\hat{r}_\theta(x)) = O\left(\frac{1}{nh}\right),$$

the O being uniform on $x \in S$ because of the compactness of S. On the other hand, the first part of (4.67) insures, once again uniformly in $x \in S$, that:
(4.69)

$$E\hat{r}_\theta(x) - r_\theta(x) = O(h^\beta).$$

(p. 100) Both results (4.68) and (4.69) can be summarized in the following:
(4.70)

$$\int_E E\left(\hat{r}_\theta(x) - r_\theta(x)\right)^2 dP_X(x) = O(h^{2\beta}) + O\left(\frac{1}{nh}\right).$$

(2) Secondly, we look at the second step of the procedure linked with the cross-validated estimation $\hat{\theta}$ of θ_0. By proceeding as we did earlier for bandwidth selection, we can show by following standard arguments on fractional delta-sequence estimators (see Marron and Härdle (1986)) that the integrated error *MISE* is asymptotically equivalent (uniformly on $\theta \in \Theta$) to its empirical version:

$$A_1(\theta) = \frac{1}{n}\sum_{j=1}^{n}\left(r_{\theta_0}(X_j) - \hat{r}_\theta^{-j}(X_j)\right)^2.$$

Hence, for some quantity A_3 independent of $\theta \in \Theta$ we can write:
(4.71)

$$CV(\theta) = A_1(\theta) + \frac{2}{n}\sum_{j=1}^{n}\left(Y_j - r_{\theta_0}(X_j)\right)\left(r_{\theta_0}(X_j) - \hat{r}_\theta^{-j}(X_j)\right) + A_3,$$

and we can show that (once again uniformly in $\theta \in \Theta$) that
(4.72)

$$\frac{2}{n}\sum_{j=1}^{n}\left(Y_j - r_{\theta_0}(X_j)\right)\left(r_{\theta_0}(X_j) - \hat{r}_\theta^{-j}(X_j)\right) = o(MISE(\hat{r}_\theta)).$$

It is worth noting that, because of the structure of the terms of *MISE*, *CV*, and A_1, the uniform consistency result (4.67) will be used repeatedly both for proving the equivalence betwen *MISE* and A_1 and for checking (4.72). Finally one arrives at the following asymptotic optimality property:
(4.73)

$$\lim_{n\to\infty}\frac{MISE(\hat{r}_\theta)}{\inf_{\theta\in\Theta} MISE(\hat{r}_\theta)} = 1, \ a.s.$$

Now, because $\theta_0 \in \Theta$, we get from (4.73) that
(4.74)

$$MISE(\hat{r}_\theta) \leq MISE(\hat{r}_{\theta_0}),$$

and result (4.66) follows directly from (4.70). □

Comments

It is worth noting that the main point in the proof of Theorem 4.9 is the statement of the asymptotic optimality property (4.73) for the cross-validated functional index $\hat{\theta}$. This result may prove interesting for many other purposes than that of the statement of (4.66), and we note that it has been stated without needing to use the strong condition that $\theta_0 \in \Theta$. Indeed, this condition is only needed when writing (4.74) in order to close the gap between the optimality property (4.73) and the required result (4.66). Note that this last gap could also be closed (p. 101) without the assumption that $\theta_0 \in \Theta$, but the rate of convergence would be more complicated since it would include an additional term depending on the quantity $\Delta = \inf_{\theta\in\Theta} \| \theta - \theta_0 \|_E$. Clearly, consistency without the condition $\theta_0 \in \Theta$ would require (at least) that Δ tends to zero as n grows to infinity. These empirical ideas still need to be precisely formulated.

To highlight the interest of the optimality property (4.73), one may note that it can be used for stating asymptotic properties of the index estimate $\hat{\theta}$ itself. The following result is just an example of this point:

Corollary 4.3 *If the operator* $\theta \to r_\theta$ *is a one-to-one correspondance on* Θ*, and if the conditions of Theorem* 4.9 *are satisfied, then we have*:

(4.75)

$$\| \hat{\theta} - \theta_0 \|_E \to 0, \quad \text{in probaility.}$$

Proof of Corollary 4.3 To save space, this proof is just outlined here; a complete version can be found in Ait-Saïdi *et al*. (2008). First of all we note that the one-to-one correspondance between θ and r_θ insures that

(4.76)

$$\forall\, \theta \in \Theta, \forall \epsilon > 0, \exists \eta > 0, \ \| \theta - \theta_0 \|_E > \epsilon \Rightarrow I(\theta) > \eta,$$

where

$$I(\theta) = \int \left(r_\theta(x) - r_{\theta_0}(x) \right)^2 dP_X(x).$$

Now, by again using standard arguments on fractional delta-sequence estimates as described in Marron and Härdle (1986) and as we did earlier in the proof of Theorem 4.9 when checking that *MISE* was equivalent to its empirical version A_1, it can be shown that *MISE* is asymptotically equivalent (uniformly over $\theta \in \Theta$) to its stochastic version $I(\theta)$. Finally, this equivalence together with (4.66) and (4.76) will be enough to prove Corollary 4.3. □

By following the same route, and subject to extra technical conditions linked with the differentiability of the operator *MISE*, we could expect to obtain the rates of convergence given in Corollary 4.3, but this result still remains to be stated precisely. In fact, many open questions remain since, as far as we know, the literature on the single functional index model is restricted to three papers. The oldest one, Ferraty *et al*. (2003), studies the identifiability of the model (4.52) and states a pointwise version of (4.67). The most complete paper is that by Ait-Saïdi *et al*. (2008), from which all the results presented earlier are taken. The preliminary step (4.67) has been extended to dependent variables by Ait-Saïdi *et al*. (2005). Interesting open theoretical questions concern the use of alternative nonparametric functional estimates (such as, for instance, kNN or local linear estimates as described earlier in Definitions 4.2 and 4.3), or the use of alternative ways of estimating the functional parameter θ_0, or the question of simultaneously choosing a data-driven bandwidth $\hat{h}$ and a data-driven parameter $\hat{\theta}$ etc. From a practical point of view, the main open issue is the construction of the set of possible directions Θ.

4.3.3 Partial linear functional regression

This section is devoted to the presentation of recent advances in the partial linear modeling of functional variables. We wish to acknowledge the assistance of German Aneiros-Pérez, who is an international expert in this field, and who kindly agreed to review and correct a preliminary draft of this section.

Presentation of the model

Partial linear modeling ideas are developed in situations in which the explanatory variable X is composed of two parts, one acting nonparametrically on the response Y and the other one acting in a linear way. These ideas have been widely studied in multivariate settings, that is, when $X \in R^{p_1} \times R^{p_2}$. A selection of recent references includes Speckman (1988), Chen (1988), Härdle *et al*. (2000), Schick (1996), Aneiros-Pérez and Quintela del Rio (2001, 2002), Aneiros-Pérez (2002), Aneiros-Pérez *et al*. (2004), and Tong *et al*. (2008). The aim of this section is to present recent extensions of these ideas to situations in which X is functional.

Let us consider the standard regression model with functional covariate, as defined in (4.1), but we now assume that the functional covariate can be broken down into a functional component V and a multivariate one $\mathbf{U}$. The partial linear functional model consists in assuming that the variable $X = (\mathbf{U}, V)$ acts on the response Y in such a way that the action of the finite-dimensional component $\mathbf{U}$ is linear and the action of the infinite-dimensional component is nonparametric. In what follows we assume that $\mathbf{U} = (U_1, \ldots, U_p)^t \in \mathbb{R}^p$ with $\mathbb{R}^p$ endowed with the Euclidean norm $||\cdot||$, and that V takes values in an abstract space F endowed with a semi-metric d_F.

Definition 4.6 *The single partial linear functional regression model is defined by assuming that the regression model defined in* (4.1) *can be rewritten as*

(4.77)

$$Y = r(X) + \in = \gamma^t \mathrm{U} + m(V) + \in,$$

where $\gamma = (\gamma_1, \ldots, \gamma_p)^t$ *is an unknown vector in* $\mathbb{R}^p$*, where* m *is an unknown functional* (*not neceessarily linear*) *operator acting on the space* F*, and where* $E(\epsilon | X) = 0$.

This is clearly a semiparametric model (see Definition 1.5 of Chapter 1), with a nonparametric component and a p-dimensional linear one. The statistical problem consists in estimating the operator m and the multivariate parameter $(\gamma_1, \ldots, \gamma_p)$ together, from a sequence of n independent variables $(X_i, Y_i) = (\mathrm{U}^i, V_i, Y_i)$ each having the same distribution as (X, Y). We will also need to use the following vectors:

$$\mathrm{U}_i = \left(U_{i,1}, \ldots, U_{i,p}\right)^t, \gamma = \left(\gamma_1, \ldots, \gamma_p\right)^t, \mathrm{Y} = \left(Y_1, \ldots, Y_n\right)^t,$$

(p. 103) and the following $n \times p$ matrix:

$$\mathrm{U}=\left(\mathrm{U}_1, \ldots, \mathrm{U}_n\right)^{\mathrm{t}}.$$

Kernel estimation

Exactly as before for the single functional index model, model (4.77) can be studied in two stages. The first stage consists in estimating the linear (finite-dimensional) component parameter γ by means of standard linear regression techniques (let us say that $\hat{\gamma}$ is such an estimate). Then a nonparametric functional technique can be used to estimate the nonlinear infinite-dimensional component m simply by performing kernel regression of the residual (one-dimensional) variable $Y-\hat{\gamma}^{\mathrm{t}}\mathrm{U}$ on the functional explanatory variable V.

The nonparametric functional methods presented above in Section 4.2 (and, more specifically, the kernel regression estimate proposed in Definition 4.1) will clearly be those of principal interest for dealing with the functional variable V. Let us first set out some general notation linked with kernel smoothing. Let us denote by I_{n} the $n \times n$ identity matrix, and by $\mathrm{W}_{\mathrm{n,ker}}$ the $n \times n$ weighting matrix

$$\mathrm{W}_{\mathrm{n,ker}}=\left(\frac{W_{n,\mathrm{ker}}\left(V_i, V_j\right)}{\sum_{k=1}^{n} W_{n,\mathrm{ker}}\left(V_i, V_k\right)}\right)_{i,j=1,\ldots,n}.$$

Recall that the weights $W_{n,\mathrm{ker}}$ are given (see Definition 4.1) by

$$\forall\left(\upsilon, \upsilon'\right) \in F \times F,\ \mathrm{W}_{n,\mathrm{ker}}\left(\upsilon, \upsilon'\right)=K\left(\frac{d_F\left(\upsilon, \upsilon'\right)}{h}\right).$$

We will make use of the random elements

$$\widetilde{Y}=\left(I_n-W_{n,\mathrm{ker}}\right)Y \text{ and } \widetilde{U}=\left(I_n-W_{n,\mathrm{ker}}\right)U.$$

The linear finite-dimensional component γ is estimated by means of

(4.78)

$$\tilde{\gamma}=\left(\tilde{U}^{\mathrm{t}}\widetilde{U}\right)^{-1}\tilde{U}^{\mathrm{t}}\widetilde{Y}.$$

Then the functional operator m is estimated by using kernel regression of the residual

$$E_i=\left(Y_i-\mathrm{U}_i^{\mathrm{t}}\hat{\gamma}\right),$$

on the functional covariate V_i. More precisely, this idea yields the following kernel-type estimate

(4.79)

$$\hat{m}_{\ker}(\upsilon) = \frac{\sum_{i=1}^{n} W_{n,\ker}(\upsilon, V_i)(Y_i - U_i^t\hat{\gamma})}{\sum_{i=1}^{n} W_{n,\ker}(\upsilon, V_i)}.$$

Finally, the estimate of the regression operator r under model (4.77) is defined by

(4.80)

$$\forall\, x = (u,\upsilon) \in E = R^p \times F,\ \hat{r}_{PLFR}(x) = \hat{\gamma}^t u + \hat{m}_{\ker}(\upsilon).$$

General assumptions

The first set of conditions that we need is composed of standard hypothesis to deal with the linear multivariate component γ. More precisely, one needs a link between the multivariate variable $\mathbf{U}$ and the functional variable V. It is assumed that

(4.81)

$$\eta = (\eta_1, \ldots, \eta_p)^t \text{ is independent of } \in,$$

where, for any $j = 1, \ldots, p$, the variable η_j is defined through the regression model

$$U_j = g_j(V) + \eta_j = E(U_j|V) + \eta_j.$$

The following standard condition is required:

(4.82)

$$E_{\eta\eta^t} \text{ is a positive definite matrix.}$$

The second set of conditions concerns the estimation of the functional component m and is, roughly speaking, composed of all the hypotheses which are necessary in order to make use of results on uniform consistency such as those in Theorem 4.7 and Corollary 4.2 stated earlier. More precisely, we need a nonparametric model for each functional component, that is, for $f \in \{m, g_1, \ldots, g_p\}$ we assume that:

(4.83)

$$\exists\, \beta > 0, (\upsilon, \upsilon') \in S_F \times S_F, |f(\upsilon) - f(\upsilon')| \leq C d_F(\upsilon, \upsilon')^{\beta},$$

where S_F is a subset of F such that

(4.84)

$$P(V \in S_F) = 1.$$

Recall that d_F is the semi-metric on F , and denote by ϕs_F the corresponding small-ball probability function (such as that defined in (4.34)) and by ϕ_{F_S} the corresponding entropy

function (such as that defined in (4.33)). The following conditions need to be satisfied in order to make possible the use of results of the same kind as in Theorem 4.7:

(4.85)

$$\text{Conditions (4.34), (4.35), and (4.36) hold for } d_F,\ S_F,\ \phi_{S_F},\ \text{and } \psi_{S_F}.$$

Similarly, regarding the parameters of the estimate (that is, the bandwidth h and the kernel K) one requires that

(4.86)

$$\text{Conditions (4.5) and (4.6) hold for any } x \in S.$$

The conditional moment condition (4.7) has to be adapted by assuming existence and continuity of:

(4.87)

$$\forall j = 1,\ \ldots,\ p,\ \forall\, m \geq 2,\ E\left(\left|U_j\right|^m \middle| V = \cdot\right) \text{ and } E\left(\left|Y - \gamma^t U\right|^m \middle| V = \cdot\right).$$

Finally, because the construction of the linear estimate $\hat{\gamma}$ involves a preliminary nonparametric step (through the term $\widetilde{Y}$), we need a last condition insuring that the error appearing in the nonparametric estimation does not alter the rate of (p. 105) convergence of the linear component. More precisely, one requires

(4.88)

$$\max\left\{nh^{4\beta};\ \frac{\log^4 n}{n\phi_{S_F}^2(h)};\ \frac{\psi_{S_F}^2\left(\frac{\log n}{n}\right)}{n\phi_{S_F}^2(h)};\ \frac{\log^2 n\,\psi_{S_F}\left(\frac{\log n}{n}\right)}{n\phi_{S_F}^2(h)}\right\} \to 0,$$

and

(4.89)

$$\exists\, \tau > 1 \left/ \right. 2,\ \frac{\phi_{S_F}(h)\log n^2}{n^{\tau-1}} \to \infty.$$

Some asymptotics

The following result is extracted from Aneiros-Pérez *et al*. (2006). We present it here in a slightly different form than in the above-mentioned paper, simply to avoid the introduction of additional technical conditions and to retain a basic common structure in all of this chapter. The next result concerns uniform almost-sure consistency (with rates) over some set $S = C \times S_F \subset \mathbb{R}^p \times F$, C being a compact subset of $\mathbb{R}^p$.

Theorem 4.10 *Assume that model* (4.77) *holds. Under conditions* (4.4), (4.81), (4.82), (4.83), (4.84), (4.85), (4.86), (4.87), (4.88), and (4.89), *we have*:

(4.90)

$$\sup_{x\in C\times S_F}\left|\hat{r}_{\mathrm{PLFR}}(x)-r(x)\right| = O\left(\sqrt{\frac{\log\log n}{n}}\right)+O\left(h^{\beta}\right) + O\left(\sqrt{\frac{\psi_S\left(\frac{\log n}{n}\right)}{n\phi_S(h)}}\right), \text{ a.s.}$$

Proof of Theorem 4.10 We just give the main steps of the proof, emphasizing aspects that are linked with the estimation of the functional operator m. A complete proof can be found in Aneiros-Pérez *et al*. (2006) under slightly different conditions.

(1) The first step of the proof consists in showing that the linear coefficient γ has the same asymptotic properties in the partial linear model (4.77) as it does when it is estimated in the simple linear multivariate model. More precisely, one can show by standard techniques that for any $j = 1, \ldots, p$ and for some finite real constants C_j we have
(4.91)

$$\limsup_{n\to\infty}\sqrt{\frac{n}{\log\log n}}\left|\hat{\gamma}_j-\gamma_j\right| = C_j.$$

We note that the proof of this result is rather long and technically difficult, but it is not surprising since it follows (globally) the same route as described, for instance, in Liang (2000) in the standard partial linear model when V is also finite dimensional. (p. 106) The functional feature consists in using the uniform consistency result (4.42) to deal with the term $\tilde{Y}$ that appears in the construction of the estimate γ, and in using condition (4.86) to make sure that this does not alter the $\sqrt{n}$ consistency of the parametric estimate.
(2) The second step consists in introducing the following estimate:

$$\tilde{m}(v)=\frac{\sum_{i=1}^{n}W_{n,\mathrm{ker}}(v,V_i)\left(m(V_i)+\epsilon_i\right)}{\sum_{i=1}^{n}W_{n,\mathrm{ker}}(v,V_i)}.$$

This is a kernel-type estimator (see Definition 4.1) with the new response variable $m(V) + \epsilon$. We are now able to apply Theorem 4.7. More precisely, the result (4.37) becomes:
(4.92)

$$\sup_{v\in S_F}|\widetilde{m}(v)-m(v)|=O\left(h^{\beta}\right)+O\left(\sqrt{\frac{\psi_{S_F}\left(\frac{\log n}{n}\right)}{n\phi_{S_F}(h)}}\right),\ a.s.$$

On the other hand, we have the following inequality:

(4.93)

$$\begin{aligned}\sup_{v\in S_F}|\hat{m}_{\ker}(v)-m(v)| \ &\leq \sup_{v\in S_F}|\widetilde{m}(v)-m(v)| \\ &\quad + \|\hat{\gamma}-\gamma\|\sup_{v\in S_F}\|W(v)\|,\end{aligned}$$

where

$$W(v)=\frac{\sum_{i=1}^{n}W_{n,\ker}(v,V_i)(U_i^t)}{\sum_{i=1}^{n}W_{n,\ker}(v,V_i)}.$$

Finally, the proof of Theorem 4.10 follows directly from (4.91), (4.92), (4.93), and from the definition of $\hat{m}_{\mathrm{PLFR}}$ (see (4.80)). □

Comments

Once again, the main point to be emphasized in the above proof is the important role of the uniform results stated in Section 4.2.4, which are not only used to deal with the estimation of the functional component m in the statement of (4.92) but are also used for the statement of asymptotic results on the estimation of γ (such as for instance (4.91)).

The literature on partial linear functional models is rather restricted. As far as we know there are only two papers in this field. The first one is by Aneiros-Pérez and Vieu (2006), and includes not only Theorem 4.10 but also the asymptotic distribution for the multivariate estimate γ^t. Note that in this paper, result (4.90) was obtained under slightly weaker conditions (for example, on the moments of the error terms ϵ and η) but was restricted to a rather specific compact set S_F. In Aneiros-Pérez and Vieu (2008) an extension of Theorem 4.10 to dependent samples is presented. This last result allows for time-series applications.

(p. 107) Many interesting open problems still remain. They include, for instance, extensions to other nonparametric estimates of the functional component m (including, for instance, kNN or local linear estimates as described before in Definitions 4.2 and 4.3), or the question of choosing a data-driven bandwidth $\hat{h}$. Alternative models in which the linear component is also functional, and many other results already known in the multivariate setting, also need suitable adaptation to functional variables …

4.4 Using kernels for additive functional regression

This section discusses some recent advances in the additive modeling of functional variables. Unlike the other parts of this chapter in which the results have been taken from the existing literature, this section contains a theorem (Theorem 4.11) that has not yet been published elsewhere.

4.4.1 Presentation of the model

The additive model has been developed in situations when the explanatory variable X is composed of various different parts, each of which has to be modeled in a nonparametric way. These ideas have been widely developed in the multivaiate setting, that is, when $X \in \mathbb{R}^p$. A selection of references here includes Stone (1985, 1986, 1994), Härdle and Hall (1993), Hastie and Tibshirani (1986, 1990), Schimek and Turlach (2000), Mammen and Park (2006), and Horowitz *et al*. (2006). The aim of this section is to present a recent extension of additive ideas to situations in which X is functional.

Let us consider the standard regression model with functional covariate, as defined in (4.1), but assume that the functional covariate can be decomposed into $X = (X^1, \ldots, X^p)$. Each component X^j takes values in some abstract space E^j endowed with a semi-metric d_j. So the space E in which the explanatory variable X takes its values is the product space $E = E^1 \times \cdots \times E^p$. Recall that E is endowed with a semi-metric d that can be (but is not necessarily) constructed from the d_j.

Definition 4.7 *The additive functional regression model is defined by assuming that the regression model defined in* (4.1) *can be rewritten as*

(4.94)

$$Y = r(X) + \in = \sum_{j=1}^{p} r^j(X^j) + \in,$$

(p. 108) *where each r^j is an unknown functional* (*not necessarily linear*) *operator acting on the marginal space E^j, and where $E(\epsilon|X) = 0$.*

It is clear that the formulation (4.94) is not unique. It is therefore necessary to put some restrictions on the additive components to deal with identifiability of the model. This is achieved by introducing the jth stage residuals

$$\forall\ j = 1, \ \ldots, p, \ \in^j = Y - \sum_{k=1}^{j} r^k(X^k),$$

and by assuming that

(4.95)

$$r^1(X^1) = E(Y|X^1) \text{ and } \forall\ j \geq 2,\ r^j(X^j) = E(\in^{j-1}|X^j).$$

This is clearly a dimensionality-reduction model (see Definition 1.4 of Chapter 1). The statistical problem consists in estimating the operators r^j from a sequence of n independent variables $(X_i, Y_i) = (X_i^1, \ldots, X_i^p, Y_i)$. We will also use the notation, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$:

$$\in_i^j = Y_i - \sum_{k=1}^{j} r^k(X_i^k).$$

Kernel estimation

Exactly as before for the single functional index model or for the partial linear functional model, the model (4.94) can be studied in series of stages. The idea here is to use nonparametric functional kernel ideas to estimate each nonlinear operator r^j. Let us introduce some kernel functions $K^1, \ldots, K^p$ and bandwidths $h_1, \ldots, h_p$. As in Definition 4.1 we also introduce the following local weights:

$$\forall\ j = 1, \ldots, p, \forall\ (\upsilon, \upsilon') \in E^j \times E^j,\ W_{n,\text{ker}}^j(\upsilon, \upsilon') = K^j\left(\frac{d_j(\upsilon, \upsilon')}{h_j}\right).$$

In a first approximation, we estimate the first additive component by means of the following kernel estimate:

(4.96)

$$\forall\ x^1 \in E^1,\ \hat{r}_{\text{ker}}^1(x^1) = \frac{\sum_{i=1}^n W_{n,\text{ker}}^1(x^1, X_i^1) Y_i}{\sum_{i=1}^n W_{n,\text{ker}}^1(x^1, X_i^1)}.$$

Then, in an iterative way, we construct the other estimates by performing, for any $j = 2, \ldots, p$, the regression of the $(j-1)$th-order estimated residuals

$$\hat{\in}_i^{j-1} = Y_i - \sum_{k-1}^{j-1} \hat{r}_{\text{ker}}^k(X_i^k),\ i = 1, \ldots, n,$$

(p. 109) on the next variable $X_i^j, i = 1, \ldots, n$. This leads, for $j \geq 2$, to the kernel estimates:

(4.97)

$$\hat{r}_{\text{ker}}^j(x^j) = \frac{\sum_{i=1}^n W_{n,\text{ker}}^j(x^j, X_i^j) \hat{\in}_i^{j-1}}{\sum_{i=1}^n W_{n,\text{ker}}^j(x^j, X_i^j)},\ x^j \in E^j.$$

Finally, under model (4.4), the additive kernel-type estimate of the regression operator r is defined by

(4.98)

$$\forall\, x = (x^1, \ldots, x^p) \in E,\ \hat{r}_{\mathrm{Add}}(x) = \sum_{j=1}^{p} \hat{r}^j_{\mathrm{ker}}(x^j).$$

General assumptions

Because the additive estimate (4.98) is just a combination of p standard kernel regression estimates, the conditions required for the statement of asymptotic results are not really surprising. Indeed, these conditions are just those required to apply repeatedly the general uniform consistency results for kernel functional regressors as stated earlier. In what follows, we introduce the following subsets of the functional spaces

$$S^j \subset E^j,\ j = 1, \ldots, p.$$

The nonparametric modeling consists here in assuming, for each $j = 1, \ldots, p$, some Lipschitz-type condition on the operator r^j to be estimated:

(4.99)

$$\exists\, \beta^j > 0, \exists\, C_j < +\infty, \forall\, (x, x') \in E^j \times E^j,\ |r^j(x) - r^j(x')| \leq C_j d_j(x, x')^{\beta^j}.$$

We need general conditions on the small-ball probability functions, whose definition we recall:

$$\forall\, j = 1, \ldots, p, \forall\, x \in S^j,\ \phi_{j,x}(\epsilon) = P_{X^j}(d_j(x, \epsilon)).$$

We assume that there exist functions ϕ_{S^j} and constants α_1^j, α_2^j and α_3^j such that for any $\epsilon > 0$ and for any $j = 1, \ldots, p$:

(4.100)

$$0 < \alpha_1^j \phi_{S^j}(\epsilon) \leq \inf_{x \in S^j} \phi_{j,x}(\epsilon) \leq \alpha_2^j \phi_{S^j}(\epsilon) < \infty,$$

and

(4.101)

$$\phi'_{S^j}(\epsilon) \leq \alpha_3^j.$$

We also need some conditions on the entropy ψ_{S^j} of each subset S^j (see Definition 4.4). Assume that for any $j = 1, \ldots, p$:

(4.102)

$$\exists\, n_0, \forall\, n > n_0,\ \frac{(\log n)^2}{n\phi_{S^j}(h_j)} < \psi_{S^j}\left(\frac{\log n}{n}\right) < \frac{n\phi_{S^j}(h_j)}{\log n},$$

(p. 110) and

(4.103)

$$\exists\, \delta > 0, \sum_{i=1}^{n} e^{-\delta \psi_{S^j}\left(\frac{\log n}{n}\right)} < \infty.$$

Each kernel K^j and each bandwidth h_j is assumed to satisfy the same general conditions as in Section 4.2. It is assumed that for any $j = 1, \ldots, p$:

(4.104)

$$K^j \geq 0, \quad \int_R K^j = 1, K^j \text{ is Lipschitz on } [0, 1), \\ \text{and support}\left(K^j\right) \subset [0, 1).$$

(4.105)

$$\forall \epsilon \leq \epsilon_0, \forall\, x \in S^j, C_1 \phi_{S^j}(\epsilon) \leq E K^j\left(\frac{d^j(x, X)}{\epsilon}\right) \leq C_2 \phi_{S^j}(\epsilon),$$

and

(4.106)

$$\lim_{n\to\infty} h_j = 0 \text{ and } \lim_{n\to\infty} \frac{n\phi_{S^j}\left(h_j\right)}{\log n} = \infty.$$

Some asymptotic results

In Theorem 4.11 below we give the uniform rate of convergence of the additive estimate $\hat{r}_{\text{Add}}$. Because such a result has not been published before and because its proof is rather simple, we will provide a complete proof. The uniformity is stated over a subset $S = S^1 \times \cdots \times S^p \subset E$, such that

(4.107)

$$\forall\, j = 1, \ldots, p, P\left(X^j \in S^j\right) = 1.$$

Theorem 4.11 *Assume that the model defined by* (4.94), (4.95), and (4.7) *holds. Assume that conditions* (4.99), (4.100), (4.101), (4.102), (4.103), (4.104), (4.105), (4.106), and (4.107) *hold. Then we have*:

(4.108)

$$\sup_{x \in \times S} \left|\hat{r}_{\text{Add}}(x) - r(x)\right| = O\left(\sum_{j=1}^{p} h_j^{\beta^j}\right) + O\left(\sum_{j=1}^{p} \sqrt{\frac{\psi_{S_j}\left(\frac{\log n}{n}\right)}{n\phi_{S^j}\left(h_j\right)}}\right), \text{ a.s.}$$

Proof of Theorem 4.11 The proof is based on the decomposition below. For $x = (x^1, \ldots, x^p) \in E$, we have:

(4.109)

$$|r(x) - \hat{r}_{\text{Add}}(x)| \leq \sum_{j=1}^{p} |r^j(x^j) - \hat{r}^j_{\text{ker}}(x^j)|.$$

(1) The first component of the sum in (4.109) can be treated directly, because the estimate $\hat{r}^1_{\text{ker}}$ is a special type of kernel estimate (see Definition 4.1) with response (p. 111) Y and explanatory variable X^1. Therefore we have, by a direct application of Theorem 4.7:
(4.110)

$$\sup_{x^1 \in S^1} |r^1(x^1) - \hat{r}^1_{\text{ker}}(x^1)| = O\left(h_n^{\beta^1}\right) + O\left(\sqrt{\frac{\psi_{S^1}\left(\frac{\log n}{n}\right)}{n\phi_{S^1}(h_1)}}\right), \text{ a.s.}$$

(2) Let us now consider the case $j = 2$. We can split the estimate $\hat{r}^2_{\text{ker}}$ into two parts:

$$\hat{r}^2_{\text{ker}} = \hat{r}^{2,1}_{\text{ker}} + \hat{r}^{2,2}_{\text{ker}},$$

with

$$\forall\, x^2 \in E^2,\ \hat{r}^{2,1}_{\text{ker}}(x^2) = \frac{\sum_{i=1}^{n} W^2_{n,\text{ker}}(x^2, X_i^2) \in_i^1}{\sum_{i=1}^{n} W^2_{n,\text{ker}}(x^2, X_i^2)},$$

and

$$\forall\, x^2 \in E^2,\ \hat{r}^{2,2}_{\text{ker}}(x^2) = \frac{\sum_{i=1}^{n} W^2_{n,\text{ker}}(x^2, X_i^2)\left(r^1(X_i^1) - \hat{r}^1_{\text{ker}}(X_i^1)\right)}{\sum_{i=1}^{n} W^2_{n,\text{ker}}(x^2, X_i^2)}.$$

The treatment of $\hat{r}^{2,1}_{\text{ker}}$ is easy because it is a special type of kernel estimate (see Definition 4.1) with response ϵ^1 and explanatory variable X^2. Therefore we have, by a direct application of Theorem 4.7:
(4.111)

$$\sup_{x^2 \in S^2} |r^2(x^2) - \hat{r}^{2,1}_{\text{ker}}(x^2)| = O\left(h_2^{\beta^2}\right) + O\left(\sqrt{\frac{\psi_{S^2}\left(\frac{\log n}{n}\right)}{n\phi_{S^2}(h_2)}}\right), \text{ a.s.}$$

To deal with the other term involving the estimate $\hat{r}^{2,2}_{\text{ker}}$ it suffices to use condition (4.107) together with Corollary 4.2 to see that
(4.112)

$$\forall\, x^2 \in E^2,\ \hat{r}^{2,2}_{\text{ker}}(x^2) = O\left(h_1^{\beta^j}\right) + O\left(\sqrt{\frac{\psi_{S^1}\left(\frac{\log n}{n}\right)}{n\phi_{S^1}(h_1)}}\right),\ a.s.$$

Finally, we have almost surely:

(4.113)

$$\begin{aligned}\sup_{x^2 \in S^2} \left| r^2(x^2) - \hat{r}^2_{\text{ker}}(x^2) \right| &= O\left(h_2^{\beta^2}\right) + O\left(\sqrt{\frac{\psi_{S^2}\left(\frac{\log n}{n}\right)}{n\phi_{S^2}(h_2)}}\right) a.s. \\ &= O\left(h_1^{\beta^1}\right) + O\left(\sqrt{\frac{\psi_{S^1}\left(\frac{\log n}{n}\right)}{n\phi_{S^1}(h_1)}}\right) a.s.\end{aligned}$$

(p. 112) **(3)** It is clear that the previous step can easily be iterated, and we arrive for any $j = 2, \ldots, p$ at:

(4.114)

$$\begin{aligned}\sup_{x^j \in S^j} \left| r^j(x^j) - \hat{r}^j_{\text{ker}}(x^j) \right| &= \sum_{k=1}^{j} \left(O\left(h_k^{\beta^k}\right) \right) \\ &\quad + \sum_{k=1}^{j} \left(O\left(\sqrt{\frac{\psi_{S^k}\left(\frac{\log n}{n}\right)}{n\phi_{S^k}(h_k)}} \right) \right) a.s.\end{aligned}$$

Finally the result (4.108) follows directly from (4.109), (4.110), and (4.114). □

Comments

Once again, and as has already been pointed out for other multistage models such as the single functional index (see Section 4.3.2) or partial linear functional (see Section 4.3.3) models, the main point to be highlighted in the above proof is the important role of the uniform results stated in Section 4.2.4. Note that, even if one wishes to study pointwise consistency properties for each additive component, the statement of results like (4.112) needs to control for random terms such as $r^1(X_i^1) - \hat{r}^1_{\text{ker}}(X_i^1)$ here Corollary 4.2 is a key tool.

To conclude this section, let us just note that the literature on additive functional modeling is not very developed. As far as we know, the first theoretical paper was that by Ferraty and Vieu (2009), in which complementary asymptotic results are given (in terms of squared prediction errors). See also Müller (2008) for recent developments (but in a slightly different context), and Aneiros-Pérez *et al*. (2006) for a real environmetrical curves dataset application. Of course, many open questions remain. The most important to be dealt with may concern data-driven bandwidth selection procedures, but it would be worth adapting many other points already well known in standard multivariate nonparametric statistics to the functional framework.

4.5 On testing functional regression models

This section is devoted to the presentation of the few very recent results that exist for testing the procedures for structural regression. Here we wish to acknowledge Laurent Delsol, who is a pioneer in this field, and who kindly agreed to review and improve a preliminary draft of this section.

4.5.1 Introduction

Despite the fact that the literature on nonparametric functional regression estimation is relatively new, it can be seen from the previous sections of this chapter that we now have a wide range of models available in this area. They include purely nonparametric models (see Section 4.2), dimensionality-reduction models such as the single functional index model (Section 4.3.2), partial linear functional models (Section 4.3.3) and additive models (Section 4.4), and purely parametric models such as the linear functional regression model discussed in Chapter 2 of this book. A natural question is therefore to decide whether one of these models is more adapted than others to some given practical situation. This question has not yet been much studied in the literature, but a few recent advances exist on testing procedures for checking the validity of some model. The aim of this section is to describe the main approaches of the testing procedures proposed in Delsol (2008*b*, 2008*c*) and Delsol *et al*. (2010). A larger bibliographical discussion can be found in this last paper by Delsol *et al*. (2010).

We will not give details of the theoretical developments nor of the technical assumptions, because this would give rise to long and tedious notation. In Section 4.5.2, we will set out a general framework for structural regression test procedures and we will briefly describe various specific situations in which these procedures can be helpful. Then in Section 4.5.3 we will discuss how the various kernel-type estimates studied earlier in this chapter can be used to construct broad families of statistics tests. Some asymptotics will be briefly described in Section 4.5.4.

4.5.2 A general structural testing problem

Within the general framework of the functional regression problem (4.1), the question is to decide whether the true regression operator r belongs to some fixed specific family $\mathscr{F}_0$ of operators or not. In other words, we wish to test the null hypothesis

(4.115)

$$H_0 : \{\exists\, R \in F_0,\, P(r(X) = R(X)) = 1\}$$

against some alternative that says that the operator r is *sufficiently far* from $\mathscr{F}_0$. For instance if $||\cdot||$ is some norm on the space of operators, we can measure the distance between r and the class $\mathscr{F}_0$ by looking at the distance between r and its closest approximation (let us say r_0) in $\mathscr{F}_0$:

$$\Delta(r, F_0) = \inf_{r \in F_0} \| r - r' \|.$$

For clarity, we will assume that there is an unique element r_0 such that

(4.116)

$$\Delta(r, F_0) = \| r - r_0 \|,$$

(p. 114) but this condition can be weakened, as indicated in Delsol *et al*. (2010). This leads us, from some given real sequence η_n, to consider the following alternative hypothesis:

(4.117)

$$H_{1,n} : \{\Delta(r, F_0) \geq \eta_n\}.$$

We will see later how a general test statistic can be constructed for this structural problem (see Section 4.5.3), for which an asymptotic normality distribution can be stated under H_0, and divergence can be observed under $H_{1,n}$ as soon as η_n is sufficiently large (see Section 4.5.4).

Before going into these details, let us first discuss a few specific examples of families $\mathscr{F}_0$ in order to show the high degree of generality of this approach.

- *Testing linearity*. One may wish to know whether the true operator is linear, and in this case one uses the family

$$F_{0,\text{lin}} = \{r, \text{linear and continuous operators}\}.$$

- *Testing non-effect*. One may wish to know whether the functional variable X has an effect on the response Y, and in this case one uses the family

$$F_{0,\text{ne}} = \{r, \text{constant operators}\}.$$

More generally in situations when $X = (X^1, X^2)$ one may wish to see whether some component, let us say X^2, has an effect on Y. In this case one uses the family

$$F_{0,\text{ne2}} = \{r, \exists\, r^1, \forall\, x = (x^1, x^2),\ r(x) = r^1(x^1)\}.$$

• *Testing additivity*. In the situation in which the response X is multiple, that is, when $X = (X^1, \ldots, X^p)$, one may wish to decide whether the additive model studied in Section 4.4 is accurate or not. In this case one uses the family

$$F_{0,\text{add}} = \left\{ r, \exists\, r^1, \ldots, r^p, \forall\, x = (x^1, \ldots, x^p), r(x) = \sum_{j=1}^{p} r^j(x^j) \right\}.$$

• *Testing a specific semiparametric model*. One may wish to decide whether the effect of the variable X can be reduced to the effect of one single projection. In other words, one may wish to test the validity of the single functional index model studied in Section 4.3.2. For this one can use the family

$$F_{0,\text{SFIM}} = \{ r, \exists\, g, \exists\, \theta, \forall\, x, r(x) = g(\langle x, \theta \rangle) \}.$$

The same kind of question may also be posed when $X = (U, V)$ with $\mathbf{U} \in \mathbb{R}^p$ to check the validity of the partial functional index model studied in Section 4.3.3. In this case one uses the family

$$F_{0,\text{PLFR}} = \{ r, \exists\, \gamma, \exists\, m, \forall\, x = (u, v), r(x) = m(v) + \gamma^t u \}.$$

(p. 115) • *Testing an unfunctional model*. Other types of questions may concern the functional features of the model. For instance, if X is a curve

$$X = \{ X(t), t \in (a, b) \},$$

one may wish to check whether X acts on Y only through a few values $X(t_1), \ldots, X(t_p)$. In this case one uses the family

$$F_{0,\text{Unf}} = \{ r, \exists\, g, \exists\, t_1, \ldots, t_p, \forall\, x, r(x) = g(x(t_1), \ldots, x(t_p)) \}.$$

These are just a few of the possible situations that could be modeled by hypotheses of the form (4.115), and one can easily imagine many other problems of this kind.

4.5.3 Construction of kernel-based test statistics

In the setting of standard finite-dimensional statistics, structural testing problems have been widely investigated in the literature (see, for instance, the monograph by Hart (1997)). One way of doing this is to build test statistics based on the difference of a purely nonparametric estimate and another estimate which is specific to the model $\mathscr{F}_0$ that one wishes to test. This idea has been popularized by Härdle and Mammen (1993) and has been used in many further papers to deal with a large variety of situations. Given the recent advances in estimation procedures for functional regression (see the previous sections of this chapter), it is natural to think of adapting Härdle–Mammen's ideas to functional data.

More precisely, we will use as a pilot estimate the kernel estimate $\hat{r}_{\text{ker}}$ (see Definition 4.1), and the first approach is to construct test statistics by looking at quantities such as

$$\int \left(\hat{r}_{\text{ker}}(x) - \hat{r}_0(x)\right)^2 \omega(x) dP_X(x),$$

where $\hat{r}_0$ is an estimate of r under the specific model defined by the null hypothesis (4.115), and w is a known weight function. For technical reasons, and once again following the ideas in Härdle and Mammen (1993), it is more convenient to use the folllowing statistics:

(4.118)

$$T_{0,n} = \int \left(\sum_{i=1}^{n} \left(Y_i - \hat{r}_0(X_i)\right) W_{n,\text{ker}}(x, X_i) \right)^2 \omega(x) dP_X(x),$$

where we recall that the local weights are defined from a kernel K and a bandwidth h by:

$$W_{n,\text{ker}}(x, X_i) = K\left(\frac{d(x, X_i)}{h}\right).$$

(p. 116) The principal advantages of this statistic are both to suppress the bias and to overcome the technical problems linked with the random denominator appearing in the kernel estimate $\hat{r}_{\text{ker}}$. In the functional setting, this statistic has been introduced by Delsol (2008*b*). We will briefly present its asymptotic behavior in the next section, but before that we would like to illustrate it through the few specific situations discussed above at the end of Section 4.5.2.

- *Testing linearity*. To test the family $\mathscr{F}_{0,\text{lin}}$, one can use any of the various existing linear functional regressors (see Chapter 2) as a linear estimate $\hat{r}_0$.
- *Testing non-effect*. To test the family $\mathscr{F}_{0,\text{ne}}$, one can use the naive constant estimator $\hat{r}_0(x) = \frac{1}{n}\sum_{i=1}^{n} Y_i, \forall\, x$. To test the family $\mathscr{F}_{0,\text{ne2}}$ one can use a functional estimate based only on the second covariate, such as for instance the kernel estimate $\hat{r}_0 = \hat{r}^1_{\text{ker}}$.
- *Testing additivity*. To test the family $\mathscr{F}_{0,\text{add}}$, one can use the additive estimate $\hat{r}_0 = \hat{r}_{\text{Add}}$ as defined in (4.98).
- *Testing a specific semiparametric model*. To test the family $\mathscr{F}_{0,\text{SFIM}}$ one can use the estimate $\hat{r}_0 = \hat{r}_{\text{SFIM}}$ as defined in (4.59), and to test $\mathscr{F}_{0,\text{PLFR}}$ one can use the estimate $\hat{r}_0 = \hat{r}_{\text{PLFR}}$ as defined in (4.80).
- *Testing an unfunctional model*. To test the family $\mathscr{F}_{0,\text{Unf}}$ the choice is rather extensive since as an estimate $\hat{r}_0$ one can use any of the wide range of well-known p-dimensional smoothers (kernel, splines, local polynomial, kNN, ...)

Of course all these applications will be possible only if, in each situation, the estimate $\hat{r}_0$ can be shown to satisfy the technical conditions required (see discussion below). Once again, these are just a few of the possible applications of the general methodology presented here. As long as our knowledge of functional regression estimation is growing,

more applications for testing could be (and certainly will be) developed directly. Indeed, to apply this general methodology to any submodel $\mathscr{F}_0$ one needs to have at hand some estimate $\hat{r}_0$ for which the rates of convergence can be controlled under the null hypothesis H_0. More precisely, recalling that $\hat{r}_0$ is defined by (4.116), one requires (for instance) conditions on $\hat{r}_0$ similar to the following:

(4.119)

$$\text{Under } H_0,\ \delta(\hat{r}_0, r_0) = o_p(r_n),$$

where δ is some measure of accuracy for estimation under the model H_0. Delsol *et al*. (2010) propose various specific choices of δ. These authors also extend the methodology to the situation in which the unicity condition (4.116) is not satisfied. They also relax (4.119) into a condition saying that, in some sense, $\hat{r}_0$ *is not too far* from the family $\mathscr{F}_0$.

4.5.4 Some theoretical advances

It is not our aim here to state precisely the various conditions needed to obtain asymptotic behavior of the statistic $T_{0,n}$. All the results presented below are stated under quadratic-type measures of accuracy, for the δ appearing in the null hypothesis (see (4.119)) as well as for the norm defining the alternative hypothesis (see (4.117)). All the details can be found in Delsol (2008*b*, 2008*c*) and Delsol *et al*. 2010. Our goal here is simply to highlight the main ideas.

First of all, by following the general approach as in multivariate settings, we may note that the statistic $T_{0,n}$ can be split into 6 terms in the following way:

$$T_{0,n} = T_1 + T_2 + T_3 + T_4 + T_5 + T_6,$$

with

$$\begin{aligned}
T_1 &= \int \sum_{i=1}^{n} \big(Y_i - r(X_i)\big)^2 W_{n,\mathrm{ker}}^2(x, X_i)\omega(x)dP_X(x),\\
T_2 &= \int \sum_{i\neq j} \big(Y_i - r(X_i)\big)\big(Y_j - r(X_j)\big) W_{n,\mathrm{ker}}(x, X_i) W_{n,\mathrm{ker}}(x, X_j)\omega(x)dP_X(x),\\
T_3 &= \int \sum_{i=1}^{n} \big(r(X_i) - \hat{r}_0(X_i)\big)^2 W_{n,\mathrm{ker}}^2(x, X_i)\omega(x)dP_X(x),\\
T_4 &= \int \sum_{i\neq j} \big(r(X_i) - \hat{r}_0(X_i)\big)\big(r(X_j) - \hat{r}_0(X_j)\big) \times W_{n,\mathrm{ker}}(x, X_j)\omega(x)dP_X(x),\\
T_5 &= \int \sum_{i\neq j} \big(r(X_i) - \hat{r}_0(X_i)\big)\big(Y_i - r(X_i)\big) W_{n,\mathrm{ker}}^2(x, X_i)\omega(x)dP_X(x),
\end{aligned}$$

and

$$T_6 = 2\int \sum_{i\neq j} \big(Y_i - r(X_i)\big)\big(r(X_j) - \hat{r}_0(X_j)\big) \times W_{n,\mathrm{ker}}(x, X_i) W_{n,\mathrm{ker}}(x, X_j)\omega(x)dP_X(x).$$

It is worth noting that the terms T_1 and T_2 have the same behavior under both hypotheses H_0 and $H_{1,n}$. The bias of the statistic $T_{0,n}$ will be provided by the term T_1, while its variance will be provided by the term T_2. This leads us naturally to consider the following statistic:

(4.120)

$$\mathrm{T}_n = \frac{T_{0,n} - ET_1}{\sqrt{\mathrm{var}(T_2)}}.$$

The first result that we state below concerns the asymptotic distribution of the centered and reducted statistic T_n under the null hypothesis. Roughly speaking (see Delsol (2008*c*) or Delsol *et al*. (2010)), if the rate of convergence of the estimate $\hat{r}_0$ is *sufficiently fast*, that is, if the sequence r_n defined in (4.119) is *small enough*, the terms T_j, $j = 3, \ldots, 6$ can be shown to be negligible. This results from the (p. 118) fact that each of these terms T_j, $j = 3, \ldots, 6$ involves quantities like $r(X) - \hat{r}_0(X)$, which are exactly equal under H_0 to $r_0(X) - \hat{r}_0(X)$. Once these terms are shown to be negligible, one can look at the precise behavior of both the previous terms T_1 and T_2 and establish, by means of some appropriate central limit theorem, results such as:

(4.121)

$$\text{Under } H_0, \ \mathrm{T}_n \xrightarrow{L} \mathrm{N}(0, 1).$$

The second result we will give concerns the asymptotic behaviour of T_n under the alternative hypothesis $H_{1,n}$. As mentioned before, the behavior of T_1 and T_2 is the same under $H_{1,n}$ as under H_0, and therefore the divergence of the test under the alternative hypothesis will depend on the remaining terms T_j, $j = 3, \ldots, 6$. Because these involve quantities like $r(X) - \hat{r}_0(X)$, they can be shown to be *as large as is* desired subject to conditions on $r - r_0$, that is, by assuming that the sequence η_n defining the null hypothesis (4.117) is *large enough* (see Delsol (2008*c*) or Delsol *et al*. (2010) for details). This leads to results of the form

(4.122)

$$\textit{Under } H_{1,n}, \ \mathrm{T}_n \xrightarrow{P} \infty.$$

4.5.5 Comments

The statements of the results (4.121) and (4.122) mentioned above are technically difficult. While the main conditions needed are those discussed before to control the alternative hypothesis $H_{1,n}$ and also to control the quality of estimation under the null hypothesis H_0, various other conditions are now necessary. On the one hand, they include the same kinds of assumptions as those given in Section 4.2 in order to insure the good behavior of the pilot estimate $\hat{r}_{\mathrm{ker}}$. On the other hand, they also include additional

conditions on the small-ball probability of the pairs (X_i, X_j), because a U-statistics methodology is used for the treatment of the leading terms T_1 and T_2. Also, for technical reasons, it is necessary to construct the null estimate $\hat{r}_0$ using a second statistical sample independent of the original sample (X_i, Y_i), $i = 1, \ldots, n$. As discussed in Delsol (2008*c*) and Delsol *et al*. (2010), the high technical level of these conditions reflects the broad generality of the method without being too restrictive.

It is worth noting that exact asymptotic expansions of the leading terms ET_1 and $\mathrm{var}(T_2)$ are also obtained in Delsol (2008*c*) and Delsol *et al*. (2010). It turns out that the rather complicated form of these expansions makes a direct use of the statistics T_n quite unrealistic in practice. While a naive point of view would lead us to conclude that these results are not very useful, it is worth noting that theoretical results like (4.121) and (4.122) are indispensable preliminary practical tools. This is already the case in standard multivariate situations, where such theoretical results are used for proving the asymptotic validity of bootstrapping procedures. We may (p. 119) conclude by saying that in our functional framework, the full power of the results (4.121) and (4.122) will be revealed only when the validity of resampling procedures can be checked. As far as we know, this is still an open question (see, however, Delsol (2008*c*) for preliminary empirical ideas on testing procedures by functional resampling that show the good behavior of bootstrapping ideas in the case of finite samples).

4.6 Comments

4.6.1 On rates of convergence in functional regression

Topological effects on pointwise results

The common feature of the various asymptotic results presented previously is the great importance of the topological structure that exists on the functional space E. We will now briefly discuss a few instances where these topogical effects appear clearly.

The influence of the topology appears directly in all the pointwise rates of convergence, through the small-ball probability function

$$\phi_x(\cdot) = \{y \in E, d(x, y) \leq \cdot\}.$$

This is true for all kinds of estimates in nonparametric modeling (see Theorems 4.1, 4.2, and 4.3). It is also true for any alternative model such as the additive model (see Theorem 4.11), the single functional model (see Theorem 4.9), or the partial linear functional model (see Theorem 4.10). And it is also true independently of the mode of convergence, since it appears for pointwise consistency (see Theorem 4.1) as well as for L_2 consistency (see Theorem 4.4) and for asymptotic normality (see Theorem 4.5). This remark has a direct practical impact (see Section 4.6.2 below) since the choice of topology (that is, the choice of the semi-metric d) will influence the behavior of any non/semiparametric functional regression estimates directly. The reader may find more comments about the statistical effects of topological structure, as well as various standard examples of possible choices of semi-metric, in Ferraty and Vieu (2006*a*).

Let us now look more precisely at the rates of convergence obtained in semiparametric modeling (see Section 4.3) or in additive modeling (see Section 4.4). Undoubtedly, the rates of convergence obtained for these models (see Theorems 4.9, 4.10, and 4.11) are much better than the rates obtained for nonparametric models (see Theorem 4.1). This is due to the fact that all these models are of dimensionality-reduction type (see Definition 1.4 of Chapter 1), and so the rates of convergence (p. 120) are controlled by the small-ball probability function of a variable lying in a lower-dimensional space (even if this new variable can also be infinite dimensional) than the original explanatory functional variable X.

We conclude these comments by mentioning some open questions. Indeed, while the standard method for stating rates of convergence in non- or semiparametric functional problems is through small-ball probability considerations, this is not the case in linear functional regression (see Chapter 2), where we usually control rates of convergence by means of conditions on the eigenvalues of the covariance operator of the functional variable X. While the probabilistic literature on links between small-ball probabilities and eigenvalues of covariance operators has been developed for various specific continuous-time Gaussian processes and specific topologies (see Nazarov and Nikitin (2004), Bronski (2003) and Gao *et al*. (2003) for a few recent references in this area), some additional work is still needed in order to allow for general statistical comparison between the results presented here in non/semiparametric functional statistics and those given in Chapter 2 in parametric functional statistics.

Topological effects on uniform results

If we consider uniform consistency results, topology is even more important. For instance, in Theorem 4.7, the topological structure acts directly on the rates of convergence through the entropy function ψ_S which measures the complexity of the set S on which uniformity is stated. In Ferraty *et al*. (2009) one can find various examples for which the entropy function can be (exactly or asymptotically) expressed, and for which the rate of convergence obtained in Theorem 4.7 is slower than the pointwise one. This confirms what the very form of functional regression estimates based on local weighting (see (4.3)) leads us to expect about the major importance of local features in the functional space E. This point will also be of great importance in practical situations (see Section 4.6.2 below).

On optimality of the rates of convergence

An interesting open question is whether the various rates of convergence presented earlier are optimal or not. More precisely, it would be nice to discover in the near future whether the results on optimal rates of convergence obtained by Stone (1982) in the multivariate setting can be extended in some sense to the functional setting. Even though we may remark (see below for more details) that, if we restrict all the results presented before to the special situation in which $E = \mathbb{R}^p$, we get back to Stones's rates, the general question of optimal rates is still completely open in the functional regression setting.

On density of the functional variable

It is worth noting that another advantage of small-ball probability considerations is that they avoid imposing conditions on the density of the functional variable X. This is a key point for our procedure because it is not clear what we could take, in a general abstract semi-space, as a reasonable measure of reference for defining the notion of density (see, however, the recent contribution by Delaigle and Hall (2010), which proposes a new concept of density for infinite-dimensional variables). On the one hand, this issue makes the proofs much more technical than they would be if we had assumed the existence of a density for X, but from another point of view it allows us to state all the results in a much more general setting (an even more general setting than the normal one in the standard multivariate literature).

Links with the standard multivariate literature

To fix our ideas even further, it may be helpful to look at how the previous results behave in the standard multivariate situation (that is, in the special case when $E = \mathbb{R}^p$ and d is Euclidean distance).

It is easy to see that, in this case, as long as X has a density with respect to the Lebesgue measure on $E = \mathbb{R}^p$, one can show that

$$\phi_x(\epsilon) \sim C_x \epsilon^p,$$

in such a way that all the pointwise results stated before match those in the earlier multivariate literature. This is true in nonparametric estimation (see Theorems 4.1, 4.2, 4.3, 4.4, and 4.5) as well as for additive models (see Theorem 4.11) and for semiparametric modeling (see Theorems 4.9 and 4.10). Furthermore, as discussed just above in Section 4.6.1, all the results presented in this chapter extend the standard multivariate literature to situations in which X has no density with respect to the Lebesgue measure.

We can make a similar observation for the uniform result stated in Theorem 4.7, since it is also very easy to see that in this situation the entropy function can be shown, for any compact set $S \subset \mathbb{R}^p$, to be

$$\psi_S(\epsilon) = \log\left(\frac{C_S}{\epsilon^p}\right).$$

Therefore the leading dispersion term in Theorem 4.7 can be rewritten as

$$\sqrt{\frac{\psi_S\left(\frac{\log n}{n}\right)}{n\phi_S(h)}} \sim C\sqrt{\frac{\log n}{nh^p}},$$

(p. 122) leading to the same rate of convergence in the pointwise and the uniform cases (compare Theorems 4.1 and 4.7), as is well known in standard multivariate nonparametric statistics.

4.6.2 On practical issues in functional regression

It is not our purpose here to provide a long discussion of practical issues. It is worth noting, however, that the nonparametric methodologies have been succesfully applied in many situations. The reader may find a free online package, including S/R+ routines, guidelines for users, and various real case studies in Ferraty and Vieu (2006*b*). These practical studies confirm the general theoretical comments previously given, at least on two points. The first point is the importance of the semi-metric; routines for various families of semi-metrics, as well as empirical ideas for chosing them, are also given in the package. The second point concerns the local features of functional data and the importance of data-driven location-adaptive selection procedures for the bandwidth. The various case studies presented in this package highlight the very good behavior of the local cross-validation procedure presented earlier in Section 4.2.4, as well as the kNN estimate presented there, because both of these approaches fully take into acount local features of the problem.

With regard to the other models (additive, partial linear functional, and single functional index model) the situation is not as developed and no similar package is yet available, but the reader may find many case studies in the various papers cited before on these topics (see Sections 4.3.2, 4.3.3, and 4.4). The same applies for the testing methodologies

described in Section 4.5, for which there remain quite a lot of open problems before a completely automatic statistical package can be presented.

4.6.3 Using kernels in other functional problems

As in standard multivariate nonparametric estimation, kernel smoothing ideas are not only of importance in regression but are also important in any other problem that involves the nonparametric estimation of some functional object. This chapter has focused on regression problems, but it is worth noting that recent work has been devoted to the study of other kernel-type estimators involving functional variables. Kernel estimation of the conditional distribution function of a real variable Y given a functional one X (with a natural application to conditional functional quantile estimation) has been investigated previously in Ferraty *et al*. (2006) and revisited in Ezzahrioui and Ould-Saïd (2008*a*) (see also Ferraty and Vieu (2006, Chapter 6)). Kernel estimation of the conditional density function of a real variable Y given a functional one X (with a natural application to conditional functional mode (p. 123) estimation) has been investigated by Ferraty *et al*. (2006) and revisited in several papers, including Dabo-Niang and Laksaci (2007) and Ezzahrioui and Ould-Saïd (2008*b*) (see also Ferraty and Vieu (2006, Chapter 6)). Kernel estimation of the conditional hazard function of a real variable Y given a functional one X has been investigated by Quintela del Rio (2008) and Ferraty *et al*. (2008). Note that kernel ideas have also been investigated for estimating the density of functional variables by Dabo-Niang (2004) and Delaigle and Hall (2010), while unsupervised classification problems involving functional variables are considered by Dabo-Niang *et al*. (2006) (see also Ferraty and Vieu (2006, Chapter 9)).

To conclude this discussion, let us mention that all the techniques that use kernel estimation can be investigated by relaxing the independence conditions on the statistical sample into some kind of dependence structure, with the principal aim of mating the methodology directly usable in time-series analysis. For functional variables, this approach began with the contribution by Ferraty *et al*. (2002), and much recent work has been developed with this goal. The dependent extensions of all the work presented in this chapter are the topic of Chapter 5.

4.7 Conclusion

This chapter has presented, through selected asymptotic results, the main recent developments in kernel methods for regression analysis with a functional covariate, including nonparametric, semiparametric, and additive kernel estimation. It has also presented a survey of the literature in this field. Except for the additive modeling section, which contains a few new results, all the other results presented here are taken from the

existing literature, some of them being slightly modified in order to maintain a common framework for all the various sections.

This chapter is presented in the context of the current infatuation with the development of statistical methods for functional data analysis; this is attested not only by the wide variety of other contributions in this book, but also by recent special issues of various top-level statistical journals (see, for instance, Davidian *et al*. (2004), González-Manteiga and Vieu (2007), Valderrama (2007), and Ferraty (2010), and by the recent books of Ramsay and Silverman (1997, 2002, 2005) and Ferraty and Vieu (2006*a*). The reader may also refer to the volume edited by Dabo Niang and Ferraty (2008), which is especially devoted to contributions given at the first international workshop on functional and operatorial statistics (IWFOS08) held in Toulouse in June 2008. We hope that this contribution will help readers to discover the main specificities of functional kernel regression and that it will contribute to motivating further advances in this challenging and active field of modern statistics.

Acknowledgements

We wish to acknowledge all those who have actively participated in the activities (through meetings, seminars, or just informal discussions) of the working group STAPH in Toulouse. This group intends to develop all the functional features of modern statistics, and its activities have obviously played an important role in our own research and have therefore had a great impact on this chapter. All the activities of this group are available online (see Staph (2009)).

Finally we wish to address specific thanks to German Aneiros-Pérez and Laurent Delsol. German Aneiros-Pérez is a top-level specialist on partial linear models and he has helped with great efficiency in the writing of Section 4.3.3 on this topic. Similarly, Laurent Delsol is a pioneer in the field of structral regression testing with functional variables and he has been of great help in producing Section 4.5 of this chapter.

References

Abraham, C., Biau, G., Cadre, B. (2006). On the kernel rule for function classification. *Ann. Inst. Statist. Math.*, **58**, 619–33.

Ait Saidi A., Ferraty, F., Kassa, R. (2005). Single functional index model for time series. *Rom. J. Pure & Applied Math.*, **50**, 321–330.

Ait-Saïdi, A., Ferraty, F., Kassa, R. and Vieu, P. (2008). Cross-validated estimations in the single-functional index model. *Statistics*, **42**, 475–494.

Aneiros-Pérez, G. (2002). On bandwidth selection in partial linear regression models under dependence. *Statist. Probab. Lett.*, **57**, 393–401.

Aneiros-Pérez, G., Cao, R., Vilar-Fernadez, J. (2008). Functional method for time series prediction: a nonparametric approach. In *Proceedings of IASC-2008* (Mizuta, M., Nakaro, J. eds), 91–100. Yokohama, Japan.

Aneiros-Pérez, G., Cardot, H., Estévez-Pérez, G., Vieu, P. (2006). Maximum ozone concentration forecasting by functional nonparametric approaches. *Environmetrics*, **15**, 675–85.

Aneiros-Pérez, G., Gonzàlez-Manteiga,W., Vieu, P. (2004). Estimation and testing in a partial linear regression model under long-memory dependence. *Bernoulli*, **10**, 49–78.

Aneiros-Pérez, G., Quintela del rio, A. (2001). Asymptotic properties in partial linear models under dependence. *Test*, **10**, 333–55.

Aneiros-Pérez, G., Quintela del rio, A. (2002). Plug-in bandwidth choice in partial linear models with autoregressive errors. *J. Statist. Plann. Inference*, **100**, 23–48.

Aneiros-Pérez, G., Vieu, P. (2006). Semi-functional partial linear regression. *Statist. Probab. Lett.*, **76**, 1102–10.

Aneiros-Pérez, G., Vieu, P. (2008). Nonparametric time series prediction: a semi-functional partial linear modeling. *J. Multivariate Anal.*, **99**, 834–57.

Baíllo, A., Grané, A. (2008). Local linear regression for functional predictor and scalar response. In *Functional and Operatorial Statistics* (Dabo-Niang, S., Ferraty, F., eds), 47–52. Physica-Verlag, Heidelberg.

(p. 125) Barrientos-Marin, J. (2007). Some practical problems of recent nonparametric procedures: testing, estimation and application. PhD Thesis, Univ. of Alicante, Spain.

Barrientos-Marin, J., Ferraty, F., Vieu, P. (2010). Locally modelled regression and functional data. *J. Nonparametr. Statist.*, **22**, 617–632.

Benhenni, K., Ferraty, F., Rachdi,M., Vieu, P. (2007). Locally smoothing regression with functional data. *Computational Statistics*, **22**, 353–70.

Benhenni, K., Hedli-Griche, S., Rachdi, M., Vieu, P. (2008). Consistency of the regression estimator with functional data under long memory conditions. *Statist. Probab. Lett*, **78**, 1043–9.

Biau, G., Bunea, F., Wegkamp, M.H. (2005). Functional classification in Hilbert spaces. *IEEE Trans. Inform. Theory*, **51**, 2163–72.

Bickel, P.J., Klaassen, C.A., Ritov, Y., Wellner, J.A. (1993) *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore.

Boj, E., Delicado, P., Fortiana, J. (2008). Local linear functional based on weighted-based regression. In *Functional and Operatorial Statistics* (Dabo-Niang, S., Ferraty, F., eds), 57–64. Physica-Verlag, Heidelberg.

Bronski, J. (2003). Small ball constants and tight eigenvalue asymptotics for fractional Brownian motions. *J. Theoret. Probab.*, **16**, 87–100.

Burba, F., Ferraty, F., Vieu, P. (2008). Convergence de l'estimateur des k plus proches voisins en régression pour variables fonctionnelles. *C.R.A.S., Série I*, **336**, 339–42.

Burba, F., Ferraty, F., Vieu, P. (2009). k-nearest neighbour method in functional nonparametric regression. *J. Nonparametr. Stat.*, **21** (4), 453–469.

Cérou, F., Guyader, A. (2006). Nearest neighbor classification in infinite dimension. *ESAIM: Probab. Stat.*, **10**, 340–355.

Chen, H. (1988). Convergence rates for parametric components in a partially linear model. *Ann. Statist.*, **16**, 136–147.

Collomb, G. (1976). Estimation non-paramétrique de la régression. PhD Thesis, Univ. Paul Sabatier, Toulouse, France.

Collomb, G. (1980). Estimation de la régression par la méthode des k points les plus proches avec noyau: quelques propriétés de convergence ponctuelle. In *Nonparametric Asymptotic Statistics*, Lecture Notes in Mathematics, **821**, 159–75. Springer, Berlin.

Collomb, G. (1981). Estimation non-paramétrique de la régression: revue bibliographique. *Internat. Statist. Rev.*, **49**, 75–93.

Collomb, G. (1984). Propriétés de convergence presque complète du prédicteur la régression. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, **66**, 441–60.

Collomb, G. (1985). Nonparametric regression: an up-to-date bibliography. *Statistics*, **16**, 309–24.

Crambes, C., Delsol, L., Laksaci, A. (2008). Robust nonparametric estimation for functional data. *J. Nonparametr. Stat.*, **20**, 573–98.

Dabo-Niang, S. (2004). Kernel density estimator in an infinite dimensional space with a rate of convergence in the case of diffusion process. *Appl. Math. Lett.*, **17**, 381–6.

Dabo-Niang, S., Ferraty, F. (2008). *Functional and Operatorial Statstics*. Springer, New York.

Dabo-Niang, S., Ferraty, F., Vieu, P. (2006). Mode estimation for functional random variable and its application for curves classification. *Far East J. Theor. Stat.*, **18**, 93–119.

Dabo-Niang, S., Laksaci., A. (2007). Estimation non paramétrique du mode conditionnel pour variable explicative fonctionnelle. *C. R. Math. Acad. Sci. Paris.*, **344**, 49–52.

(p. 126) Davidian, M., Lin, X., Wang, J.L. (2004). Introduction to the emerging issues in longitudinal and functional data analysis (with discussion). *Statistica Sinica*, **14**, 613–19.

Delaigle, A., Hall, P. (2010). Defining probability density for a distribution of random functions. *Annals of Statistics*, **38**, 1171–93.

Delecroix, M., Hristache, M., Patilea, V. (2006). Semiparametric *M*-estimation in single-index regression. *J. Statist. Plann. Inference*, **136**, 730–69.

Delsol, L. (2007). Régréssion non-paramétrique fonctionnelle: expressions asymptotiques des moments. *Revue de l'Inst. Statist. Univ Paris*, **LI**(3), 43–67.

Delsol, L. (2008*a*). Advances on asymptotic normality in nonparametric functional time series analysis. *Statistics*, **43**(1), 13–33.

Delsol, L. (2008*b*). Tests de structure en régression sur variable fonctionnelle. *C. R. Math. Acad. Sci. Paris*, **346**, 343–6.

Delsol, L. (2008*c*). Régression sur variable fonctionnelle: estimation, tests et applications. PhD Thesis, Université Paul Sabatier, Toulouse, France.

Delsol L., Ferraty F., and Vieu P. (2010). Structural test in regression on functional variables (submitted).

Ezzahrioui, M. and Ould-Saïd, E. (2008*a*). Asymptotic normality of the kernel estimator of conditional quantiles in a normed space. *Far East J. Theor. Stat.*, **25**, 15–38.

Ezzahrioui, M. and Ould-Saïd, E. (2008*b*). Asymptotic results of a nonparametric conditional quantile estimator for functional time series. *Comm. Statist. Theory Methods*, **37**, 2735–2759.

Ferraty, F., Goia, A., Vieu, P. (2002). Functional nonparametric model for time series: a fractal approach for dimension reduction. *Test*, **11**, 317–44.

Ferraty, F., Laksaci, A., Vieu, P. (2006). Estimating some characteristics of the conditional distribution in nonparametric functional models. *Statistical Inference for Stochastic Processes*, **9**, 47–76.

Ferraty, F., Laksaci, A., Vieu, P., Tadj, A. (2009). Rate of uniform consistency for nonparametric estimates with functional variables. *J. Stat. Plan. Infer.*, **140**, 335–352.

Ferraty, F., van Keilegom, I. and Vieu, P. (2010). On the Validity of the Bootstrap in Non-Parametric functional regression. *Scand. J. Statist.*, **37**, 286–306.

Ferraty, F., Mas, A., Vieu, P. (2007). Nonparametric regression on functional data: inference and practical aspects. *Austral. New Zealand J. Statist.*, **49**(3), 267–86.

Ferraty, F., Peuch, A., Vieu, P. (2003). Modèle à indice fonctionnel simple. *Comptes Rendus Académie Sciences Paris*, **336**, 1025–8.

Ferraty, F., Rabhi, A., Vieu, P. (2008). Estimation non-paramétrique de la fonction de hasard avec variable explicative fonctionnelle. *Rom. J. Pure & Applied Math.*, **52**, 1–18.

Ferraty, F., Vieu, P. (2000). Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés. *Comptes Rendus Académie Sciences Paris*, **330**, 139–42.

Ferraty, F., Vieu, P. (2002). The functional nonparametric model and application to spectrometric data. *Computational Statistics*, **17**, 545–64.

Ferraty, F., Vieu, P. (2003). Curves discrimination: a nonparametric functional approach. *Comp. Statist. and Data Anal.*, **44**, 161–73.

Ferraty, F., Vieu, P. (2004). Nonparametric models for functional data, with application in regression, time-series prediction and curve discrimination. *J. Nonparametr. Stat.*, **16**, 111–125.

(p. 127) Ferraty, F., Vieu, P. (2006*a*). *Nonparametric Functional Data Analysis. Theory and Practice*. Springer, New York.

Ferraty, F., Vieu, P. (2006*b*). *NPFDA: R/S+ routines*. Free access online at **http://www.math.univ-toulouse.fr/staph/npfda/**.

Ferraty, F., Vieu, P. (2008) Erratum to: “Nonparametric models for functional data, with application in regression, time-series prediction and curve discrimination” (*J. Nonparametr. Stat*. (2004), 16 (1–2), 111–25). *J. Nonparametr. Stat.*, **20**, 187–9.

Ferraty, F., Vieu, P. (2009). On functional regression modelling. *Comp. Statist. Data Analysis*, **53**, 1400–13.

Ferraty, F. (2010). Special Issue: Statistical Methods and Problems in Infinite-dimensional Spaces (Ed.). *J. Mult. Anal.*, **101**, 305–490.

Gao, F., Hannig, J., Torcaso, F. (2003). Integrated Brownian motions and exact L_2-small balls. *Ann. Probab.*, **31**, 1320–37.

González Manteiga, W., Vieu, P. (2007). Introduction to the special issue on statistics for functional data. *Comp. Statist. Data Analysis*, **51**, 4788–92.

Györfi, L., Kohler, M., Krzyzak, A., Walk, H. (2002). *A distribution-free Theory of Nonparametric Regression.* Springer, New York.

Härdle, W. (1990). *Applied Nonparametric Regression*. Oxford University Press, Oxford.

Härdle, W., Bowman, A.W. (1988). Bootstrapping in nonparametric regression: local adaptive smoothing and confidence bands. *J. Amer. Statist. Assoc.*, **83**, 102–10.

Härdle, W., Hall, P. (1993). On the backfitting algorithm for additive regression models. *Statist. Neerlandica*, **47**, 43–57.

Härdle, W., Hall, P., Ichimura, H. (1993). Optimal smoothing in single index models. *Ann. Statist.*, **21**, 157–78.

Härdle, W., Liang, H., Gao, J. (2000). *Partially Linear Models*. Physica, Heidelberg.

Härdle, W., Marron, J.S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.*, **13**, 1465–81.

Härdle, W., Marron, J.S. (1991). Bootstrap simultaneous error bars in nonparametric regression. *Ann. Statist.*, **19**, 778–96.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.*, **21**, 1926–1947.

Härdle, W., Müller, M., Sperlich, S., Werwatz, A. (2004). *Nonparametric and Semi-parametric Models.* Springer, New York.

Hart, J. (1997). Nonparametric smoothing and lack-of-fit tests. Springer Series in Statistics. Springer-Verlag, New York.

Hastie, T., Tibshirani, R. (1986). Generalized additive models (with discussion). *Statist. Sci.*, **1**, 297–318.

Hastie, T., Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Horowitz, J. (1998). *Semiparametric Methods in Econometrics*. Lecture Notes in Statistics, **131**. Springer, New York.

Horowitz, J., Klemelä, J., Mammen, E. (2006). Optimal estimation in additive regression models. *Bernoulli*, **12**, 271–98.

Hristache,M., Juditsky, A., Spokoiny, V. (2001). Direct estimation of the index coefficient in the single index model. *Ann. Statist.*, **29**, 595–623.

(p. 128) Kolmogorov, A., Tikhomirov, V. (1959). ϵ-entropy and ϵ-capacity. (in Russian). *Uspekhi Mat. nauk.*, **14**, 3–86.

Liang, H. (2000). Asymptotic normality of parametric part in partly linear model with measurement errors in the nonparametric part. *J. Statist. Plann. Inference*, **86**, 51–62.

Mammen, E. (2000). Resampling methods in nonparametric regression. In *Smoothing and Regression. Approaches, Computation, and Application* (M. Schimek, ed.), 425–50. John Wiley & Sons, New York.

Mammen, E., Park, B. (2006). A simple smooth backfitting method for additive models. *Ann. Statist.*, **34**, 2252–71.

Marron, J.S., Härdle, W. (1986). Random approximations to some measures of accuracy in nonparametric curve estimation. *J. Multivar. Anal.*, **20**, 91–113.

Masry, E. (2005). Nonparametric regression estimation for dependent functional data: asymptotic normality. *Stochastic Process. Appl.*, **115**, 155–77.

Müller, H.G. (2008). Functional additive modelling. In *Abstracts of ISNI'2008 Meeting, Vigo, Spain*. Available at **http://www.isni2008.com**.

Nazarov, A., Nikitin, Y. (2004). Exact L_2-small ball behavior of integrated Gaussian processes and spectral asymptotics of boundary value problems. *Probab. Theory Related Fields*, **129**, 469–94.

Pelletier, B. (2006). Non-parametric regression estimation on closed Riemannian manifolds. *J. of Nonparametric Statistics*, **18**, 57–67.

Quintela del Rio, A. (2008). Hazard function given a functional variable: non-parametric estimation under strong mixing conditions. *J. Nonparametr. Stat.*, **20**, 413–30.

Rachdi, M., Vieu, P. (2007). Nonparametric regression for functional data: automatic smoothing parameter selection. *J. Statist. Plann. Inference*, **137**(9), 2784–801.

Ramsay, J.O., Silverman, B.W. (1997). *Functional Data Analysis.* Springer, New York.

Ramsay, J.O., Silverman, B.W. (2002). *Applied Functional Data Analysis: Methods and Case Studies.* Springer, New York.

Ramsay, J.O., Silverman, B.W. (2005). *Functional Data Analysis* (second edition). Springer, New York.

Sarda, P., Vieu, P. (2000). Kernel regression. In *Smoothing and Regression. Approaches, Computation, and Application* (M. Schimek, ed.), 43–70. John Wiley & Sons, New York.

Schick, A. (1996). Root-n consistent estimation in partly linear regression models. *Statist. Probab. Lett.*, **28**, 353–358.

Schimek, M. (2000). *Smoothing and Regression. Approaches, Computation, and Application*. John Wiley & Sons, New York.

Schimek, M., Turlach, B. (2000). Additive and generalized additive models. In *Smoothing and Regression. Approaches, Computation, and Application* (M. Schimek, ed.), 277–328. John Wiley & Sons, New York.

Staph. (2009). Groupe de travail en statistique fonctionnelle et oṕeratorielle, Toulouse, France. Activities online at **http://www.math.univ-toulouse.fr/staph/**.

Stone, C. (1982). Optimal global rates of convergence for nonparametric estimates. *Ann. Statist.*, **10**, 1040–53.

Stone, C. (1985). Additive regression and other nonparametric models. *Ann. Statist.*, **13**, 689–705.

Stone, C. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.*, **14**, 590–606.

(p. 129) Stone, C. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.*, **22**, 118–84.

Sperlich, S., Härdle, W., Aydinh, G. (2006). *The Art of Semiparametrics*. Physica, Heidelberg.

Speckman, P. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. Ser. B*, **50**, 413–436.

Tong, X.W., Cui, H.J., Yu, P. (2008). Consistency and normality of Huber-Dutter estimators for partial linear model. *Sci. China Ser. A*, **51**, 1831–42.

Valderrama, M. (2007). Introduction to the special issue on modelling functional data in practice. *Comput. Statist.*, **22**, 331–4.

Vieu, P. (1991). Nonparametric regression: optimal local bandwidth choice. *J. R. Statist. Soc.*, **53**, 453–64.

Frédéric Ferraty

Frédéric Ferraty, Institut de Mathématiques de Toulouse, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex 9, France, frederic.ferraty@math.univ-toulouse.fr

Philippe Vieu

Philippe Vieu, Institut de Mathématiques de Toulouse, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex 9, France, philippe.vieu@math.univtoulouse.fr