HAL
open science

Nonparametric estimation of the average growth curve with general nonstationary error process

Karim Benhenni, Mustapha Rachdi

To cite this version:

Karim Benhenni, Mustapha Rachdi. Nonparametric estimation of the average growth curve with general nonstationary error process. Communications in Statistics - Theory and Methods, 2007, 36 (6), pp.1173-1186. 10.1080/03610920601076586 . hal-00023332

HAL Id: hal-00023332

https://hal.science/hal-00023332

Submitted on 4 Aug 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Nonparametric estimation of the average growth curve with general nonstationary error process

K. Benhenni * and M. Rachdi

Université de Grenoble, UFR SHS, BP. 47, F38040 Grenoble Cedex 09, France.

Abstract

The non-parametric estimation of the growth curve has been extensively studied in both stationary and some nonstationary particular situations.
In this work, we consider the statistical problem of estimating the average growth curve for a fixed design model with nonstationary error process. The nonstationarity considered here is of a general form, and this note may be considered as an extension of previous results. The optimal bandwidth is shown to depend on the singularity of the autocovariance function of the error process along the diagonal. A Monte Carlo study is conducted in order to assess the influence of the number of subjects and the number of observations per subject on the estimation.

Key words: Growth curve, Nonparametric estimation, Non-stationary errors

1 Introduction

The growth curve model is useful especially for growths of animals and plants and is applied extensively to biostatistics, medical research and epidemiology, and was considered by many authors such as Geisser (1980) and Von Rosen (1991). In pharmacokinetic research, it is a useful problem for the estimation of the concentration-time curve based on the drug's concentration at different sites within the organism.

* Corresponding author
Email address: `Karim.Benhenni@upmf-grenoble.fr` , `Tel: (33)476825769`, `Fax: (33)476825665` (K. Benhenni).

In nonparametric regression, the correlation of the errors can have some important consequences on the statistical properties of the curve function estimator and on the selection of the smoothing parameter.

The nonparametric regression model with correlated errors was considered by many authors. For instance, Severini and Staniswalis (1994), Zeger and Diggle (1994), Wild and Yee (1996), Chiang and Hoover (1998), Lin and Carroll (2000) among others. These authors considered different modifications of kernel construction of the nonparametric regression estimator in order to improve the efficiency over the standard kernel estimator when correlated observations are introduced. In practical situations when we only have access to quantized and correlated data, Benhenni and Rachdi (2005) constructed a consistent estimator of the growth curve when repeated observations are available.

The focus of this paper is to look at the problem of estimating the mean function $f(.)$ in the presence of correlation not that of estimating the correlation function itself.

The emphasis of Zimmerman and Núñez-Antón (2001)'s paper is put in modeling the covariance structure of growth curve data, where antedependence models are considered, that is, a transformation of the time scale that can produce nonstationary covariance. They also provide an overview of the nonparametric regression literature that deals with the correlated errors case.

We consider the statistical problem of estimating the average growth curve for a fixed design model. We consider m experimental units, each of them having n measurements of the response:

$$Y_j(x_i) = f(x_i) + \varepsilon_j(x_i) \text{ where } j = 1, \ldots, m \text{ and } i = 1, \ldots, n$$

where f is the unknown average growth curve and (ε_j) is the error process. The sampling points $\{x_i,\, i = 1, \ldots, n\}$ are usually taken equally spaced in time series data, but other type of sampling designs can also be considered such as deterministic regular (non uniform) designs and random designs. Although repeated measurements can naturally arise in practical situations, they can make the estimators of the curve f asymptotically consistent, as was pointed out by Hart and Wehrly (1986) and the comments of Härdle (1989).

In this paper, we estimate the growth curve using the kernel methods under a very general nonstationary error process where the autocovariance function does not have any specific form. This include, as a special case, processes with stationary autocovariance function such as the Ornstein-Uhlenbeck process, and some specific nonstationary class of parametric autocovariance structure considered by Ferreira et al. (1997) and introduced by Núñez-Antón and Woodworth (1994) known as *antedependance* models. The popular brownian motion is also an example of a nonstationary error process.

We show that the asymptotic properties of the kernel regression estimators and the optimal bandwidth are governed by the behavior of the autocovariance function along the diagonal. This paper is organized as follows. In section 2, we present the average growth curve estimation method for correlated observations with the autocovariance function satisfying the required assumptions and then the corresponding results. Section 3 is devoted to the case when the autocovariance function is smooth. Some examples and simulations are given in Section 4 in order to verify the theoretical results of this paper. Finally, Section 5 serves for the proofs of the results.

2 Estimation of the average growth curve with non-regular error process

The estimator of f based on the observations $\{Y_j(x_i),\, i = 1, \ldots, n,\ \ j = 1, \ldots, m\}$, when the x_i's are known constants such that $0 \leq x_1 < x_2 < \cdots < x_n \leq 1$, with

$$\max_i |x_i - x_{i-1}| = O\left(\frac{1}{n}\right)$$

is given for $x \in [0, 1]$, by (see, Gasser and Müller (1984), Hart and Wehrly (1986)):

$$\widehat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} W_{h,i}(x)\, \bar{Y}(x_i)$$

where

$$\bar{Y}(x) = \frac{1}{m} \sum_{j=1}^{m} Y_j(x) \text{ and } W_{h,i}(x) = n \int_{m_{i-1}}^{m_i} K_h(x-u)\, du$$

and the midpoints $\{m_i,\ i = 0, \ldots, n\}$ are defined by

$$m_0 = 0,\ m_i = (x_i + x_{i+1})/2,\ \text{ for } i = 1, \ldots, n-1 \text{ and } m_n = 1$$

with $K_h(x) = 1/h\, K(x/h)$, where the kernel K is a function satisfying the following assumptions:

K is Hölder continuous, with support $[-1, 1]$

and $h = h(n, m)$ is the bandwidth, such that: $h \geq 0$ and $\lim_{n,m \to +\infty} h = 0$.

In the following theorem, we establish the asymptotic behavior of the estimator $\widehat{f}_h$ through the mean squared error. Then, we derive the optimal asymptotic bandwidth h^*. A global optimal bandwidth is also derived from the integrated mean squared error.

Assumptions :

(i) The autocovariance function ρ exits and is continuous on the square $[0,1]^2$.

(ii) $\rho(x,y)$ has left and right first order derivatives at the diagonal $x=y$, that is:

$$\rho^{(0,1)}(x,x^-) = \lim_{y \nearrow x} \frac{\partial \rho}{\partial y}(x,y) \text{ and } \rho^{(0,1)}(x,x^+) = \lim_{y \searrow x} \frac{\partial \rho}{\partial y}(x,y)$$

exist and are continuous.
The jump function along the diagonal $\alpha(x) = \rho^{(0,1)}(x,x^-) - \rho^{(0,1)}(x,x^+)$ is assumed to be continuous and not identically equal to zero.

(iii) $\rho(x,y)$ is assumed to have continuous mixed partial derivatives up to order two off the diagonal $x \neq y$ in the unit square and satisfies:

$$\sup_{0 \leq x \neq y \leq 1} \left| \rho^{(i,j)}(x,y) \right| < \infty \text{ for all integers } i,j \text{ such that } 0 \leq i+j \leq 2.$$

(iv) The covariance function ρ has at least 2 continuous mixed partial derivatives such that

$$\rho^{(i,j)}(x,x) \neq 0, \; i+j=2, \; \forall x \in [0,1]$$

The assumptions *(i)*, *(ii)* and *(iii)* (respectively assumption *(iv)*) are satisfied by the error processes defined in (1), (2), (3) and (4) (respectively (5)).

Example 1

(1) The Wiener error process with covariance $\rho(x,y) = \sigma^2 \min(x,y)$. The jump function $\alpha(x) = \sigma^2 > 0$ and $\rho^{(i,j)}(x,y) = 0$, for all integers i, j such that $i+j=2$ and $x \neq y$.

(2) The Ornstein-Uhlenbeck process with stationary covariance $\rho(x,y) = \sigma^2 \exp(-\lambda|x-y|/2)$ for $\sigma > 0$ and $\lambda > 0$. The jump function is constant $\alpha(x) = 2\sigma^2\lambda$.

(3) A general class of covariance functions with non constant jump functions can be found by considering the class of processes of the form:

$$\varepsilon(x) = \theta(x) W(\gamma(x)), \; \forall x \in [0,1]$$

where $W(x)$ is a Wiener process, $\theta(x)$ is a continuously differentiable function on $[0,1]$ and $\gamma(x)$ a continuously differentiable function and strictly increasing on $[0,1]$. Then, the covariance function of the error process $\varepsilon(x)$ is of the form:

$$\rho(x,y) = \theta(x) \min(\gamma(x), \gamma(y)) \theta(y)$$

and the corresponding jump function is $\alpha(x) = \theta^2(x)\gamma'(t)$. In particular, when $\theta(x) = \gamma(x) = x$, then $\rho(x,y) = xy \min(x,y)$ and $\alpha(x) = x^2$. When $\theta(x) = 1$ and $\gamma(x) = x$, then the error process has a Wiener covariance, and when $\theta(x) = \exp(-x)$ and $\gamma(x) = \exp(2x)$ then it has a Gauss-Markov covariance function.

(4) Consider a transformation of the time scale that can produce nonstationary covariance of the form

$$\rho(x,y) = \sigma^2 \rho^{|x^\lambda - y^\lambda|/\lambda} \text{ for } (x,y) \in [0,1]^2,\ \sigma^2 > 0,\ 0 < \rho < 1 \text{ and } \lambda > 0$$

(see, Nuñez-Anton and Woodworth (1994) and Ferreira et al. (1997)). In particular when $\lambda = 1$, we obtain a Ornstein-Uhlenbeck error process with a stationary covariance function with $\alpha(x) = -2\sigma^2 \ln(\rho)$. When $\lambda \neq 1$, then the covariance is nonstationary with jump function $\alpha(x) = -2\sigma^2 \ln(\rho) x^{\lambda-1}$.

(5) Consider the error process with stationary covariance function:

$$\rho(x,y) = (1 + \alpha|x-y|)\exp(-\alpha|x-y|),\ \alpha > 0$$

and spectral density $\varphi(\lambda) = \frac{2\alpha^3}{\pi(\alpha^2+\lambda^2)^2}$.
Then this covariance satisfies the above assumption with

$$\rho^{(i,j)}(x,x) = \rho''(0) = -\alpha^2 \neq 0, \text{ for } i+j = 2.$$

Theorem 2 *If the covariance function ρ satisfies Assumptions (i), (ii) and (iii), and f is a twice differentiable continuous function on $[0,1]$ with $f''(x) \neq 0$ for $0 < x < 1$, and $m/n = O(1)$, then as $n, m \to +\infty$*

$$\mathbb{E}\left(\widehat{f}_h(x) - f(x)\right)^2 = \frac{1}{m}\left(\rho(x,x) - \frac{1}{2}\alpha(x)C_K h\right) + \frac{h^4}{4} d_K^2 (f''(x))^2 + O\left(\frac{1}{mn} + \frac{h^2}{n}\right) + o\left(h^4 + \frac{h}{m}\right)$$

where

$$C_K = \int_{-1}^{1}\int_{-1}^{1} |u-v| K(u)K(v)du\,dv = 2\int_{-1}^{1}\int_{u}^{1} (v-u)K(u)K(v)du\,dv \text{ and } d_K = \int u^2 K(u)\,du$$

The asymptotic mean squared error is minimized by taking the bandwidth as in the following corollary.

Corollary 3 *Under the hypotheses of Theorem 2, if $m/n = O(1)$ as $n, m \to +\infty$, then the asymptotic mean squared error is minimized by taking the bandwidth*

$$h_x^* = \left(\frac{\alpha(x)C_K}{2d_K^2(f''(x))^2}\right)^{1/3} m^{-1/3}$$

Remark 4

(1) In example 1 of a Wiener error process with nonstationary covariance $\rho(x,y) = \sigma^2 \min(x,y)$, the asymptotic optimal bandwidth

$$h_x^* = \left(\frac{\sigma^2 C_K}{2 d_K^2 (f''(x))^2} \right)^{1/3} m^{-1/3}$$

And the global optimal bandwidth is then:

$$h^* = \left(\frac{C_K \int_0^1 \alpha(x)\, dx}{2 d_K^2 \int_0^1 (f''(x))^2\, dx} \right)^{1/3} m^{-1/3}$$

(2) For a stationary error process with covariance $\rho(x,y) = \rho(x-y)$ such as the Ornstein-Uhlenbeck process, the jump function

$$\alpha(x) = \rho^{(1)}(0^-) - \rho^{(1)}(0^+) = 2\rho^{(1)}(0^-)$$

is constant. Then

$$h_x^* = \left(\frac{\rho^{(1)}(0^-) C_K}{d_K^2 (f''(x))^2} \right)^{1/3} m^{-1/3}$$

which corresponds to the bandwidth given by Hart and Wehrly (1986).

(3) For the nonstationary covariance function $\rho(x,y) = \sigma^2 \rho^{|x^\lambda - y^\lambda|/\lambda}$, where the jump function $\alpha(x) = -2\ln(\rho)$, the optimal bandwidth is

$$h_x^* = \left(-\frac{x^{\lambda-1} \ln(\rho) C_K \sigma^2}{d_K^2 (f''(x))^2} \right)^{1/3} m^{-1/3}$$

which corresponds to the bandwidth obtained by Ferreira et al. (1997).

It can be seen from the Corollary that the optimal bandwidth h^* depends on the covariance function along the diagonal through its derivative whereas for the stationary error process, h^* depends only on a single parameter which represents the behavior of the derivative of the covariance function at the origin. Therefore if the covariance is unknown or misspecified the effect on the value of the optimal bandwidth would be more significant in the nonstationary case than in the stationary case.

The asymptotic global bandwidth can be obtained by using a global error measure of estimation such as the integrated mean squared error:

$$IMSE = \int_0^1 \mathbb{E} \left(\widehat{f}_h(x) - f(x) \right)^2 dx$$

Likewise, The asymptotic IMSE is minimized for:

$$h^* = \left(\frac{C_K \int_0^1 \alpha(x)dx}{2d_K^2 \int_0^1 (f''(x))^2 dx} \right)^{1/3} m^{-1/3}$$

3 Smooth nonstationary error process

In this case the jump function $\alpha(x) = 0$, $\forall x \in [0,1]$ The following theorem gives the asymptotic expression for the mean squared error for smoother error process.

Theorem 5 *Assume that ρ satisfies the above assumption and f is a twice differentiable continuous function on $[0,1]$ with $f''(x) \neq 0$, for $0 < x < 1$, then as n, $m \to +\infty$*

$$\mathbb{E}\left(\widehat{f}_h(x) - f(x)\right)^2 = \frac{1}{m}\left(\rho(x,x) - \rho^{(0,2)}(x,x)\, d_K h^2\right) + \frac{h^4}{4} d_K^2 (f''(x))^2 + O\left(\frac{1}{mn} + \frac{h^2}{n}\right) + o\left(h^4 + \frac{h^2}{m}\right)$$

The asymptotic optimal bandwidth is given by the following corollary.

Corollary 6 *Under the hypotheses of Theorem 5, if $m/n = O(1)$ as n, $m \to +\infty$, then the asymptotic mean squared error is minimized by taking the bandwidth*

$$h_x^* = \left(\frac{2\rho^{(0,2)}(x,x)}{d_K (f''(x))^2} \right)^{1/2} m^{-1/2}$$

In particular for a stationary error process $\rho^{(0,2)}(x,x) = -\rho''(0)$, then

$$h_x^* = \left(\frac{-2\rho''(0)}{d_K (f''(x))^2} \right)^{1/2} m^{-1/2}$$

The asymptotic optimal global bandwidth using the integrated mean square criteria is:

$$h^* = \left(\frac{2 \int_0^1 \rho^{(0,2)}(x,x)dx}{d_K \int_0^1 (f''(x))^2 dx} \right)^{1/2} m^{-1/2}$$

4 Simulation study

The data are generated from three different growth curves for different values of the sample size n, the number of replications m, with 100 simulations for each case.

- Linear growth curve: $f(x) = 10(1 + x)$, $0 < x < 1$.
- Cubic growth curve: $f(x) = 10x^3 - 15x^4 + 6x^5$, $0 < x < 1$.
- The growth curve studied by Ferreira et al. (1997): $f(x) = x - 0.5\ \exp(-80(x - 0.5)^2)$, $0 < x < 1$.

where the sampling points are taken to be $x_i = i/(n + 1)$, $i = 1, \ldots, n$ in the interval $[0, 1]$.

We consider the stationary error process with covariance function

$$\rho(t, s) = \exp(-\lambda|t - s|),\ for \lambda > 0$$

The data are generated from the function f and the noise is a brownian motion with unit variance and with variance equal 2.

The Epanechnikov kernel is used and is defined by:

$$K(u) = 0.75\,(1 - u^2)\,1_{[-1,1]}(u)$$

First we study the performance of the estimator $\widehat{f}_h$ by considering the mean and the median of all the estimations obtained from these 100 simulations for estimating the three growth curves with different values of (n, m). For space reasons, we only show the results for the median estimator. However, the same conclusions are also true for the mean estimator.
The cross-validation procedure is used in each case in order to generate the optimal bandwidth h, as explained in Remarks **??** and **??**.
It can be seen in figures **??**, 10, **??** that where there are no replications ($m = 1$), the estimator has a very poor performance even for a sample size $n = 100$. When there are sufficient number of replications available ($m = 10$, $m = 50$), the estimator has a better performance in estimating the three curves f. Indeed, when the number of replications m increases, they become more consistent, as shown in Theorems 2 and 5.

It should be noticed that this estimator suffers from boundary effects. To avoid this, one can use for instance local polynomials, see Fan (1992) and Fan and Gijbels (1996) instead of the ordinary kernel type of estimators.

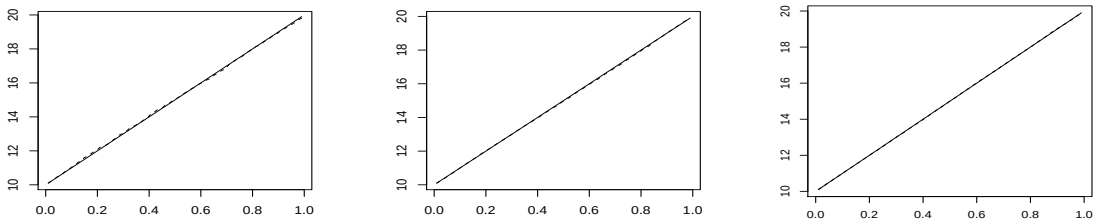

Fig. 1. Linear growth curve f is in plain line, the estimator $\widehat{f}_h$ is in dashed line for $n = 100$, $m \in \{1, 10, 50\}$ with unit variance of the error process

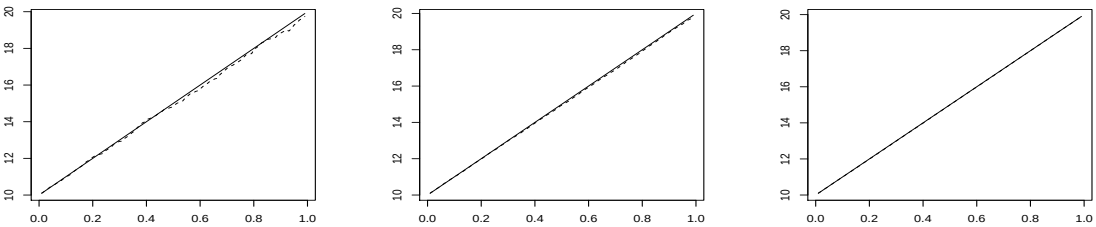

Fig. 2. Linear growth curve f is in plain line, the estimator $\widehat{f}_h$ is in dashed line for $n = 100$, $m \in \{1, 10, 50\}$ with 2 as the variance of the error process

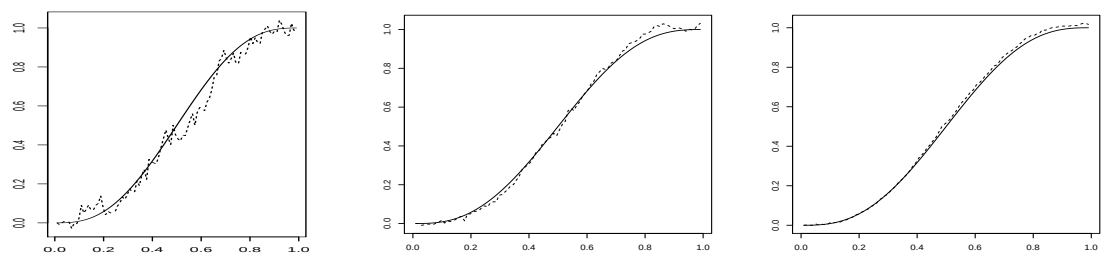

Fig. 3. Quadratic growth curve f is in plain line, the estimator $\widehat{f}_h$ is in dashed line for $n = 100$, $m \in \{1, 10, 50\}$ with unit variance of the error process

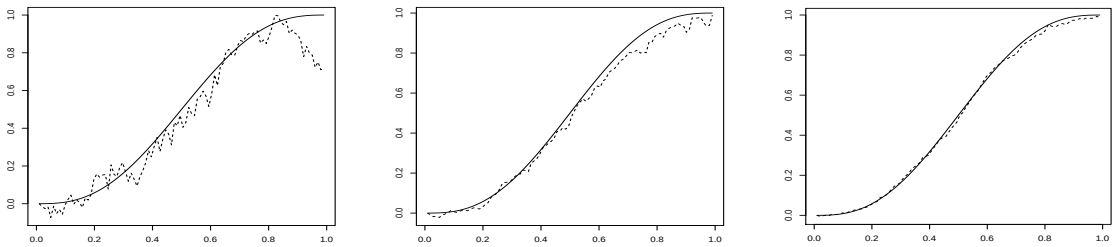

Fig. 4. Quadratic growth curve f is in plain line, the estimator $\widehat{f}_h$ is in dashed line for $n = 100$, $m \in \{1, 10, 50\}$ with 2 as the variance of the error process

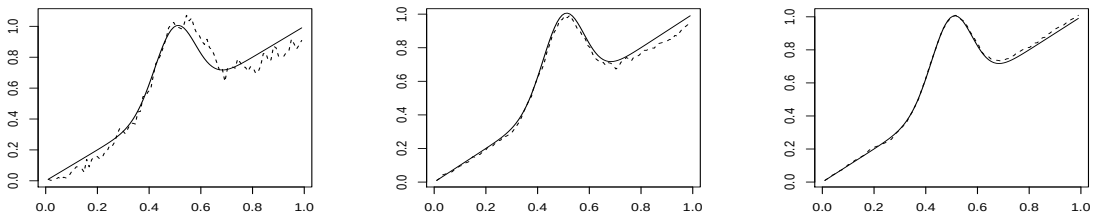

Fig. 5. Cubic growth curve f is in plain line, the estimator $\widehat{f}_h$ is in dashed line for $n = 100$, $m \in \{1, 10, 50\}$ with unit variance of the error process

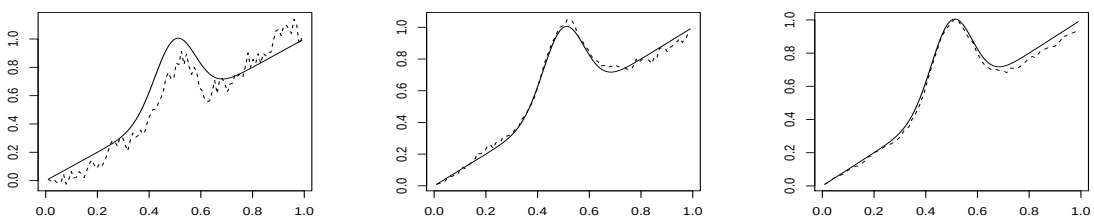

Fig. 6. Cubic growth curve f is in plain line, the estimator $\widehat{f}_h$ is in dashed line for $n = 100$, $m \in \{1, 10, 50\}$ with 2 as the unit variance of the error process

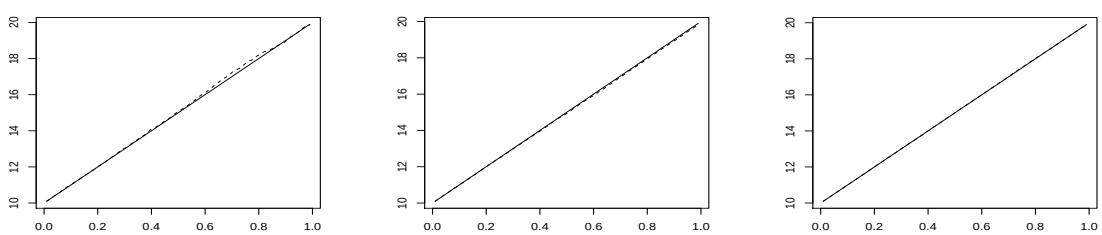

Fig. 7. Linear growth curve f is in plain line, the estimator $\widehat{f}_h$ is in dashed line for $n = 100$, $m \in \{1, 10, 50\}$ with fractional brownian error process

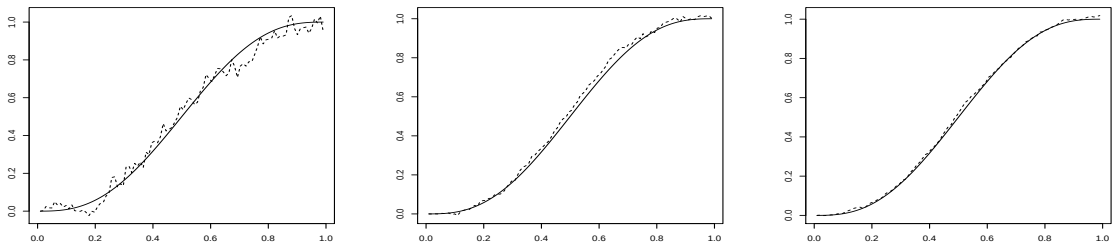

Fig. 8. Quadratic growth curve f is in plain line, the estimator $\widehat{f}_h$ is in dashed line for $n = 100$, $m \in \{1, 10, 50\}$ with fractional brownian error process

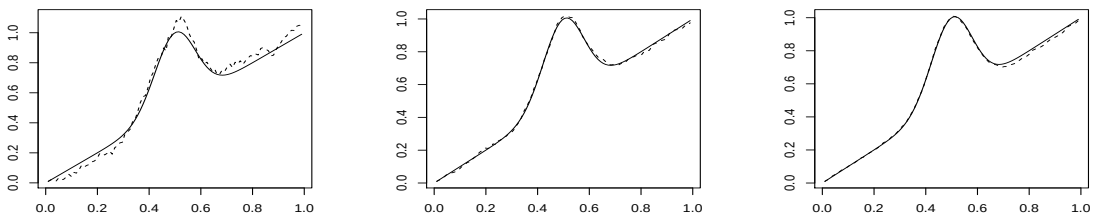

Fig. 9. Cubic growth curve f is in plain line, the estimator $\widehat{f}_h$ is in dashed line for $n = 100$, $m \in \{1, 10, 50\}$ with fractional brownian error process

5 Proofs

Proof of Theorem 2

The mean squared error can be decomposed as usual in a bias term and a variance term:

$$\mathbb{E}\left(\widehat{f}_h(x) - f(x)\right)^2 = Bias^2(\widehat{f}_h(x)) + var(\widehat{f}_h(x)). \tag{1}$$

First, we study the asymptotic behaviour of the bias term. For this, we have that

$$\mathbb{E}(\widehat{f}_h(x)) = \frac{1}{n}\sum_{i=1}^{n} W_{h,i}(x)\, \mathbb{E}(\bar{Y}(x_i))$$

where

$$\mathbb{E}(\bar{Y}(x_i)) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}(Y_j(x_i)) = f(x_i)$$

Then, from the Mean Value Theorem, and by a Taylor expansion of f around x up to order 2

$$\begin{aligned}
\mathbb{E}(\widehat{f}_h(x)) &= \frac{1}{n}\sum_{i=1}^{n} W_{h,i}(x)\, f(x_i) \\
&= h^{-1}\sum_{i=1}^{n}\int_{m_{i-1}}^{m_i} K\left(\frac{x-u}{h}\right)\, du\, f(x_i) \\
&= h^{-1}\sum_{i=1}^{n} K\left(\frac{x-\xi_i}{h}\right)\, \Delta mi\, f(x_i) \text{ where } \Delta m_i = m_i - m_{i-1} \\
&= h^{-1}\int K\left(\frac{x-u}{h}\right)\, f(u)\, du + O\left(\frac{1}{n}\right) \\
&= \int K(v)\; f(x-hv)\, dv + O\left(\frac{1}{n}\right) \\
&= f(x)\int K(v)\; dv - h\, f'(x)\int v\, K(v)\, dv \\
&\quad + \frac{h^2}{2}\, f''(x)\int v^2\, K(v)\, dv + o(h^2) + O\left(\frac{1}{n}\right) \\
&= f(x) + \frac{h^2}{2}\, f''(x)\, d_K + o(h^2) + O\left(\frac{1}{n}\right)
\end{aligned}$$

Thus

$$Bias(\widehat{f}_h(x), f(x)) = \frac{h^2}{2}\, f''(x)\, d_K + O(h^2) + \left(\frac{1}{n}\right)$$

For the variance term, from the hypothesis that Y_l are uncorrelated with the same autocovariance, we have that

$$\begin{aligned}
var(\widehat{f}_h(x)) &= \frac{1}{n^2m^2}\sum_{i,j=1}^{n} W_{h,i}(x)W_{h,j}(x)\sum_{k,l=1}^{m} cov\left(Y_k(x_i), Y_l(x_j)\right) \\
&= \frac{1}{n^2m^2}\sum_{i,j=1}^{n} W_{h,i}(x)W_{h,j}(x)\sum_{k=1}^{m} cov\left(Y_k(x_i), Y_k(x_j)\right) \\
&= \frac{1}{n^2m}\sum_{i,j=1}^{n} W_{h,i}(x)W_{h,j}(x)\rho\left(x_i, x_j\right) \\
&= h^{-2}n^2(n^2m)^{-1}\sum_{i,j=1}^{n}\int_{m_{i-1}}^{m_i}\int_{m_{j-1}}^{m_j} K\left(\frac{x-u}{h}\right) K\left(\frac{x-v}{h}\right) du\, dv\, \rho\left(x_i, x_j\right) \\
&= (mh^2)^{-1}\sum_{i,j=1}^{n}\int_{m_{i-1}}^{m_i}\int_{m_{j-1}}^{m_j} K\left(\frac{x-u}{h}\right) K\left(\frac{x-v}{h}\right) du\, dv\, \rho\left(x_i, x_j\right)
\end{aligned}$$

In another hand, we have that

$$\int_{-1}^{1}\int_{-1}^{1} \rho(x-hu, x-hv)\, K(u)\, K(v)\, du\, dv$$
$$= h^{-2} \sum_{i,j=1}^{n} \int_{m_{i-1}}^{m_i} \int_{m_{j-1}}^{m_j} K\left(\frac{x-u}{h}\right) K\left(\frac{x-v}{h}\right) \rho(u,v)\, du\, dv$$

For n sufficiently large, we obtain that

$$\left| var(\widehat{f}_h(x)) - m^{-1} \int_{-1}^{1}\int_{-1}^{1} \rho(x-hu, x-hv)\, K(u)\, K(v)\, du\, dv \right| \tag{2}$$
$$\leq (mh^2)^{-1} \sum_{i,j=1}^{n} \int_{m_{i-1}}^{m_i} \int_{m_{j-1}}^{m_j} |\rho(x_i,x_j) - \rho(u,v)|\, K\left(\frac{x-u}{h}\right) K\left(\frac{x-v}{h}\right) du\, dv$$

We decompose the double summation as follows:

$$\sum_{i,j=1}^{n} \int_{m_{i-1}}^{m_i} \int_{m_{j-1}}^{m_j} |\rho(x_i,x_j) - \rho(u,v)|\, K\left(\frac{x-u}{h}\right) K\left(\frac{x-v}{h}\right) \tag{3}$$
$$= \sum_{i=1}^{n} \int_{m_{i-1}}^{m_i} \int_{m_{i-1}}^{m_i} |\rho(x_i,x_i) - \rho(u,v)|\, K\left(\frac{x-u}{h}\right) K\left(\frac{x-v}{h}\right) \tag{4}$$
$$+ \sum_{\substack{i,j=1 \\ i\neq j}}^{n} \int_{m_{i-1}}^{m_i} \int_{m_{j-1}}^{m_j} |\rho(x_i,x_j) - \rho(u,v)|\, K\left(\frac{x-u}{h}\right) K\left(\frac{x-v}{h}\right) \tag{5}$$

Concentrating first on the diagonal term (4), we have

$$(4) = \sum_{i=1}^{n} \left(\int_{m_{i-1}}^{m_i} \int_{m_{i-1}}^{v} + \int_{m_{i-1}}^{m_i} \int_{v}^{m_i} \right) |\rho(x_i,x_i) - \rho(u,v)|\, K\left(\frac{x-u}{h}\right) K\left(\frac{x-v}{h}\right)$$

For $m_{i-1} < u < v < m_i$, the Taylor expansion of the covariance function ρ around (x_i, x_i) up to order 1 gives:

$$\rho(u,v) = \rho(x_i,x_i) + (v-x_i)\rho^{(0,1)}(x_i,\eta_i) + (u-x_i)\rho^{(1,0)}(\xi_i,v)$$

where the intermediate points ξ_i, η_i are between x_i, u and x_i, v respectively. Then

$$|\rho(x_i,x_i) - \rho(u,v)| \leq (|u-x_i| + |v-x_i|) \sup_{u\neq v} \left|\rho^{(0,1)}(u,v)\right|$$

$$
\begin{aligned}
&\leq \sup_k(\Delta m_k) \sup_{u\neq v}\left|\rho^{(0,1)}(u,v)\right| \\
&\leq \frac{C}{n}
\end{aligned}
$$

for some positive constant C.
Likewise for $m_{i-1} < v < u < m_i$, we have

$$
|\rho(x_i, x_i) - \rho(u,v)| \leq \frac{C}{n}
$$

For the off-diagonal term (5), we use the Taylor expansion of the covariance ρ around (x_i, x_j), $i \neq j$ up to order two since $\rho(u,v)$ is assumed to have twice continuous derivatives for $u \neq v$.

$$
\begin{aligned}
\rho(u,v) = &\rho(x_i,x_j) + (u-x_i)\rho^{(1,0)}(x_i,x_j) + (u-x_i)(v-x_j)\rho^{(1,1)}(x_i,\eta_j) \\
&+\frac{1}{2}(u-x_i)^2\rho^{(2,0)}(\xi_i,v)
\end{aligned}
$$

where the intermediate points ξ_i, η_i are between (x_i, u) and (x_j, v) respectively. Then

$$
|\rho(x_i,x_j) - \rho(u,v)| \leq |u-x_i|A^{(1,0)} + |u-x_i||v-x_j|A^{(1,1)} + \frac{1}{2}(u-x_i)^2 A^{(2,0)}
$$

where

$$
A^{(i,j)} = \sup_{u\neq v}\left|\rho^{(i,j)}(u,v)\right|,\ 0 \leq i+j \leq 2.
$$

Let $B = \max_{0\leq i+j\leq 2} A^{(i,j)}$, then

$$
\begin{aligned}
|\rho(x_i,x_j) - \rho(u,v)| &\leq B\left(\frac{1}{2}\sup_k(\Delta m_k) + \frac{3}{8}\sup_k(\Delta m_k)^2\right) \\
&\leq C\left(\frac{1}{n} + o\left(\frac{1}{n}\right)\right)
\end{aligned}
$$

It follows from the diagonal and the off-diagonal terms that

$$
\begin{aligned}
(3) &\leq \frac{C}{n}(mh^2)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_{m_{i-1}}^{m_i}\int_{m_{j-1}}^{m_j} K\left(\frac{x-u}{h}\right)K\left(\frac{x-v}{h}\right)du\,dv \\
&\leq \frac{C}{mn}\left[h^{-1}\int_0^1 K\left(\frac{x-u}{h}\right)du\right]^2 \\
&= O\left(\frac{1}{mn}\right)
\end{aligned}
$$

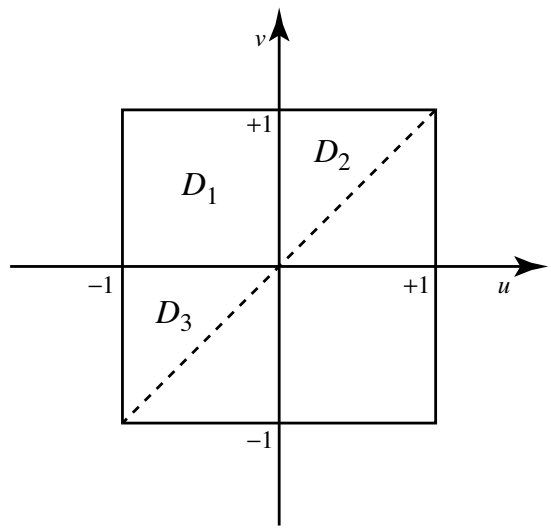

Fig. 10.

Therefore

$$
\begin{aligned}
var(\widehat{f}_h(x)) &= m^{-1} \int_{-1}^{1} \int_{-1}^{1} \rho(x-hu, x-hv) K(u) K(v) du dv + O\left(\frac{1}{mn}\right) \\
&= 2m^{-1} \int_{-1}^{1} \int_{u}^{1} \rho(x-hu, x-hv) K(u) K(v) du dv + O\left(\frac{1}{mn}\right) \\
&= 2m^{-1} \sum_{i=1}^{3} \int \int_{D_i} \rho(x-hu, x-hv) K(u) K(v) du dv + O\left(\frac{1}{mn}\right)
\end{aligned}
$$

where

$$D_1 = \{1 > v > 0 > u\},\ D_2 = \{1 > v > u > 0\} \text{ and } D_3 = \{0 > v > u > -1\}$$

as shown in the figure

We expand the covariance function ρ along the diagonal (x, x) in each area D_i, $i = 1, 2, 3$.

- *When* $(u, v) \in D_1$, we have:

$$\rho(x-hu, x-hv) = \rho(x, x) - hu\rho^{(1,0)}(\eta_x, x) - hv\rho^{(0,1)}(x-hu, \xi_x)$$

when the intermediate points are such that: $x < \eta_x < x - hu$ and $x - hv < \xi_x < x - hu$. Moreover, using the assumptions on ρ, we have:

$$\rho^{(0,1)}(x - hu, \xi_x) = \rho^{(0,1)}(x, \xi_x) - hu\rho^{(1,1)}(\xi'_x, \xi_x)$$

where $x - hv < \xi_x < x < \xi'_x < x - hu$, so that

$$\rho(x-hu, x-hv) = \rho(x,x) - hu\rho^{(0,1)}(x, \eta_x) - hv\rho^{(0,1)}(x, \xi_x) + huv\rho^{(1,1)}(\xi'_x, \xi_x)$$

Therefore as $h \to 0$

$$\begin{aligned}\frac{\rho(x - hu, x - hv) - \rho(x, x)}{h} &= -u\rho^{(0,1)}(x, \eta_x) - v\rho^{(0,1)}(x, \xi_x) + uv\rho^{(1,1)}(\xi'_x, \xi_x) \\ &\to -u\rho^{(0,1)}(x, x^+) - v\rho^{(0,1)}(x, x^-)\end{aligned}$$

Using that $A^{(1,1)} = \sup_{u \neq v} \left|\rho^{(1,1)}(u, v)\right| < \infty$

- *When* $(u, v) \in D_2$: using appropriate expansions of ρ and skipping the details, we obtain:

$$\begin{aligned}\rho(x - hu, x - hv) &= \rho(x, x) - hu\rho^{(0,1)}(x, \eta_x) - hv\rho^{(0,1)}(x, \xi_x + hv) \\ &\quad + h^2uv\left[\rho^{(1,1)}(\xi_x + hv, \eta'_x) + \rho^{(2,0)}(\xi'_x, x - hv)\right] \\ &\to -u\rho^{(0,1)}(x, x^+) - v\rho^{(0,1)}(x, x^-)\end{aligned}$$

where the intermediate points satisfy:

$$x - hv < \xi_x < \xi'_x < \xi_x + hv \text{ and } x - hv < \eta'_x < x < \eta_x + hv.$$

And thus using $A^{(i,j)} = \sup_{u \neq v} \left|\rho^{(i,j)}(u, v)\right| < \infty$, we have as $h \to 0$

$$\frac{\rho(x - hu, x - hv) - \rho(x, x)}{h} \to -v\rho^{(0,1)}(x, x^-) - u\rho^{(0,1)}(x, x^+)$$

- Finally, *when* $(u, v) \in D_3$, we obtain:

$$\begin{aligned}\rho(x - hu, x - hv) &= \rho(x, x) - hu\rho^{(0,1)}(x, \eta_x) - hv\rho^{(0,1)}(x, \xi_x + hu) \\ &\quad + huv\left[\rho^{(1,1)}(\xi'_x, \xi_x + hu) + \rho^{(0,2)}(x - hu, \xi'_x)\right] \\ &\to -u\rho^{(0,1)}(x, x^+) - v\rho^{(0,1)}(x, x^-)\end{aligned}$$

and thus, as $h \to 0$

$$\frac{\rho(x - hu, x - hv) - \rho(x, x)}{h} \to -v\rho^{(0,1)}(x, x^-) - u\rho^{(0,1)}(x, x^+)$$

where the intermediate points are such that:

$$\xi_x + hu < x < \xi'_x < \xi_x < x - hu \text{ and } x + hu < x < \eta'_x < x - hu.$$

Therefore

$$\begin{aligned}
&\int_{-1}^{1}\int_{u}^{1} \rho(x-hu, x-hv)K(u)K(v)dudv \\
&= h\int_{-1}^{1}\int_{u}^{1} \frac{\rho(x-hu, x-hv)-\rho(x,x)}{h}K(u)K(v)dudv \\
&= -\rho^{(0,1)}(x,x^-)\sum_{i=1}^{3}\int\int_{D_i} vK(u)K(v)dudv - \rho^{(0,1)}(x,x^+)\sum_{i=1}^{3}\int\int_{D_i} uK(u)K(v)dudv + o(h) \\
&= -\rho^{(0,1)}(x,x^-)A^- - \rho^{(0,1)}(x,x^+)A^+ + o(h)
\end{aligned}$$

Lemma 7

$$A^- = -A^+ = \frac{1}{4}C_K$$

where

$$C_k = \int_{-1}^{1}\int_{-1}^{1} |u-v|K(u)K(v)dudv$$

From the Lemma, we obtain:

$$\begin{aligned}
&\int_{-1}^{1}\int_{u}^{1} \rho(x-hu, x-hv)K(u)K(v)dudv \\
&= \frac{1}{2}\rho(x,x) - h\left(\rho^{(0,1)}(x,x^-) - \rho^{(0,1)}(x,x^+)\right)\frac{1}{4}C_K + o(h)
\end{aligned}$$

and thus

$$var(\widehat{f}_h(x)) = \frac{1}{m}\left(\rho(x,x) - \frac{1}{2}\alpha(x)C_K h\right) + o\left(\frac{h}{m}\right) + O\left(\frac{1}{mn}\right)$$

and the result of Theorem 2 is established from the decomposition (3).

Proof of Theorem 5

The covariance function ρ is assumed to have at least twice continuous mixed partial derivatives with $\rho^{(i,j)}(x,x) \neq 0$, $i+j=2$. Then we can expand ρ around (x,x) up to order 2 as follows:

$$\rho(x-hu, x-hv) = \rho(x,x) - hu\rho^{(1,0)}(x,x) - hv\rho^{(0,1)}(x,x)$$

$$+h^2uv\rho^{(1,1)}(\xi_x^1,x)+\frac{1}{2}h^2u^2\rho^{(2,0)}(\xi_x^2,x)$$
$$+\frac{1}{2}h^2v^2\rho^{(0,2)}(x-hu,\eta_x)$$

when the intermediate points ξ_x^1, ξ_x^2 are between x and $x-hu$, and η_x is between x and $x-hv$.
Since $\sup_{u\neq v}\left|\rho^{(i,j)}(u,v)\right|<\infty$, $i+j=2$, we have from the dominated convergence Theorem that:

$$\int_{-1}^{1}\int_{-1}^{1}\rho(x-hu,x-hv)K(u)K(v)dudv$$
$$=\rho(x,x)-2h\rho^{(0,1)}(x,x)\int_{-1}^{1}\int_{-1}^{1}uK(u)K(v)dudv$$
$$+h^2\rho^{(1,1)}(x,x)\int_{-1}^{1}\int_{-1}^{1}uvK(u)K(v)dudv$$
$$+\frac{1}{2}h^2\rho^{(2,0)}(x,x)\int_{-1}^{1}\int_{-1}^{1}u^2K(u)K(v)dudv$$
$$+\frac{1}{2}h^2\rho^{(0,2)}(x,x)\int_{-1}^{1}\int_{-1}^{1}v^2K(u)K(v)dudv+o(h^2)$$

Since the kernel K is symmetric around 0, we obtain

$$\int_{-1}^{1}\int_{-1}^{1}\rho(x-hu,x-hv)K(u)K(v)dudv=\rho(x,x)+h^2\rho^{(0,2)}(x,x)d_K+o(h^2)$$

Therefore we have

$$var\left(\widehat{f}_h(x)\right)=\frac{1}{m}\left(\rho(x,x)+h^2\rho^{(0,2)}(x,x)d_K\right)+o\left(\frac{h^2}{m}\right)+O\left(\frac{1}{mn}\right)$$

and the result of Theorem 5 follows from the decomposition (3).

References

[1] N. Altman. Kernel regression estimation of data with correlated errors. *Journal of the American Statistical Association*, 85:749–759, 1990.

[2] T. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley., 1984.

[3] K. Benhenni and M. Rachdi. Nonparametric estimation of the average growth curve from quantized observations and correlated errors. *C. R. Acad. Sci. Paris, Sér. I337:?-?, 2005.*

[4] K. Benhenni and M. Rachdi. Nonparametric estimation of the regression function from quantized observations. *Comput. Stat. Data Anal. ?, No.?, ?–?, 2005.*

[5] J. Boularan, L. Ferré and P. Vieu. Growth curves: a two-stage nonparametric approach. *J. Stat. Plann. Inference*, (38)3:327-350, 1994.

[6] J. Fan. Design-adaptive nonparametric regression. *J. Am. Stat. Assoc.* (87)420:998–1004, 1992.

[7] J. Fan and I. Gijbels. *Local polynomial modelling and its applications.* Monographs on Statistics and Applied Probability. 66. London: Chapman and Hall. xvi, 1996.

[8] E. Ferreira, V. Núñez-Antón and J. Rodríguez-Póo. Kernel regression estimates of growth curves using nonstationary correlated errors. *Stat. Probab. Lett.* (34)4:413-423, 1997.

[9] T. Gasser and M. G. Müller. Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, 11:171–185, 1984.

[10] S. Geisser. Growth curve analysis. *Analysis of variance, Handb. Stat.*, 1:89–115, 1980.

[11] W. Härdle. *Applied nonparametric regression*, volume 19. Cambridge, University Press, 1989.

[12] J Hart. Kernel regression with time series errors. *Journal of the Royal Statistical Society – Series B*, 53:173–187, 1991.

[13] J. Hart and T. Wehrly. Kernel regression estimation using repeated measurements data. *Journal of the American Statistical Association*, 81:1080–1088, 1986.

[14] X. Lin and R.J. Carroll. Nonparametric function estimation for clustered data when the predictor is measured without/with error. *J. Am. Stat. Assoc.*, (95)450:520–534, 2000.

[15] S. P. Lloyd. Least squares quantization in pcm. *IEEE Trans. Inform. Theory*, IT-28:129–137, 1982.

[16] J. Max. Quantizing for minimum distortion. *IRE. Trans. Inform. Theory*, IT-6:7–12, 1960.

[17] E. A. Nadaraya. On estimating regression. *Theory Probab. Appl.*, 9:141–142, 1964.

[18] V. Núñez-Antón and G. Woodworth. Analysis of longitudinal data with unequally spaced observations and time dependent correlated errors. *Biometrics*, (50):445–456, 1994.

[19] P. Sarda and P. Vieu. *Kernel regression. Smoothing and regression: Approaches, Computation, and Application.* Wiley Series in Probability and Statistics, ed. m. g. schimek edition, 2000.

[20] G. S. Watson. Smooth regression analysis. *Sankhya Ser. A*, 26:359–372, 1964.

[21] T. Severini and J. Staniswalis. Quasi-likelihood estimation in semiparametric models. *J. Am. Stat. Assoc.*, 89(426):501–511, 1994.

[22] D. Von Rosen. The growth curve model: A review. *Commun. Stat., Theory Method*, 20(9):2791–2822, 1991.

[23] C. Wild and T. Yee. Additive extensions to generalized estimating equation methods. *J. R. Stat. Soc., Ser. B*, 58(4):711–725, 1996.

[24] O. Wu, C-T. Chiang, and D. Hoover. Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *J. Am. Stat. Assoc.*, 93(444):1388–1402, 1998.

[25] S. Zeger and P. Diggle. Semiparametric models for longitudinal data with application of CD4 cell numbers in HIV seroconverters. *Biometrics*, 50(3):689–699, 1994.

[26] D. L. Zimmerman and V. Núñez-Antón. Parametric modelling of growth curve data: An overview. (With comments). *Test*, (10)1:1–73, 2001.