Statistics Surveys
Vol. 5 (2011) 30–43
ISSN: 1935-7516
DOI: 10.1214/09-SS049

Curse of dimensionality and related issues in nonparametric functional regression*

Gery Geenens

Department of Mathematics and Statistics,
The University of Melbourne,
Melbourne, 3010 (Australia)
and
School of Mathematics and Statistics,
The University of New South Wales,
Sydney, 2052 (Australia)
e-mail: ggeenens@unsw.edu.au

Abstract: Recently, some nonparametric regression ideas have been extended to the case of functional regression. Within that framework, the main concern arises from the infinite dimensional nature of the explanatory objects. Specifically, in the classical multivariate regression context, it is well-known that any nonparametric method is affected by the so-called "curse of dimensionality", caused by the sparsity of data in high-dimensional spaces, resulting in a decrease in fastest achievable rates of convergence of regression function estimators toward their target curve as the dimension of the regressor vector increases. Therefore, it is not surprising to find dramatically bad theoretical properties for the nonparametric functional regression estimators, leading many authors to condemn the methodology. Nevertheless, a closer look at the meaning of the functional data under study and on the conclusions that the statistician would like to draw from it allows to consider the problem from another point-of-view, and to justify the use of slightly modified estimators. In most cases, it can be entirely legitimate to measure the proximity between two elements of the infinite dimensional functional space via a semi-metric, which could prevent those estimators suffering from what we will call the "curse of infinite dimensionality".

AMS 2000 subject classifications: Primary 62G08; secondary 62M40.
Keywords and phrases: Nonparametric regression, functional regression, curse of dimensionality, Nadaraya-Watson estimator, semi-normed space.

Received June 2009.

Contents

*This paper was accepted by Jianqing Fan, Associate Editor for the IMS.

1. Introduction

Consider the classical multivariate regression model

$$Y = m(X) + \varepsilon, \tag{1.1}$$

where Y is a scalar response, X a p-dimensional vector of continuous covariates, $m(\cdot)$ a smooth function from $\mathbb{R}^p$ to $\mathbb{R}$, and ε a random disturbance such that $\mathbb{E}(\varepsilon|X) = 0$ almost surely. In this context, the estimation of $m(\cdot)$ from a sample of observations, say $\{(X_k, Y_k), k = 1, \ldots, n\}$, is of great interest as this function, easily seen to be the conditional mean of Y given the value of X, captures the effect of the regressors on the response. Nonparametric methods have been proposed for many decades in literature, as an alternative basing this estimation on hazardous prior parametric assumptions on the shape of m, and granting the model maximal flexibility. See [18] for a complete overview of the nonparametric regression theory. Historically, the first nonparametric univariate regression estimator was independently proposed by [21] and [28], and is given by

$$\hat{m}(\cdot) = \frac{\sum_{k=1}^{n} K((\cdot - X_k)/h)Y_k}{\sum_{k=1}^{n} K((\cdot - X_k)/h)},$$

where K, called the kernel function, is usually a univariate density function, symmetric, supported on $[-1, 1]$ such that $xK'(x) \leq 0$, and h is a positive number depending on n known as the bandwidth. For any x in the domain of interest, the estimation $\hat{m}(x)$ therefore turns out to be the weighted average of the observed responses $\{Y_k\}$, with weights depending upon the distance between x and X_k: the kernel K defines the way the weights decrease with the distance, while the bandwidth h quantifies the notion of closeness between two points of $\mathbb{R}$. Thereby, only those observations such that X_k lies within $]x - h, x + h[$ are used in the estimation of m at x. This estimator is readily adapted to the multivariate setting as

$$\hat{m}(\cdot) = \frac{\sum_{k=1}^{n} \mathbf{K}(H^{-1}(\cdot - X_k))Y_k}{\sum_{k=1}^{n} \mathbf{K}(H^{-1}(\cdot - X_k))}, \tag{1.2}$$

where the kernel $\mathbf{K}$ is now a p-variate density and H a $(p \times p)$ bandwidth matrix, with respective roles similar to previously. Very often, in order to keep the symmetric nature of the kernel function (up to a suitable normalization of the different components of X), H is chosen as the diagonal matrix

$$H = hI_p \tag{1.3}$$

for some scalar bandwidth h, and $\mathbf{K}$ is taken as

$$\mathbf{K}(u) = K(\|u\|), \tag{1.4}$$

with K a univariate density supported and decreasing on $[0,1]$, and $\|\cdot\|$ the usual Euclidean norm of $\mathbb{R}^p$. Then, only those observations with X_k belonging to the hypersphere of radius h centred at x are used in the computation of $\hat{m}(x)$. Although many other nonparametric regression estimators have been subsequently developed, and sometimes shown to outperform it (see [8]), the Nadaraya-Watson (NW) estimator remains a reference in nonparametric regression theory, as well as in lots of practical studies, mainly due to its simplicity of derivation, interpretation and implementation. Therefore, plenty of theoretical results about it are available. For instance, it is well-known that, under appropriate mild assumptions,

$$\mathbb{E}(\hat{m}(x)) - m(x) = O(h^2) \tag{1.5}$$

and

$$\mathbb{V}\mathrm{ar}(\hat{m}(x)) = O((nh^p)^{-1}) \tag{1.6}$$

for x fixed in the interior of the domain under study, so that pointwise consistency of the estimator requires the following two conditions: $h \to 0$ and $nh^p \to \infty$. This is understood intuitively: when estimating m at x, h must tend to zero in order to keep only relevant information, that is observations as close as possible to x to avoid bias, but at the same time the number of observations used in the estimation i.e., the observations in the concerned neighborhood of x, needs to grow to infinity so as to ensure that the variance of the estimator tends to zero. Note that, from a basic binomial argument, this number of observations effectively used in the estimation is asymptotically equivalent to $n\varphi_x(h)$, where $\varphi_x(h)$ is the small ball probability

$$\varphi_x(h) = \mathbb{P}(\|X - x\| \leq h).$$

If X admits a density f bounded away from zero, which is usually assumed, then it is readily seen that

$$\varphi_x(h) \sim f(x)\ h^p \frac{2\pi^{p/2}}{\Gamma(p/2)}$$

as $h \to 0$, where $h^p(2\pi^{p/2})/\Gamma(p/2)$ is the volume of the hypersphere of radius h in $\mathbb{R}^p$. Therefore, $n\varphi_x(h) \sim nh^p$, and this must tend to infinity. Note that the condition can also be strengthened to $(\log n)^{-1}(nh^p) \to \infty$ to get the consistency uniformly in x. Also, [27] showed that, if m is r times differentiable, the fastest achievable rate of uniform convergence of $\hat{m}$ to m is $(\log n/n)^{r/(2r+p)}$, which clearly emphasizes the role of the dimension of the regressor vector in the quality of the estimator: increasing the number of regressors dramatically decreases its performance. This phenomenon, essentially due to the sparsity of data in higher dimensional spaces, is known as the curse of dimensionality and probably represents the main drawback of nonparametric techniques.

On the other hand, an interesting extension of the model (1.1) is to let p become infinitely large, which introduces the functional regression problem. A random element is said to be functional if it takes its values in an infinite dimensional space. To fix ideas, only random curves will be considered in the sequel, although many other random elements enter this general definition (vectors of random curves, random surfaces, random fields of any dimension,...). Concretely, the regression problem is no longer about linking the response Y to just a set of p characteristics $(X^{(1)}, \ldots, X^{(p)})$, but rather to a whole curve observed on a domain T

$$\mathcal{X} \doteq \{\mathcal{X}(t) : t \in T\},$$

as is often the case required in various applications, from quantum mechanics to econometrics, communications, medicine, musicology and climatology. Thereby, the functional regression model we consider is

$$Y = \mu(\mathcal{X}) + \varepsilon,$$

with $Y \in \mathbb{R}$ and ε a scalar random disturbance such that $\mathbb{E}(\varepsilon|\mathcal{X}) = 0$ almost surely. The random curve $\mathcal{X}$ is assumed to belong to an appropriate set of functions, say the set $L^2(T)$ of all square-integrable functions on T, so that $\mu(\cdot)$ is now an operator from $L^2(T)$ to $\mathbb{R}$ satisfying some necessary regularity conditions. For a long time, only parametric models were developed for estimating μ in this context, see [24, 25, 26, 5, 9] or [19]. However, in the functional case there is no visual guide available since any graphical representation is inconceivable in an infinite-dimensional space. As graphical techniques like scatter-plots and residual plots are usually the primary tools to define and validate a suitable parametric regression model, it is not surprising that the risk of model misspecification is even higher in the functional regression problem than in the vectorial case. The flexibility guaranteed by nonparametric methods therefore led the authors in [10], following some pioneer papers, to recently propose the generalization of the Nadaraya-Watson estimator to the case of functional regression. Section 2 describes the basic extension, which we call the naive functional Nadaraya-Watson estimator, and points out the main problem that arises from this. In Section 3, some thoughts and examples motivate the use of another concept of proximity rather than a classical norm, whilst in Section 4, this concept is properly defined and adapted to the NW estimator. Section 5 illustrates these ideas on a real data example, and Section 6 concludes.

2. The naive functional Nadaraya-Watson estimator and the curse of infinite dimensionality

From the brief reminder of the ideas underlying the Nadaraya-Watson estimator given in Section 1, an extension to the functional case is straightforward. The natural distance between two elements of $L^2(T)$, say χ_1 and χ_2, is measured in term of the L^2-norm of their difference, defined by

$$\|\chi_1 - \chi_2\|_2 = \left(\int_T (\chi_1(t) - \chi_2(t))^2 dt \right)^{1/2},$$

so that, from a sample $\{(\mathcal{X}_k, Y_k) \in L^2(T) \times \mathbb{R}, k = 1, \dots, n\}$, the NW estimator of the operator μ applied to some fixed element χ of $L^2(T)$ is given by

$$\hat{\mu}(\chi) = \frac{\sum_{k=1}^n K(\|\chi - \mathcal{X}_k\|_2/h) Y_k}{\sum_{k=1}^n K(\|\chi - \mathcal{X}_k\|_2/h)}, \tag{2.1}$$

similarly to (1.2), (1.3) and (1.4), with K a univariate kernel function and h a scalar bandwidth. In the same spirit as the development which yields (1.5) and (1.6), [12] showed that estimator (2.1) is such that, under appropriate conditions,

$$\mathbb{E}(\hat{\mu}(\chi)) - \mu(\chi) = O(h)$$

and

$$\mathbb{V}\mathrm{ar}(\hat{\mu}(\chi)) = O((n\varphi_\chi(h))^{-1}),$$

where $\varphi_\chi(h)$ is the small ball probability associated to the random functional $\mathcal{X}$ in $L_2(T)$, that is

$$\varphi_\chi(h) = \mathbb{P}(\|\mathcal{X} - \chi\|_2 \leq h).$$

The rate of convergence of $\hat{\mu}$ toward μ ought to be directly influenced by its variance, and therefore by this small ball probability. Although it is not really necessary, assume there exists an operator ϕ from $L^2(T)$ to $\mathbb{R}^+$ and an absolutely continuous function ν from $\mathbb{R}^+$ to $\mathbb{R}^+$ with $\nu(0) = 0$, such that

$$\varphi_\chi(h) \sim \nu(h)\phi(\chi) \tag{2.2}$$

as $h \to 0$. This allows to draw clear parallels with the vectorial situation, seeing ϕ as the functional probability density of $\mathcal{X}$ at χ, while ν, called the concentration function, measures how densely packed are the considered elements of $L^2(T)$ in an infinite-dimensional ball of radius h. Furthermore, how this function precisely behaves represents the main difference between the functional and the vectorial cases. In the Euclidean space $\mathbb{R}^p$, the concentration is only dictated by the volume of the p-dimensional sphere of radius h, so that $\nu(h) \sim h^p$, while in the functional context of interest, it also depends upon the structure of the stochastic process regulating the behavior of $\mathcal{X}$ and on the considered topology. Although this can hardly be stated in general, it appears that any continuous time random process which has been studied so far (see [11] for a comprehensive discussion), for instance in the common classes of Gaussian processes or diffusion processes, admits a concentration function associated with any usual norm (such as any L^p-norm) of the form

$$\nu(h) \sim h^{-\alpha} \exp(-ch^{-\beta}) \tag{2.3}$$

as $h \to 0$, for some positive constants α, β and c, so that it can be conjectured that $\nu(h)$ typically decreases to 0 exponentially quickly as $h \to 0$ in our context. In light of the usual observations related to the curse of dimensionality in the multivariate setting, this extreme sparsity of data in the functional space of interest is expected to yield poor theoretical properties for the estimator (2.1),

and indeed, it can be shown that the rate of convergence of $\hat{\mu}$ toward μ is of order $(\log n)^{-\gamma}$, for some $\gamma > 0$. This unacceptable logarithmic rate, consequence of what we could call the *curse of infinite dimensionality*, has brought many authors to forget about nonparametric estimators for functional regression. Nevertheless, results of [10] show that the use of another wisely chosen proximity measure in place of $\|\cdot\|_2$ in the previous development can get around this curse. This idea is explored and justified in the next section.

3. A closer look at functional data

Clearly, the failure of the naive functional Nadaraya-Watson estimator described in the previous section arises from the dramatically fast decrease of the concentration function (2.3) when h tends to zero. Concretely, for a fixed χ, there are too few $\mathcal{X}_k$ close enough to χ in the sense of $\|\cdot\|_2$, even when the sample size grows to infinity, and hence insufficient to make the estimation of $\mu(\chi)$ reliable. Therefore, an interesting idea would then be to use a "looser" proximity measure between functions, which would result in another concentration function, that hopefully decreases less rapidly. In other words, a proximity measure which would find more curves close to each other, and therefore would allow more observations to enter the computation of $\hat{\mu}(\chi)$.

Actually, the L^2-norm is the strict generalization of the Euclidean norm to functions. As such, it just considers functions as infinite-dimensional vectors, only focusing on the actual values taken by the functions, and by doing so it fails to capture many of the features proper to functions, such that their general appearance, the way they vary over short or long range, etc. This fact is illustrated by the following two examples.

Example 1. Consider the two functions χ_1 and χ_2 represented in Figure 1.

If you ask people if they are similar or not, without specifying how "similar" is to be understood, many would answer in the affirmative. Indeed, the striking point when looking at this graph is that their general appearance is almost identical (actually it is identical), and that only a vertical shift makes them

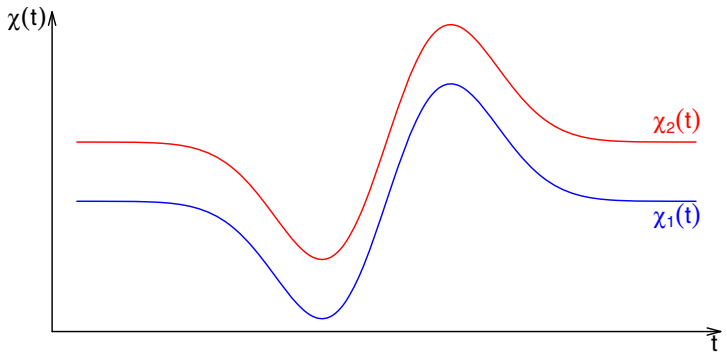

FIG 1. *Example 1.*

different. Nevertheless, the L^2-norm, quantifying in a sense the area between the curves, only focuses on that shift, and concludes that they are totally different. This would probably be suitable if the actual values taken by the functions were of importance in the problem under study, but if they are not, e.g. if only the shape of the functions is related to the response Y, this may lead to wrong conclusions. In this latter case, the proximity between χ_1 and χ_2, say $(\!(\chi_1 - \chi_2)\!)$, could be based on the norm of the difference between their first derivatives, which obviously cancels out in the case of a vertical shift:

$$(\!(\chi_1 - \chi_2)\!) = \left(\int_T (\chi_1'(t) - \chi_2'(t))^2 dt \right)^{1/2}. \tag{3.1}$$

With the two curves represented in Figure 1, we would have $(\!(\chi_1 - \chi_2)\!) = 0$.

Example 2. Consider the oscillating functions χ_1, χ_2 and χ_3 represented in Figure 2, and again one can ask people if they are similar or not. Probably most would answer would that χ_1 and χ_2 are similar, but both very different to χ_3. Nevertheless, $\|\chi_1 - \chi_2\|_2$ is certainly not small, owing to the differing amplitude and phase of the respective oscillations. What make them similar is the general trend they seem to share, and which is totally different to the general trend of χ_3. Suppose that the oscillations are just noise, and that only the general trend of the curves is relevant in the regression model. Then, the proximity measure could be based for instance on the first term(s) in some suitable polynomial expansions, in order to have $(\!(\chi_1 - \chi_2)\!) \simeq 0$, $(\!(\chi_1 - \chi_3)\!) \gg 0$ and $(\!(\chi_2 - \chi_3)\!) \gg 0$. On the other hand, if the response is mainly related to the oscillating behavior of the functional predictors and not at all to their trend, it would be interesting to work with a proximity measure such that $(\!(\chi_2 - \chi_3)\!) \simeq 0$ and $(\!(\chi_1 - \chi_2)\!) \gg 0$, for example by basing it on the Fourier transform of the considered curves.

The previous two simple examples emphasize that the natural norm $\|\cdot\|_2$ is probably too exacting in most situations, as it fails to leave out some irrelevant features of the curves in the analysis, whereas other looser proximity measures

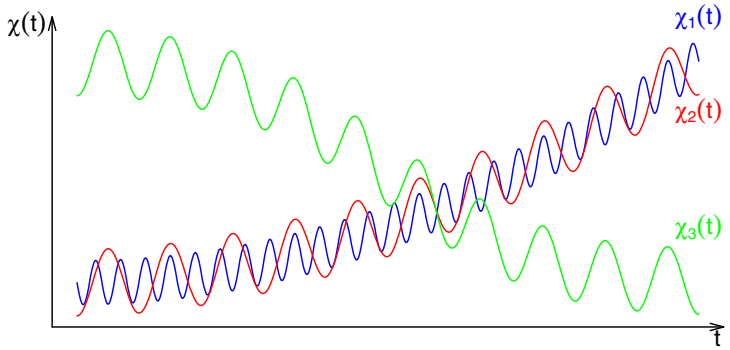

FIG 2. *Example 2.*

might do a better job. Besides, using such measures rather than $\|\cdot\|_2$ would kill two birds with one stone in that: *(i)* we have the liberty to focus on some particular characteristics of the curves that we know to be directly related to the response, and *(ii)* as some differences between curves are smoothed over, the associated concentration function should (hopefully) decrease much slower than the one associated to $\|\cdot\|_2$. Proximity measures such as (3.1) are actually semi-norms, and some results of [10] fully support assertion *(ii)*. The next section summarizes those results.

4. Semi-norms and related results

A measure $(\!(\cdot)\!)$ is a semi-norm on $L^2(T)$ if *(i)* for all $\lambda \in \mathbb{R}$ and $\chi \in L^2(T)$, $(\!(\lambda\chi)\!) = |\lambda|(\!(\chi)\!)$, and *(ii)* for all $\chi_1, \chi_2 \in L^2(T)$, $(\!(\chi_1+\chi_2)\!) \leq (\!(\chi_1)\!) + (\!(\chi_2)\!)$. The important point is that it is not required that $(\!(\chi)\!) = 0$ if and only if $\chi = 0$, as it should for $(\!(\cdot)\!)$ to be a genuine norm. Note that this is well in accord with the ideas put forward in the previous section: the proximity measure flattens some differences out, as you could find distinct elements χ_1 and χ_2 with $(\!(\chi_1-\chi_2)\!) = 0$. The general idea is therefore to work in a semi-normed functional space, that is a functional space endowed with a semi-norm $(\!(\cdot)\!)$, rather than in the usual normed spaces, such as Banach or Hilbert spaces. Obviously, the selected semi-norm has to be able to extract the whole information needed to link the response to the functional predictor, which can be formalized by the property

$$\mu(\chi) = \mathbb{E}\big(\mu(\mathcal{X})|(\!(\mathcal{X}-\chi)\!) = 0\big). \tag{4.1}$$

This assertion is therefore a key for the consistency of the estimation procedure and has consequently to be rigorously verified.

Now, once a semi-norm $(\!(\cdot)\!)$ has been selected, the functional Nadaraya-Watson estimator naturally becomes

$$\hat{\mu}(\chi) = \frac{\sum_{k=1}^n K((\!(\chi-\mathcal{X}_k)\!)/h)Y_k}{\sum_{k=1}^n K((\!(\chi-\mathcal{X}_k)\!)/h)}. \tag{4.2}$$

Naturally, the performance of this estimator again strongly depends on the concentration function associated with the considered semi-norm. Although it is not possible to give general results about concentration functions for continuous random processes in semi-normed spaces, Lemma 13.6 of [10] provides some interesting results. We particularize them to our case, working with the inner product $\langle\chi_1,\chi_2\rangle = \big(\int_T \chi_1(t)\chi_2(t)dt\big)^{1/2}$ in $L^2(T)$, via the following two lemmas.

Lemma 4.1. *Suppose that* $\{e_j, j = 1, 2, \cdots\}$ *forms an orthonormal basis of* $L^2(T)$*, so that any element* χ *of* $L^2(T)$ *can be written* $\chi = \sum_{j=1}^\infty \langle\chi, e_j\rangle e_j$*. Then, the function defined as*

$$(\!(\chi)\!)_p = \sqrt{\sum_{j=1}^p \langle\chi, e_j\rangle^2}, \tag{4.3}$$

with p *a fixed positive integer, is a semi-norm on* $L^2(T)$*.*

This kind of semi-norm is often called 'projection-type', as it essentially returns the L^2-norm of the projection of the element of interest on a p-dimensional subset of $L^2(T)$. The main asset of using such a semi-norm is highlighted by the following result.

Lemma 4.2. *Let $\mathcal{X}$ be a random element of $L^2(T)$, and write $\mathcal{X} = \sum_{j=1}^{\infty} X^{(j)} e_j$ with $X^{(j)} = \langle \mathcal{X}, e_j \rangle$. If the random vector $(X^{(1)}, X^{(2)}, \ldots, X^{(p)})$ admits an absolutely continuous density on $\mathbb{R}^p$, then the concentration function associated to the semi-norm (4.3) is such that*

$$\nu(h) \sim h^p. \tag{4.4}$$

The advantage is evident: the concentration function associated to this kind of projection semi-norms decreases to zero at the same rate as for usual p-dimensional random vectors (see the lines following (2.2)), which is no real surprise as each functional object is now essentially characterized by a vector of p components $\langle \chi, e_j \rangle, j = 1 \ldots, p$. Obviously, (4.4) decays to 0 much slower than what (2.3) indicates. Note that many practical procedures make use of this result, as it applies to any semi-norm based on some expansion of the functions of interest in some basis of $L^2(T)$, for instance in Example 2 above: we have expansion in a polynomial basis or expansion in a Fourier basis. One can also think of the expansion in the orthonormal basis of eigenfunctions arising from a functional Principal Components Analysis (see a.o. Chapter 6 in [24]). Besides, the assumption made on the joint distribution of $(X^{(1)}, X^{(2)}, \ldots, X^{(p)})$ is usually fulfilled in practice. For instance, the coefficients $(X^{(1)}, X^{(2)}, \ldots, X^{(p)})$ arising in a Karhunen-Loève expansion of a Gaussian or a Wiener process form a multivariate gaussian vector. In consequence of (4.4), the rate of uniform convergence of estimator (4.2) computed from this kind of semi-norm is similar to those in the p-dimensional case, that is, under appropriate conditions, $(n^{-1} \log n)^{r/(2r+p)}$, where r quantifies the smoothness of μ with respect to $(\!(\cdot)\!)$. As expected, the use of a suitable semi-norm allows to get around the curse of infinite dimensionality. Furthermore, the problem even amounts to a univariate regression if it makes sense, in the considered context, to measure the proximity between two functions through their first components in a suitable expansion. Note that this totally concurs with an idea of single-index model in a functional framework, see [1].

Now, the problem of selecting the right semi-norm appears to be crucial. Ideally, the choice of $(\!(\cdot)\!)$ should be dictated by some prior knowledge about the underlying phenomenon. In Example 2 for instance, whether the response is mainly influenced by the general trend of the explanatory curve or by the oscillating behavior directly motivates the use of one or another semi-norm. However, in practice, it is often unknown which feature of $\mathcal{X}$ is directly related to Y and which other one is not, so an important open question remains as to how to choose a suitable semi-norm when little prior information on the phenomenon is available. A possible methodology could be based on some test for assumption (4.1), which could be seen as a kind of goodness-of-fit test.

5. A real data example

In this section we illustrate the above ideas within the framework of automatic signature recognition, a problem having attracted attention for a long time. There is a clear need for accurate and reliable recognition systems, aiming at discriminating forgeries from genuine signatures. It seems natural to tackle this problem from a functional perspective, modelling a signature as a random function

$$\mathcal{S} : \mathcal{T} \subset \mathbb{R}^+ \to \mathcal{P} \subset \mathbb{R}^2 : t \to \mathcal{S}(t) = (\mathcal{X}(t), \mathcal{Y}(t))$$

where $\mathcal{S}(t) = (\mathcal{X}(t), \mathcal{Y}(t))$ represents the position of the pen in $\mathcal{P}$, a given portion of the two-dimensional plane, at time $t \in \mathcal{T}$, the considered time domain. The functional object of interest is here a vector of functions, and we assume that it lies in an appropriate infinite-dimensional functional space, say Σ. The random nature of the so-defined object obviously accounts for the natural variability between successive signatures from one writer.

Suppose we observe a realization ς of the random object $\mathcal{S}$, and we have to make a decision as to whether this observed signature is a fake or not. This is obviously nothing else but a classification problem. The decision will be based on an estimation of the probability of ς being a fake, that is

$$\pi(\varsigma) = P(Z = 1 | \mathcal{S} = \varsigma),$$

where Z is a binary random variable, taking the value 1 if $\mathcal{S}$ is a forgery and 0 if it is a genuine signature. Note that, due to the binary nature of Z, this conditional probability can also be written

$$\pi(\varsigma) = E(Z | \mathcal{S} = \varsigma),$$

so that $\pi(\varsigma)$ can be estimated by functional regression methods. However, it is not clear why the operator π should follow any of the usual parametric models for binary regression, e.g. logit or probit. We therefore argue that a nonparametric estimation is suitable here.

We propose to measure the proximity between two signatures ς_1 and ς_2 with the semi-distance

$$(\!(\varsigma_1 - \varsigma_2)\!) \doteq \left(\int (\varsigma_1''(t) - \varsigma_2''(t))^2 \, dt \right)^{1/2}, \tag{5.1}$$

where $\varsigma''(t)$ is the tangential projection of the vector of second derivatives with respect to time of the signature-function ς, as this would account for the similarity (or dissimilarity) in tangential pen acceleration between two signing processes. It is commonly admitted that the acceleration of the pen is mainly dictated by the movement of the wrist of the person signing. Besides, it is quite clear that the "genuine" wrist movement is very hard, if not impossible, to detect and reproduce even for a skilled faker. Unlike the drawing itself, which a usual distance between ς_1 and ς_2 would focus on, this movement and the acceleration it induces are consequently unique to every single person and should be very

efficient discriminatory elements. Moreover, working with second derivatives obviates the need for an important pre-processing of the recorded signatures, as the second order differentiation cancels out any location or size effect. We can therefore assume that $\mathcal{S}$ belongs to Σ, the space of functions from $\mathbb{R}^+$ to $\mathbb{R}^2$, both of whose components are twice differentiable, dotted with the semi-distance (5.1). Now, assume that we have a sample $\{(\mathcal{S}_k, Z_k), k = 1, \ldots, n\}$ of i.i.d. replications of $(\mathcal{S}, Z) \in \Sigma \times \{0, 1\}$. Then, observing a signature ς, a Nadaraya-Watson-like estimator for the conditional probability $\pi(\varsigma)$ is given by the estimator (4.2) adapted to this context, i.e.

$$\hat{\pi}(\varsigma) = \frac{\sum_{k=1}^{n} K(((\varsigma - \mathcal{S}_k))/h) Z_k}{\sum_{k=1}^{n} K(((\varsigma - \mathcal{S}_k))/h)}, \tag{5.2}$$

where K is a nonnegative kernel function, supported and decreasing on $[0, 1]$, and h is the bandwidth, as discussed in the previous section. The decision then directly follows by comparing $\hat{\pi}(\varsigma)$ with a given threshold, say c: if $\hat{\pi}(\varsigma) > c$, the observed signature is likely to be a fake and is therefore rejected. If $\hat{\pi}(\varsigma) \leq c$, the signature is accepted. The usual Bayes rule would set c to $1/2$, however, depending on the application, this threshold value can be adjusted to match the required standards.

We illustrate this idea with the freely available signature data set used for the First International Signature Verification Competition (SVC2004), see [29]. This database consists of 100 sets of signatures data, each set containing 20 genuine signatures from one signature contributor and 20 skilled forgeries from at least four other contributors. The validity of using the semi-distance (5.1) in this application is illustrated in Figure 3, which shows five tangential acceleration functions for genuine signatures and a 'fake' tangential acceleration function, for the first user of the database. The consistency of the tangential acceleration over the genuine signatures is clear, in contrast to what is shown for the forgery.

For each user, we decided to split the 40 available signatures in two: 10 genuine signatures and 10 forgeries would be utilized as the training set, with the other 20 (again, 10 genuine signatures and 10 forgeries) being our test set. For each signature of the test set we estimated the fake probability with (5.2), using a Gaussian kernel and a (user-dependent) bandwidth of type k-nearest neighbor determined by least-squares cross-validation, and computed the equal error rate (EER), that is, the false rejection rate plus the false acceptance rate, for each user. We observed important variations over the users, which renders the fact that some signatures are easier to reproduce than others - even in terms of wrist movement. For some users, the EER was 0 (perfect discrimination), but for some others it was around 25%. On average, the EER was 9%, with a median of 5%, which is after all quite a good result for the functional discrimination rule we set up. More details about this study can be found in [17].

6. Concluding remarks

The aim of this note was to summarize the main ideas developed in [10] concerning nonparametric functional regression, and to comment on them in order

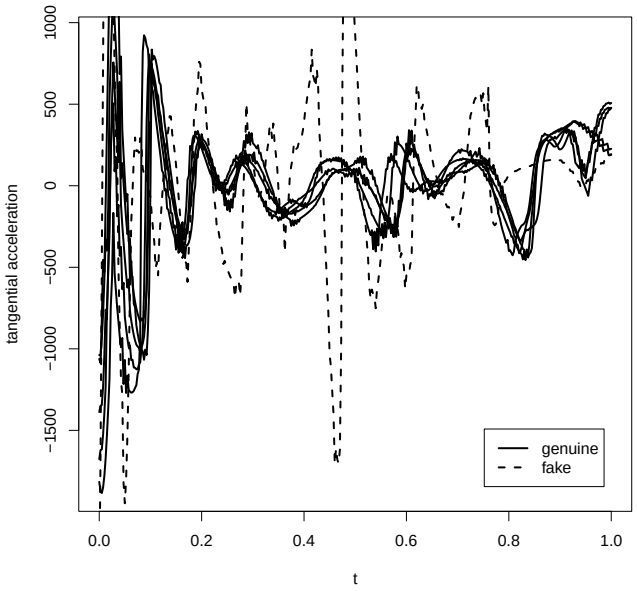

FIG 3. *Five 'genuine' tangential acceleration functions (plain line) and one 'fake' tangential acceleration function (dotted line).*

to give another viewpoint about some features of the procedure. In particular, the introduction of semi-norms as proximity measures is often presented as a technical tool used for dimension reduction purposes, in order to get around the curse of infinite dimensionality; see also the closely related work of [16]. Instead, we would like to stress that resorting to this kind of proximity measure is, in most of the situations, totally justified, and preferable to using the classical L^2-norm, not only theoretically but also intuitively. This allows a proper taking into account of the functional nature of the regressor and to exploit maximally the wealth of this kind of element, which is not possible when considering it just as an infinite-dimensional vector. The theory of nonparametric functional regression is only at its beginning, and the possibilities to improve on or discuss estimator (4.2) appear numerous, see e.g. [20, 23, 2, 22, 6, 3, 4, 7, 13] or [14] for a few paths in that direction. Note that [15] presents an up-to-date and comprehensive review of the literature in this field.

Acknowledgements

This work was partially supported by a FSR grant from the Université catholique de Louvain (Belgium), a Centre of Excellence grant to MASCOS from the Australian Research Council and an ECR grant from the University of New South Wales (Australia). Useful discussions with Francis Hui (UNSW) are also gratefully acknowledged.

References

[1] Ait-Saïdi, A., Ferraty, F., Kassa, K. and Vieu, P. (2008). Cross-validated estimations in the single-functional index model, Statistics, 42, 475–494. MR2465129

[2] Aneiros-Perez, G. and Vieu, P. (2008). Nonparametric time series prediction: A semi-functional partial linear modeling, J. Multivariate Anal., 99, 834–857. MR2405094

[3] Baillo, A. and Grané, A. (2009). Local linear regression for functional predictor and scalar response, J. Multivariate Anal., 100, 102–111. MR2460480

[4] Burba, F., Ferraty, F. and Vieu, P. (2009). k-Nearest Neighbour method in functional nonparametric regression, J. Nonparam. Stat., 21, 453–469. MR2571722

[5] Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional linear model, Stat. Probabil. Lett., 45, 11–22. MR1718346

[6] Crambes, C., Kneip, A. and Sarda, P. (2009). Smoothing splines estimators for functional linear regression, Ann. Statist., 37, 35–72. MR2488344

[7] Delsol, L. (2009). Advances on asymptotic normality in nonparametric functional time series analysis, Statistics, 43, 13–33. MR2499359

[8] Fan, J. and Gijbels, I. (1996). Local Polynomial Modelling and Its Applications, Chapman and Hall, London. MR1383587

[9] Fan, J. and Zhang, J.-T. (2000). Two-step estimation of functional linear models with application to longitudinal data, J. Roy. Stat. Soc. B, 62, 303–322. MR1749541

[10] Ferraty, F. and Vieu, P. (2006). Nonparametric Functional Data Analysis, Springer-Verlag, New York. MR2229687

[11] Ferraty, F., Laksaci, A. and Vieu, P. (2006). Estimating Some Characteristics of the Conditional Distribution in Nonparametric Functional Models, Statist. Inf. Stoch. Proc., 9, 47–76. MR2224840

[12] Ferraty, F., Mas, A. and Vieu, P. (2007). Nonparametric regression on functional data: inference and practical aspects, Aust. NZ. J. Stat., 49, 267–286. MR2396496

[13] Ferraty, F., Van Keilegom, I. and Vieu, P. (2010). On the validity of the bootstrap in nonparametric functional regression, Scand. J. Stat., 37, 286–306. MR2682301

[14] Ferraty, F., Laksaci, A., Tadj, A. and Vieu, P. (2010). Rate of uniform consistency for nonparametric estimates with functional variables, J. Stat. Plan. Inf., 140, 335–352. MR2558367

[15] Ferraty, F. and Romain, Y. (2011). Oxford handbook on functional data analysis (Eds), Oxford University Press.

[16] Gasser, T., Hall, P. and Presnell, B. (1998). Nonparametric estimation of the mode of a distribution of random curves, J. Roy. Stat. Soc. B, 60, 681–691. MR1649539

[17] Geenens, G. (2011). A nonparametric functional method for signature recognition, *Manuscript*.

[18] Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004). Nonparametric and semiparametric models, Springer-Verlag, Berlin. MR2061786

[19] James, G.M. (2002). Generalized linear models with functional predictors, J. Roy. Stat. Soc. B, 64, 411–432. MR1924298

[20] Masry, E. (2005). Nonparametric regression estimation for dependent functional data: asymptotic normality, Stochastic Process. Appl., 115, 155–177. MR2105373

[21] Nadaraya, E.A. (1964). On estimating regression, Theory Probab. Applic., 9, 141–142.

[22] Quintela-Del-Rio, A. (2008). Hazard function given a functional variable: nonparametric estimation under strong mixing conditions, J. Nonparam. Stat., 20, 413–430. MR2424250

[23] Rachdi, M. and Vieu, P. (2007). Nonparametric regression for functional data: automatic smoothing parameter selection, J. Stat. Plan. Inf., 137, 2784–2801. MR2323791

[24] Ramsay, J. and Silverman, B.W. (1997). Functional Data Analysis, Springer-Verlag, New York. MR2168993

[25] Ramsay, J. and Silverman, B.W. (2002). Applied functional data analysis; methods and case study, Springer-Verlag, New York. MR1910407

[26] Ramsay, J. and Silverman, B.W. (2005). Functional Data Analysis, 2nd Edition, Springer-Verlag, New York. MR2168993

[27] Stone, C.J. (1982). Optimal global rates of convergence for nonparametric regression, Ann. Stat., 10, 1040–1053. MR0673642

[28] Watson, G.S. (1964). Smooth regression analysis, Sankhya A, 26, 359–372. MR0185765

[29] Yeung, D.T., Chang, H., Xiong, Y., George, S., Kashi, R., Matsumoto, T. and Rigoll, G. (2004). SVC2004: First International Signature Verification Competition, Proceedings of the International Conference on Biometric Authentication (ICBA), Hong Kong, July 2004.