UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 07-61
Statistic and Econometric Series 15
August 2007

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34-91) 6249849

LOCAL LINEAR REGRESSION FOR FUNCTIONAL PREDICTOR AND SCALAR RESPONSE

Amparo Baíllo[1] and Aurea Grané[2]

Abstract

The aim of this work is to introduce a new nonparametric regression technique in the context of functional covariate and scalar response. We propose a local linear regression estimator and study its asymptotic behaviour. Its finite-sample performance is compared with a Nadayara-Watson type kernel regression estimator via a Monte Carlo study and the analysis of two real data sets. In all the scenarios considered, the local linear regression estimator performs better than the kernel one, in the sense that the mean squared prediction error and its standard deviation are lower.

Keywords: Functional data, nonparametric smoothing, local linear regression, kernel regression, Fourier expansion, cross-validation.
AMS Classification 2000: 62G08 (62G30)

[1] Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049 Madrid, Spain. E-mail: amparo.baillo@uam.es
[2] Corresponding autor. Statistics Department, Universidad Carlos III de Madrid, 28903 Getafe (Madrid), Spain. E-mail: agrane@est-econ.uc3m.es

Research partially supported by the IV PRICIT program titled *Modelización Matemática y Simulación Numérica en Ciencia y Tecnología* (SIMUMAT), by Spanish grant MTM2004-00098 and by MTM2006-09920 (Ministry of Education and Science-FEDER)

Local linear regression for functional predictor and scalar response*

Amparo Baíllo
Universidad Autónoma de Madrid
28049 Madrid (Spain)

Aurea Grané
Universidad Carlos III de Madrid
28903 Madrid (Spain)

Abstract

The aim of this work is to introduce a new nonparametric regression technique in the context of functional covariate and scalar response. We propose a local linear regression estimator and study its asymptotic behaviour. Its finite-sample performance is compared with a Nadayara-Watson type kernel regression estimator via a Monte Carlo study and the analysis of two real data sets. In all the scenarios considered, the local linear regression estimator performs better than the kernel one, in the sense that the mean squared prediction error and its standard deviation are lower.

Keywords: Functional data, nonparametric smoothing, local linear regression, kernel regression, Fourier expansion, cross-validation.
AMS Classification 2000: 62G08 (62G30)

1 Introduction

In the last years there has been an increasing interest in the analysis, modelling and use of functional data. The observation of functional variables has become usual due, for instance, to the development of measuring instruments that allow to observe variables (time or space dependent) at finer and finer resolution. Then it seems natural to assume that the data are actually observations from a random variable taking values in a functional space.
There is nowadays a large number of fields where functional data are collected: environmetrics, medicine, finance, pattern recognition,... This has led to the extension of finite dimensional statistical techniques to the infinite dimensional data setting. A classical statistical problem is that of regression: studying the relationship between two observed variables with the aim to predict the value of the response variable when a new value of the auxiliary one is observed.

*Research partially supported by the IV PRICIT program titled *Modelización Matemática y Simulación Numérica en Ciencia y Tecnología* (SIMUMAT), by Spanish grant MTM2004-00098 and by MTM2006-09920 (Ministry of Education and Science-FEDER).

In this work we consider the regression problem with functional auxiliary variable X taking values in $L^2[0,T]$ and scalar response Y. Without loss of generality from now on we assume that $T=1$. A sample of random elements $(X_1,Y_1),\ldots,(X_n,Y_n)$ is observed, where the X_i are independent and identically distributed as X and only recorded on an equispaced grid $t_1,t_2,\ldots,t_N$ of $[0,1]$ whose internodal space is $w=1/N$. It is assumed that the response variable Y has been generated as

$$Y_i = m(X_i) + \epsilon_i, \quad i=1,\ldots,n \tag{1}$$

and that the errors ϵ_i are independent, with zero mean and finite variance σ^2_ϵ, and are also independent from any of the X_j.

In the context of regression with functional data a common assumption is that $m(x)$ is a linear function of x. The linear model has been studied in a large number of works: see, e.g., Cardot, Ferraty and Sarda (2003), Ramsay and Silverman (2005), Cai and Hall (2006), Hall and Horowitz (2007) and references therein. Extensions of the linear model have been considered, for instance, by James (2002), Ferré and Yao (2003), Cardot and Sarda (2005) or Müller and Stadtmüller (2005). However, when dealing with functional data, it is difficult to gain an intuition on whether the linear model is adequate at all or which is the parametric model that would best fit the data, since graphical techniques are of scarce use here. Nonparametric techniques come in naturally in this situation.

Here we are interested in estimating the regression function m in a nonparametric fashion. Nonparametric functional regression estimation has already been considered, for instance, by Ferraty and Vieu (2000, 2006), who study a kernel estimator of Nadaraya-Watson type

$$\hat{m}_K(x) := \frac{\sum_{i=1}^n Y_i K_h(\|X_i - x\|)}{\sum_{i=1}^n K_h(\|X_i - x\|)}, \tag{2}$$

where $K_h(\cdot) := h^{-1}K(\cdot/h)$, $h=h_n$ is a positive smoothing parameter and $\|\cdot\|$ denotes the $L^2[0,1]$ norm. From now on K is assumed to be an asymmetrical decreasing kernel function. Observe that the estimator $\hat{m}_K(x)$ is the value of a minimizing the weighted squared error

$$\mathrm{WSE}_0(x) = \sum_{i=1}^n (Y_i - a)^2 K_h(\|X_i - x\|).$$

Thus the kernel estimator given by (2) is locally approximating m by a constant (a zero-degree polynomial). However, in the context of nonparametric regression with finite-dimensional auxiliary variables, local polynomial smoothing has become the "golden standard" (see Fan 1992, Fan and Marron 1993, Wand and Jones 1995). Local polynomial smoothing at a point x fits a polynomial to the pairs (X_i,Y_i) for those X_i falling in a neighbourhood of x determined by a smoothing parameter h. In particular, the local linear regression estimator locally fits a polynomial of degree one. Here we plan to extend the ideas of local linear smoothing to the functional data setting, giving a first answer to the *open question 5* in Ferraty and Vieu (2006): "How can the local polynomial ideas be adapted to infinite dimensional settings?"

Section 2 contains our proposal for obtaining the local linear regression estimator in the context of functional auxiliary variable and scalar response. Section 3 is devoted to the study of the

asymptotic behaviour of this estimator. In Section 4 we compare the finite-sample performance of the kernel and the local linear regression estimators via a Monte Carlo study. In Section 5 this comparison is carried out through the analysis of two real data sets. Finally the Appendix contains some technical results together with the proof of the theorem stated in Section 3.

2 Local linear smoothing for functional data

Local polynomial smoothing is based on the assumption that the regression function m is smooth enough to be locally well approximated by a polynomial. Thus from now on we will assume that m is differentiable in a neighbourhood of x and, consequently, for every z in this neighbourhood we may approximate $m(z)$ by a polynomial of degree 1, that is, $m(z) \simeq m(x) + \langle b, z - x \rangle$, where $b = b(x) \in L^2[0,1]$ and $\langle\,,\,\rangle$ denotes the $L^2[0,1]$ inner product (see Cartan 1967 for a comprehensive review on this subject). In particular, we have $m(X_i) \simeq m(x) + \langle b, X_i - x \rangle$ for every sample point in a neighbourhood of x. Then the weighted squared error

$$\mathrm{WSE}(x) := \sum_{i=1}^{n} (Y_i - m(X_i))^2 \, K_h(\|X_i - x\|)$$

may be approximated by

$$\sum_{i=1}^{n} \left(Y_i - (m(x) + \langle b, X_i - x \rangle)\right)^2 \, K_h(\|X_i - x\|).$$

A first naive answer to the question posed by Ferraty and Vieu (2006) would come from optimizing, in a and $b \in L^2[0,1]$, the following error expression

$$\mathrm{WSE}_1(x) = \sum_{i=1}^{n} (Y_i - (a + \langle b, X_i - x \rangle))^2 \, K_h(\|X_i - x\|). \tag{3}$$

Once the value $\hat{a}$ of a minimizing (3) were found, we would take $\hat{m}_{LL}(x) = \hat{a}$ as the local linear estimator of $m(x)$, the regression function at x (see Fan 1992).

2.1 Smoothing the functional parameter b

The minimization of WSE_1 may be achieved by a "wiggly" $\hat{b}$ that forces $\hat{m}_{LL}(x)$ to adapt to all the data points in a neighbourhood of x (see Chapter 15 of Ramsay and Silverman 2005 for a similar reasoning in the context of linear regression). Cai and Hall (2006) express the same idea stating that optimizing in b is an infinite-dimensional problem. In order to reduce the dimension of parameter b it is necessary an intermediate step of smoothing or regularization. A standard approach in the functional linear regression setting is to expand b and X_i using an orthonormal basis $\{\phi_j\}_{j \geq 1}$ of $L^2[0,1]$,

$$b = \sum_{j=1}^{\infty} b_j \phi_j \qquad \text{and} \qquad X_i - x = \sum_{j=1}^{\infty} c_{ij} \phi_j \tag{4}$$

with $b_j = \langle b, \phi_j \rangle$ and $c_{ij} = \langle X_i - x, \phi_j \rangle$. The system $\{\phi_j\}_{j \geq 1}$ can be, for example, the Fourier trigonometric basis (see Ramsay and Silverman 2005) or the eigenfunctions of the covariance operator of X (see Cai and Hall 2006). If we substitute (4) in expression (3), Parseval's theorem yields

$$\text{WSE}_1 = \sum_{i=1}^{n} \left(Y_i - \left(a + \sum_{j=1}^{\infty} b_j c_{ij} \right) \right)^2 K_h(\|X_i - x\|).$$

The regularization step consists in truncating the series at a certain cut-off J. Thus we will minimize the following approximation to WSE_1

$$\text{AWSE}_1 := \sum_{i=1}^{n} \left(Y_i - \left(a + \sum_{j=1}^{J} b_j c_{ij} \right) \right)^2 K_h(\|X_i - x\|). \tag{5}$$

Adding a penalization term that prevents b from oscillating too much is another possible regularization procedure (see Ramsay and Silverman 2005). However simulation studies analogous to the ones presented in Section 4 reveal that, in the context of this work, the penalization procedure performs worse than the truncation one. The question of how to choose (in an automatic way) the optimal or at least a "good" J in practice is addressed in Section 4.

2.2 Estimating the regression function

In order to find the values of a and b_j, for $j = 1, \ldots, J$, minimizing AWSE_1, we differentiate the expression given in (5) with respect to these parameters and equate the derivatives to zero. As a result, assuming that $\mathbf{C'WC}$ is a nonsingular matrix, we obtain

$$\begin{pmatrix} \hat{a} \\ \hat{b}_1 \\ \vdots \\ \hat{b}_J \end{pmatrix} = (\mathbf{C'WC})^{-1}\mathbf{C'WY},$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, $\mathbf{W} = \text{diag}(K_h(X_1 - x), \ldots, K_h(X_n - x))$ and

$$\mathbf{C} = \begin{pmatrix} 1 & c_{11} & \ldots & c_{1J} \\ 1 & c_{21} & \ldots & c_{2J} \\ \vdots & & & \vdots \\ 1 & c_{n1} & \ldots & c_{nJ} \end{pmatrix}.$$

Finally, our proposal for the local linear estimator of $m(x)$ is

$$\hat{m}_{LL}(x) = \hat{a} = \mathbf{e}_1'(\mathbf{C'WC})^{-1}\mathbf{C'WY}, \tag{6}$$

where $\mathbf{e}_1$ is the $(J+1) \times 1$ vector having 1 in the first entry and 0's in the rest.

3 Asymptotic behaviour

The aim of this section is to state a consistency result for the local linear estimator introduced in expression (6) of Section 2. More concretely we are interested in conditions under which the mean squared error

$$E((\hat{m}_{LL}(x) - m(x))^2 | X_1, \ldots, X_n) \tag{7}$$

converges to 0 as $n \to \infty$ and $J \to \infty$. From now on we will denote by $\mathbf{X}$ the sample $X_1, \ldots, X_n$ appearing in the conditional expectation and variance.
Let us first state some hypotheses to be used in this section.

(A1) The kernel $K : \mathbb{R} \to \mathbb{R}^+$ satisfying $\int K = 1$ is a kernel of type I if there exist two real constants $0 < c_I < C_I < \infty$ such that $c_I 1_{[0,1]} \leq K \leq C_I 1_{[0,1]}$.

The following condition states that the probability of observing X in any neighbourhood of x is not null (see Ferraty and Vieu 2006).

(A2) For any $\epsilon > 0$, the small ball probability $\varphi_x(\epsilon) := P\{\|X - x\| < \epsilon\}$ is strictly positive.

Conditions (A3) and (A4) are used to bound the error made when approximating X_i and x by the trigonometric series (4) truncated at the cut-off J (see Zygmund 1988).

(A3) With probability one, any trajectory $X(\cdot, \omega)$ of X has derivative of ν-th order which is uniformly bounded on $[0, 1]$ by a constant independent of ω.

(A4) The element x has derivative of ν-th order which is uniformly bounded on $[0, 1]$.

The asymptotic behaviour of the local linear regression estimator defined in formula (6) is studied in the following result, whose proof is detailed in the Appendix. From this proof we can see that if m is a linear functional then the local linear estimator is unbiased. This fact was already observed by Fan (1992) and Ruppert and Wand (1994) in the case that X has finite dimension.

Theorem: *Let the assumptions (A1)–(A4) hold. Assume also that $h \to 0$ and $n\varphi_x(h) \to \infty$ as $n \to \infty$. If the regression function m is differentiable in a neighbourhood of x and twice differentiable at x with continuous second derivative, then*

$$E((\hat{m}_{LL}(x) - m(x))^2 | \mathbf{X}) = \left(O(J^{-\nu}) + O_P(h^2)\right)^2 + O_P\left((n\varphi_x(h))^{-1}\right).$$

In the following corollary we obtain rates of convergence to 0 (in probability) for the mean squared error, under an assumption on the fractal dimensionality of the probability measure of X. More precisely, the random element $X \in L^2[0, 1]$ is said to be of fractal order τ with respect to $\|\cdot\|$ if $\varphi_x(\epsilon) = O(\epsilon^\tau)$ as $\epsilon \to 0$. This definition was introduced in the context of kernel regression estimation by Ferraty and Vieu (2000) and further explored by Ferraty and Vieu (2006).

Corollary: *Let the assumptions of the theorem hold. If X is of fractal order τ with respect to $\|\cdot\|$, $h = O\left(n^{-1/(4+\tau)}\right)$ and $J = O(h^{-2/\nu})$, then*

$$E((\hat{m}_{LL}(x) - m(x))^2|\mathbf{X}) = O_P\left(n^{-4/(4+\tau)}\right).$$

The rates obtained in the corollary agree with the asymptotic results for the kernel estimator appearing in Ferraty and Vieu (2006), p. 208, in the sense that the more concentrated X is around x (as measured by the small ball probability $\varphi_x(h)$), the faster the local linear estimator will converge to the true regression function.

4 Simulations

In this section we compare the finite-sample behaviour of the local linear and the kernel regression estimators, $\hat{m}_{LL}$ and $\hat{m}_K$ respectively, via a Monte Carlo study. The performance of each regression estimator $\hat{m}$ is described by the squared prediction error $\mathrm{SE}(X) := (\hat{m}(X) - m(X))^2$. More concretely, for $b = 1, \ldots, B$ Monte Carlo trials we generate a sample $X_1^{(b)}, \ldots, X_n^{(b)}$ and a test observation $X^{(b)}$ from X. For each estimator $\hat{m}$ we compute $\hat{m}^{(b)}$, the regression estimator constructed from $X_1^{(b)}, \ldots, X_n^{(b)}$, and $\mathrm{SE}^{(b)}(X^{(b)}) := \left(\hat{m}^{(b)}(X^{(b)}) - m(X^{(b)})\right)^2$. The regression estimators are compared using the mean and standard deviation of $\{\mathrm{SE}^{(b)}(X^{(b)}),\, b = 1, \ldots, B\}$. In the simulations displayed below we have taken $B = 2000$ and $n = 100, 200$.

4.1 Models for the simulations

In this subsection we specify the models used to generate (X, Y). In all the cases considered we have used the same distribution to generate X

$$X(t) = \sum_{j=1}^{50} Z_j 2^{1/2} \cos(j\pi t),$$

where $\{Z_j\}_{1 \le j \le 50}$ are independent random variables and each Z_j follows a normal distribution with mean 0 and variance $\sigma_z^2\, j^{-2}$. To generate the response variable Y we have used the three models described below. In Model 1 and Model 2 we have taken $\sigma_z = 2$ and in Model 3, $\sigma_z = 1$.

Model 1 (Linear regression function): $Y = \langle \beta, X \rangle + \epsilon$, where the "slope" is given by

$$\beta(t) = \sum_{j=1}^{50} j^{-4} 2^{1/2} \cos(j\pi t)$$

and the error ϵ is normally distributed with mean 0 and standard deviation $\sigma_\epsilon = 2$. This linear model was used in the simulation study of Cai and Hall (2006). (See also Baíllo 2007).

Model 2 (Piecewise linear regression function):

$$Y = \begin{cases} \alpha_1 + \langle \beta^{(1)}, X \rangle + \epsilon, & \text{if } \|X\|^2 \le 10, \\ \alpha_2 + \langle \beta^{(2)}, X \rangle + \epsilon, & \text{otherwise,} \end{cases}$$

where $\alpha_1 = 2$, $\alpha_2 = 0$, $\beta_1(t) = \sum_{j=1}^{50} j^{-4} 2^{1/2} \cos(j\pi t)$, $\beta_2(t) = \sum_{j=1}^{50} j^{-5} 2^{1/2} \cos(j\pi t)$ and the error ϵ is normally distributed with mean 0 and standard deviation $\sigma_\epsilon = 2$.

Model 3 (Strictly non-linear regression function): $Y = \|X\|^2 + \epsilon$, where the error ϵ is normally distributed with mean 0 and standard deviation $\sigma_\epsilon = 1$.

4.2 Choosing the cut-off J and the bandwidth h

For both the kernel and the local linear regression estimators we use the asymmetrical Gaussian kernel $K(t) := \sqrt{2/\pi} \exp(-t^2/2)$ for $t \in (0, \infty)$. The kernel bandwidth is chosen via the following cross-validation procedure described in Ferraty and Vieu 2006, p. 101,

$$h_K = \arg\min_h \mathrm{CV}_K(h), \tag{8}$$

where $\mathrm{CV}_K(h) := \sum_{i=1}^n (Y_i - \hat{m}_{K,-i}(X_i))^2$ is a sum of squared residuals and

$$\hat{m}_{K,-i}(x) := \frac{\sum_{j=1, j\neq i}^n Y_j K_h(X_j - x)}{\sum_{j=1, j\neq i}^n K_h(X_j - x)}.$$

Let us now turn to the practical aspects of the local linear estimator. Concerning the basis used in the series expansion we choose the (orthonormal) trigonometric basis

$$\phi_1(t) = 1, \quad \phi_{2r}(t) = 2\sin(2\pi r t), \quad \phi_{2r+1}(t) = 2\cos(2\pi r t), \quad r = 1, 2, \ldots. \tag{9}$$

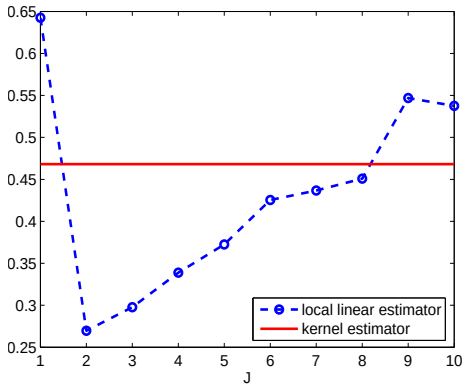
Regarding the cut-off J, it is clear that, as Ramsay and Silverman (2005) point out, the value of J should be low in order to avoid the curse of dimensionality. As a first step, to check how things work, we fix some values $J = 1, 2, 3, \ldots, 10$. Once J is fixed, we choose h via a cross-validation procedure analogous to the one proposed for the kernel estimator, $h_{LL}(J) := \arg\min_h \mathrm{CV}_{LL}(J, h)$, where $\mathrm{CV}_{LL}(J, h) := \sum_{i=1}^n (Y_i - \hat{m}_{LL,-i}(X_i))^2$ and $\hat{m}_{LL,-i}$ is the local linear estimator of m constructed using the sample $\{(X_j, Y_j)\}_{1 \leq j \leq n,\, j \neq i}$ and the first J terms of (9).
In Table 1 and Figure 1 we reproduce the simulation results for Model 1 and $B = 1000$ Monte Carlo samples of size $n = 100$. We have computed the sample mean and the standard deviation (in brackets) of the resulting squared prediction error $\{\mathrm{SE}^{(b)}(X^{(b)}),\, b = 1, \ldots, B\}$ for different values of J. In the graph we have only represented the mean squared error. We have included, just for better comparison, the squared prediction error for the $\hat{m}_K$ estimator (as a solid line), which is computed independently of J. In both, the table and the figure, we can see that certainly the optimal cut-off J should be low: even under the assumption of a linear model, the performance of the local linear estimator is markedly better than that of the kernel estimator only for J =2, 3, 4 or 5.
It is clear, then, the convenience of developing an automatic, data-based procedure of choosing simultaneously the number of terms in the series expansion, J, and the window width, h. In this work we choose both the optimal J and h as follows

$$J_{LL} = \arg\min_J \mathrm{CV}_{LL}(J, h_{LL}(J)) \quad \text{and} \quad h_{LL} = h_{LL}(J_{LL}).$$

Table 1: Error for $\hat{m}_K$ and $\hat{m}_{LL}$ over $B = 1000$ simulations of size $n = 100$ from Model 1.

	$\hat{m}_K$	$\hat{m}_{LL}$								
		$J=1$	$J=2$	$J=3$	$J=4$	$J=5$	$J=6$	$J=7$	$J=8$	$J=9$	$J=10$
Mean SE	0.468	0.643	0.270	0.298	0.339	0.372	0.425	0.437	0.451	0.547	0.538
	(1.310)	(1.726)	(0.850)	(0.641)	(1.159)	(0.933)	(1.357)	(1.173)	(1.146)	(1.719)	(1.162)

Figure 1: Error for $\hat{m}_K$ and $\hat{m}_{LL}$ over $B = 1000$ simulations of size $n = 100$ from Model 1.



4.3 Simulation results

Table 2 contains the mean squared error and the standard deviation (in brackets) for Models 1–3, taking $B = 2000$ Monte Carlo samples of size $n = 100$. The auxiliary variable X was evaluated on $N = 50$ equispaced nodes. The last line in the table displays the resulting mean of the optimal cut-off's J. Table 3 contains analogous results for $B = 2000$ Monte Carlo samples of size $n = 200$ and evaluated at $N = 100$ nodes.

Table 2: Simulation results for $B = 2000$ samples of size $n = 100$ from Models 1–3.

	Model 1		**Model 2**		**Model 3**	
	$\hat{m}_K$	$\hat{m}_{LL}$	$\hat{m}_K$	$\hat{m}_{LL}$	$\hat{m}_K$	$\hat{m}_{LL}$
Mean SE	0.4311	0.3429	0.6043	0.4459	0.5503	0.3636
	(0.9054)	(0.8081)	(1.0456)	(0.9638)	(2.2255)	(0.8922)
Mean J		3.2		3.1		2.9

Observe that the local linear regression estimator performs better than the kernel one: both the mean and the standard deviation are smaller for the former. Other choices for the parameters of the models yield similar results, favourable to the local linear smoother.

On the other hand, Table 2 and Table 3 suggest that, as n increases, the value of the optimal J slightly increases too. This agrees with the theoretical results stated in Section 3, particularly with the corollary, where the optimal J was a slowly increasing function of n.

Table 3: Simulation results for $B = 2000$ samples of size $n = 200$ from Models 1–3.

	Model 1		**Model 2**		**Model 3**	
	$\hat{m}_K$	$\hat{m}_{LL}$	$\hat{m}_K$	$\hat{m}_{LL}$	$\hat{m}_K$	$\hat{m}_{LL}$
Mean SE	0.3017	0.1807	0.4246	0.2553	0.4255	0.2449
	(0.6967)	(0.3923)	(0.9115)	(0.6637)	(2.5469)	(0.8330)
Mean J		3.4		3.3		3.0

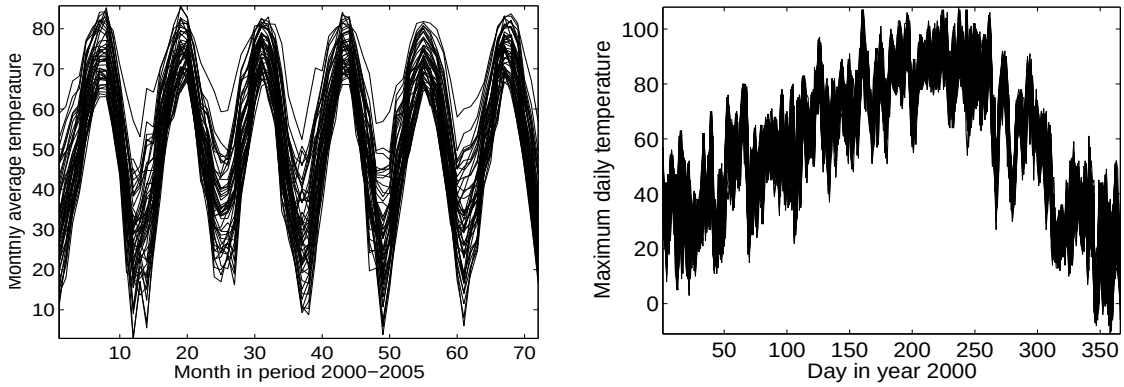
5 Analysis of real data

The aim of this section is to compare the performance of the kernel and local linear regression estimators via the analysis of two real climate data sets from the U.S. National Climatic Data Center web-site (www.ncdc.noaa.gov).

In the first group of data the response variable Y_i is logarithm of the total number of tornados in each U.S. state ($i = 1, \ldots, 48$) along the period 2000-2005. The predictor variable X_i is the monthly average temperature (measured in ºF) in state i in the same period of time. This is of interest, for instance, when assessing the possible consequences, like an increase in the number of extreme climatic events, of an overall increase in the temperatures due to the climatic change. Figure 2 (a) depicts the evolution of the temperature curves.

In the second data set the predictor is the daily maximum temperature (in ºF) recorded in $n = 80$ weather stations from South Dakota in year 2000 (see Figure 2 (b)). The response variable Y_i is the logarithm of the total precipitation in each of the stations during the same year.

Figure 2: (a) Average monthly temperatures in U.S.A. states from 2000 to 2005 and (b) Daily maximum temperatures along year 2000 in 80 stations from South Dakota.



(a) (b)

In Table 4 we have computed (via a cross-validation procedure) the mean squared prediction error and the corresponding standard deviation (in brackets) attained by each of these estimators. Observe that, in this study with real data, the differences between the performance of both regression estimators are stressed, with a considerable reduction of the prediction error when using the local linear estimator.

Table 4: Squared prediction error for (a) tornado data and (b) South Dakota data.

	$\hat{m}_K$	$\hat{m}_{LL}$
Mean SE	1.5127	0.9283
	(2.4302)	(1.1038)
Mean J		3

(a)

	$\hat{m}_K$	$\hat{m}_{LL}$
Mean SE	0.0354	0.0142
	(0.0513)	(0.0189)
Mean J		4

(b)

Appendix

Here we state some auxiliary results that are used throughout the proof of the theorem stated in Section 3. The last part of this appendix contains the proof of the theorem. First we reproduce Bernstein's inequality as appearing in Ferraty and Vieu (2006).

Bernstein's inequality: *Let $Z_1, \ldots, Z_n$ be independent identically distributed random variables with zero mean. If for all $m \geq 2$ there exists a constant $C_m > 0$ such that $E|Z_1^m| \leq C_m\, a^{2(m-1)}$, we have that*

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right| > \epsilon\right\} \leq 2\,\exp\left(-\frac{\epsilon^2\, n}{2\,a^2\,(1+\epsilon)}\right), \quad \forall \epsilon > 0.$$

Below we state a technical lemma together with its proof.

Lemma: *Let $X_1, \ldots, X_n$ be independent random elements identically distributed as X and $\Delta_i := K(h^{-1}\|X_i - x\|)/E(K(h^{-1}\|X - x\|))$, for $i = 1, \ldots, n$, where K is an asymmetrical decreasing kernel function satisfying assumption (A1). If $h \to 0$ and $n\varphi_x(h) \to \infty$ as $n \to \infty$, then*

(i) $n^{-1}\sum_{i=1}^{n} \Delta_i = 1 + o_p((n\varphi_x(h))^{-1/2})$,

(ii) $n^{-1}\sum_{i=1}^{n} c_{ij}\,\Delta_i = O_p(h)$ *and* $n^{-1}\sum_{i=1}^{n} c_{ij}\,c_{ik}\,\Delta_i = O_p(h^2)$, *for* $j, k = 1 \ldots J$.

Proof of the lemma:

(i) For each $\epsilon > 0$, we bound $P\{|n^{-1}\sum_{i=1}^{n}\Delta_i - 1| > \epsilon\}$ using the Bernstein-type inequality introduced above. Since $E(\Delta_1) = 1$, we define $Z_i = \Delta_i - 1$, for $i = 1, \ldots, n$. In order to bound $E|Z_1|^m = E|\Delta_1 - E\Delta_1|^m$ for $m \geq 2$, remark first that

$$Z_1^m = (\Delta_1 - E\Delta_1)^m = \sum_{k=0}^{m}\binom{m}{k}\Delta_1^k(-1)^{m-k}.$$

Then

$$E|Z_1|^m \leq \sum_{k=0}^{m}\binom{m}{k}E(\Delta_1^k) \leq C\,\max_{k=0,\ldots,m} E(\Delta_1^k),$$

where C denotes a generic positive constant. Due to assumption (A1) we have that, for $k \geq 2$, $E(\Delta_1^k) = O(\varphi_x(h))^{-(k-1)}$. For $k = 0$ or 1, $E(\Delta_1^k) = 1$. Since, by assumption,

$\varphi_x(h) \to 0$ as $n \to \infty$, we conclude that $\max_{k=0,\ldots,m} E(\Delta_1^k) = O(\varphi_x(h))^{-(m-1)}$. By applying the exponential inequality with $a^2 = (\varphi_x(h))^{-1}$ we obtain that, for all $\epsilon > 0$ small enough,

$$P\left\{\left|n^{-1}\sum_{i=1}^{n}\Delta_i - 1\right| > \epsilon\right\} \leq 2\exp\left(-C\epsilon^2 n\varphi_x(h)\right),$$

which yields the desired result.

(ii) Note that each c_{ij} is multiplied by $K(h^{-1}\|X_i - x\|)$ for any value of j. Assumption (A1) implies that, if $K(h^{-1}\|X_i - x\|) \neq 0$, then $\|X_i - x\|^2 = \sum_{j=1}^{\infty} c_{ij}^2 \leq h^2$, for all i, and this in turn implies that $0 \leq |c_{ij}| \leq h$, for $i = 1, \ldots, n$ and $j = 1, \ldots, J$. Consequently $E|c_{1j}\Delta_1| = O(h)$ and $E|c_{1j}c_{1k}\Delta_1| = O(h^2)$ for $j, k = 1, \ldots, J$. Markov inequality finally yields $n^{-1}\sum_{i=1}^{n} c_{ij}\Delta_i = O_P(h)$ and $n^{-1}\sum_{i=1}^{n} c_{ij}c_{ik}\Delta_i = O_P(h^2)$ for $j, k = 1, \ldots, J$. □

We now proceed to prove the theorem stated in Section 3.

Proof of the theorem: Observe that the mean squared error (7) can be decomposed as

$$E((\hat{m}_{LL}(x) - m(x))^2|\mathbf{X}) = \text{Bias}^2(\hat{m}_{LL}(x)|\mathbf{X}) + \text{Var}(\hat{m}_{LL}(x)|\mathbf{X}),$$

where

$$\text{Var}(\hat{m}_{LL}(x)|\mathbf{X}) = E\left((\hat{m}_{LL}(x) - E(\hat{m}_{LL}(x)|\mathbf{X}))^2|\mathbf{X}\right)$$

and

$$\text{Bias}(\hat{m}_{LL}(x)|\mathbf{X}) = E(\hat{m}_{LL}(x)|\mathbf{X}) - m(x).$$

Let us first prove that the bias term is $O(J^{-\nu}) + O_P(h^2)$. Using the expression for the local linear estimator $\hat{m}_{LL}(x)$ given in (6) we get

$$E(\hat{m}_{LL}(x)|\mathbf{X}) = \mathbf{e}_1'(\mathbf{C}'\mathbf{\Delta}\mathbf{C})^{-1}\mathbf{C}'\mathbf{\Delta}\mathbf{M},$$

where $\mathbf{\Delta} = \text{diag}(\Delta_1, \ldots, \Delta_n)$, $\Delta_i := K(h^{-1}\|X_i - x\|)/E(K(h^{-1}\|X - x\|))$ for $i = 1, \ldots, n$ and $\mathbf{M} := E(\mathbf{Y}|\mathbf{X}) = (m(X_1), \ldots, m(X_n))'$.

We start with the term $\mathbf{C}'\mathbf{\Delta}\mathbf{M}$. Since K has support in $[0, 1]$, the X_i's for which $K(h^{-1}\|X_i - x\|) \neq 0$ are in $B(x, h)$, the ball of center x and radius h. Then, using that $h \to 0$, the following Taylor expansion is valid (see Cartan 1967)

$$m(X_i) = m(x) + m_x'(X_i - x) + m_x''(X_i - x)^2 + o(\|X_i - x\|^3),$$

where m_x' and m_x'' are lineal continuous operators on $L^2[0, 1]$ and $L^2[0, 1] \times L^2[0, 1]$, respectively, and $(X_i - x)^2$ denotes $(X_i - x, X_i - x)$. Using this expansion, we get

$$\begin{aligned}
\mathbf{\Delta}\,\mathbf{M} &= (m(X_1)\Delta_1, \ldots, m(X_n)\Delta_n)' \\
&= \begin{pmatrix} (m(x) + m_x'(X_1 - x) + m_x''(X_1 - x)^2 + o(\|X_1 - x\|^3))\Delta_1 \\ \vdots \\ (m(x) + m_x'(X_n - x) + m_x''(X_n - x)^2 + o(\|X_n - x\|^3))\Delta_n \end{pmatrix} \\
&= \begin{pmatrix} (m(x) + m_x'(X_1 - x))\Delta_1 \\ \vdots \\ (m(x) + m_x'(X_n - x))\Delta_n \end{pmatrix} + \begin{pmatrix} (O_P(h^2) + o_P(h^3))\Delta_1 \\ \vdots \\ (O_P(h^2) + o_P(h^3))\Delta_n \end{pmatrix}.
\end{aligned}$$

To derive the last approximation we have used (see Cartan 1967) that, if m_x'' is continuous, then $\|m_x''\| < \infty$ and $|m_x''(z)^2| \leq \|m_x''\| \, \|z\|^2$ if $z \in B(0,1)$. By the assumption that $h \to 0$, if $K(h^{-1}\|X_i - x\|) \neq 0$, we have that $X_i - x \in B(0,1)$ for n sufficiently large, and $|m_x''(X_i - x)^2| \leq \|m_x''\| \, \|X_i - x\|^2 = O(h^2)$ a.s.

Observe that we may approximate $m_x'(X_i - x)$ by $\sum_{j=1}^{J} m_{x,j}' \, c_{ij}$, where $m_{x,j}' := \langle m_x', \phi_j \rangle$ are the Fourier coefficients of m_x' (we use the fact that the space of lineal operators on $L^2[0,1]$ is isometric to $L^2[0,1]$). More precisely, by assumptions (A3) and (A4), we have $\max_{i=1,\ldots,n} |m_x'(X_i - x) - \sum_{j=1}^{J} m_{x,j}' \, c_{ij}| = O(J^{-\nu})$ (see Zygmund 1988). Consequently

$$\mathbf{\Delta M} = \mathbf{\Delta C} \begin{pmatrix} m(x) \\ m_{x,1}' \\ \vdots \\ m_{x,J}' \end{pmatrix} + \begin{pmatrix} (O(J^{-\nu}) + O_P(h^2))\Delta_1 \\ (O(J^{-\nu}) + O_P(h^2))\Delta_2 \\ \vdots \\ (O(J^{-\nu}) + O_P(h^2))\Delta_n \end{pmatrix}$$

and thus

$$E(\hat{m}_{LL}(x)|\mathbf{X}) = m(x) + \mathbf{e}_1'(\mathbf{C}'\mathbf{\Delta C})^{-1}\mathbf{C}' \begin{pmatrix} (O(J^{-\nu}) + O_P(h^2))\Delta_1 \\ (O(J^{-\nu}) + O_P(h^2))\Delta_2 \\ \vdots \\ (O(J^{-\nu}) + O_P(h^2))\Delta_n \end{pmatrix}. \tag{10}$$

In order to study the asymptotic behaviour of the bias, we multiply and divide the second term in the right-hand side of (10) by n. Applying the previous Lemma, we may express the bias of the local linear estimator as follows

$$\text{Bias}(\hat{m}_{LL}(x)|\mathbf{X}) = \mathbf{e}_1' \, (n^{-1}\mathbf{C}'\mathbf{\Delta C})^{-1} \begin{pmatrix} O(J^{-\nu}) + O_P(h^2) \\ O_P(h)(O(J^{-\nu}) + O_P(h^2)) \\ \vdots \\ O_P(h)(O(J^{-\nu}) + O_P(h^2)) \end{pmatrix}. \tag{11}$$

Now let us examine the components in matrix

$$\begin{aligned} n^{-1}\mathbf{C}'\mathbf{\Delta C} &= \begin{pmatrix} n^{-1}\sum_{i=1}^n \Delta_i & n^{-1}\sum_{i=1}^n c_{i1}\Delta_i & \ldots & n^{-1}\sum_{i=1}^n c_{iJ}\Delta_i \\ n^{-1}\sum_{i=1}^n c_{i1}\Delta_i & n^{-1}\sum_{i=1}^n c_{i1}^2\Delta_i & \ldots & n^{-1}\sum_{i=1}^n c_{i1}c_{iJ}\Delta_i \\ \vdots & \vdots & & \vdots \\ n^{-1}\sum_{i=1}^n c_{iJ}\Delta_i & n^{-1}\sum_{i=1}^n c_{i1}c_{iJ}\Delta_i & \ldots & n^{-1}\sum_{i=1}^n c_{iJ}^2\Delta_i \end{pmatrix} \\ &= \begin{pmatrix} 1 + o_P((n\varphi_x(h))^{-1/2}) & O_P(h) & \ldots & O_P(h) \\ O_P(h) & O_P(h^2) & \ldots & O_P(h^2) \\ \vdots & \vdots & & \vdots \\ O_P(h) & O_P(h^2) & \ldots & O_P(h^2) \end{pmatrix}. \end{aligned}$$

To derive the last equality we have used again the same lemma. Note that we can express $n^{-1}\mathbf{C}'\mathbf{\Delta C}$ as a block matrix

$$n^{-1}\mathbf{C}'\mathbf{\Delta C} = \left(\begin{array}{c|c} 1 + o_P((n\varphi_x(h))^{-1/2}) & O_P(h)\,\mathbf{b}' \\ \hline O_P(h)\,\mathbf{b} & O_P(h^2)\,\mathbf{B} \end{array} \right),$$

where $\mathbf{b}$ is a $J \times 1$ vector and $\mathbf{B}$ is a $J \times J$ nonsingular matrix. Using a well-known formula to invert a block nonsingular symmetric matrix (see Seber 1984), we can write

$$(n^{-1}\mathbf{C}'\mathbf{\Delta}\mathbf{C})^{-1} = \left(\begin{array}{c|c} r & -r\, O_P(h^{-1})\, \mathbf{b}'\, \mathbf{B}^{-1} \\ \hline -r\, O_P(h^{-1})\, \mathbf{B}^{-1}\, \mathbf{b} & O_P(h^{-2})\, (\mathbf{B}^{-1} + r\, \mathbf{B}^{-1}\, \mathbf{b}\, \mathbf{b}'\, \mathbf{B}^{-1}) \end{array}\right), \tag{12}$$

where $r = 1/(1 + o_P((n\varphi_x(h))^{-1/2}) - \mathbf{b}'\, \mathbf{B}^{-1}\, \mathbf{b}$. Finally, substituting formula (12) into expression (11), we conclude that $\text{Bias}(\hat{m}_{LL}(x)|\mathbf{X}) = O_P(h^2)$.
Let us now analize the variance term

$$\begin{aligned} \text{Var}(\hat{m}_{LL}(x)|\mathbf{X}) \quad &= E\left((\hat{m}_{LL}(x) - E(\hat{m}_{LL}(x)|\mathbf{X}))^2|\mathbf{X}\right) \\ &= E\left(\mathbf{e}_1'(\mathbf{C}'\mathbf{\Delta}\mathbf{C})^{-1}\mathbf{C}'\mathbf{\Delta}(\mathbf{Y}-\mathbf{M})(\mathbf{Y}-\mathbf{M})'\mathbf{\Delta}\mathbf{C}(\mathbf{C}'\mathbf{\Delta}\mathbf{C})^{-1}\mathbf{e}_1|\mathbf{X}\right) \\ &= \mathbf{e}_1'(\mathbf{C}'\mathbf{\Delta}\mathbf{C})^{-1}\mathbf{C}'\mathbf{\Delta}\mathbf{V}\mathbf{\Delta}\mathbf{C}(\mathbf{C}'\mathbf{\Delta}\mathbf{C})^{-1}\mathbf{e}_1, \end{aligned} \tag{13}$$

where $\mathbf{V} := \text{Var}(\mathbf{Y}|\mathbf{X}) = \text{diag}(\text{Var}(Y_1|X_1), \ldots, \text{Var}(Y_n|X_n)) = \text{diag}(\sigma_\epsilon^2, \ldots, \sigma_\epsilon^2)$. Multiplying and dividing (13) by $n^{-2}E(K^2(h^{-1}\|X-x\|))$ it is easy to check that

$$\begin{aligned} &\text{Var}(\hat{m}_{LL}(x)|\mathbf{X}) \\ &\quad = \frac{\sigma_\epsilon^2}{n}\, \frac{E(K^2(h^{-1}\|X-x\|))}{E^2(K(h^{-1}\|X-x\|))}\, \mathbf{e}_1'(n^{-1}\mathbf{C}'\mathbf{\Delta}\mathbf{C})^{-1} n^{-1}\mathbf{C}'\mathbf{\Lambda}\mathbf{C}(n^{-1}\mathbf{C}'\mathbf{\Delta}\mathbf{C})^{-1}\mathbf{e}_1, \end{aligned}$$

where $\mathbf{\Lambda} = \text{diag}(\Lambda_1, \ldots, \Lambda_n)$ and $\Lambda_i = K^2(h^{-1}\|X_i - x\|)/E(K^2(h^{-1}\|X-x\|))$ for $i = 1, \ldots, n$. Observe that, by assumption (A1), $E(K^2(h^{-1}\|X-x\|))/E^2(K(h^{-1}\|X-x\|)) = O(\varphi_x^{-1}(h))$ and that

$$n^{-1}\mathbf{C}'\mathbf{\Lambda}\mathbf{C} = \begin{pmatrix} n^{-1}\sum_{i=1}^n \Lambda_i & n^{-1}\sum_{i=1}^n c_{i1}\Lambda_i & \ldots & n^{-1}\sum_{i=1}^n c_{iJ}\Lambda_i \\ n^{-1}\sum_{i=1}^n c_{i1}\Lambda_i & n^{-1}\sum_{i=1}^n c_{i1}^2\Lambda_i & \ldots & n^{-1}\sum_{i=1}^n c_{i1}c_{iJ}\Lambda_i \\ \vdots & \vdots & & \vdots \\ n^{-1}\sum_{i=1}^n c_{iJ}\Lambda_i & n^{-1}\sum_{i=1}^n c_{i1}c_{iJ}\Lambda_i & \ldots & n^{-1}\sum_{i=1}^n c_{iJ}^2\Lambda_i \end{pmatrix}.$$

Following the same steps as with the bias term we obtain the asymptotic behaviour of the components in the previous matrix. More concretely, $n^{-1}\sum_{i=1}^n \Lambda_i = 1 + o_P((n\varphi_x(h))^{-1/2})$, $n^{-1}\sum_{i=1}^n c_{ij}\Lambda_i = O_P(h)$ and $n^{-1}\sum_{i=1}^n c_{ij}c_{ik}\Lambda_i = O_P(h^2)$ for any $j, k = 1, \ldots, J$. Thus

$$n^{-1}\mathbf{C}'\mathbf{\Lambda}\mathbf{C} = \left(\begin{array}{c|c} 1 + o_P((n\varphi_x(h))^{-1/2}) & O_P(h)\, \tilde{\mathbf{b}}' \\ \hline O_P(h)\, \tilde{\mathbf{b}} & O_P(h^2)\, \tilde{\mathbf{B}} \end{array}\right).$$

Then the variance term can be expressed as follows

$$
\begin{aligned}
&\mathrm{Var}(\hat{m}_{LL}(x)|\mathbf{X}) \\
&\quad = \frac{\sigma_\epsilon^2}{n}\,\frac{E(K^2(h^{-1}\|X-x\|))}{E^2(K(h^{-1}\|X-x\|))}\,\mathbf{e}_1' \\
&\quad\quad \begin{pmatrix} r & -r\,O_P(h^{-1})\,\mathbf{b}'\,\mathbf{B}^{-1} \\ -r\,O_P(h^{-1})\,\mathbf{B}^{-1}\,\mathbf{b} & O_P(h^{-2})(\mathbf{B}^{-1}+r\,\mathbf{B}^{-1}\,\mathbf{b}\,\mathbf{b}'\,\mathbf{B}^{-1}) \end{pmatrix} \\
&\quad\quad \begin{pmatrix} 1+o_P((n\varphi_x(h))^{-1/2}) & O_P(h)\,\tilde{\mathbf{b}}' \\ O_P(h)\,\tilde{\mathbf{b}} & O_P(h^2)\,\tilde{\mathbf{B}} \end{pmatrix} \\
&\quad\quad \begin{pmatrix} r & -r\,O_P(h^{-1})\,\mathbf{b}'\,\mathbf{B}^{-1} \\ -r\,O_P(h^{-1})\,\mathbf{B}^{-1}\,\mathbf{b} & O_P(h^{-2})(\mathbf{B}^{-1}+r\,\mathbf{B}^{-1}\,\mathbf{b}\,\mathbf{b}'\,\mathbf{B}^{-1}) \end{pmatrix} \mathbf{e}_1 \\
&\quad = \frac{C}{n\varphi_x(h)}\left[1+o_P((n\varphi_x(h))^{-1/2}) - \tilde{\mathbf{b}}'\mathbf{B}^{-1}\mathbf{b} - \mathbf{b}'\mathbf{B}^{-1}\tilde{\mathbf{b}} + \mathbf{b}'\mathbf{B}^{-1}\tilde{\mathbf{B}}\mathbf{B}^{-1}\mathbf{b}\right] \\
&\quad = O_P\left((n\varphi_x(h))^{-1}\right).
\end{aligned}
$$

□

References

[1] Baíllo, A. (2007). A note on functional linear regression. Manuscript.

[2] Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *Ann. Statist.*, 34, 2159–2179.

[3] Cardot, H., Ferraty, F. and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica*, 13, 571–591.

[4] Cardot, H. and Sarda, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis*, 92, 24–41.

[5] Cartan, H. (1967). *Calcul Differentiel*. Hermann, Paris.

[6] Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, 87, 998–1004.

[7] Fan, J. and Marron, J. S. (1993). Local Regression: Automatic Kernel Carpentry: Comment. *Statistical Science*, 8, 129–134.

[8] Ferraty, F. and Vieu, P. (2000). Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés. *Compte Rendus Acad. Sci. Paris*, 330, Série I, 139-142.

[9] Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis*. Springer, New York.

[10] Ferré, L. and Yao, A. F. (2003). Functional sliced inverse regression analysis. *Statistics*, 37, 475–488.

[11] Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35, 70–91.

[12] James, G. M. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society, series B*, 64, 411–432.

[13] Müller, H.-G. and Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics*, 33, 774–805.

[14] Ramsay, J. O. and Silverman, B. (2005). *Functional Data Analysis*. Second edition. Springer-Verlag, New York.

[15] Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.*, 22, 1346–1370.

[16] Seber, G. A. F. (1984). *Multivariate observations*. John Wiley & Sons.

[17] Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall.

[18] Zygmund, A. (1988). *Trigonometric Series*. Cambridge University Press, Cambridge.