HAL open science

No effect tests in regression on functional variable and some applications to spectrometric studies

Laurent Delsol

▶ To cite this version:

Laurent Delsol. No effect tests in regression on functional variable and some applications to spectrometric studies. Computational Statistics, 2013, 28 (4), pp.1-37. 10.1007/s00180-012-0378-1 . hal-00745267

HAL Id: hal-00745267

https://hal.science/hal-00745267v1

Submitted on 25 Oct 2012

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.

Distributed under a Creative Commons Attribution 4.0 International License

Computational Statistics manuscript No.
(will be inserted by the editor)

No effect tests in regression on functional variable and some applications to spectrometric studies

Laurent DELSOL

the date of receipt and acceptance should be inserted later

Abstract Recent advances in structural tests for regression on functional variable are used to construct test of no effect. Various bootstrap procedures are considered and compared in a simulation study. These tests are finally applied on real world datasets dealing with spectrometric studies using the information collected during this simulation study. The results obtained for the Tecator dataset are relevant and corroborated by former studies. The study of a smaller dataset concerning corn samples shows the efficiency of our method on small size samples. Getting information on which derivatives (or which parts) of the spectrometric curves have a significant effect allows to get a better understanding of the way spectrometric curves influence the quantity to predict. In addition, a better knowledge of the structure of the underlying regression model may be useful to construct a relevant predictor.

Keywords · no effect test · regression · functional variable · bootstrap · spectrometric curves.

1 Introduction

Many real world issues involve functional type phenomena (evolution of a quantity over time, spectrometric curves, sound records, images, ...). Recent advances in computerized measuring devices now allow to collect, stock and treat data discretized on thinner grids which enable to reflect their functional nature. In order to avoid the classical drawbacks of a multivariate modelization, it is often more relevant to adopt a more general point of view and consider these data as the discretization of functional random variables (i.e. random variables taking values in an infinite dimensional space). More precisely, we consider in this paper functional random variables taking values in a semi-metric space $(\mathcal{E}, d)$, following the definition introduced in Ferraty and Vieu (2000). Functional statistics have become an important topic of modern statistics. A more complete overview of recent advances in statistics for functional data is given in the monographs by Ramsay and Silverman (1997, 2002, 2005), Bosq (2000), Ferraty and Vieu (2006) and in the reviews by Davidian *et al.* (2004), González Manteiga and Vieu (2007), Valderama (2007), or Ferraty (2010) (see also for additional references Ferraty and Romain (2011) and the web site of the working

Laurent DELSOL
Institut de Statistique, U.C.L., 20 voie du roman pays, 1348 Louvain-la-Neuve, Belgique.
Tel.: +32-10-479403
Fax: +32-10-473032
E-mail: laurent.delsol@uclouvain.be

presently at
MAPMO, Université d'Orléans, Bâtiment de mathématiques - Rue de Chartres. B.P. 6759 - 45067 Orléans cedex 2, France.
Tel.: +33238492696
Fax: +33238417205
E-mail: laurent.delsol@univ-orleans.fr

group STAPH of Toulouse: $http://www.lsp.ups-tlse.fr/staph/$). In this paper, we consider more precisely regression models of the form

$$Y = r(X) + \epsilon, \tag{1}$$

where Y is a real random variable, X is a functional random variable taking values in a semi-metric space $(\mathcal{E}, d)$, and the residual ϵ fulfills $\mathbb{E}[\epsilon \mid X] = 0$. Many estimation results have been proposed for such models in the linear case (see for instance Ramsay and Dalzell, 1991, Cardot *et al.*, 1999, Ramsay and Silverman, 2005, and Crambes *et al.*, 2009) or in the nonparametric case with kernel methods (see Ferraty and Vieu, 2006 for a review). Other alternative methods have also been considered by James and Silverman (2005), Rossi and Conan Guez (2005), Laloë (2007), or Hernandez *et al.* (2008), among others.

Testing the validity of structural assumptions made on the underlying regression model is a different issue. They may be interesting by themselves (testing the validity of an a priori model, of a given kind of modelization, ...). However, structural tests are also complementary tools to estimation methods. On the one hand, they may be used before any estimation to test if X has an effect on Y and more generally if this effect has a specific nature. Because many estimation methods are based on structural hypotheses (linear model, single-index model,...), it seems relevant to check the validity of these assumptions. On the other hand, estimation results may lead to the formulation of new structural assumptions whose validity has to be considered. An interesting question is for instance the detection of the features of a functional data (for instance the portions of a curve) that have a significant effect on the response variable. The construction of no effect tests is a way to answer to such issues. Recently, a kernel method to construct general structural tests in regression on functional variable has been proposed by Delsol *et al.* (2011), extending the ideas of Härdle and Mammen (1993). The aim of this paper is to present how this general approach can be used to construct no effect tests. The considered testing procedure completes former no effect tests proposed in the specific case of functional linear models by Cardot *et al.* (2003, 2004) or Müller and Stadtmüller (2005) and a no effect test based on projection methods introduced by Gadiaga and Ignaccolo (2005). The no effect test presented in this paper allows to consider non linear alternatives and does not depend on the choice of the projection basis.

The remainder of the paper is organized as follows. In the next section, we consider the issue of no effect test, give the expression of the test statistic and state its asymptotic normality under the null hypothesis and its divergence under the alternative. Then Corollary 1 focuses on specific local alternatives. To improve the performances of our approach, various residual based bootstrap methods are introduced to compute efficiently the threshold value. Section 3 is devoted to simulation studies to compare the performances of our bootstrap methods. The issue of the choice of the smoothing parameter and the bootstrap iterations number is also discussed. Finally, the application of our testing procedures to Tecator and Corn spectrometric datasets is presented in Section 4. The proofs of the main results of this work are given in Section 6.

2 No effect tests in regression on functional variable

2.1 Problem presentation and test statistic

This paper does not directly focus on the classical issues of estimating the regression operator or predicting a value of the response variable. The aim is to get a better understanding of the underlying regression model from the use of no effect tests. Such tests may be used as a preliminary step to check if the explanatory curve (or some of its features) has a significant effect on the response, what may be also useful to construct a relevant prediction method. We want to test if the explanatory functional variable X has an effect on the variable of interest Y by checking if the regression operator r is constant.

Assume now we have three independent i.i.d. samples $D : (X_i, Y_i)_{1 \leq i \leq n}$, $D_0 : (X_{0,i}, Y_{0,i})_{1 \leq i \leq m_n}$ and $D_1 : (X_{1,i}, Y_{1,i})_{1 \leq i \leq l_n}$ of respective lengths n, m_n and l_n corresponding to the same regression model (1) (as explained below, the regression operator [and hence the response] of this model may

depend on n under the alternative hypothesis). In practice these samples may come from an original sample of size $N_n = n + m_n + l_n$. Then, for any function $f : \mathcal{E} \mapsto \mathbb{R}$ such that $\mathbb{E}[|f(X)|] < +\infty$, we use the notation $\int f(x) dP_X(x) := \mathbb{E}[f(X)]$ and denote P_X the law of X. Finally, by simplicity, we introduce the notation $\overline{Y_0} := \frac{1}{m_n} \sum_{i=1}^{m_n} Y_{0,i}$ (similar notations are used later to make reference to empirical means of other variables).

The aim of this paper is to present a theoretical and practical way to test the null hypothesis

$$\mathcal{H}_0 : \{\exists C \in \mathbb{R}, \mathbb{P}(r(X) = C) = 1\}$$

against the set of local alternatives

$$\mathcal{H}_1 : \left\{ \forall n \in \mathbb{N}^*, \eta_n := \inf_{C \in \mathbb{R}} \|r_n - C\|_{\mathbb{L}^2(wdP_X)} > 0 \right\},$$

where r_n are the regression operators corresponding to a sequence of local alternative models $Y^n = r_n(X) + \epsilon$, w is a known weight function (see the definition of the test statistic below), and $(\eta_n)_{n \in \mathbb{N}}$ is a sequence of positive numbers which may tend to 0 (see assumption (7)). The alternative hypothesis $\mathcal{H}_1$ allows to consider the case of a fixed alternative model $Y = r(X) + \epsilon$ with $\mathcal{H}_1 : \{\inf_{C \in \mathbb{R}} \|r - C\|_{\mathbb{L}^2(wdP_X)} > 0\}$. But $\mathcal{H}_1$ is defined in such a way that one may focus on a sequence of local alternative models $Y^n = r_n(X) + \epsilon$ (from which, for each n, D, D_0, and D_1 are three independent i.i.d. samples) for which the distance between the true regression operator r_n and the family of constant operators may tend to 0 when n goes to infinity. Note the law of explanatory variables and residuals do not depend on n (only the law of the responses depends on n). Such alternatives are usually considered to put in relief the way the test statistic is able to detect smaller and smaller differences when n grows. Consider for instance a sequence of regression models defined by $r_n = C + \eta_n \Delta(x)$ (common local alternative, see e.g. Stute, 1997, Lavergne and Patilea, 2007), with $\eta_n \to 0$ and fulfilling assumption (6). Even if the sequence $(r_n)_{n \in \mathbb{N}}$ tends to a constant operator when n grows, the power of the test still tends to one (see Corollary 1). Stating the consistency under such sequences of local alternatives is relevant to investigate from a theoretical point of view the power of the test. However, in simulations studies and data set examples we usually deal with fixed sample sizes and fixed alternatives. In real world studies the underlying regression model usually does not change if more observations are made. Considering local alternative instead of fixed ones mainly have a theoretical interest: giving a better understanding of the test behavior in limit situations.
By simplicity, we forget the dependence on n in our notations, write Y_i instead of Y_i^n and r instead of r_n under the set of local alternatives $\mathcal{H}_1$. We consider in this paper the following test statistic:

$$T_n = \int \left(\sum_{i=1}^{n} \left(Y_i - \overline{Y_0}\right) K\left(\frac{d(x, X_i)}{h_n}\right) \right)^2 w(x)\, dP_X(x),$$

in which K and d respectively stand for a kernel function and a semimetric on $\mathcal{E}$. It corresponds to the test statistic proposed in Delsol *et al.* (2011) in the specific case of no effect tests. The use of a weight function w is a standard tool in structural testing procedures (see for instance [32], [8]]) as an alternative to the assumption that the law of X has a bounded support (see for instance [37]). In our simulations and applications, the weight function w is the indicator of a ball $B(0, M)$, with M large enough to ensure all X_i's are in this ball. However, it is possible to consider other choices of the weight function (for instance to remove outliers). The test statistic T_n also depends on the nature of the kernel function K. It is usual to use a quadratic kernel (see for instance [25] for its definition) in practice. The use of an other kernel function may be more relevant in some specific situations. However, one may expect that as in estimation, the impact of the choice of the kernel is small with respect to the influence of the choice of the smoothing parameter.

In order to explain more the effect of the choice of the semi-metric, let us assume that there exist $\tilde{X}$ (respectively $(\tilde{X}_i)_{1 \le i \le n}$, $(\tilde{X}_{0,i})_{1 \le i \le m_n}$, and $(\tilde{X}_{1,i})_{1 \le i \le l_n}$) such that $d(X, \tilde{X}) = 0$ a.s. (respectively $\forall 1 \le i \le n$, $1 \le j \le m_n$, $1 \le k \le l_n$, $d(\tilde{X}_i, X_i) = d(\tilde{X}_{0,j}, X_{0,j}) = d(\tilde{X}_{1,k}, X_{1,k}) =$

0 *a.s.*). By definition of T_n one gets

$$T_n = \mathbb{E}[(\sum_{i=1}^{n}(Y_i - \overline{Y_0})K(\frac{d(X_i, X)}{h_n}))^2 w(X)]$$
$$= \mathbb{E}[(\sum_{i=1}^{n}(Y_i - \overline{Y_0})K(\frac{d(\tilde{X}_i, \tilde{X})}{h_n}))^2 w_0(\tilde{X})],$$

with $w_0(\tilde{X}) = \mathbb{E}[w(X)|\tilde{X}]$. Hence, T_n actually considers assumptions $\mathcal{H}_0$ and $\mathcal{H}_1$ on the model $Y = r(\tilde{X}) + \tilde{\epsilon}$, with $\mathbb{E}[\tilde{\epsilon}|\tilde{X}] = 0$. If d is a metric, $d(x_1, x_2) = 0 \Leftrightarrow x_1 = x_2$ and there is no ambiguity on the model. However, when a projection semi-metric is used, the null hypothesis and the alternative one concern the effect of the selected components of the explanatory variable. When a semi-metric based on derivative is considered, we actually test the no effect of the derivative of the curve. Consequently, when d is a semi-metric, the test actually focuses on the null and alternative hypotheses for any model $Y = r(\tilde{X}) + \tilde{\epsilon}$, with $d(X, \tilde{X}) = 0$ a.s. and $\mathbb{E}[\tilde{\epsilon}|\tilde{X}] = 0$. Moreover, the main results of this work may be obtained for these new hypothesis if assumptions are changed in consequence (X and ϵ replaced by $\tilde{X}$ and $\tilde{\epsilon}$). By simplicity we keep the notations used for model (1).

2.2 Assumptions and theoretical results

In order to obtain the asymptotic normality of T_n, we need some assumptions. These conditions were initially introduced and discussed in Delsol *et al.* (2011) where various sets of alternative assumption are also considered. We start with assumptions on the statistic T_n:

w is nonnegative, not P_X a.s. null, has a bounded support W, and is bounded. (2)

K has a compact support [0, 1], is nonincreasing and $\mathcal{C}^1$ on $]0, 1[$ and $K(1) > 0$. (3)

For all $\gamma > 0$ the γ neighborhood of W is defined by

$$W_\gamma := \{u \in \mathcal{E}, \exists s \in W, d(u, s) \leq \gamma\}$$

and some assumptions are made on the regression model:

Under $\mathcal{H}_1$, $\exists \gamma_0 > 0$, $\exists C_0 > 0$, $\exists \beta > 0$, $\forall x, y \in W_{\gamma_0}$, $\forall n \in \mathbb{N}^*$, $|r_n(x) - r_n(y)| \leq C_0 d^\beta(x, y)$, (4)

$\exists M > 0$, $\mathbb{E}\left[\epsilon^4|X\right] \leq M$ a.s. and $\mathbb{E}\left[\epsilon^2 \mid X\right] = \sigma_\epsilon^2 > 0$. (5)

Finally, we introduce some notations for key elements and sets that appear in our study:

$$F_x(s) = \mathbb{P}(d(x, X) \leq s), \; F_{x,y}(s, t) = \mathbb{P}(d(x, X) \leq s, \, d(y, X) \leq t),$$
$$\Omega_4(s) = \int_{E \times E} F_{x,y}^2(s, s) \, w(x) \, w(y) \, dP_X(x) \, dP_X(y).$$

The following assumptions are linking the sequence η_n, the smoothing parameter h_n and small ball probabilities:

$\exists C_1, C_2$, $\exists \Phi : \mathbb{R}_*^+ \to \mathbb{R}_*^+$, $\forall x \in W_{\gamma_0}$, $\forall n \in \mathbb{N}$, $C_1 \Phi(h_n) \leq F_x(h_n) \leq C_2 \Phi(h_n)$, and $\Phi(h_n) \to 0$ (6)

$\exists v_n \to +\infty$, $\theta_n := v_n \left(\frac{1}{n^{\frac{1}{2}} \Phi^{\frac{1}{4}}(h_n)} + h_n^\beta \right) \to 0$ and $\eta_n \geq \theta_n$, (7)

$\exists C_3 > 0$, $\Omega_4(h_n) \geq C_3 \Phi^{3+l}(h_n)$ with $l < \frac{1}{2}$ and $n\Phi^{1+2l}(h_n) \to +\infty$. (8)

The sequence r_n is uniformly bounded on W_{γ_0}, $\frac{1}{m_n} \sum_{i=1}^{m_n} r_n(X_{0,i}) = O_p(1)$ and $n\Phi^{\frac{1-l}{2}}(h_n) \, m_n^{-1} \to 0$. (9)

Let us make some comments on previous conditions (see Delsol *et al.*, 2010, for a deeper discussion). Firstly, assumption (3) is common for functional kernel smoothing methods (see for instance [25] or [20]). The assumption $K(1) > 0$ may be replaced by any assumption allowing to state there exists two positive constants C and C' such that $\mathbb{E}[K(\frac{d(X,x)}{h_n})] \geq CF_x(h_n)$ and $\mathbb{E}[K(\frac{d(X,x)}{h_n})K(\frac{d(X,y)}{h_n})] \geq C'F_{x,y}(h_n)$. Assumption (4) requires some smoothness of the local alternatives with respect to the semi-metric d. This notably implies that $r_n(x_1) = r_n(x_2)$ for any x_1, x_2 such that $d(x_1, x_2) = 0$. If d is a semi-metric, this is not trivial and means the effect of X is reduced the effect of any $\tilde{X}$ such that $d(\tilde{X}, X) = 0$ a.s.. Hence the test is only able to detect alternative corresponding to an effect of $\tilde{X}$. In fact the test considers hypothesis on the regression model with explanatory variable $\tilde{X}$ (see comments at the end of Section 2.1). Then, assumption (5) can be extended to take into account heteroscedastic errors. However, this would require to change the remaining assumptions and make them less readable. It is also possible to make a weaker assumption on η_n if one considers local alternatives of the form $r(X) = C + \eta_n \Delta_n$ (see Corollary 1). Then, assumptions made on the law of X mainly concern the nature of the small ball probabilities $F_x(h)$ and $F_{x,y}(h)$. The choice of the semi-metric is crucial because it has a direct influence on the regularity assumption made on the regression operator (and hence the null and alternative hypotheses) but also on the nature of these small ball probabilities. It is for instance possible to show that assumption (6) holds for some processes when one uses a projection semi-metric or that assumption (8) is fulfilled with $l = 0$ in the case of fractal processes. Finally, assumption (9) holds for instance if $m_n = n$ but it is possible to consider datasets of different sizes.
We now introduce the variables $T_{1,n}$ and $T_{2,n}$ that provide respectively bias and variance dominant terms. Their law does not depend on the nature of the regression operator.

$$T_{1,n} = \int \sum_{i=1}^{n} K^2\left(\frac{d(X_i, x)}{h_n}\right) \epsilon_i^2 w(x)\, dP_X(x),$$
$$T_{2,n} = \int \sum_{1 \leq i \neq j \leq n} K\left(\frac{d(X_i, x)}{h_n}\right) K\left(\frac{d(X_j, x)}{h_n}\right) \epsilon_i \epsilon_j w(x)\, dP_X(x),$$

We are now able to state the following theorem dealing with the asymptotic normality of T_n under the null hypothesis and its divergence under the alternative.

Theorem 1 *Under assumptions (2)-(9) one gets:*

- *Under ($\mathcal{H}_0$),* $\frac{1}{\sqrt{Var(T_{2,n})}}(T_n - \mathbb{E}[T_{1,n}]) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$,
- *Under ($\mathcal{H}_1$),* $\frac{1}{\sqrt{Var(T_{2,n})}}(T_n - \mathbb{E}[T_{1,n}]) \xrightarrow{P} +\infty$.

It is usual to consider local alternatives of the form $\mathcal{H}_1' : \{\forall n \in \mathbb{N}^*, r_n(x) = C + \eta_n \Delta_n(x)\}$ where C is a constant, η_n a sequence of positive numbers, and Δ_n a uniformly bounded sequence of centered operators. In this case, the lower bound for the conditional mean in Lemma 3 may be precised and we are able to state a more precise result under $\mathcal{H}_1'$ when the following assumptions hold:

$$\exists \gamma_0 > 0,\ \exists C_0 > 0,\ \exists \beta > 0,\ \forall x, y \in W_{\gamma_0},\ \forall n \in \mathbb{N}^*, |\Delta_n(x) - \Delta_n(y)| \leq C_0 d^\beta(x,y), \tag{10}$$
$$\mathbb{E}[\Delta_n(X)] = 0,\ \exists C_\Delta > 0,\ \forall n \in \mathbb{N}^* \Delta_n(X) \leq C_\Delta,\ \inf_{n,C} \|\Delta_n - C\|_{\mathbb{L}^2(wdP_X)} > 0, \text{ and } h_n \to 0. \tag{11}$$

The next corollary give a better understanding of the behavior of the test under these local alternatives.

Corollary 1 *Let* $Z_n = \frac{1}{\sqrt{Var(T_{2,n})}}(T_n - \mathbb{E}[T_{1,n}])$. *Under assumptions (2)-(3), (5)-(6) and (8)-(11) one gets:*

- *Under ($\mathcal{H}_0$),* $Z_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$,
- *Under ($\mathcal{H}_1'$),*
  1. *if* $n\Phi^2(h_n)\eta_n^2 \Omega_4^{-\frac{1}{2}}(h_n) \to 0$, *then* $Z_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$,
  2. *if* $n\Phi^2(h_n)\eta_n^2 \Omega_4^{-\frac{1}{2}}(h_n) \to +\infty$, *then* $Z_n \xrightarrow{P} +\infty$.

3. *if* $\exists \mu_1, \mu_2 > 0$, $\forall n \in \mathbb{N}^*$, $\mu_1 \leq n\Phi^2(h_n)\eta_n^2\Omega_4^{-\frac{1}{2}}(h_n) \leq \mu_2$ *, then* Z_n *is asymptotically Gaussian with positive mean* B_n *(uniformly bounded w.r.t. n) and variance 1 (i.e.* $Z_n - B_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$*) with*

$$B_n \sim n\Phi^2(h_n)\eta_n^2\Omega_4^{-\frac{1}{2}}(h_n)\frac{\int(\Delta_n(x))^2(\mathbb{E}\left[K\left(\frac{d(X,x)}{h_n}\right)\right]\Phi^{-1}(h_n))^2 w(x)dP_X(x)}{\sigma_\epsilon^2\sqrt{2\int\int\left(\mathbb{E}[K(\frac{d(X,x)}{h_n})K(\frac{d(X,y)}{h_n})]\right)^2 w(x)w(y)dP_X(x)dP_X(y)\Omega_4^{-1}(h_n)}}.$$

In practice one has to estimate the critical value of the test. The most natural way would be to estimate the bias and variance dominant terms and use directly the quantiles of the asymptotic law. However, the estimation of bias and variance terms seems difficult and it is often irrelevant to use directly the quantiles of the asymptotic law to estimate the threshold. Instead of doing so, bootstrap methods are introduced to generate N_{boot} datasets, with similar distribution as the original dataset, for which the null hypothesis approximately holds. Then, we compute on each datasets the test statistic and, for a given level α, as the critical value we take the $1-\alpha$ quantile of these values.

Finally, the test statistic T_n is based on an integral with respect to dP_X. Because this integral can be regarded as a conditional expectation, we propose to use the following Monte Carlo approximation:

$$T_n \approx V_n := \frac{1}{l_n}\sum_{k=1}^{l_n}\left(\sum_{i=1}^{n}\left(Y_i - \overline{Y_0}\right)K\left(\frac{d\left(X_{1,k}, X_i\right)}{h_n}\right)\right)^2 w\left(X_{1,k}\right).$$

2.3 Residual based bootstrap procedures

From the pioneer work of Efron (1979), bootstrap methods have encountered a strong interest. They have been extensively studied in the context of nonparametric regression with scalar or multivariate covariate. They allow for example to improve the performances of confidence bands (see for instance Härdle and Marron, 1990, Cao, 1991, Hall, 1992) and testing procedures (see for instance Hall and Hart,1990, Härdle and Mammen, 1993, Stute *et al.*, 1998) and can also be useful to choose the smoothing parameter (see for instance Hall, 1990, Gonzalez Manteiga *et al.*, 2004). In the functional context, bootstrap methods have been less developed (see Cuevas and Fraiman, 2004, Fernández de Castro *et al.*, 2005 and Cuevas *et al.*, 2006). Our approach follows ideas introduced for the functional nonparametric regression model in the recent work of Ferraty *et al.* (2010) which focuses on the use of residual-based bootstrap methods to estimate confidence bands and provides both theoretical and practical interesting results.

Direct methods that consist in making bootstrap directly on the pairs (X_i, Y_i) are not adapted to our situation. Indeed, it is obvious that if in the original dataset r is not constant, this will be the same for bootstrap datasets obtained from such procedures. We propose to keep X_i unchanged and apply bootstrap methods on the estimated residuals. Then, we construct bootstrap responses making as if the null hypothesis is true.

We first introduce some notations to make easier the understanding of the bootstrap procedure. If F is the cumulative distribution function of a given law $\mathcal{L}$, and $U \sim \mathcal{U}([0;1])$, then $F^{-1}(U) \sim \mathcal{L}$, where F^{-1} denote the generalized inverse of F. Consequently, it is natural to propose to generate bootstrap samples from an estimation of this cumulative distribution obtained from the original dataset. A first estimation is the empirical cumulative distribution function, whose use leads to resampling methods (bootstrap values are drawn with replacement from the original sample). Let $(Z_1, \ldots, Z_q)$ be q real random variables and denote $(Z_{(1)}, \ldots, Z_{(q)})$ the corresponding order statistics. We define the function $G_{\{(Z_1,\ldots,Z_q)\}} : [0;1] \mapsto \mathbb{R}$ by

$$G_{\{(Z_1,\ldots,Z_q)\}}(u) = \begin{cases} Z_{(i)} + (u(q+1)-i)(Z_{(i+1)} - Z_{(i)}) \text{ if } u \in [\frac{i}{q+1}; \frac{i+1}{q+1}[, 1 \leq i < q \\ Z_{(1)} - 0.1|Z_{(1)}| \text{ if } u \in [0; \frac{1}{q+1}] \\ Z_{(q)} + 0.1|Z_{(q)}| \text{ if } u \in [\frac{q}{q+1}; 1] \end{cases}$$

G is the general inverse of kind of estimator of F (see the next lines). Let $\tilde{F}_q$ be the linear interpolation between the points $(Z_{(i)}, \frac{i}{q})$ (piecewise affine and continuous approximation of F between $Z_{(1)}$ and $Z_{(q)}$). Now, we may consider

$$F_1(t) = \begin{cases} 0 \text{ if } t < Z_{(1)}, \\ 1 \text{ if } t > Z_{(n)}, \\ \tilde{F}_q(t) \text{ elsewhere} \end{cases}$$

to be able to generate bootstrap values in $[Z_{(1)}, Z_{(q)}]$ that are not in the original sample. Because we would like to be able to generate bootstrap values out of the range of the original sample, we finally consider

$$F_2(t) = \begin{cases} -1/q \text{ if } t \leq Z_{(1)} - 0.1|Z_{(1)}|, \\ 1 \text{ if } t \geq Z_{(n)} + 0.1|Z_{(n)}|, \\ \frac{q}{q+1} F_1(t) \text{ elsewhere} \end{cases}$$

The function G is the generalized inverse of F_2 which is an approximation of the common cumulative distribution of $Z_i's$ if q is large.

We propose the procedure presented hereafter in which $\hat{r}$ stands for the functional kernel estimator constructed from the three samples D, D_0 and D_1.

Bootstrap Procedure:

Pre-treatment:

1. Compute estimated residuals: $\hat{\epsilon_i} = Y_i - \hat{r}(X_i), 1 \leq i \leq n$ and $\hat{\epsilon}_{0,i} = Y_{0,i} - \hat{r}(X_{0,i}), 1 \leq i \leq m_n$.
2. Center estimated residuals: $\hat{\hat{\epsilon}}_i = \hat{\epsilon_i} - \bar{\hat{\epsilon}}, 1 \leq i \leq n$, and $\hat{\hat{\epsilon}}_{0,i} = \hat{\epsilon}_{0,i} - \bar{\hat{\epsilon}}_0, 1 \leq i \leq m_n$.

Repeat for $1 \leq b \leq N_{boot}$ steps 3-5 to generate N_{boot} bootstrap values of T_n:

3. Generate bootstrap residuals (three alternative methods):
   a) Resampling or Naive bootstrap: $\left(\epsilon_i^{*,b}\right)_{1 \leq i \leq n}$, respectively $\left(\epsilon_{0,i}^{*,b}\right)_{1 \leq i \leq m_n}$, are drawn with replacement from $\left\{\hat{\hat{\epsilon}}_i, 1 \leq i \leq n\right\}$, respectively from $\left\{\hat{\hat{\epsilon}}_{0,i}, 1 \leq i \leq m_n\right\}$.
   or
   b) Smooth Naive bootstrap: $\epsilon_i^{*,b} = G_{\{(\hat{\hat{\epsilon}}_i)_{1 \leq i \leq n}\}}(U_i)$, $1 \leq i \leq n$, and $\epsilon_{0,i}^{*,b} = G_{\{\left(\hat{\hat{\epsilon}}_{0,i}\right)_{1 \leq i \leq m_n}\}}(U_{0,i})$, $1 \leq i \leq m_n$, where $((U_i)_{1 \leq i \leq n}, (U_{0,j})_{1 \leq j \leq m_n}) \overset{i.i.d.}{\sim} \mathcal{U}([0;1])$.
   or
   c) Wild bootstrap: $\epsilon_i^{*,b} = \hat{\hat{\epsilon}}_i U_i, 1 \leq i \leq n$ and $\epsilon_{0,i}^{*,b} = \hat{\hat{\epsilon}}_{0,i} U_{0,i}, 1 \leq i \leq m_n$, where the variables $((U_i)_{1 \leq i \leq n}, (U_{0,i})_{1 \leq i \leq m_n}) \overset{i.i.d.}{\sim} P_B$, are independent of $(X_i, Y_i)_{1 \leq i \leq N}$ and fulfill $\mathbb{E}[U_1] = 0$, $\mathbb{E}\left[U_1^j\right] = 1, j = 2, 3$.
4. Generate bootstrap responses "under $\mathcal{H}_0$": $Y_i^{*,b} = \overline{Y_0} + \epsilon_i^{*,b}, 1 \leq i \leq n$ and $Y_{0,i}^{*,b} = \overline{Y_0} + \epsilon_{0,i}^{*,b}, 1 \leq i \leq m_n$.
5. Compute bootstrap test statistic: $T_n^{*,b}$ computed from the sample $\left(Y_i^{*,b}, X_i\right)_{1 \leq i \leq n}$, $\left(Y_{0,i}^{*,b}, X_{0,i}\right)_{1 \leq i \leq m_n}$ and $(X_{1,i})_{1 \leq i \leq l_n}$.

Compute the empirical threshold value:

6. Take as threshold the empirical $(1-\alpha)$-quantile of the family $\left(\tilde{T}_n^{*,b}\right)_{1 \leq b \leq N_{boot}}$, where α is the nominal level of the test. Denote τ_α its value.

Finally, we reject assumption $\mathcal{H}_0$ if our test statistic T_n takes a value $t_n \geq \tau_\alpha$. This is equivalent to reject $\mathcal{H}_0$ if the value empirical signification degree $\frac{1}{N_{boot}} \sum_{b=1}^{N_{boot}} 1_{T_n^{*,b} > T_n}$ is smaller than the nominal level α.

Remark: We study three wild bootstrap procedures constructed from three distributions P_B^1, P_B^2, P_B^3 initially introduced in Mammen (1993):

- $P_B^1 = \frac{\sqrt{5}+1}{2\sqrt{5}} \delta_{\frac{1-\sqrt{5}}{2}} + \frac{\sqrt{5}-1}{2\sqrt{5}} \delta_{\frac{1+\sqrt{5}}{2}}$ where δ is the Dirac function.
- P_B^2 is the law of the random variable U defined by $U = \frac{V_1}{\sqrt{2}} + \frac{(V_2^2-1)}{2}$, where $V_1, V_2 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$.
- P_B^3 is the law of the variable U defined by $U = \left(\zeta_1 + \frac{V_1}{\sqrt{2}}\right)\left(\zeta_2 + \frac{V_2}{\sqrt{2}}\right) - \zeta_1\zeta_2$, where V_1 and V_2 are independent $\mathcal{N}(0,1)$ random variables, $\zeta_1 = \sqrt{\frac{3}{4} + \frac{\sqrt{17}}{12}}$ and $\zeta_2 = \sqrt{\frac{3}{4} - \frac{\sqrt{17}}{12}}$.

However, it is possible to apply the previous algorithm with any law P_B.

In the remainder of the paper we use the following notations to make reference to the various bootstrap methods we want to compare:

Res.	Resampling procedure
S.N.B.	Smooth Naive Bootstrap procedure
*W.B.*1	Wild Bootstrap procedure with P_B^1
*W.B.*2	Wild Bootstrap procedure with P_B^2
*W.B.*3	Wild Bootstrap procedure with P_B^3

3 Simulation studies: Nonparametrically generated growth curves

The aim of the present section is to compare empirical level and power properties of these bootstrap procedures. It is also important to study how many bootstrap iterations are necessary to get relevant results and to focus on the problem of choosing the smoothing parameter. The usual nominal level $\alpha = 0.05$ is used in our simulations and applications unless another value is precised.

In many simulation studies, the simulated curves are generated parametrically and hence only depend on few parameters. To avoid this drawback, one focuses on datasets in which the curves are simulated in a nonparametric way. To simulate each functional random variable X_i, we propose the following procedure:

1. Simulate 1000 standard Gaussian random variables $(\epsilon_j)_{1 \leq j \leq 1000}$,
2. Compute $U_t = \sum_{j=1}^{t} \epsilon_j$ for $1 \leq t \leq 1000$,
3. Compute $U_t^+ = U_t + |\min(\min_{1 \leq t \leq 1000} U_t, 0)|$, for $1 \leq t \leq 1000$
4. Compute $X_{i,t} = \frac{\sum_{j=1}^{t} U_j^+}{1000}$ for $1 \leq t \leq 1000$.

The values $\left(\frac{U_t}{\sqrt{1000}}\right)_{1 \leq t \leq 1000}$ may be viewed as the discretization $\left(B_i\left(\frac{t}{1000}\right)\right)_{1 \leq t \leq 1000}$ of a Brownian motion defined on $[0;1]$ with $B_i(0) = 0$. Then $W_i(t) = B_i(t) + |\min\left(\inf_{s \in [0;1]} B_i(s), 0\right)|$ corresponds to a vertical translation of $B_i(t)$ which only takes nonnegative values. Finally $X_i(t)$ is defined as the integral of $\sqrt{1000} W_i$ between 0 and t. Each value $X_{i,t}$ corresponds to approximation of the value of X_i at the discretization point $t/1000$. The way simulated curves are defined implies they all start from zero. The functional variables $(X_i)_{1 \leq i \leq n}$ may be viewed as growth curves (see Figure 1).

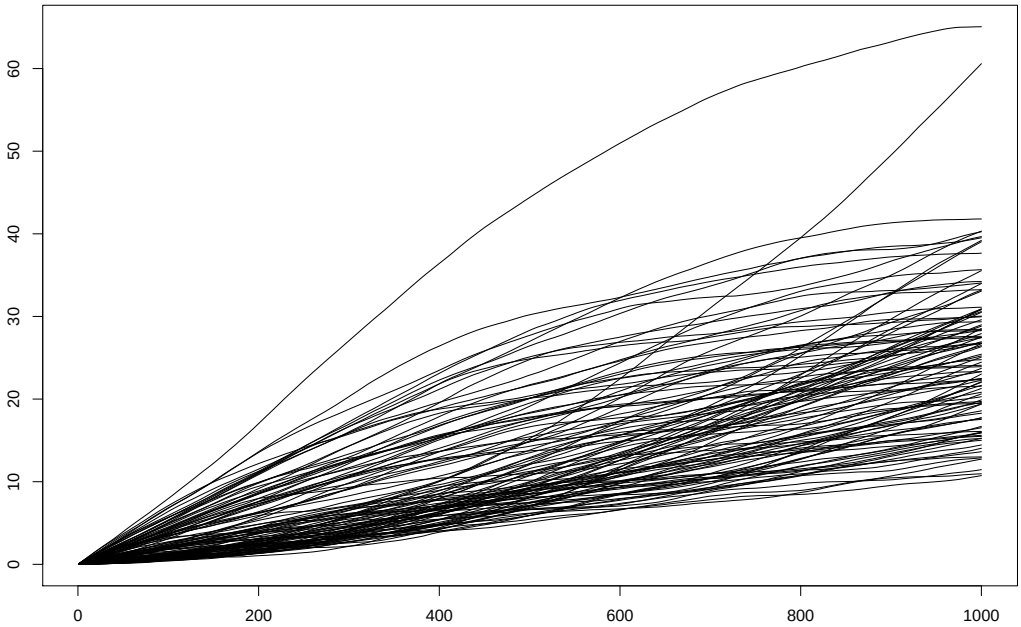

Fig. 1 Sample of 100 nonparametrically simulated growth curves X_i's.

Each simulated dataset contains 300 independent pairs (X_i, ϵ_i), where $\epsilon_i \sim \mathcal{N}(0,1)$. For each dataset, we consider various linear models (respectively non linear models) of the form $Y = k\int_0^1 X(t)\, cos\,(7.5t)\, dt + \epsilon$ (respectively $Y = k\exp(-\int_0^1 X(t)\, cos\,(7.5t)\, dt) + \epsilon$) and construct for each model three independent sub-datasets D, D_0, and D_1 containing 100 pairs (X_i, Y_i). Each sub-dataset is then used in a similar way as in the previous section. In this simulation one uses the $\mathbb{L}^2([0;1])$ metric, a quadratic kernel $K(t) = (1-t^2)1_{[0;1]}(t)$ and the smoothing parameter $h_n = 4$ (see Table 4 to get a study of the effect of the choice of the smoothing parameter). Table 1 gathers the empirical probabilities of rejecting the no-effect assumption for various values of k computed on 10000 samples with $N_{boot} = 100$. R represents the empirical signal-to-noise ratio. The parameter k quantifies the effect of the explanatory variable X on the response variable Y. When $k = 0$, the variable X has no effect on Y, we are under the null hypothesis hence the empirical rejection probability corresponds to the empirical level of the test. On the contrary, when $k > 0$ we are under the alternative hypothesis and the empirical rejection probability represents the empirical power of the test. Furthermore, the greater k is, the more the effect of X on Y is important (for each family of models), and the more the empirical power of the test grows (see Table 1).

The first line of Table 1 hence corresponds to the empirical level of our no effect testing procedures while the other lines contain the empirical power of our testing procedures for various alternatives. The results of the proposed methods are fairly good and have a similar nature. However the resampling procedure and the second and third wild bootstrap procedures seem a little better in terms of level while wild bootstrap methods seem more powerful than smooth naive bootstrap and give similar results than resampling. In addition, as discussed below, wild bootstrap procedures are by nature more robust to the heteroscedasticy of the errors (see the discussion of Table 3). Testing procedures are globally relevant to respect the nominal level but some methods (resampling, second and third wild bootstrap for instance) seem better. Table 2 presents empirical levels obtained on the same 10000 samples (those used to get Table 1) for various values of α.

We have only considered homoscedastic errors until now. In the multivariate case, it is well-known that wild bootstrap methods are adapted to the presence of heteroscedastic errors. It would be interesting to study the behavior of our methods in the heteroscedastic context. Let $a : t \mapsto \cos(7.5t)$, $b : t \mapsto sin\,(7.5t)$ and define heteroscedastic errors : $\epsilon_i^k \sim \mathcal{N}\left(0, \left(1 + k|\int_0^1 X(t)\, b(t)\, dt|\right)\right)$.

Table 1 Comparison of empirical level and power properties

	Res	SNB	WB1	WB2	WB3	R
$Y=\epsilon$	0.051	0.043	0.054	0.050	0.051	1
$Y=\frac{1}{3}\int_0^1 X(t)\,cos\,(7.5t)\,dt+\epsilon$	0.204	0.179	0.211	0.206	0.204	1.151
$Y=\frac{2}{3}\int_0^1 X(t)\,cos\,(7.5t)\,dt+\epsilon$	0.653	0.606	0.673	0.661	0.662	1.605
$Y=\int_0^1 X(t)\,cos\,(7.5t)\,dt+\epsilon$	0.933	0.913	0.951	0.946	0.945	2.360
$Y=\frac{4}{3}\int_0^1 X(t)\,cos\,(7.5t)\,dt+\epsilon$	0.990	0.983	0.996	0.994	0.994	3.417
$Y=\frac{5}{3}\int_0^1 X(t)\,cos\,(7.5t)\,dt+\epsilon$	0.998	0.995	0.999	0.999	0.999	4.776
$Y=5\exp\left(-\int_0^1 X(t)\,cos\,(7.5t)\,dt\right)+\epsilon$	0.182	0.156	0.183	0.178	0.179	1.145
$Y=10\exp\left(-\int_0^1 X(t)\,cos\,(7.5t)\,dt\right)+\epsilon$	0.618	0.570	0.621	0.608	0.605	1.582
$Y=15\exp\left(-\int_0^1 X(t)\,cos\,(7.5t)\,dt\right)+\epsilon$	0.907	0.881	0.905	0.904	0.906	2.310
$Y=20\exp\left(-\int_0^1 X(t)\,cos\,(7.5t)\,dt\right)+\epsilon$	0.980	0.968	0.978	0.976	0.977	3.330
$Y=25\exp\left(-\int_0^1 X(t)\,cos\,(7.5t)\,dt\right)+\epsilon$	0.993	0.988	0.992	0.991	0.991	4.641

Table 2 Comparison of empirical level for various values of α

Model	α	Res	SNB	WB1	WB2	WB3
$Y=\epsilon$	0.01	0.012	0.007	0.012	0.009	0.011
	0.05	0.051	0.043	0.054	0.050	0.051
	0.1	0.102	0.085	0.102	0.099	0.100
	0.2	0.196	0.175	0.197	0.198	0.196

Table 3 Level and power properties in the case of heteroscedastic errors

k	0	0.25	0.5	0.75	1	1.25
$Y=\epsilon^k$ Res	0.061	0.037	0.041	0.037	0.027	0.029
$Y=\epsilon^k$ WB3	0.050	0.054	0.049	0.055	0.050	0.049
$Y=\int_0^1 X(t)\,a(t)\,dt+\epsilon^k$ Res	0.935	0.857	0.740	0.610	0.491	0.399
$Y=\int_0^1 X(t)\,a(t)\,dt+\epsilon^k$ WB3	0.942	0.878	0.806	0.701	0.604	0.530
R	2.363	1.956	1.697	1.526	1.409	1.326

We compare the results obtained by resampling and the third wild bootstrap methods on 1000 samples, with $N_{boot}=100$ and $\alpha=0.05$, when the heteroscedasticity effect (i.e. k) grows. The way we construct our alternative model implies that when k grows, the signal to noise ratio R decreases. Hence the power of the test should decrease when k grows. As expected, wild bootstrap performs better than resampling one because it does not destroy the heteroscedastic structure. The main idea of bootstrap is to construct artificial samples whose distribution mimics the distribution of the original sample in case of a no effect model to get a good approximation of the distribution of T_n under the null hypothesis. Consequently, the resampling method may be unadapted in case of heteroscedasticity because if $\mathbb{E}[(r-\hat{r})^2(X_1)]=o(1)$ (which holds under general assumptions, see for instance Ferraty and Vieu, 2006), $\mathbb{E}[(\epsilon_j^b)^2|X_j]=\frac{1}{n}(\sum_{1\leq i\neq j\leq n}\mathbb{E}[\sigma^2(X_1)]+\sigma^2(X_j))+o_p(1)$ but $\mathbb{E}[\epsilon_j^2|X_j]=\sigma^2(X_j)$. Such differences explain resampling method are not relevant (under the null hypothesis and under alternatives) in the heteroscedastic case while wild bootstrap ones are still adapted (because under the same assumption $\mathbb{E}[(\epsilon_j^b)^2|X_j]=\sigma^2(X_j)+o_p(1)$). We observe in Table 3 a significant difference between the results obtained by the third wild bootstrap method and those obtained by the resampling procedure. The empirical power of the wild bootstrap method is significantly greater than the one obtained by the resampling procedure when the heteroscedasticity is important. Moreover, the empirical level of the wild bootstrap stays near the theoretical level 0.05 while the resampling procedure leads to much worse results. As a conclusion, if one knows that the model under study is homoscedastic, the use of both wild bootstrap or resampling will lead to comparable results. However, if one has no information on the heteroscedasticity or homoscedasticity of the residuals, one would rather use a wild bootstrap procedure.

Consequently, we pay more attention to wild bootstrap methods in the remainder of the paper and choose the third wild bootstrap procedure because it seems to make the balance between good

level and power properties.

We are now interested in exploring how the choice of the smoothing parameter has an influence on the level and the power of our testing procedures. We run our no effect test for various values (fixed or data-driven) of the smoothing parameter on the same simulated samples (constructed as before) in order to understand the impact of the smoothing parameter. Table 4 presents the empirical rejection probabilities obtained with $N_{boot} = 100$ and $\alpha = 0.05$ from these 1000 samples $(X_i, \epsilon_i)_{1 \leq i \leq 300}$. For each sample we consider two models

$$M_0 : \ Y = \epsilon \text{ and } M_1 : \ Y = 10 \exp\left(- \int_0^1 X(t)\, a(t) dt \right) + \epsilon$$

corresponding to null and alternative hypothesis respectively. It appears clearly that a wrong choice of the smoothing parameter (too big or too small) lead to a loss in empirical power and may also have a negative effect on empirical level. The choice of the smoothing parameter has hence to be considered with care (it is common to statistical procedures based on kernel smoothing). However, the empirical level of the test does not seem to be dramatically affected by the use of a large smoothing parameter (over-smoothing $r - \mathbb{E}[Y]$ is not an issue under the null hypothesis).

Cross-validation criterion is commonly used to get an appropriate value h_{CV} of the smoothing parameter for estimation purposes. However, nothing guarantees this data-driven value of the smoothing parameter is still relevant for testing. Because K has a compact support, taking a small smoothing parameter makes the integrand null. As explained at the end of Section 2, the integral over dP_X is estimated using Monte-Carlo approximation on a third subsample (D_1 : $\{(X_{1,k}, Y_{1,k}),\ 1 \leq k \leq l_n\}$). We introduce h_{min} as the infimum of the smoothing parameter values for which the integrand is strictly positive at any curve of D_1 that belongs to W. In other words, h_{min} corresponds to the infimum of the values of the smoothing parameter h for which

$$\min_{1 \leq k \leq l_n,\, X_{1,k} \in W} \sum_{i=1}^{n} K\left(\frac{d(X_i, X_{1,k})}{h} \right) > 0.$$

Because K is decreasing on $[0;1]$ and $K(1) > 0$ (or more generally $K(t) > 0$ for $0 \leq t < 1$) this infimum is defined by

$$h_{min} := \max_{1 \leq k \leq l_n,\, X_{1,k} \in W} \left(\min_{1 \leq i \leq n} \left(d(X_i, X_{1,k}) \right) \right).$$

For each sample, we compute the value of $h_0 = E(h_{min} * 10000 + 1)/10000$, where E stands for the integer part function. This choice of the smoothing parameter ensures the integrand is positive at any curve of D_1 belonging to W.

The results presented in Table 4 illustrate that these data-driven choices of the smoothing parameter lead to relevant empirical level and power properties even if they are not designed for testing. Figure 2.a presents the results of Table 4 in a graphical way to make easier the understanding of the way empirical level (dotted curve) and power (solid curve) depends on the choice of the smoothing parameter (the horizontal line represents the nominal level $\alpha = 0.05$). Then, Figure 2.b presents boxplots of the data-driven values of h_0, h_{CV0} (model M_0), and h_{CV1} (model M_1) computed on simulated samples in order to give an idea of their position with respect to fixed values. Some values of h_{CV0} are greater than 50 and does not appear because we have restricted our plot to a vertical range of $[0; 50]$ in which fixed values of the smoothing parameter have been considered. We can observe h_{min} and h_{CV1} take value concentrated around values for which we have obtained good empirical level and power properties. Under the null hypothesis, the cross-validation criterion may lead to take huge values of the smoothing parameter (the largest one was more than 171), However, this does not seem to lead to non relevant empirical level properties.

Table 4 Effect of the choice of the smoothing parameter on level and power.

h	h_0	h_{CV}	1	3	5	7	9	11	16	21	30	50
M_0	0.058	0.056	0.003	0.064	0.057	0.056	0.060	0.048	0.066	0.054	0.057	0.055
M_1	0.524	0.483	0.398	0.753	0.454	0.279	0.183	0.137	0.111	0.091	0.085	0.079

Table 5 Effect of the choice of the bootstrap iteration number on level and power.

N_{boot}	20	50	100	200	500	1000
Level	0.044	0.056	0.044	0.047	0.047	0.047
Power	0.529	0.644	0.613	0.619	0.628	0.632

This work aims to present a new no-effect test, study its performances on simulations, and illustrate its practical use on spectrometric data. The automatic choice of an optimal smoothing parameter for testing purposes is a relevant challenge for the future, but I think it is out of the scope of this work. The seminal paper by Gao and Gijbels (2009) in the multivariate context is a relevant starting point.

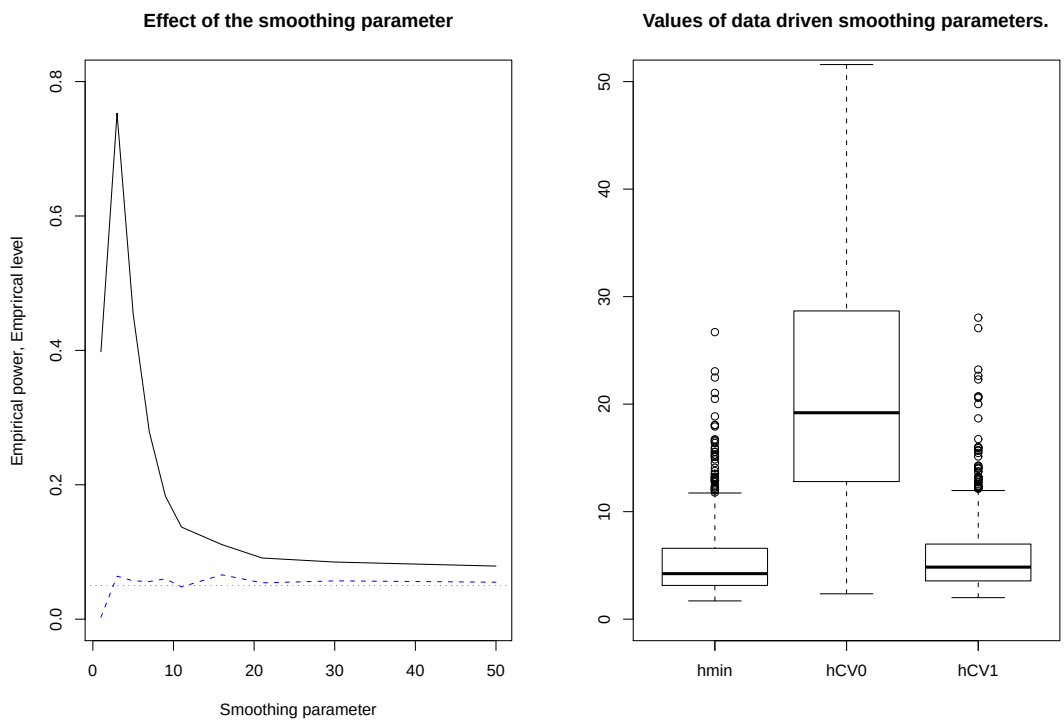

Fig. 2 a) Effect of the smoothing parameter on empirical level and power. b) Repartition of the values taken by data-driven smoothing parameters.

Our testing procedures also depends on the value of the bootstrap iteration number that has to be chosen by the user. Table 5 illustrates how empirical level and power may be sensitive to the number of bootstrap iterations N_{boot}. The rejection probabilities given in this table have been obtained on 10000 samples, on which null and alternative hypotheses have been considered, for various values of N_{boot}. As expected, empirical level and power may change a lot and be irrelevant for small values of N_{boot}. A number of bootstrap iterations (N_{boot}) around 100 or 200 seems enough to have a good approximation of the quantiles. Taking a higher number of bootstrap iterations leads to similar results and takes more time.

Table 6 No effect tests for the original curve X (unobserved), the denoised curve Z and the residual part R. Empirical rejection probabilities obtained on 1000 samples with $N_{boot} = 100$, $\alpha = 0.05$ and $h = h_0$.

	Explanatory variable used for the test		
Model	X	Z	R
M_0	0.056	0.055	0.054
M_1	0.502	0.505	0.056

Table 7 No effect tests for BM, BM_3 and R_3. Empirical rejection probabilities obtained on 1000 samples with $N_{boot} = 100$, $\alpha = 0.05$ and $h = h_0$.

	Explanatory variable used for the test		
Model	BM	BM_3	R_3
$M_1' :\ Y = \epsilon$	0.050	0.043	0.062
$M_2' :\ Y = 5\exp\left(-\int_0^1 BM_3\,(t)\,cos\,(7.5t)\,dt\right) + \epsilon$	0.493	0.509	0.050
$M_3' :\ Y = 24\exp\left(-\int_0^1 R_3\,(t)\,cos\,(7.5t)\,dt\right) + \epsilon$	0.065	0.046	0.523

Let us now conclude this section with two simulations in which only some features of the curves have an effect on the response variable to observe if our testing procedures are relevant to detect them. By simplicity, we consider situations where the informative and non-informative features of the curves are independent to check if the test is relevant to detect their respective nature. However, the features of a curve (derivatives, parts) may be dependent. In this case, the detection of informative features is more complex and require to consider variable selection tests. An heuristic use of no effect tests on residuals is also discussed in Section 4.

Assume first explanatory curves are observed with an additive independent white noise η, that is to say we observe $\tilde{X}_i(t) = X_i(t) + \eta(t)$ instead of $X_i(t)$. A spline approximation (with three knots and splines of order 3) Z of each curve X is used to remove the independent noise (see Fig 3.a). The empirical rejection probabilities presented in Table 7 show our no effect testing procedures is able to detect the effect of the de-noised curve Z and does not detect any significant effect of the residual curve $R = \tilde{X} - Z$.

A similar use of no effect test is finally considered to check if the effect of a Brownian motion BM starting from 0 may be reduced to the effect of its first three principal components scores (explaining more than 90 percent of the variability). For each simulated sample, three models M_1', M_2', and M_3' have been introduced to cover no-effect of BM, effect of BM reduced to BM_3 (projection of BM on its first three components), and effect of BM reduced to $R_3 = BM - BM_3$. As expected, the effect of BM_3 (respectively R_3), which may be regarded as the global shape (see Figure 3.b) of BM (respectively the deviation of BM from its global shape), is well detected as significant for model M_2' (respectively model M_3') and not significant elsewhere. However, even if signal to noise ratios in models M_2' and M_3' are similar, a significant effect of BM itself is only detected for M_2'. The use of the $\mathbb{L}^2$ metric gives a lot of importance to the first components of BM (explaining a great part of variability of BM) and is hence not relevant when the effect only comes from the residual part of the trajectory. The use of the semi metric induced by the three first PC scores is equivalent to consider BM_3 as explanatory variable, while the use of a metric based on remaining scores would lead to consider R_3 (see the comment at the end of Section 2.1). These no effect tests may be seen as tools to check if $r_d(x) = \mathbb{E}[Y|d(X,x) = 0]$ is constant or not. The use of various semi-metrics may be relevant to detect the effect of some features of a curve (see Section 4).

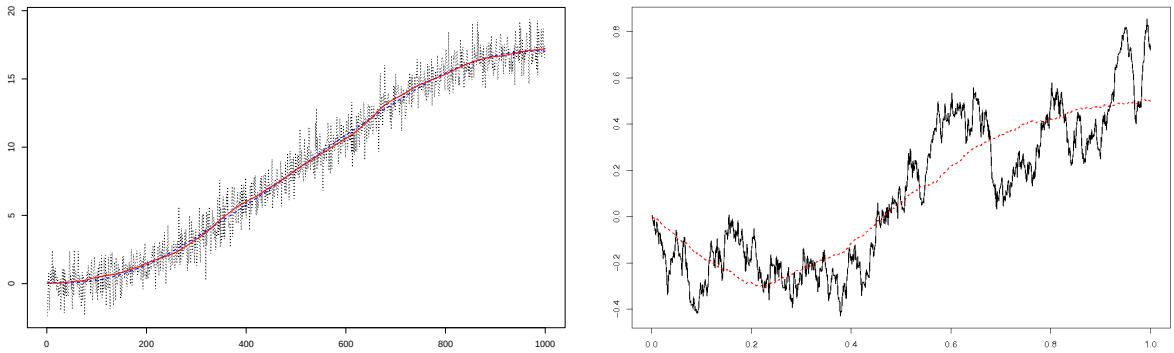

Fig. 3 a) An example of noisy curve $\tilde{X}$ (dotted line), its denoised version Z (dashed line) and the true curve X (solid line). b) An example of simulated Brownian motion trajectory BM (solid line) and its projection BM_3 (dotted line)

4 Application to spectrometric datasets

In many practical situations one is interested in getting the chemical composition of a given substance. Such situations may for instance appear in domains like chemistry, medicine or food industry. Health related and nutritional considerations have a strong impact on food industry since people want to have more and more informations and guarantees on the quality of the products they consume. Chemical analysis allows to get the exact composition of an aliment but costs time and money. To avoid these drawbacks, industrials often prefer to give an estimation of this chemical composition from the observation of spectrometric curves that can be obtained more easily. During a spectrometric study, one emits a light of a given wavelength on the substance under study and one measures how much it absorbs the light emitted. This operation is repeated for various wavelengths, what allows to construct a spectrometric curve that represents the absorbance in function of the wavelength of the light emitted. This kind of data is by nature functional. See for instance the discussion in Leurgans *et al.* (1993): "[...] the spectra observed are to all intents and purposes functional observations". The functional representation of spectrometric data is a common practice in the chemometrics community since the seminal paper by Alsberg (1993). A lot of work has been done on multivariate feature selection in this field with the aim of selecting wavelength ranges (see Leardi, 2003, and the references therein). The no effect tests discussed in this paper offer an interesting way to test if some functional features have a significant effect.

We consider in this paper two real world datasets coming from food industry where the use of spectrometric curves to predict the chemical composition of the aliment can be considered. Previous studies of spectrometric samples have allowed to observe that derivatives of spectrometric curves play an important role in the prediction of the chemical composition of the substance. This fact is also corroborated by empirical procedures classically used in chemistry in the same context (see the discussion in Ferraty and Vieu, 2002). Two questions arise from these considerations. Which derivatives have a significant effect to predict the quantity of a given chemical element contained into the substance? Which parts of the spectrometric curve have a significant effect to make this prediction? These two questions can be investigated by means of no effect testing procedures.

We propose to use our no effect tests taking as explanatory variables the successive derivatives (or some parts) of spectrometric curves. We will use in the remaining of the manuscript the $\mathbb{L}^2$ metric (unless more precision is given). Note that the choice of the semi-metric has a direct influence on the nature of the null and alternative hypothesis of our tests (see the comment at the end of Section 2.1).

4.1 Tecator dataset

The first dataset is related to the issue of predicting the moisture, fat and protein contents of finely chopped meat pieces.

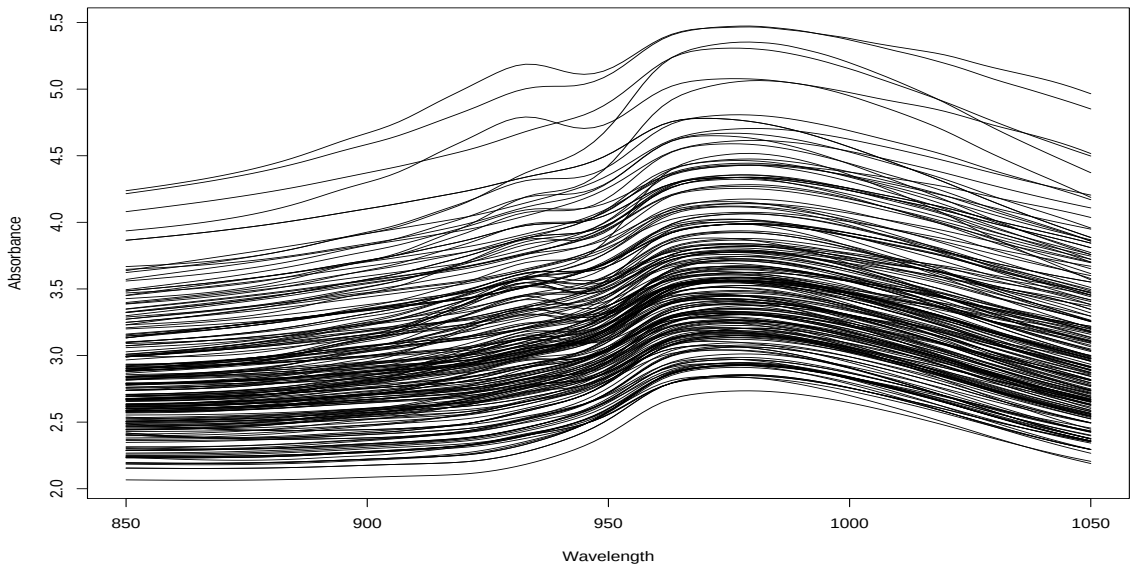

Fig. 4 Sample of spectrometric curves

For each of the 215 meat pieces, we have at our disposal the spectrometric curve X_i (with a wavelength range of 850 - 1050 nm and 100 measures per curve) and the exact values of moisture (Y_{Mi}), fat (Y_{Fi}) and protein (Y_{Pi}) contents (obtained by chemical analysis). This dataset, called Tecator dataset in the literature, is available on Statlib web site (http://lib.stat.cmu.edu/datasets/tecator). It can be regarded as a reference dataset in functional statistics. Indeed, since the original study of Borggaard and Thodberg (1992) based on neural networks, many authors have considered the issue of predicting the fat content from the spectrometric curve (see for instance Ferraty and Vieu, 2002, Ferré and Yao, 2005, Aneiros-Pérez and Vieu, 2006, Ferraty *et al.*, 2006, Ferraty and Vieu, 2006, Ferré and Villa, 2006, Ferraty *et al.*, 2007, or Mas and Pumo, 2007, for some recent references). This dataset has also been considered through other issues like curve classification (see Dabo-Niang *et al.*, 2006) or common structure detection (see Ferraty *et al.*, 2007). Most of these works were oriented towards estimation or classification issues while in the present paper we focus on no effect tests. The aim of structural testing procedures is clearly different and is interesting by itself (testing some a priori model or some theoretical assumptions on the model) but also as a complementary tool to these methods (check the validity of structural assumptions used to construct an estimator, test some questions arising from estimation results, ...).

Before discussing our results let us explain how they have been obtained. For each set of no effect tests, the original sample is randomly splitted into three datasets of respective length $n = 90$, $m_n = 90$, and $l_n = 35$. The empirical signification degree (defined above in Section 2.3) of each test is computed from these three subsamples with the third wild bootstrap method, the smoothing parameter h_0, and $N_{boot} = 100$. The values of empirical signification degrees obtained for 100 repetitions of this procedure are presented through boxplots (see for instance Fig 5). In this plots, an horizontal line stands for the nominal level $\alpha = 0.05$. A significant effect is detected if the values of the empirical signification degree are smaller, what means their boxplot is concentrated below this line. Because the value of the empirical signification degree depends on the way the original sample is splitted, in some cases it is not clear if we should decide that the effect is significant or not (some values are smaller than α but others are greater). The construction of testing procedures

avoiding the issue of splitting the original dataset is an important aim for further improvements of our testing procedure but is out of the scope of this work.

At first glance, spectrometric curves have a similar shape and seem to differ mainly by a vertical shift (see Figure 4) which may be artificial or irrelevant in our context. It is usual in chemometrics to transform the original spectrometric curves to remove this shift and eventually consider derivatives (see the discussion in Ferraty and Vieu, 2002). However, one may wonder if these shifts do not contain any useful information on fat, protein or moisture content. Denote M the pointwise average mean of the spectrometric curves. Each curve X_i of the dataset is decomposed in the following way $X_i = A_i + R_i$, where A_i is the average curve M plus the vertical shift of X_i and R_i stands for the remainder. Please note $A_i - A_j$ (respectively $R_i - R_j$) corresponds to the difference between the averages m_i and m_j of the curves X_i and X_j (respectively to the difference $(X_i - m_i) - (X_j - m_j)$). A first set of no effect tests is made to check if the original curve (explanatory variables X_i's), its vertical shift (explanatory variables A_i's), or the remainder (explanatory variables R_i's) have a significant effect on moisture, fat or protein content.

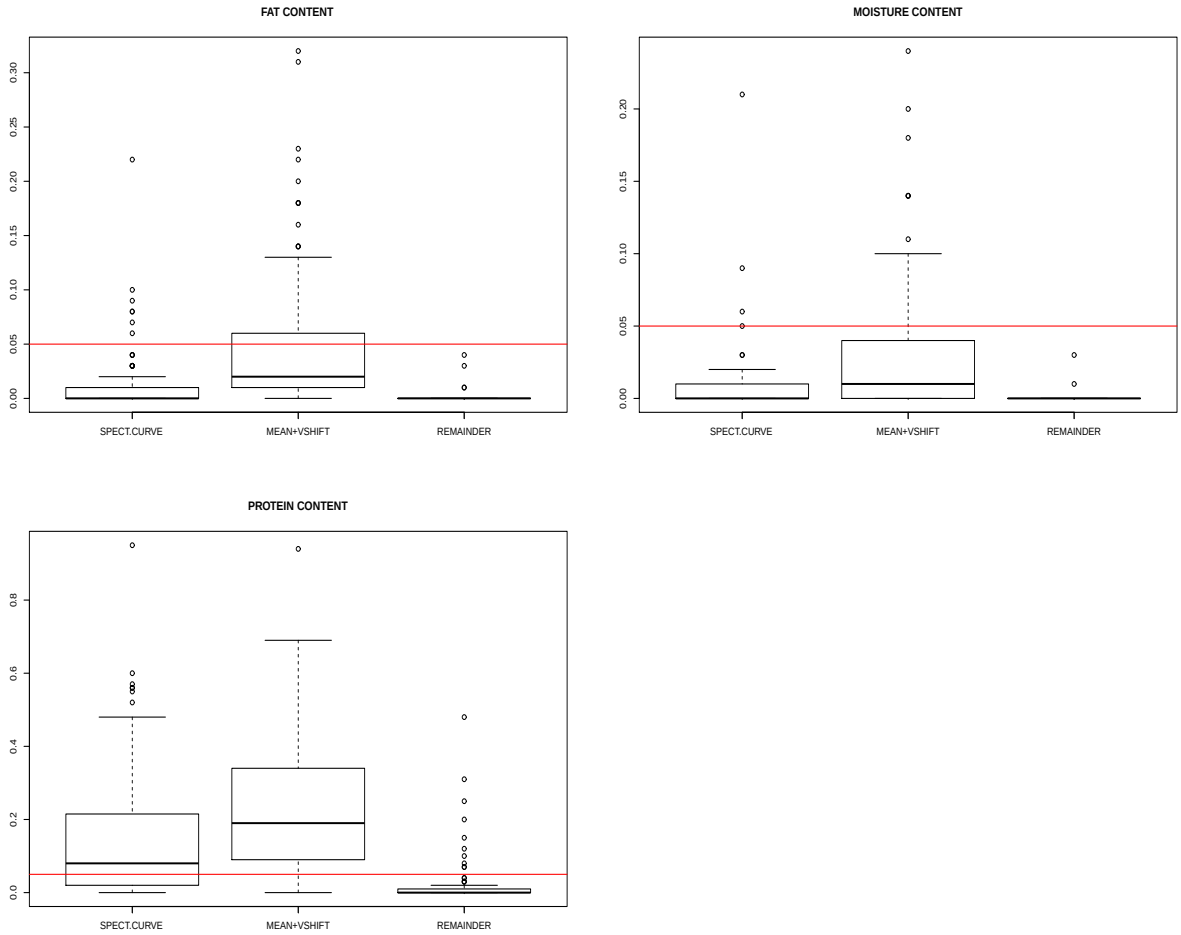

Fig. 5 Empirical signification degrees of no effect tests for the original curve, the vertical shift and the remainder.

The test detects a significant effect of the remainder on protein, fat and moisture contents. The effect of the spectrometric curve is detected as significant only for fat and moisture contents and we globally observe greater values of the empirical signification degree than those obtained with the remainder. This may come from the fact the $\mathbb{L}^2$ of the differences between curves gives a lot of importance to differences between their average values which may have only a small (or inexistent) effect on the chemical content. Indeed, the effect of the vertical shift appears to be non significant

on protein content and is not actually clearly significant for fat content (and moisture content in a smaller proportion). However, a significant effect of this vertical shift on fat and moisture content is often detected. As explained above, this vertical shift is usually removed from the explanatory variable. Is there a loss of information by doing so or is the information provided by the vertical shift also present in the residual part? The average m of the curve X depends significantly on the remainder R (see results of no effect tests using m_i as response variable and R_i as explanatory variable in Figure 6.a). To understand if this vertical shift provides its own information one should in principle use variable selection tests. However, because this paper focuses on no effect tests, an heuristic approach is considered. The information on fat, moisture and protein contents contained in the remainder R is first removed by estimating the corresponding residuals R_F, R_M and R_P (using an additive model with several derivatives of R whose successive order is chosen by cross-validation and k-nearest-neighbor functional kernel estimators [see Ferraty an Vieu, 2006, Burba *et al.*, 2009 for some references], where the number of neighbors k is chosen locally by cross-validation). No effect tests do not detect any significant effect of the vertical shift (using as explanatory variables A_i's) on estimated residuals (see Figure 6.b)) and allow to check the effect of the remainder has been removed (see Figure 6.c)).

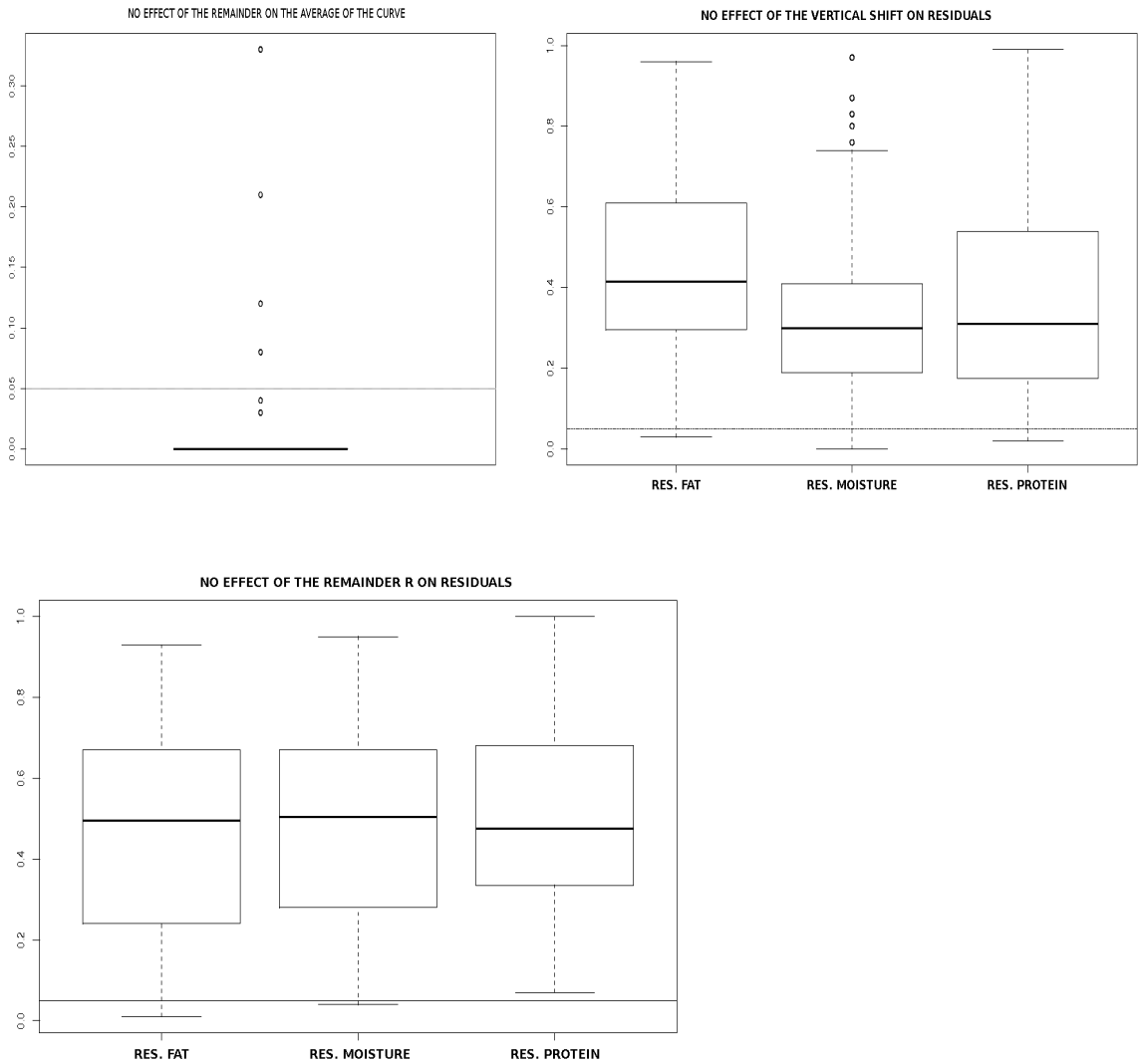

Fig. 6 Results of No effect tests a) of the remainder R on the average of the curve m, b) of the vertical shift on the residuals R_F, R_M, and R_P, c) of the remainder on the residuals R_F, R_M, and R_P.

As discussed above the derivatives of the spectrometric curves (see Figures 4 and 7) are usually considered to put in relief relevant features of the curves (see Ferraty and Vieu, 2002, 2006). No effect tests constructed from the spectrometric curve and semi-metrics based on $\mathbb{L}^2$ norm of derivatives detect a significant effect of the derivatives of order 1,2,3 and 4 on fat, moisture and protein content (see Figure 8).

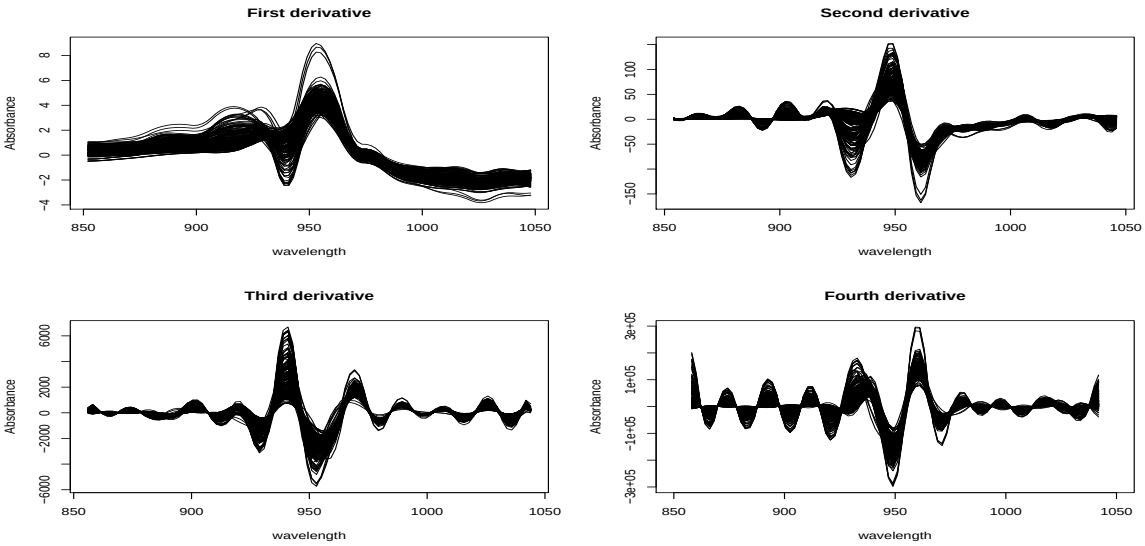

Fig. 7 Sample of derivatives of spectrometric curves

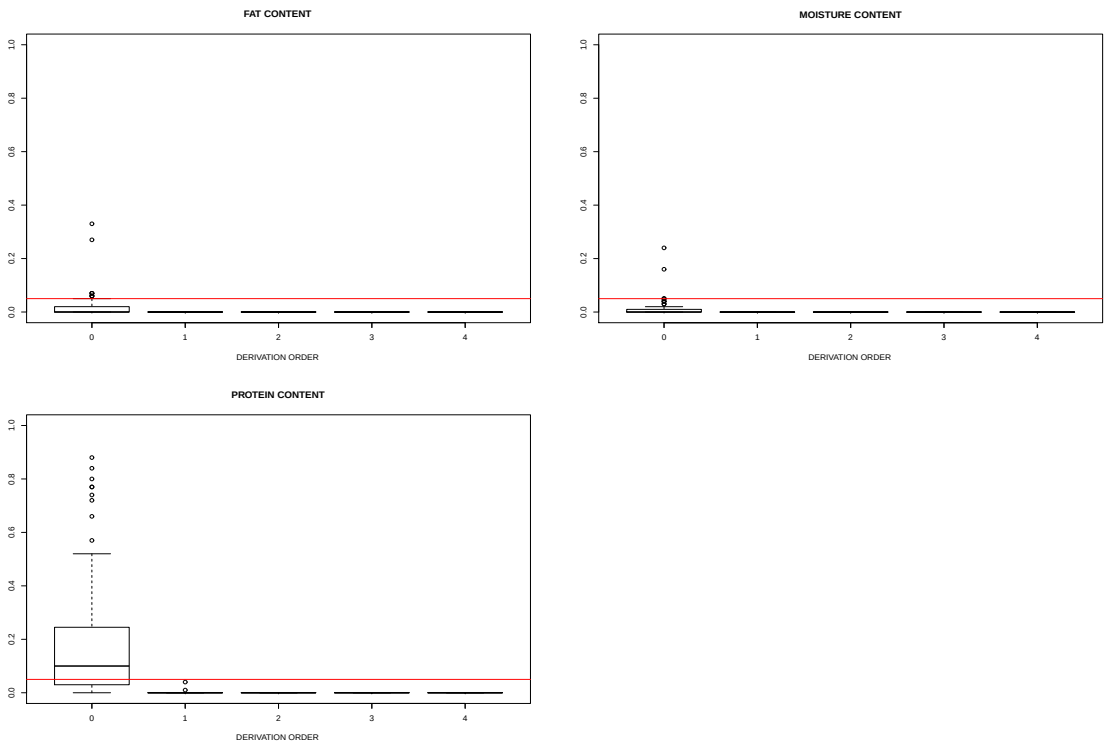

Fig. 8 No effect tests of the successive derivatives of order 0,1,2,3, and 4 on moisture, fat, or protein content.

Table 8 Leave one out mean squared errors

Derivation order	0	1	2	3	4
Moisture	28.361	5.311	**3.009**	6.191	3.947
Fat	48.938	6.685	**2.731**	6.747	3.967
Protein	4.523	**1.450**	2.100	1.579	1.628

Table 9 Boosting and leave one out mean squared errors

Number of derivatives used	1	2	3	4	Successive orders
Moisture	3.009	1.827	1.483	1.366	2,1,4,3
Fat	2.731	1.834	1.520	1.423	2,4,1,3
Protein	1.450	1.156	1.081	1.074	1,3,4,2

Moreover, the empirical signification degrees obtained with the original spectrometric curves seem to be globally greater than those observed for their derivatives (the $\mathbb{L}^2$ norm gives a lot of importance to vertical shifts which contain few relevant information, see the discussion above). The successive derivatives of spectrometric curves are linked by nature. Hence the result of our test does not ensure every derivative provide its own relevant information and one can wonder if the effect of all the derivatives is summarized by the effect of the best predictor. Former studies of this sample (see for instance Ferraty and Vieu, 2002, 2006) have shown the second derivative of the spectrometric curve is the best predictor for fat content. The detection of a significant effect of other derivatives does not allow to understand if this effect is contained in the effect of the second derivative. Such questions are relevant to detect informative features of the curve and should be considered through variable selection tests. However, the recent work Ferraty and Vieu (2009) shows the regression of estimated residuals on other derivatives of the spectrometric curve may be relevant to improve the quality of the estimation (see also Table 9 for an application of these boosting ideas to improve leave one out mean squared error). Hence it seems the estimated residuals still depend on the other derivatives of the spectrometric curve. In this paper we focus on no effect tests and propose in a heuristic way to use no effect tests on estimated residuals to see if these additional effects are significant.
The efficiency of the successive derivatives to estimate the chemical content is investigated through leave one out mean squared errors $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{r}^{-i}(X_i))^2$ in which $\hat{r}^{-i}$ stand for the k-nearest neighbor functional kernel estimator constructed from the sample $(X_j, Y_j)_{1\leq j\neq i\leq 215}$ (the number of neighbors k is chosen by local cross-validation). The second derivative appears to be the best predictor for fat and moisture contents while the first derivative is more relevant to predict protein content (see Table 8). The quality of the estimation is illustrated in Figure 9.

Estimated residuals

$$R_{Mi} := Y_{Mi} - \hat{r_M}\left(X_i^{(2)}\right),\ R_{Fi} := Y_{Fi} - \hat{r_F}\left(X_i^{(2)}\right) \text{ and } R_{Pi} := Y_{Pi} - \hat{r_P}\left(X_i^{(1)}\right)$$

are obtained using k-nearest-neighbor kernel estimators (the number of neighbors k is chosen by local cross-validation each time such estimators are used in the remaining pages).

No effect tests detect a significant effect (for a level $\alpha = 0.05$) of the third derivative on fat content residuals for 53 random decompositions of the sample. The test detects a significant effect of the first and fourth derivatives on fat content residuals only for a smaller number of decompositions (30 and 39). No significant effect is detected for the other cases (or the test is significant for few decompositions of the sample). This does not mean the considered derivative has no effect on the estimated residuals, but only that this effect is not strong enough to be detected as significant by our testing procedure (for a nominal level $\alpha = 0.05$). A bad estimation of the residuals may decrease signal to noise ratio and make more difficult the detection of additional effects. Moreover, the construction of more powerful testing procedures might lead to the detection of other significant effects.

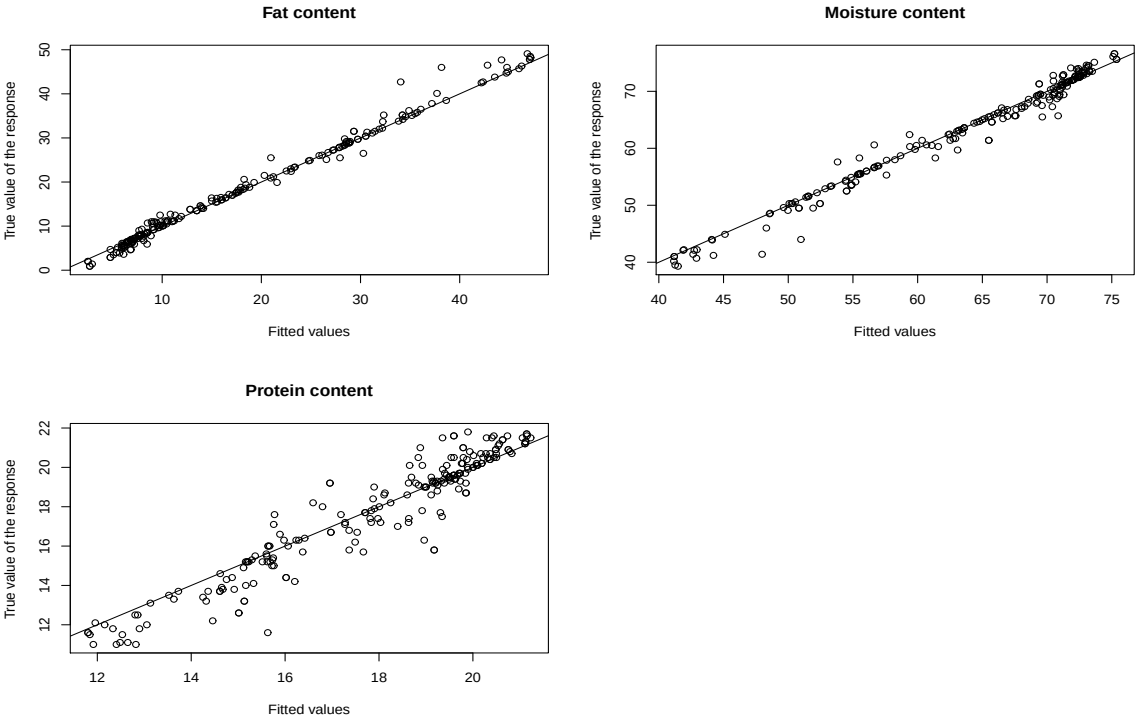

Fig. 9 Quality of the estimation of fat, moisture and protein contents using respectively twice the second derivative and the first one.

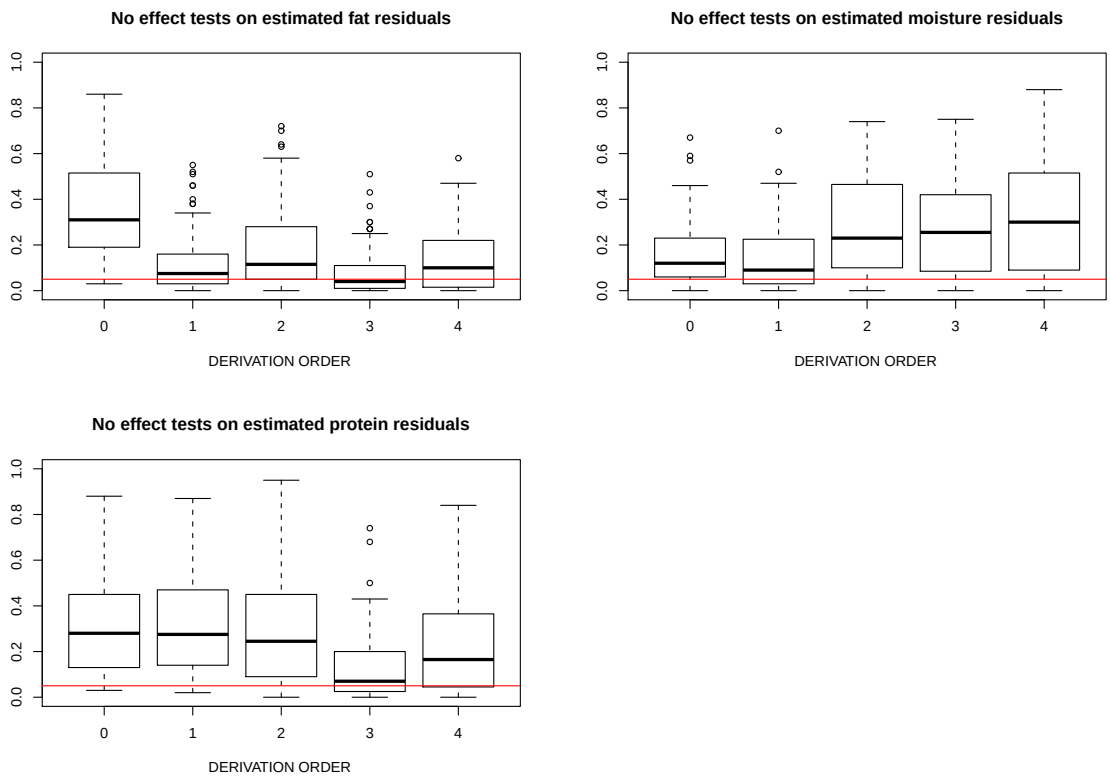

Fig. 10 No effect tests on estimated residuals (with respect to the best predictor) of fat, moisture an protein content.

In this study of tecator data, no effect tests have been used to detect informative features (vertical shift, deviation to the mean curve plus vertical shift, derivatives) of the spectrometric curve. Because these features may be dependent an heuristic method using no effect tests on residuals (w.r.t. the best predictive feature) is used to look for additional effect of the other features. No evidence is given that removing the vertical shift leads to a loss of information. And the dependence of moisture content residuals (obtained from the second derivative) on other derivatives observed in former studies is significantly detected.

4.2 Corn dataset

We now consider a smaller dataset coming from the spectrometric study of 80 corn samples. This dataset has a fairly small size. However, some simulation studies on small datasets (not presented in this manuscript) show our testing procedures still have fairly good level and power properties in such cases.

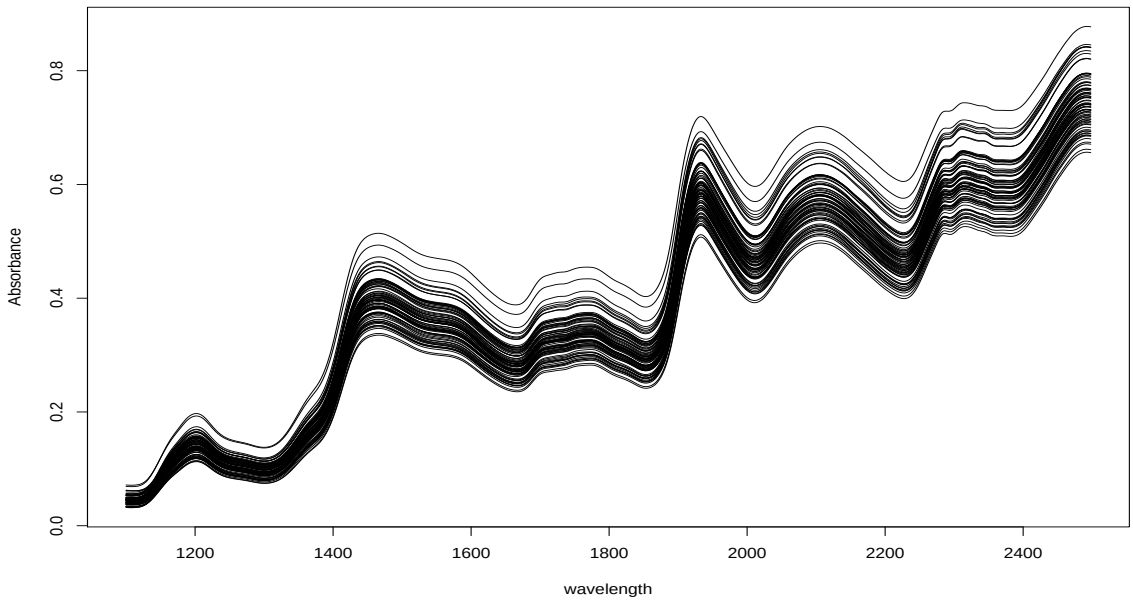

Fig. 11 Sample of spectrometric curves

For each corn sample we have at our disposal a spectrometric curve (with a wavelength range of 1100 - 2498 nm with 700 measures per curve, see Figure 11) and the exact values of moisture, oil, protein and starch contents (obtained by chemical analysis). This dataset is available at the following address : $http://software.eigenvector.com/Data/Corn/index.html$ (where a thorough description of the data is given). The issue related to this dataset still is to study how the chemical composition of the corn samples is explained by the corresponding spectrometric curve. However, the small size of this dataset constitutes an additional interesting challenge. Of course, our results could be improved if we had at our disposal a more important dataset.

A first analysis is conducted as for Tecator dataset to detect significant effect of the derivatives. The original sample is splitted into three sub-datasets of respective length $n = 40$, $m_n = 20$ and $l_n = 20$.

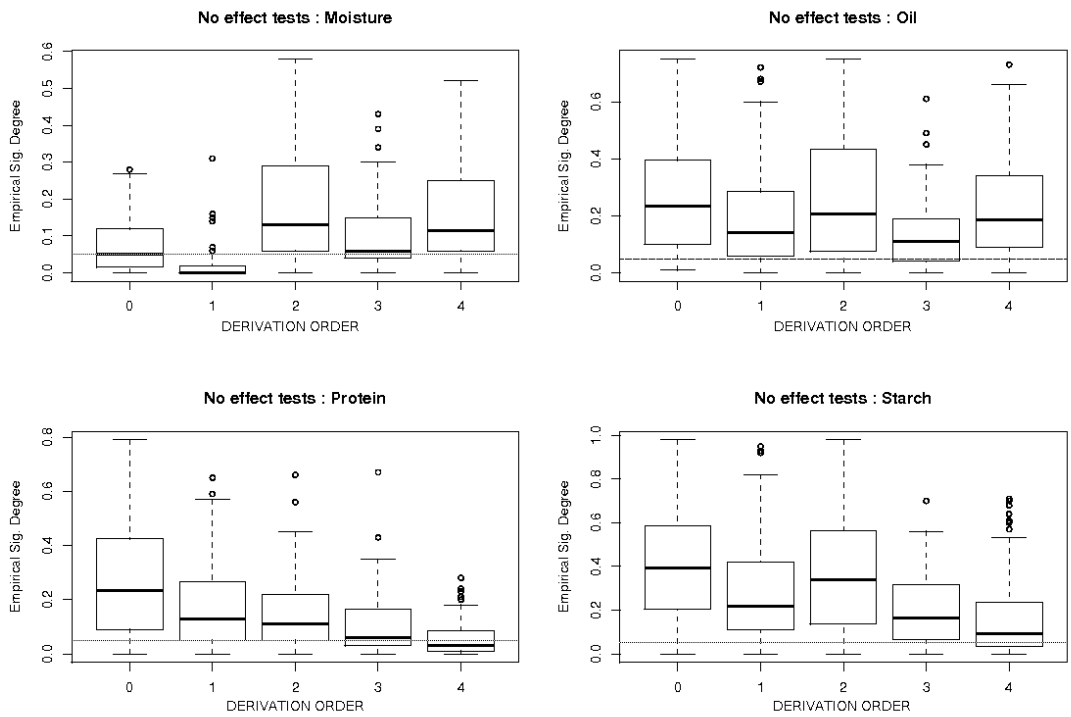

Fig. 12 No effect tests in moisture, oil, protein and starch content prediction

The first derivative has a significant effect on moisture content and the fourth derivative has significant effect on protein content for 61 random decompositions. The significant effect of the original curve (and the third derivative in smaller proportion) on moisture content and of the third derivative on protein content is less clear. No significant effect is detected for oil and starch content.

To end this study, let us focus now on moisture and protein content prediction. The whole spectrometric curve is splitted into seven consecutive curves $X_a, X_b, X_c, X_d, X_e, X_f,$ and X_g (see Fig. 13) and the question is to detect parts which have a significant effect on moisture or protein content. This decomposition has been chosen in an arbitrary way and other decompositions could be considered. Semi-metrics based on derivatives are considered to help detect significant effect of different features.

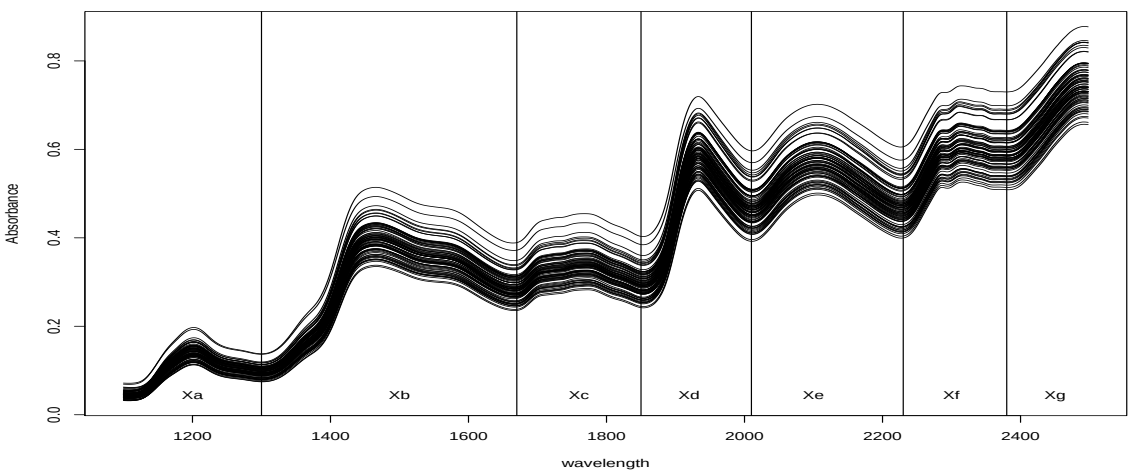

Fig. 13 Spectrometric curves decomposition

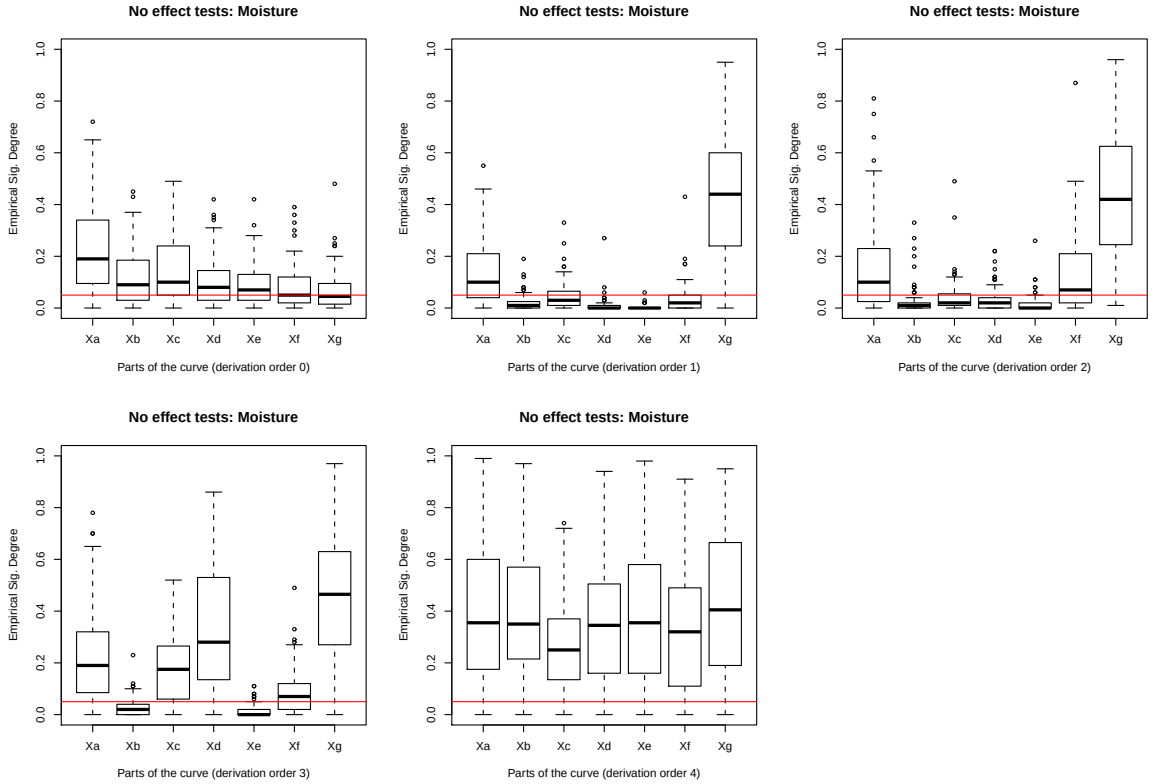

Fig. 14 No effect tests and spectrometric curves decomposition with various semi-metrics (based on derivatives of order 0, 1, 2,3, and 4)

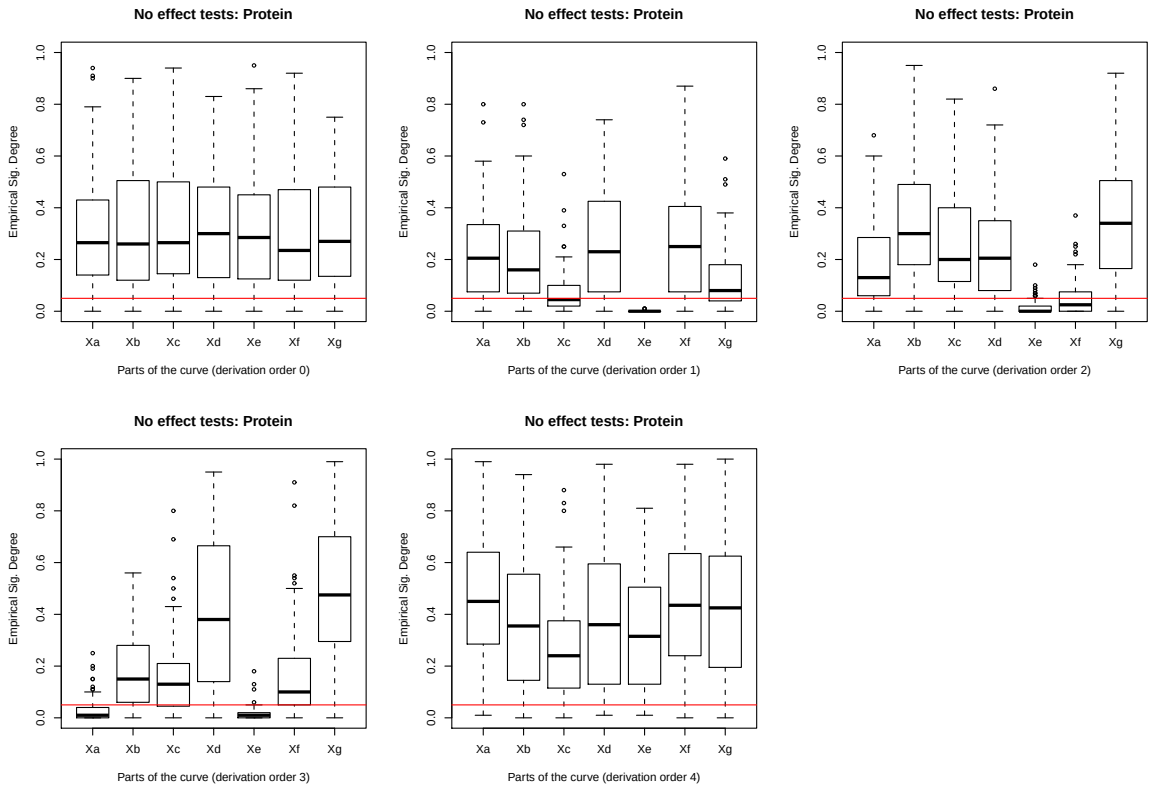

Fig. 15 No effect tests and spectrometric curves decomposition with various semi-metrics (based on derivatives of order 0, 1, 2, 3, and 4).

On the one hand, the tests put in relief a significant effect of X_b, X_d and X_e on moisture content for various semi-metrics (see Figure 14). The effect of X_b and X_f may be seen as significant for some semi-metrics (derivative of order one and two respectively). However, no clear evidence of the effect of extreme parts X_a and X_g is given. On the other hand, X_e has a significant effect on protein content for several semi-metrics (see Figure 15). The significant effect of parts X_a, X_f (and in a smaller proportion X_c) may be discussed from the results obtained with third, second (and first) derivative. No clear evidence of the effect of X_b, X_d and X_g is observed.

These testing procedures have been used to detect portions of the spectrometric curve which are "informative" on moisture or protein content. Some differences in terms of selected parts and derivation orders are observed between these two situations. This illustrates the information on moisture and protein contents may come from different features of the original curve. However, the successive parts (and their derivative) of the original spectrometric curve clearly depend one from the other. Hence, the information provided by a portion of the curve may be included into the effect of an other. Consequently, it would be interesting to use variable selection tests. We may propose, as in the study of Tecator data, to estimate the residuals obtained from the best predictive part and apply no effect tests on estimated residuals in an heuristic way. Residuals are obtained using a linear spline estimator (see for instance Crambes *et al.*, 2009) computed from X_b for moisture content and X_e for protein content (the best predictive parts in terms of leave one out mean squared error). See estimation quality in Figure 16. The same no effect testing procedures are used on these estimated residuals. No significant effect is detected (see Figure 17 and 18). What only leads to the conclusion the test does not allow to detect as significant the effect of other parts of the curve on estimated residuals. A better estimation of residuals, a larger sample or a more powerful testing procedure might lead to detect significant effects on residuals. Moreover, variable selection tests would be more adapted to actually test if the information contained in the whole spectrometric curve concerning moisture content (respectively protein content) is given by X_b (respectively X_e).

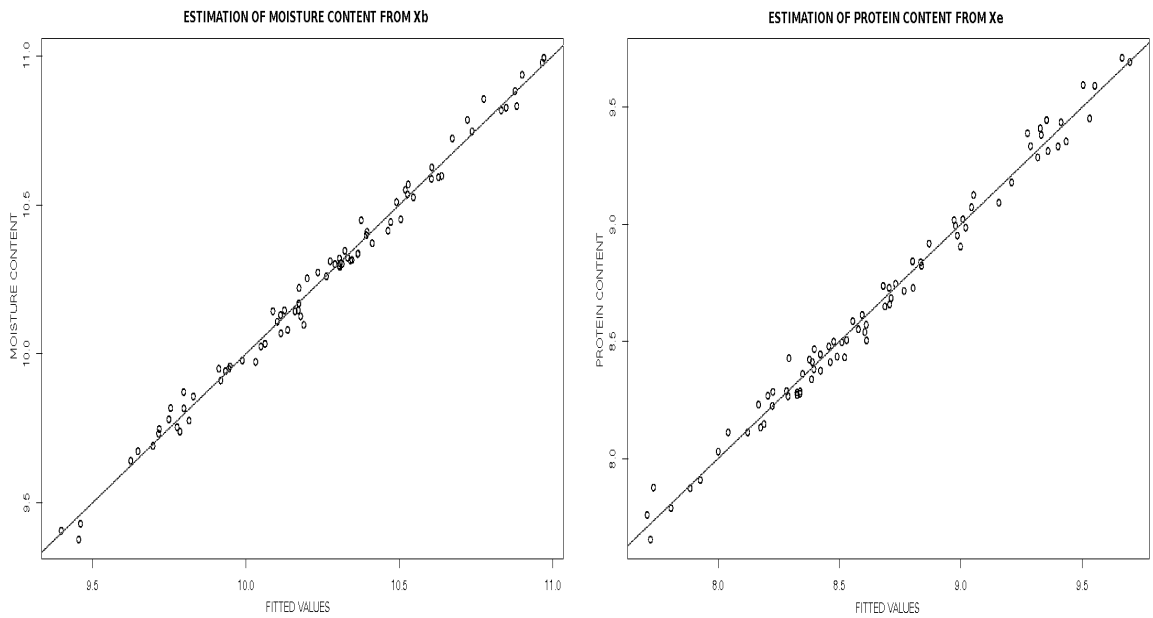

Fig. 16 Linear estimation of moisture and protein contents from X_b and X_e

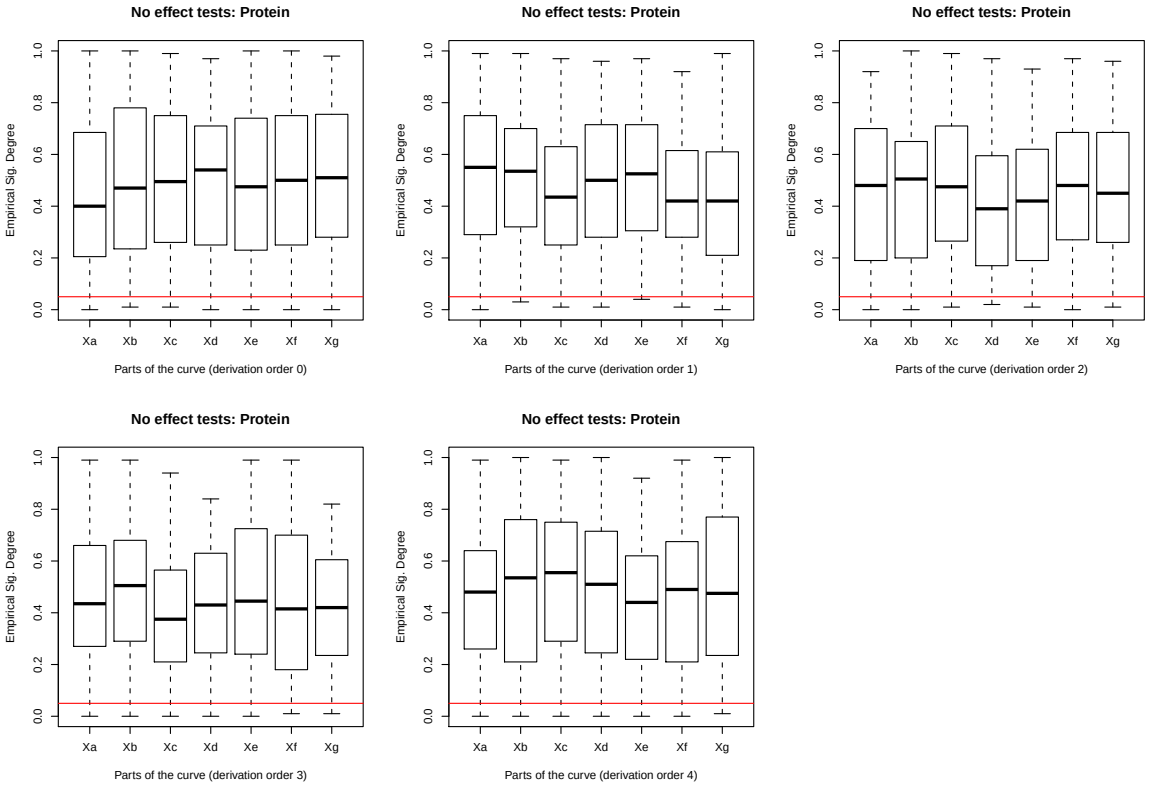

Fig. 17 No effect tests on estimated residuals of protein content with various semi-metrics (based on derivatives of order 0, 1, 2, 3, and 4).

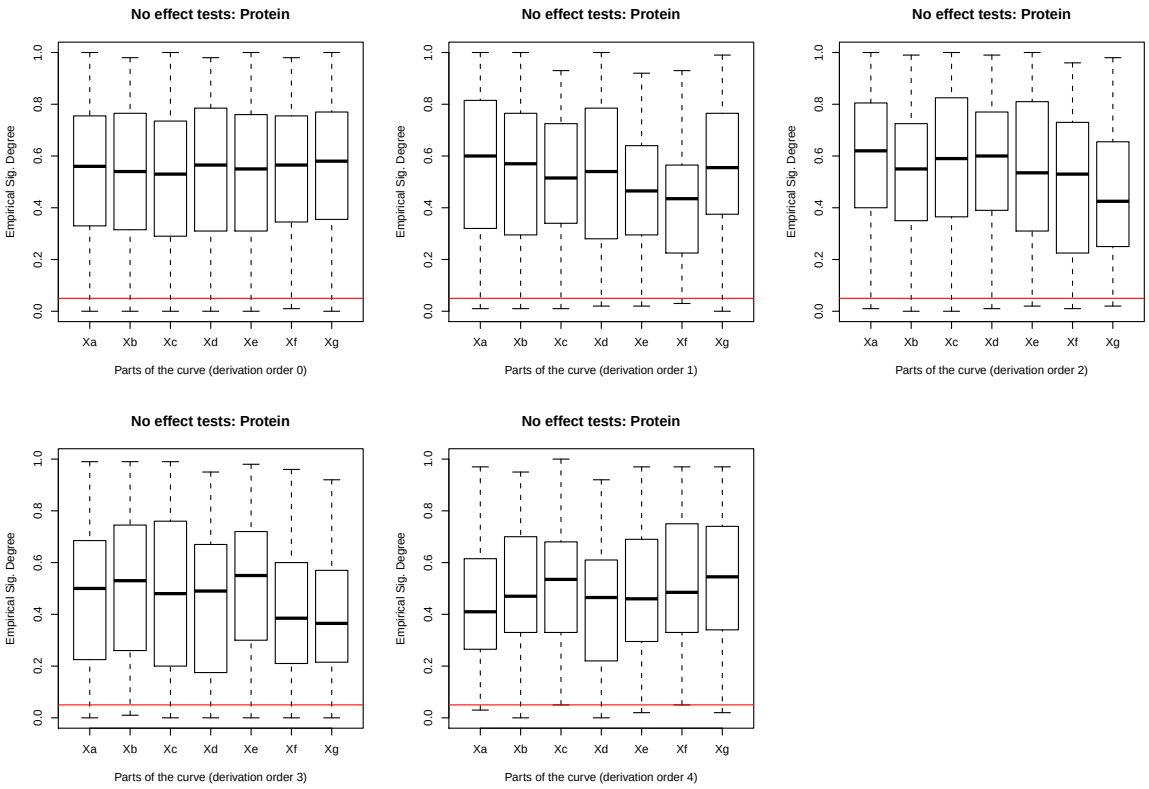

Fig. 18 No effect tests on estimated residuals of moisture content with various semi-metrics (based on derivatives of order 0, 1, 2, 3, and 4).

In a more general context, one can imagine to develop an automatic testing procedure to detect which part of a spectrometric curve has an effect to predict the proportion of a specific substance. This would allow to adjust the necessary wavelength range we have to study to predict this proportion. Such an automatic procedure to detect functional features provides an interesting alternative to the classical multivariate feature selection methods considered in chemometrics (see Leardi, 2003, for more references).

5 Conclusion

A new way to construct no effect tests in regression on functional variable is introduced and various bootstrap no effect testing procedures to compute the threshold are proposed. An extensive simulation study has been made to compare these methods and to consider the issue of the choice of the smoothing parameter and the number of bootstrap iterations. As expected, wild bootstrap methods appear to be more adapted to study models with heteroscedastic errors. This paper also focuses on practical considerations arising from two real world spectrometric datasets. The issue related to such datasets is often to predict the quantity of a given chemical element contained in the substance we study. Many authors have noted that the derivatives of spectrometric curves are good predictors. Because of this observation one may wonder which derivatives and which parts of spectrometric curves have a significant effect on the variable of interest. The results obtained on the tecator dataset are relevant and corroborated by former studies of this "reference" dataset. Moreover, the last application on Corn dataset illustrates that our testing procedure leads to interesting and efficient results even in the case of small datasets. The main theoretical result presented in this work is relevant to ensure the good behavior of the test statistic T_n. However, in practice, the direct use of the asymptotic distribution to fix the threshold value may lead to irrelevant results. That is why we propose to use bootstrap procedures to compute this threshold. The theoretical proof of the validity of these bootstrap procedures is actually in progress. The asymptotic normality result stated in this manuscript is used as a preliminary step and the remaining of the proof follows similar ideas as those used to state our result. Of course, this point is an important challenge for the future and should be considered in an other work.

6 Proofs:

Here are the proofs of the main results of this manuscript. The letter C stand for any positive constant (which may change from one line to the other).
Proof of Theorem 1: Theorem 1 is a direct application of Theorem 3.2 in Delsol *et al.* (2011). Here are the main ideas of its proof.

Lemma 1 *Under assumptions of Theorem 1, $Var(T_{1,n}) = O(n\Phi^2(h_n))$ and $T_{2,n}$ is asymptotically Gaussian with asymptotic variance V_n such that $C_1 n^2 \Phi^{3+l}(h_n) \leq V_n \leq C_2 n^2 \Phi^3(h_n)$ for two positive constants C_1 and C_2.*

Lemma 2 *Under $\mathcal{H}_0$ and assumptions of Theorem 1, one gets*

$$T_n - T_{1,n} - T_{2,n} = O_p(n^2\Phi^2\mathbb{E}[(r(X) - \overline{Y_0})^2]) + O_p(n^{\frac{3}{2}}\Phi^2(h_n)\mathbb{E}[(r(X) - \overline{Y_0})^2]^{\frac{1}{2}})$$

Lemma 3 *Under $\mathcal{H}_1$ and assumptions of Theorem 1, one gets*

$$T_n - T_{1,n} - T_{2,n} = \mathbb{E}[T_n - T_{1,n} - T_{2,n}|D_0] + R_n,$$

with $\mathbb{E}[T_n - T_{1,n} - T_{2,n}|D_0] \geq Cn^2\phi^2(h_n)\|r_n - \overline{Y_0}\|^2_{\mathbb{L}^2(wdP_X)}(1 + o_p(1))$ for some positive constant C and $R_n = o_p(\mathbb{E}[T_n - T_{1,n} - T_{2,n}|D_0])$.

The proof of these lemma (given in a more general context in Delsol *et al.*, 2010) is based on bounds for the mean and variance of these terms (or their decomposition). We omit the details of these tedious but straightforward computations to give more explanations on the main changes.

Now, under the null hypothesis $r \equiv \mathbb{E}[Y]$ and hence $\mathbb{E}[(r(X) - \overline{Y_0})^2] = O(\frac{1}{m_n})$. Consequently, Lemmas 1 and 2 together with assumptions 8 and 9 lead to

$$\begin{aligned}\frac{T_n - \mathbb{E}[T_{1,n}]}{\sqrt{Var(T_{2,n})}} &= \frac{T_{2,n}}{\sqrt{Var(T_{2,n})}} + O_p(\frac{1}{\sqrt{n\Phi^{1+l}(h_n)}}) + O_p(\frac{n\Phi^{\frac{1-l}{2}}(h_n)}{m_n}) + O_p(\frac{\sqrt{n}\Phi^{\frac{1-l}{2}}(h_n)}{\sqrt{m_n}}).\\ &= \frac{T_{2,n}}{\sqrt{Var(T_{2,n})}} + o_p(1)\end{aligned}$$

Lemma 2 and Slutsky's theorem are enough to conclude the asymptotic normality of T_n.

Under the alternative, $\|r_n - \overline{Y_0}\|^2_{\mathbb{L}^2(wdP_X)} \geq \eta_n^2$ because $\overline{Y_0}$ is a constant operator. Then, Lemma 3 lead us to the conclusion

$$(1 + \frac{R_n}{\mathbb{E}[T_n - T_{1,n} - T_{2,n}|D_0]}) = 1 + o_p(1).$$

Hence, if $A_n := \{(1 + \frac{R_n}{\mathbb{E}[T_n - T_{1,n} - T_{2,n}|D_0]}) \geq \frac{1}{2}\}$, for all $\epsilon > 0$, there exists N_ϵ such that for all $n \geq N_\epsilon$, $P(A_n) \geq 1 - \epsilon$. On the other hand, on A_n one gets:

$$\begin{aligned}\frac{T_n - \mathbb{E}[T_{1,n}]}{\sqrt{Var(T_{2,n})}} &\geq \frac{T_{2,n}}{\sqrt{Var(T_{2,n})}} + O_p(\frac{1}{\sqrt{n\Phi(h_n)}}) + \frac{C}{2}n\phi^{\frac{1}{2}}(h_n)\eta_n^2(1 + o_p(1)\\ &\geq O_p(1) + \frac{C}{2}n\phi^{\frac{1}{2}}(h_n)\eta_n^2(1 + o_p(1))\\ &\geq \frac{C}{2}n\phi^{\frac{1}{2}}(h_n)\eta_n^2(1 + o_p(1))\end{aligned}$$

This is enough to state from assumption (7) the conclusion

$$\forall A \in \mathbb{R}, \forall \epsilon > 0,\ \exists N_{A,\epsilon},\ \forall n \geq N_{A,\epsilon},\ P(\frac{T_n - \mathbb{E}[T_{1,n}]}{\sqrt{Var(T_{2,n})}} \leq A) \leq \epsilon.$$

Proof of Corollary 1: The asymptotic normality under $\mathcal{H}_0$ is given by Theorem 1.

Assume now $\mathcal{H}_1'$ holds. We first decompose $\overline{Y_0} = (C + \frac{1}{m_n}\sum_{i=1}^{m_n}\epsilon_{0,i}) + \eta_n\frac{1}{m_n}\sum_{i=1}^{m_n}\Delta_n(X_{0,i}) =: \overline{Y_{00}} + \eta_n\overline{\Delta}$ and $T_n = T_{0n} + R_n$, with T_{0n} the test statistic constructed from the variables $((X_i, C + \epsilon_i))_{1\leq i\leq N_n}$. Because T_{0n} is constructed from a no effect model, Theorem 1 can be used to state its asymptotic normality. Only remains the study of the remaining term R_n which can be decomposed in the following way (where $\overline{\epsilon_0}$ stands for the empirical mean of the residuals in D_0)

$$\begin{aligned}R_n &= 2\eta_n \int \left(\sum_{i=1}^n (\Delta_n(X_i) - \overline{\Delta})K\left(\frac{d(X_i,x)}{h_n}\right)\right)\left(\sum_{j=1}^n (\epsilon_j - \overline{\epsilon_0})K\left(\frac{d(X_j,x)}{h_n}\right)\right) w(x)dP_X(x)\\ &\quad + \eta_n^2 \int \left(\sum_{i=1}^n (\Delta_n(X_i) - \overline{\Delta})K\left(\frac{d(X_i,x)}{h_n}\right)\right)^2 w(x)dP_X(x)\\ &= 2\eta_n R_{1,n} + \eta_n^2 R_{2,n}\end{aligned} \tag{12}$$

Now, denotes $S_{1n}(x) = \sum_{i=1}^n(\Delta_n(x) - \overline{\Delta})K\left(\frac{d(X_i,x)}{h_n}\right)$ and $S_{2n}(x) = \sum_{i=1}^n(\Delta_n(X_i) - \Delta_n(x))K\left(\frac{d(X_i,x)}{h_n}\right)$. One gets easily

$$\begin{aligned}&\mathbb{E}[R_{2,n}|D_0]\\ &= \int \mathbb{E}[S_{1n}^2(x)|D_0]w(x)dP_X(x) + \int \mathbb{E}[S_{2n}^2(x)|D_0]w(x)dP_X(x)\\ &\quad + 2\int \mathbb{E}[S_{1n}(x)S_{2n}(x)|D_0]w(x)dP_X(x).\end{aligned} \tag{13}$$

For n large enough $d(x,X) \leq h_n$ and $x \in W$ implies $X \in W_{\gamma_0}$, hence assumption (10) on Δ_n gives $|S_{2n}| \leq C_0 h_n^\beta \sum_{i=1}^n K\left(\frac{d(X_i,x)}{h_n}\right)$ and

$$\int \mathbb{E}[S_{2n}^2(x)|D_0]w(x)dP_X(x) = O_p(n^2 h_n^{2\beta}\Phi^2(h_n)).$$

On the other hand, one gets

$$\begin{aligned}
&\int \mathbb{E}[S_{1n}^2(x)|D_0]w(x)dP_X(x) \\
&= \int (\Delta_n(x) - \overline{\Delta})^2 \mathbb{E}\left[\sum_{i=1}^n K^2\left(\frac{d(X_i,x)}{h_n}\right) + \sum_{1\leq i\neq j\leq n} K\left(\frac{d(X_i,x)}{h_n}\right) K\left(\frac{d(X_j,x)}{h_n}\right)|D_0\right] w(x)dP_X(x) \\
&= \int (\Delta_n(x) - \overline{\Delta})^2 n(n-1)(\mathbb{E}\left[K\left(\frac{d(X_i,x)}{h_n}\right)\right])^2 w(x)dP_X(x)(1+o_p(1)) \qquad (14)
\end{aligned}$$

Assumption (11), (14) and Cauchy Schwartz inequality lead to the conclusion

$$\int \mathbb{E}[S_{2n}^2(x)|D_0]w(x)dP_X(x) = o_p(\int \mathbb{E}[S_{1n}^2(x)|D_0]w(x)dP_X(x)) \qquad (15)$$

$$\int \mathbb{E}[S_{2n}(x)S_{1n}(x)|D_0]w(x)dP_X(x) = o_p(\int \mathbb{E}[S_{1n}^2(x)|D_0]w(x)dP_X(x)). \qquad (16)$$

Because $R_{1,n}$ is centered conditionally to D_0, it comes directly from (13)-(16) that

$$\mathbb{E}[R_n|D_0] = \eta_n^2 \int (\Delta_n(x) - \overline{\Delta})^2 n(n-1)(\mathbb{E}\left[K\left(\frac{d(X_i,x)}{h_n}\right)\right])^2 w(x)dP_X(x)(1+o_p(1)) \qquad (17)$$

Since $\overline{\Delta} \xrightarrow{P} 0$, one gets

$$\mathbb{E}[R_n|D_0] = \eta_n^2 \int (\Delta_n(x))^2 n(n-1)(\mathbb{E}\left[K\left(\frac{d(X_i,x)}{h_n}\right)\right])^2 w(x)dP_X(x)(1+o_p(1)) \qquad (18)$$

Assume now

$$R_n - \mathbb{E}[R_n|D_0] = \begin{cases} o_p(n^2\Phi^2(h_n)\eta_n^2) \text{ if } \exists m_1 > 0,\ \forall n \in \mathbb{N}^*,\ n\Phi^2(h_n)\eta_n^2\Omega_4^{-\frac{1}{2}}(h_n) \geq \mu_1, \\ o_p(\sqrt{Var(T_{2,n})}) \text{ if } n\Phi^2(h_n)\eta_n^2\Omega_4^{-\frac{1}{2}}(h_n) \to 0. \end{cases} \qquad (19)$$

Recall that there exists an explicit sequence $K^2(1) \leq \Gamma_n \leq K^2(0)$ such that $Var(T_{2,n}) = 2(\sigma_\epsilon^2)^2\Gamma_n n(n-1)\Omega_4(h_n)$ (see Delsol *et al.*, 2010). If (19) is true then one gets:

$$Z_n := \frac{T_n - \mathbb{E}[T_{1,n}]}{\sqrt{Var(T_{2,n})}} = \frac{T_{0n} - \mathbb{E}[T_{1,n}]}{\sqrt{Var(T_{2,n})}} + \frac{\mathbb{E}[R_n|D_0]}{\sigma_\epsilon^2\sqrt{2\Gamma_n n(n-1)\Omega_4(h_n)}} + \frac{R_n - \mathbb{E}[R_n|D_0]}{\sqrt{Var(T_{2,n})}}.$$

If $n\Phi^2(h_n)\eta_n^2\Omega_4^{-\frac{1}{2}}(h_n) \to 0$, then $Z_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$ because two last terms are negligible.
If $\exists \mu_1, \mu_2 > 0$, $\forall n \in \mathbb{N}^*$, $\mu_1 \leq n\Phi^2(h_n)\eta_n^2\Omega_4^{-\frac{1}{2}}(h_n) \leq \mu_2$, the last term is negligible and the second one provide an additional bias

$$B_n \sim \Phi^2(h_n)\eta_n^2\Omega_4^{-\frac{1}{2}}(h_n)\frac{\int(\Delta_n(x))^2(\mathbb{E}\left[K\left(\frac{d(X_i,x)}{h_n}\right)\right]\Phi^{-1}(h_n))^2 w(x)dP_X(x)}{\sigma_\epsilon^2\sqrt{2\Gamma_n}}.$$

Finally, if $n\Phi^2(h_n)\eta_n^2\Omega_4^{-\frac{1}{2}}(h_n) \to +\infty$, the second term diverges and the others are negligible towards this one. Hence $Z_n \xrightarrow{P} +\infty$.

To end the proof we have to show (19). We introduce the notation $\Gamma_i(x) = K\left(\frac{d(X_i,x)}{h_n}\right)$ and $D_{in} = \Delta_n(X_i) - \overline{\Delta}$ and use the following bounds for means and conditional variances:

$$\overline{\epsilon_0} = O_p(\frac{1}{\sqrt{m_n}}) \tag{20}$$

$$\mathbb{E}[|\int \sum_{1\leq i,j\leq n} D_{in}\Gamma_i(x)\Gamma_j(x)w(x)dP_X(x)|] = O(n^2\Phi^2(h_n)) \tag{21}$$

$$\mathbb{E}[(\int \sum_{1\leq i,j\leq n} D_{in}\epsilon_j\Gamma_i(x)\Gamma_j(x)w(x)dP_X(x))^2] = O(n^3\Phi^4(h_n)) \tag{22}$$

$$Var(\int \sum_{1\leq i\neq j\leq n} D_{in}D_{jn}\Gamma_i(x)\Gamma_j(x)w(x)dP_X(x)|D_0) = O_p(n^3\Phi^4(h_n)) \tag{23}$$

$$Var(\int \sum_{i=1}^{n} D_{in}^2\Gamma_i^2(x)w(x)dP_X(x)|D_0) = O_p(n\Phi^2(h_n)). \tag{24}$$

They may be obtained with similar ideas as those used in Delsol *et al.* (2011) together with the fact $\Delta_n(X_i) - \overline{\Delta}$ is almost surely uniformly (in n) bounded. Hence their proof is omitted here.

From these bounds, one gets

$$\begin{aligned} R_n - \mathbb{E}[R_n|D_0] &= \eta_n(O_p(\frac{1}{\sqrt{m_n}})O_p(n^2\Phi^2(h_n)) + O_p(\sqrt{n^3\Phi^4(h_n)})) + \eta_n^2 O_p(\sqrt{n^3\Phi^4(h_n)}) \\ &= O_p(\eta_n\frac{n^2\Phi^2(h_n)}{\sqrt{m_n}}) + O_p(\eta_n n^{\frac{3}{2}}\Phi^2(h_n)) + O_p(\eta_n^2 n^{\frac{3}{2}}\Phi^2(h_n)) \end{aligned}$$

Assumptions (8) and (9) are hence enough to state (19).

References

1. Aneiros-Perez, G. and Vieu, P. (2006) Semi-functional partial linear regression. *Statist. Probab. Lett.* 76 (11) 1102-1110.
2. Borggaard, C. and Thodberg, H.H. Optimal minimal neural interpretation of spectra *Analytical chemistry* **64** (5) pp 545 - 551.
3. Bosq, D. (2000) *Linear Processes in Function Spaces: Theory and Applications* Lecture Notes in Statistics **149** Springer-Verlag, New York.
4. Burba, F. Ferraty, F., and Vieu, P. (2009) k-Nearest Neighbor method in functional nonparametric regression. *J. of Nonparametric Statistics*, **21**, 453-469.
5. Cao, R. (1991) Rate of convergencefor the wild bootstrap in nonparametric regression *Annals Statist.* **19** 2226-2231.
6. Cardot, H., Ferraty, F.,Mas, A. and Sarda, P. (2003) Testing Hypothesys in the Functional Linear Model *Scandinavian Journal of Statistics* **30** 241-255
7. Cardot, H., Goia, A. and Sarda, P. (2004) Testing for no effect in functional linear regression models, some computational approaches. *Comm. Statist. Simulation Comput.* **33** (1) 179-199.
8. Chen, S.X. and Van Keilegom, I. (2009) A goodness-of-fit test for parametric and semiparametric models in multiresponse regression, *Bernoulli* **15** 955-976.
9. Crambes, C., Kneip, A. and Sarda, P. (2009) Smoothing splines estimators for functional linear regression *Annals of Stat.*, **37**, 35-72.
10. Cuevas, A. and Fraiman, R. (2004) On the bootstrap methodology for functional data. (English summary) *COMPSTAT 2004—Proceedings in Computational Statistics [Ed. Antoch, J.]* 127-135 Physica, Heidelberg.
11. Cuevas, A., Febrero, M. and Fraiman, R. (2006) On the use of the bootstrap for estimating functions with functional data. *Comput. Statist. & Data Anal.* **51** (2) 1063-1074.
12. Dabo-Niang, S., Ferraty, F. and Vieu, P. (2006) Mode estimation for functional random variable and its application for curves classification. *Far East J. Theor. Stat.* **18** (1) 93-119.
13. Davidian, M., Lin, X., and Wang, J.-L. (2004) Introduction [Emerging issues in longitudinal and functional data analysis]. *Statist. Sinica* **14** (3) 613-614.
14. Delsol, L. (2008) Régression sur variable fonctionnelle: Estimation, Tests de structure et Applications. PhD thesis Université de Toulouse.
15. Delsol, L., Ferraty, F., and Vieu, P. (2011) Structural test in regression on functional variables. *Journal of Multivariate Analysis.* **102**, (3), 422-447.
16. Efron, B. (1979) Bootstrap Methods: Another Look at the Jackknife. *Annals Statist.* **7** (1) 1-26.

17. Fernandez de Castro, B., Guillas, S., and Gonzalez Manteiga, W. (2005) Functional Samples and Bootstrap for Predicting Sulfur Dioxide Levels *Technometrics* **47** (2) 212-222.
18. Ferraty, F. (2010) Editorial to the special issue Statistical Methods and Problems in Infinite-dimensional Spaces *Journal of Multivariate Analysis,* **101** (2), 305-306.
19. Ferraty, F., Laksaci, A. and Vieu, P. (2006) Estimating some characteristics of the conditional distribution in nonparametric functional models. *Stat. Inference Stoch. Process* **9** (1) 47-76.
20. Ferraty, F., Mas, A. and Vieu, P. (2007) Advances on nonparametric regression for fonctionnal data. *ANZ Journal of Statistics* **49** 267-286.
21. Ferraty, F, and Romain, Y. (2011) *The Oxford handbook of functional data analysis.* Oxford University Press, Oxford.
22. Ferraty, F., Van Keilegom, I., and Vieu, P. (2010) On the validity of the bootstrap in nonparametric functionl regression. *Scand. J. Stat.*, **37**, 286-306.
23. Ferraty, F. and Vieu, P. (2000) Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés *Compte Rendus de l'Académie des Sciences* Paris, **330**, 403-406.
24. Ferraty, F. and Vieu, P. (2002) The functional nonparametric model and application to spectrometric data. *Comput. Statist.* **17** (4) 545-564.
25. Ferraty F. and Vieu P. (2006) *Nonparametric modelling for functional data.* Springer-Verlag, New York.
26. Ferraty F. and Vieu P. (2009) Additive prediction and boosting for functional data *Computational Statistics & Data Analysis*, **53** (4), 1400-1413.
27. Ferraty, F, Vieu, P. and Viguier-Pla, S. (2007) Factor-based comparison of groups of curves *Comp. Stat. & Data Anal.* **51** (10) 4903-4910.
28. Ferré, L. and Villa, N. (2006) Multi-Layer perceptron with functional inputs: an inverse regression approach. *Scandinavian Journal of Statistics* **33** (4), 807-823.
29. Ferré, L.; Yao, A.-F. (2005) Smoothed functional inverse regression. *Statist. Sinica* **15** (3) 665-683.
30. Gadiaga, D. and Ignaccolo, R.(2005) Test of no-effect hypothesis by nonparametric regression. *Afr. Stat.* **1**, (1) 67-76.
31. González-Manteiga, W., Martinez Miranda, M.D., and Perez Gonzalez, A. (2004) The choice of smoothing parameter in nonparametric regression through Wild Bootstrap *Comp. Stat. & data Analysis* **47** 487-515.
32. Gonzalez-Manteiga, W., Quintela-del-Río, A. and Vieu, P. (2002) A note on variable selection in nonparametric regression with dependent data, *Statist. Probab. 40 Lett.*, **57** 259-268.
33. González Manteiga, W. and Vieu, P. (2007) Editorial of the special issue Statistics for Functional Data *Computational Statistics & Data Analysis* **51** (10) pp 4788-4792.
34. Hall, P. (1990) Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *J. Multivariate Anal.* **32** 177-203.
35. Hall, P. (1992) On bootstrap confidence intervals in nonparametric regression. *Ann. Statist.* **20** 695-711.
36. Hall, P. and Hart, J. (1990) Bootstrap test for differene between means in nonparametric regression *J. Amer. Statist. Assoc.* **85** 1039-1049.
37. Härdle, W. and Mammen, E. (1993) Comparing Nonparametric Versus Parametric Regression Fits *The Annals of Statistics* **21**, (4) 1926-1947.
38. Härdle, W. and Marron, J.S. (1990) Semiparametric comparison of regression curves. *Ann. Statist.* **18** (1) 63-89.
39. Hernandez, N., Biscay, R.J., and Talavera, I. (2008) Support vector regression methods for functional data *Lecture Notes Comput. Sci.* **4756** 564-573.
40. James, G. and Silverman, B.W. (2005) Functional adaptative model estimation *J. Amer. Statist. Assoc* **100** 565-576.
41. Laloë, T. (2007) A k-nearest neighbor approach for functional regression. *Statist. Probab. Lett.* **78** (10) 1189-1193.
42. Leardi, R. (2003) *Nature-inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks* Elsevier.
43. Leurgans, S. E., Moyeed, R. A. and Silverman, B. W. (1993) Canonical correlation analysis when the data are curves. *J. Roy. Statist. Soc. Ser. B* **55** (3) 725-740.
44. Mammen, E. (1993) Bootstrap and wild bootstrap for high-dimensional linear models. *Ann. Statist.* **21** (1) 255-285.
45. Mas, A. and Pumo, B. (2007) Functional linear regression with derivatives, *submitted*
46. Müller, H.-G. and Stadtmüller, U. (2005) Generalized functional linear models. *Ann. Statist.*, **33**, (2), 774-805.
47. Ramsay, J. and Dalzell, C. (1991) Some tools for functional data analysis *J. R. Statist. Soc. B.* **53** 539-572.
48. Ramsay, J. and Silverman, B. (1997) *Functional Data Analysis* Springer-Verlag, New York
49. Ramsay, J. and Silverman, B. (2002) *Applied functional data analysis: Methods and case studies* Spinger-Verlag, New York
50. Ramsay, J. and Silverman, B. (2005) *Functional Data Analysis (Second Edition)* Spinger-Verlag, New York.
51. Rossi, F., Delannay, N., Conan-Guez, B. and Verleysen, M. (2005) Representation of Functional Data in Neural Networks *Neurocomputing* **64** 183-210.
52. Stute, W., Gonzalez Manteiga, W., and Presedo Quindimil, M. (1998) Bootstrap approximations in model checks for regression. *J. Amer. Statist. Assoc.* **93** (441) 141-149.
53. Valderrama, M. (2007) An overview to modelling functional data, *Comp. Stat. & Data Analysis*, **22** (3) pp. 331-334.