Available online at www.sciencedirect.com

ScienceDirect

Journal of Statistical Planning and Inference 137 (2007) 829–840

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Regression models for functional data by reproducing kernel Hilbert spaces methods

Cristian Preda*

Département de Statistique-CERIM, Faculté de Médecine, Université de Lille 2, 1, Place de Verdun, 59045 Lille Cedex, France

Available online 25 July 2006

Abstract

Non-parametric regression models are developed when the predictor is a function-valued random variable $X=\{X_t\}_{t\in T}$. Based on a representation of the regression function $f(X)$ in a reproducing kernel Hilbert space such models generalize the classical setting used in statistical learning theory. Two applications corresponding to scalar and categorical response random variable are performed on stock-exchange and medical data. The results of different regression models are compared.

MSC: Primary 62G08; secondary 62H30

Keywords: Functional data; Stochastic process; RKHS; Statistical learning

1. Introduction

Functional data analysis extends the classical multivariate methods when data are functions or curves. Examples of functional data can be found in different application fields such as medicine, economics, chemometrics and many others (see Ramsay and Silverman, 2002 for an overview). A well accepted model for this kind of data is to consider it as paths of a stochastic process $X=\{X_t\}_{t\in T}$ taking values in a Hilbert space H of functions on some set T. For example, a second order stochastic process $X=\{X_t\}_{t\in[0,1]}$, L_2-continuous with sample paths in $L_2([0,1])$ can be used as model for index stock-exchange evolution during a lapse of time or for the knee flexion angle measure over a complete gait cycle (Fig. 1(a) and (b)).

A major interest when dealing with this kind of data is to develop regression models. Let Y be a real random variable and $\{(X_i, Y_i)\}_{i=1,\dots,n}$ a sample of independent and identically distributed random variables from the distribution of (X, Y). As in the classical setting, regression models for functional data are looking for a functional f such that it minimizes the expected risk $R[f]=\mathbb{E}(\mathscr{C}(X, Y, f(X)))$, where $\mathscr{C}$ is a loss function depending on the problem (square loss, logistic loss, ε-insensitive loss, etc.).

Methods from the multivariate analysis are developed and adapted to functional data (Ramsay and Silverman, 1997). Thus, generalized linear regression models are developed in James (2002) and more recently by Cardot and Sarda (2005). Different linear approaches based on decomposition of the underlying stochastic process are proposed: principal component regression (Aguilera et al., 1997; Cardot et al., 1999), partial least squares regression (Preda

* Tel.: +33 3 20 62 69 69; fax: +33 3 20 52 10 22.
E-mail address: cpreda@univ-lille2.fr.

0378-3758/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2006.06.011

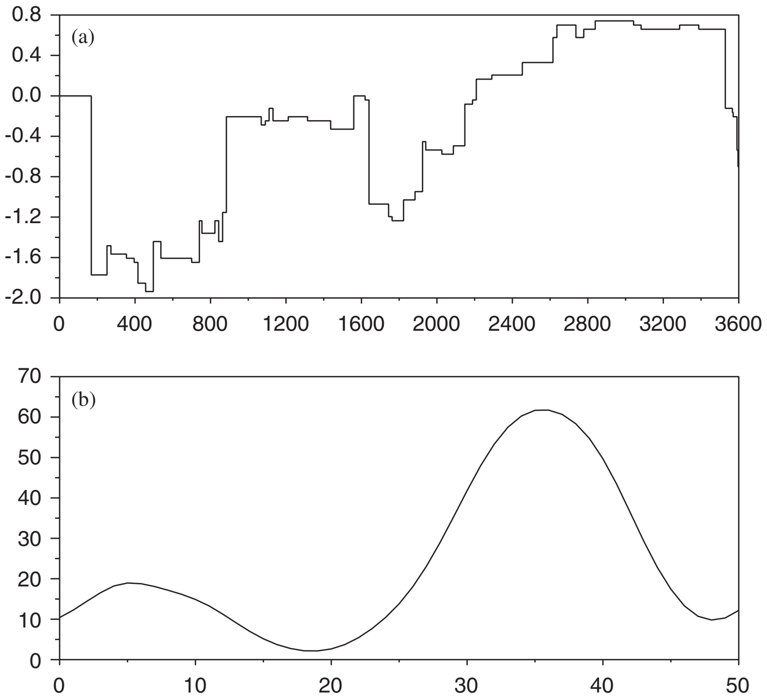

Fig. 1. Some examples of functional data. (a) Share index evolution during 1 h, (b) knee flexion angle (degree) over a complete gait cycle.

and Saporta, 2005). Ferraty and Vieu (2004) propose non-parametric models for regression on functional data using classical kernel estimators both for scalar and categorical response.

Regression models for functional data are used in several application domains: Ratcliffe et al. (2002a, b) develop functional regression models for foetal heart data in order to predict the probability of high risk birth outcome, Ferraty and Vieu (2003) are interested in classification curves from spectrometric data, Escabias et al. (2005) develop logistic models for environmental data, Preda and Saporta (2005) and Aguilera et al. (1999) use linear models for economic data based on partial least square approach and principal component analysis respectively, etc.

The point of view we adopt in this paper is inspired by techniques widely used nowadays in statistical learning theory. An estimator $\hat{f}$ for f is then obtained by minimizing, over a suitable class of functionals $\mathscr{H}$, the regularized empirical risk defined by

$$R_{\text{reg}}[f] = \frac{1}{n}\sum_{i=1}^{n}\mathscr{C}(X_i, Y_i, f(X_i)) + \lambda\Omega[f], \tag{1}$$

where $\lambda > 0$ is a regularized parameter and $\Omega[f]$ a stabilization term (Schölkopf and Smola, 2002).

In this context, the goal of the statistical learning methods based on kernels is to search an estimate of the regression function f into a class of functions which has the properties of a Hilbert space with reproducing kernel. In the finite dimensional case, a such approach allows to fit non-linear regression models and gives simple representations for $\hat{f}$ (Wahba, 1990). Conditions which guarantee consistency for $\hat{f}$ are given in Cucker and Smale (2001), Mendelson (2002) and Schölkopf and Smola (2002) and are related to the notion of covering number or VC dimension.

The aim of this paper is to point out the main results which allow to extend the results of statistical learning theory in the context of regression on functional data. After a brief introduction to reproducing kernel Hilbert spaces in Section 2, an explicit form for $\hat{f}$ is given by the representer theorem in Section 3. The consistency of such an estimator is studied. In order to compare the performances of such models, a simulation study as well as two applications on real data, stock-exchange data and medical data (knee flexion angle data), are presented in Sections 4 and 5.

2. A brief introduction to reproducing kernel Hilbert spaces

We recall some basic facts about reproducing kernel Hilbert spaces (RKHS) which will be required in our analysis. For more details on RKHS see Aronszajn (1950), Parzen (1961) or Berlinet and Thomas-Agnan (2004).

Let H be a set, for example $H \subseteq \mathbb{R}^d$ or $H \subseteq L^2_{([0,1])}$, and $\mathscr{H}$ be a Hilbert space of functions (or functionals) on H. Denote by $\langle \cdot, \cdot \rangle$ the inner product of $\mathscr{H}$. A bivariate real valued function $\mathbf{K}$ on H is said to be a reproducing kernel for $\mathscr{H}$ if

(R_1) for every $x \in H$, $\mathbf{K}(\cdot, x) \in \mathscr{H}$, and
(R_2) (reproducing property) for every $x \in H$ and $F \in \mathscr{H}$, $F(x) = \langle F, \mathbf{K}(\cdot, x)\rangle$.

If $\mathscr{H}$ admits an reproducing kernel $\mathbf{K}$, then $\mathbf{K}$ has the following properties (Aronszajn, 1950):

(K_1) $\mathbf{K}$ is the unique reproducing kernel for $\mathscr{H}$.
(K_2) $\mathbf{K}$ is symmetric and non-negative definite.
(K_3) Elements of the form $\sum_{i=1}^n a_i \mathbf{K}(t_i, \cdot)$, $n \in \mathbb{N}$, $n \geqslant 1$, $\{a_i \in \mathbb{R}, i = 1, \ldots, n\}$, $\{t_i \in H, i = 1, \ldots, n\}$, are dense in $\mathscr{H}$.

In view of (K_3) if $\mathbf{K}$ is a symmetric and non-negative definite function, one can construct a Hilbert space $\mathscr{H}_\mathbf{K}$ which is the completion of all functions on H of the form $\sum_{i=1}^n a_i \mathbf{K}(t_i, \cdot)$ under the inner product

$$\left\langle \sum_{i=1}^n a_i \mathbf{K}(t_i, \cdot), \sum_{j=1}^m b_j \mathbf{K}(s_j, \cdot) \right\rangle_{\mathscr{H}_\mathbf{K}} = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \mathbf{K}(t_i, s_j). \tag{2}$$

Thus, $\mathscr{H}_\mathbf{K}$ is a RKHS with reproducing kernel $\mathbf{K}$ and we have the well known result (Aronszajn, 1950):

Theorem 1 (*Moore–Aronszajn*). *To every non-negative definite function* $\mathbf{K}$ *on* $H \times H$ *there corresponds a unique RHKS* $\mathscr{H}_\mathbf{K}$ *of real valued functions on H and vice versa.*

Let us assume that on H we have defined an inner product $\langle \cdot, \cdot \rangle_H$. A class of kernels widely used in statistical learning applications is the radial basis function kernels which can be written in the form $\mathbf{K}(t, s) = f(u(s, t))$, where u is a metric on H and f is a function on $\mathbb{R}^+$. An example of such a kernel is the Gaussian kernel defined by $\mathbf{K}(t, s) = \mathrm{e}^{-\|t-s\|_H^2/2\sigma^2}$, $\sigma > 0$. Other useful kernels include inhomogeneous polynomial and spline kernels (Schölkopf and Smola, 2002).

Many optimization problems have explicit solutions when the class of admissible solutions is a RHKS. This is also the case for the regression problem.

3. Regression and RKHS

The regression of Y on $X = \{X_t\}_{t \in T}$ consists to find a function f such that some criterion is optimized with respect to a loss function $\mathscr{C}(Y, X, f(X))$. The square loss function $\mathscr{C}(Y, X, f(X)) = (Y - f(X))^2$ is a common choice when Y is a scalar response and defines the least squares criterion. If Y is a Bernoulli random variable we are interested in estimating $p(x) = P(Y = 1|X = x)$. Then, denoting by $f(x) = \log(p(x)/(1 - p(x)))$, the penalized likelihood criterion for estimating $p(x)$ (Wahba, 2002) is based on the loss function $\mathscr{C}(X, Y, f(X)) = -Yf(X) + \log(1 + \mathrm{e}^{f(X)})$. See Schölkopf and Smola (2002) for other examples of loss function.

A general approach to estimate f is to use a sample $\{(X_i, Y_i)\}_{i=1,\ldots,n}$ of (X, Y) and then minimize the empirical risk

$$R_{\mathrm{emp}}[f] = \frac{1}{n} \sum_{i=1}^n \mathscr{C}(X_i, Y_i, f(X_i)) \tag{3}$$

over some class of functions $\mathscr{H}$. A difficult task is to find a suitable $\mathscr{H}$ such that one can minimize $R_{\mathrm{emp}}[f]$ with respect to $\mathscr{H}$. Moreover, the minimization of $R_{\mathrm{emp}}[f]$ can lead to ill-posed problems as is well known in linear regression on a stochastic process (Cardot et al., 1999). Instead of considering $\mathscr{H}$ compact and R_{emp} continuous in f, which can be cumbersome in practice, a stabilization (regularization) term $\Omega[f]$ is added to R_{emp} to obtain the regularized risk

$$R_{\mathrm{reg}}[f] = R_{\mathrm{emp}}[f] + \lambda\Omega[f]. \tag{4}$$

In the above formula, $\lambda > 0$ is the regularized parameter which specifies the tradeoff between the minimization of $R_{\mathrm{emp}}[f]$ and the smoothness or simplicity of f, enforced by small $\Omega(f)$ (Schölkopf and Smola, 2002). Ω is a functional on $\mathscr{H}$, usually chosen convex, in order to ensure that there exists an unique minimum, provided that $R_{\mathrm{emp}}[f]$ is also convex. When $\mathscr{H}$ is a normed space, a common choice for $\Omega[f]$ is $\Omega[f] = \|f\|^2_{\mathscr{H}}$.

3.1. Representation of the estimate for the regression problem in RKHS

In the following we focus on the class of admissible solutions for the regression problem which is a RKHS associated to the kernel $\mathbf{K}$ defined on $H \times H$, denoted by $\mathscr{H}_{\mathbf{K}}$.

The explicit form of a minimizer of $R_{\mathrm{reg}}[f]$ is given by the representer theorem of Kimeldorf and Wahba (1971). Schölkopf and Smola (2002) give a simple proof of a more general version of the theorem.

Theorem 2 (*Representer theorem*). *Let $\{(x_i, y_i)\}_{i=1}^n$, $x_i \in H$, $y_i \in \mathbb{R}$, $n > 0$, $n \in \mathbb{N}$, $\lambda > 0$ and $\mathscr{C} : H \times \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$ be a convex loss function of third argument. Then, the solution to the problem: find $\hat{f} \in \mathscr{H}_{\mathbf{K}}$ to minimize*

$$\frac{1}{n}\sum_{i=1}^{n} \mathscr{C}(x_i, y_i, \hat{f}(x_i)) + \lambda\|\hat{f}\|^2_{\mathscr{H}_{\mathbf{K}}}$$

exists, is unique and admits a representation of the form

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i \mathbf{K}(x, x_i) \quad \forall x \in H, \tag{5}$$

where $\alpha_i \in \mathbb{R}$, $i = 1, \dots, n$.

Generally, $\mathscr{H}_{\mathbf{K}}$ is an infinite dimensional space containing linear combinations of kernels centered on arbitrary points of H. The significance of representer theorem is that it states that the minimizer of the regularized risk lies in the span on the n kernels centered on the sample points.

Remark 3. The original form of this theorem is due to Kimeldorf and Wahba (1971) for the particular case of squared loss, $\mathscr{C}(x_i, y_i, f(x_i)) = (y_i - f(x_i))^2$. In this case, it is only necessary to solve the linear system

$$(\lambda n\mathbf{I} + [\mathbf{K}])\alpha = \mathbf{y}, \tag{6}$$

where $\alpha = (\alpha_1, \dots, \alpha_n)'$, $\mathbf{y} = (y_1, \dots, y_n)'$, $\mathbf{I}$ is the identity matrix of order n and $[\mathbf{K}] \in \mathscr{M}_n(\mathbb{R})$ is the Gram matrix associated to kernel $\mathbf{K}$, $[\mathbf{K}]_{i,j} = \mathbf{K}(x_i, x_j)$, $1 \leqslant i, j \leqslant n$.

For other types of convex loss functions, the minimization has to be done using an iterative descent technique, such as the gradient methods. It is the case for the loss function when Y is a Bernoulli random variable.

Remark 4. The choice of the regularization parameter is important and a general well-accepted method for choosing λ, and also for other "tuning" parameters, is the cross validation. Properties of this method are established for particular cases of the loss function. Thus, the generalized cross validation (GCV) is developed by Craven and Wahba (1979) and Golub et al. (1979) for the square loss function. For Bernoulli data a generalized approximate cross validation (GACV) is proposed in Xiang and Wahba (1996). For large n, computational aspects for CGV and GACV based on randomization techniques are given in Girard (1989) and Wahba et al. (1999).

3.2. Uniform convergence and consistency

Let $f_{\mathrm{opt}} \in \mathscr{H}_{\mathbf{K}}$ be the minimizer of the expected risk $R[f] = \mathbb{E}(\mathscr{C}(X, Y, f(X)))$ over $\mathscr{H}_{\mathbf{K}}$ and $\hat{f}^n$ the minimizer of R_{reg} based on the sample $\{(X_i, Y_i)\}_{i=1,\dots,n}$ of (X, Y).

We are interested in the consistency of the minimization of R_{reg} and in particular in establishing conditions under which $R_{\text{emp}}[\hat{f}^n]$ converges in probability to $R[f^{\text{opt}}]$ as $n \to \infty$. For example, a necessary and sufficient condition is the uniform convergence in probability over $\mathscr{H}_{\mathbf{K}}$ of $R_{\text{emp}}[f]$ to $R[f]$, as stated by the Vapnik and Chervonenkis theorem (Vapnik, 1998). When H is of finite dimension and under conditions such as the compactness of $\mathscr{H}$ or infinite differentiability of $\mathbf{K}$, uniform bounds are given for $P(\sup_{f\in\mathscr{H}_{\mathbf{K}}}|R[f] - R_{\text{emp}}[f]| > \varepsilon)$. These bounds depend on quantities such as VC dimension (Schölkopf and Smola, 2002; Mendelson, 2002) or covering number (Cucker and Smale, 2001). The uniform convergence is then guaranteed when these quantities are finite, condition which is quite difficult to check in practice.

For infinite dimensional H, and in particular for functional data (e.g., $H = L_2([0, 1])$) the VC dimension and the covering number are generally infinite.

Based on the stability concept of the loss function, Schölkopf and Smola (2002, p. 363), establish bounds for $P(|R_{\text{emp}}[\hat{f}^n] - R[\hat{f}^n]| > \varepsilon)$ under the following conditions:

Theorem 5. *Let $\mathscr{C}(X, Y, f(X))$ be a convex loss function and $\hat{f}^n$ the minimizer over $\mathscr{H}_{\mathbf{K}}$ of the regularized risk*

$$\frac{1}{n}\sum_{i=1}^{n}\mathscr{C}(x_i, y_i, f(x_i)) + \lambda\|f\|^2_{\mathscr{H}_{\mathbf{K}}}$$

based on the training sample $\{(y_1, x_1), \ldots, (y_n, x_n)\}$ of (X, Y).

Assuming that

1. *$\mathscr{C}$ is differentiable and has the Lipschitz property in the third argument,*
2. *$\exists M \in \mathbb{R}$ such that $\hat{f}^n(x) \leqslant M$, $\forall x \in H$, $\forall n \geqslant 1$,*
3. *$\exists \kappa$ such that $\sqrt{\mathbf{K}(x, x)} \leqslant \kappa$, $\forall x \in H$,*

then, $\forall n \geqslant 1$

$$P(|R_{\text{emp}}[\hat{f}^n] - R[\hat{f}^n]| > \beta + \varepsilon) \leqslant 2\exp\left(-\frac{n\varepsilon^2}{2(n\beta + M)^2}\right), \tag{7}$$

where $\beta = C^2\kappa^2/n\lambda$ and C is the constant of the Lipschitz property.

Observe that these bounds have the same exponential rate of convergence as the Hoeffding's bounds.

Let us consider the cases of the square loss function, $\mathscr{C}(x, y, f(x)) = (y - f(x))^2$ and of the logistic loss function, $\mathscr{C}(x, y, f(x)) = -yf(x) + \log(1 + \mathrm{e}^{f(x)})$. Observe that condition 1 of Theorem 5 is not verified for the square loss and that condition 2 appears quite artificial. We replace these two conditions by a much simpler one, that is, we will assume that the random variable $|Y|$ is bounded from above. The following proposition states that the result of Theorem 5 holds true for both loss functions.

Preposition 6 (*Square and logistic loss*). *Let $\mathscr{C}$ be the square or logistic loss function and suppose that $\exists M_Y > 0$ such that $|Y| \leqslant M_Y$ a.s. Suppose further that $\exists\kappa$ such that $\forall x \in H$, $\sqrt{\mathbf{K}(x, x)} \leqslant \kappa$.*

Then, for the square loss we have

$$P(|R_{\text{emp}}[\hat{f}^n] - R[\hat{f}^n]| > \beta + \varepsilon) \leqslant 2\exp\left(-\frac{n\varepsilon^2}{2(n\beta + M_Y\frac{\kappa}{\sqrt{\lambda}})^2}\right), \tag{8}$$

where $\beta = 8\kappa^2 M_Y^2(1 + 1/\sqrt{\lambda})^2/n\lambda$, and for the logistic loss,

$$P\left(|R_{\text{emp}}[\hat{f}^n] - R[\hat{f}^n]| > \varepsilon + \frac{\kappa^2}{n\lambda}\right) \leqslant 2\exp\left(-\frac{2n\varepsilon^2\lambda^2}{9\kappa^4}\right). \tag{9}$$

Proof. *Square loss*: α is solution of the linear system (6). Since λn is a lower bound for the set of eigenvalues of $\lambda n I + [\mathbf{K}]$, it follows

$$\|\alpha\|_{\mathbb{R}^n} \leqslant \frac{1}{\lambda n}\|\mathbf{y}\|_{\mathbb{R}^n} \leqslant \frac{1}{\lambda\sqrt{n}} M_Y. \tag{10}$$

Therefore, for each $x \in H$,

$$|\hat{f}^n(x)| = |\langle \mathbf{K}(x,\cdot), \hat{f}^n\rangle| \leqslant \kappa \|\hat{f}^n\|_{\mathscr{H}_{\mathbf{K}}} = \kappa\sqrt{\alpha'[\mathbf{K}]\alpha}$$

and

$$\alpha'[\mathbf{K}]\alpha = \alpha'(\mathbf{y} - \lambda n\alpha) \leqslant \|\alpha\|_{\mathbb{R}^n}\|\mathbf{y}\|_{\mathbb{R}^n} \leqslant \frac{M_Y^2}{\lambda}. \tag{11}$$

Then, $|\hat{f}^n(x)| \leqslant \kappa(M_Y/\sqrt{\lambda})$. Taking $M = \kappa(M_Y/\sqrt{\lambda})$ and $2M_Y(1 + 1/\lambda)$ as upper bound for the derivative of $\mathscr{C}$ with respect to f in the proof of Theorem 5, one obtains, by the same argument,

$$\beta = \frac{8\kappa^2 M_Y^2 (1 + 1/\sqrt{\lambda})^2}{n\lambda}. \tag{12}$$

Logistic loss: Let $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)'$ be as in the representer theorem. For each $i = 1, \ldots, n$,

$$\frac{\partial R_{\text{reg}}}{\partial \alpha_i} = -\frac{1}{n}\sum_{j=1}^{n} \mathbf{K}(x_i, x_j) y_i + \frac{1}{n}\alpha_i \sum_{j=1}^{n} \frac{\mathrm{e}^{\hat{f}^n(x_j)}}{1 + \mathrm{e}^{\hat{f}^n(x_j)}} + 2\lambda \sum_{j=1}^{n} \mathbf{K}(x_i, x_j)\alpha_j. \tag{13}$$

Let

$$\gamma = \sum_{j=1}^{n} \frac{\mathrm{e}^{\hat{f}^n(x_j)}}{1 + \mathrm{e}^{\hat{f}^n(x_j)}}. \tag{14}$$

Then, $\partial R_{\text{reg}}/\partial\alpha = 0$ implies

$$(2\lambda n[\mathbf{K}] + \gamma I)\alpha = [\mathbf{K}]y. \tag{15}$$

and it follows that

$$\|\alpha\|_{\mathbb{R}^n} = \frac{\|[\mathbf{K}]y\|_{\mathbb{R}^n}}{\|2\lambda n[\mathbf{K}] + \gamma I\|} \leqslant \frac{\|y\|_{\mathbb{R}^n}}{2\lambda n} \leqslant \frac{1}{2\lambda\sqrt{n}},$$

and

$$|\hat{f}^n(x)| \leqslant \frac{\kappa^2}{2\lambda} \quad \forall x \in H.$$

Put $M = \kappa^2/2\lambda$, $C = 1$ and $\beta = \kappa^2/n\lambda$ in Theorem 5 to complete the proof. □

Remark 7. Using alternatively (15) and (5)+(14), we provide an algorithm for estimating α which requires only the resolution of linear systems. Starting with an initial values for $\alpha^{(0)}$, at the step $i \geqslant 0$, $\alpha^{(i)}$ allows to compute by (5) $\hat{f}^{n^{(i)}}(x_j)$, $j = 1, \ldots, n$, and by (14) $\gamma^{(i)}$. The step i is completed by computing $\alpha^{(i+1)}$ by (15) using $\gamma^{(i)}$.

Corollary 8. *Assuming that the conditions of Theorem 5 or of Proposition 6 hold, then*

$$\begin{aligned} &R_{\text{emp}}[f^{\text{opt}}] - R_{\text{emp}}[\hat{f}|^n] \xrightarrow{\mathrm{P}} 0,\\ &R[\hat{f}^n] - R[f^{\text{opt}}] \xrightarrow{\mathrm{P}} 0. \end{aligned} \tag{16}$$

Proof. Observe that

$$R_{\mathrm{emp}}[f^{\mathrm{opt}}] - R_{\mathrm{emp}}[\hat{f}^n] \geqslant 0 \quad \text{and} \quad R[\hat{f}^n] - R[f^{\mathrm{opt}}] \geqslant 0. \tag{17}$$

Theorem 5 and Proposition 6 state that

$$R_{\mathrm{emp}}[\hat{f}^n] - R[\hat{f}^n] \xrightarrow{\mathrm{P}} 0. \tag{18}$$

On the other hand, due to the law of large numbers ($f^{\mathrm{opt}} \in \mathscr{H}_{\mathbf{K}}$ is independent of the training sample) we have

$$R_{\mathrm{emp}}[f^{\mathrm{opt}}] - R[f^{\mathrm{opt}}] \xrightarrow{\mathrm{P}} 0. \tag{19}$$

From (18) and (19) it follows that

$$R_{\mathrm{emp}}[f^{\mathrm{opt}}] - R_{\mathrm{emp}}[\hat{f}^n] + R[\hat{f}^n] - R[f^{\mathrm{opt}}] \xrightarrow{\mathrm{P}} 0,$$

and using (17) the proof is complete. □

4. Simulation study

In this study we compare regression models based on RKHS methods and linear approaches for which we developed a software using the C language implementation.

For a scalar response, we compare the RKHS approach with linear models given by

- FPCR, a functional linear regression model using the principal components of the predictor r.v.'s set. There are two ways to introduce principal components in the model: ordered by their explained variance rate or by a stepwise method based on the correlation coefficient with the response variable (Aguilera et al., 1997). For our simulation we used the first approach, the rate of variance explained by the few first principal components being very large (more than 95%).
- FPLS, a functional linear regression model based on partial least square approach (Preda and Saporta, 2005). For our simulations we use cross validation procedure for the selection of the number of PLS components.

For categorical response, the RKHS regression models are compared with the following models:

- LDA_FPCR, linear discriminant analysis on principal components,
- LOG_FPCR, logistic regression on principal components.
  For both models, the principal components are introduced in the model using a stepwise procedure based on the ratio of within-class variance and total variance of the principal component. For functional logistic regression, other selection methods based on conditional likelihood ratio tests are proposed in Escabias et al. (2004).

For RKHS models, we use the following two types of kernels which are widely used in statistical learning:

- Gaussian (σ) : $\mathbf{K}(x_1, x_2) = \mathrm{e}^{-\|x_1 - x_2\|^2_{L_2[0,T]}/2\sigma^2}$, $\sigma > 0$.
- Inhomogeneous polynomial (IP(c, d)): $\mathbf{K}(x_1, x_2) = (c + \langle x_1, x_2 \rangle_{L_2[0,T]})^d$, $c \in \mathbb{R}$ and $d \in \mathbb{N}^*$.

For all our simulations, the choice of the regularization parameter λ is given by leave-one-out cross validation. For instance, we have not yet implemented the cross validation for the optimal values of the kernel parameters (σ, c and d). Several models are tested.

4.1. Scalar response: regression

Let $X = \{X_t\}_{t \in [0,1]}$ be the standard brownian motion and Y the zero-mean random variable defined by

$$Y = \int_0^1 \left(t - \frac{2}{3}\right) X_t^2 \, \mathrm{d}t + \varepsilon,$$

where ε is a gaussian noise, $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon)$.

Table 1
$\overline{\text{MSE}} \times 10^{-3}$ over 1000 test samples of different sizes

	$n = 50$	$n = 100$	$n = 500$
F-PCR	3.721	3.631	3.628
F-PLS	3.321	3.242	3.227
Gaussian (3)	2.112	2.214	2.232
Gaussian (4)	1.616	1.221	1.191
Gaussian (5)	2.412	2.427	2.183
IP (1, 2)	1.761	1.242	1.255
IP (1, 3)	1.558	1.253	1.238

Our simulation is based on the following conditions:

- the trajectories of X are discretized in 100 equidistant points.
- values of Y are computed using integration by trapezoidal interpolation.
- the variability of noise is such that $\sigma_\varepsilon^2/\mathbb{E}(Y^2) = 0.1$.
- the training and test sample sizes are identical, $n = 50$, 100 and 500.
- the number of simulations for each simple size is $I = 1000$.

The criterion used to compare models is the mean of square errors (MSE) averaged over the I test samples:

$$\overline{\text{MSE}} = \frac{1}{I}\sum_{i=1}^{I} \text{MSE}(i),$$

where $\text{MSE}(i) = (1/n)\sum_{j=1}^{n}(Y_j - \hat{Y}_j)^2$ is the mean of square errors on the i simulated test sample, $i \in \{1, \ldots, I\}$.

Table 1 presents the results ($\overline{\text{MSE}} \times 10^{-3}$) given by linear and non-linear models for different test sample sizes.

The first four principal components of the predictor set, $X = \{X_t\}_{t\in[0,1]}$, explain of about 95% of the total variance of X and are used by the FPCR model. The two linear models (FPCR and FPLS) give comparable results whereas the RKHS regression models improve substantially the prediction, especially the Gaussian kernel approach (Gaussian (4)). As expected, the inhomogeneous polynomial kernels give also good results.

4.2. Categorical response: curve discrimination

The simulated data we consider correspond to a binary response for which the predictor has the following form:

$$\text{Class } \{Y = 0\} : X(t) = Uh_1(t) + (1 - U)h_2(t) + \varepsilon(t),$$

$$\text{Class } \{Y = 1\} : X(t) = Uh_1(t) + (1 - U)h_3(t) + \varepsilon(t),$$

where U is a r.v. uniformly distributed on [0, 1], $\varepsilon(t)$ are uncorrelated standard normal r.v.'s and $h_1(t) = \max\{6 - |t - 11|, 0\}$, $h_2(t) = h_1(t - 4)$ and $h_3(t) = h_1(t + 4)$. As in Ferraty and Vieu (2003), we consider the observed predictor is a discretized curve with 101 equidistant points $\{t = 1, 1.2, 1.4, \ldots, 21\}$. Fig. 2 displays a sample of 100 simulated curves for each class.

We consider 100 simulated samples of size 1000, with 500 observations in each class. Each sample is randomly divided into a training sample of size 800 and a test sample of size 200, each class having the same number of observations in both samples.

The principal components of the process $\{X_t\}_{t\in[1,21]}$, as well as the observed kernel function $\mathbf{K}(x_i, x_j)$ are computed using linear interpolation and the trapezoidal integration method.

Table 2 presents the results representing error rates averaged over the 100 test samples given by different models.

All models give satisfactory results with respect to the classification error rate criterion. The RKHS regression models are well adapted to the functional discrimination problem and the Gaussian kernel seems to be very competitive.

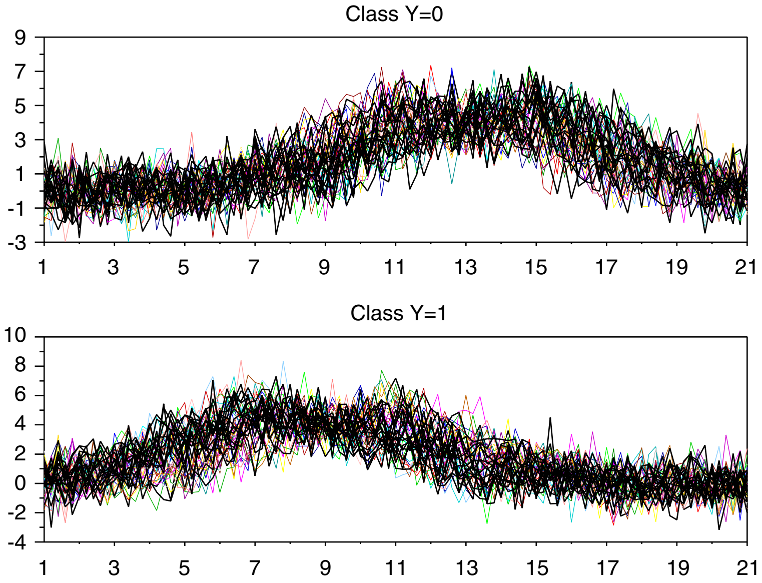

Fig. 2. Sample of 100 curves for each class.

Table 2
Error rate averaged over 100 test samples

Model	Test error rate
LDA_FPCR	0.0351
LOG_FPCR	0.0344
Gaussian (5)	0.0251
Gaussian (6)	0.0244
Gaussian (7)	0.0287
IP (1, 2)	0.0312
IP (1, 3)	0.0282

5. Applications

We present two applications of the RKHS based regression models corresponding to scalar and categorical response Y. The first application concerns stock-exchange data provided by the Groupe SBF (Bourse de Paris). The aim is to give an approximation of the way that some share has in a certain lapse of time. The second application deals with data provided by the Department of Movement Disorders, Lille University Medical Center (France). This data is described by a set of curves representing the knee flexion angle evolution over one complete gait cycle and characterizes patients from two classes of age (Duhamel et al., 2004). We are interested in predicting the class of age from the knee curve.

In both applications we assume that data represent sample paths of a stochastic process $\{X_t\}_{t\in T}$ of second order and L_2 continuous.

5.1. Stock-exchange data

We have 84 shares quoted at the Paris stock exchange for which we know the whole behavior of the growth index during one hour (between 10^{00} and 11^{00}). Note that a share is likely to change every second (see Fig. 3). We also know the evolution of the growth index of a new share quoted by 85 (Fig. 1a), between 10^{00} and 10^{55}. The aim is to approximate the way that share has between 10^{55} and 11^{00}, more precisely, we are interested in the average estimation of the growth index for each minute of the last five. The approximations obtained will then match the average level of

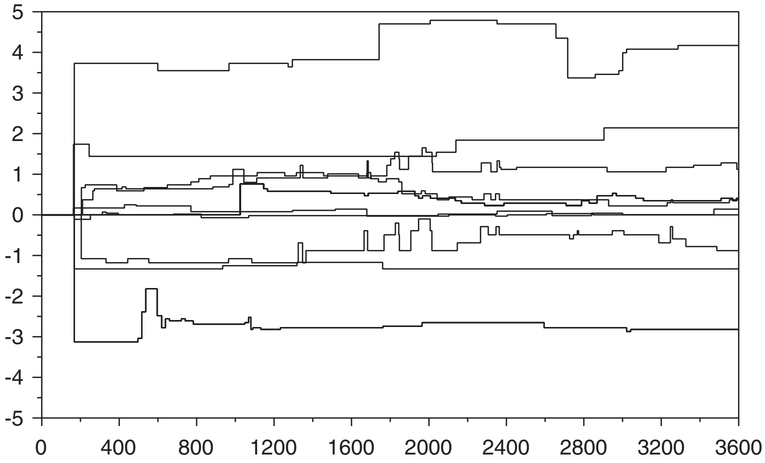

Fig. 3. Subset of shares. Growth index evolution during 1 h.

Table 3
Approximations for share 85

	$\hat{m}_{56}(85)$	$\hat{m}_{57}(85)$	$\hat{m}_{58}(85)$	$\hat{m}_{59}(85)$	$\hat{m}_{60}(85)$	*SSE*
Observed	0.700	0.678	0.659	0.516	−0.233	—
FPLS (2)	0.312	0.355	0.377	0.456	0.534	0.926
FPCR (3)	0.613	0.638	0.669	0.825	0.963	1.535
Gaussian (1)	0.620	0.685	0.665	0.454	0.062	0.097
λ	0.17	0.16	0.20	0.09	0.15	
IP (1, 2)	0.535	0.583	0.497	0.442	0.025	0.134
λ	0.085	0.12	0.18	0.11	0.12	
IP (1, 3)	0.624	0.660	0.545	0.332	0.004	0.109
λ	0.04	0.08	0.12	0.05	0.14	

the growth index of share 85 considered on each interval (expressed in seconds) $[60 \cdot (i-1), 60 \cdot i)$, $i = 56, \ldots, 60$. For each $i \in \{56, \ldots, 60\}$ let us denote by $m_i = \frac{1}{60} \int_{60 \times (i-1)}^{60 \times i} X_t \, \mathrm{d}t$. Observe that, since the paths of $\{X_t\}_{t \in [0,3600]}$ are completely known and constant on sub-intervals, the integral in the definition of m_i is reduced to a finite sum.

The results given by the RKHS based models are compared (Table 3) with those obtained by Preda and Saporta (2005) on the same data using linear models such as principal component regression (FPCR(k)) and partial least squares regreassion (FPLS(k)), where k is the number of components in the model. Notice that the three first principal components explain 98.5% of total variance. For the PLS model k is determined by cross validation. The comparison criterion is the sum of square errors,

$$\mathrm{SSE} = \sum_{i=56}^{60} (\hat{m}_i - m_i)^2.$$

Except PLS, each regression model was independently fitted for each response variable m_i.

5.2. Gait data

Two groups of 30 subjects were studied: 30 young students (mean age 27 years and standard deviation 4 years) and 30 healthy senior citizens (mean age 64 years and standard deviation 6 years). For each subject the observed data represent the flexion angle for the right knee measured during one complete gait cycle. Each curve represents a gait cycle and is given by a set $\{(x_{t_i}, t_i)\}_{i=1,\ldots,50}$ of 50 values corresponding to an equidistant discretization of the cycle. Since data

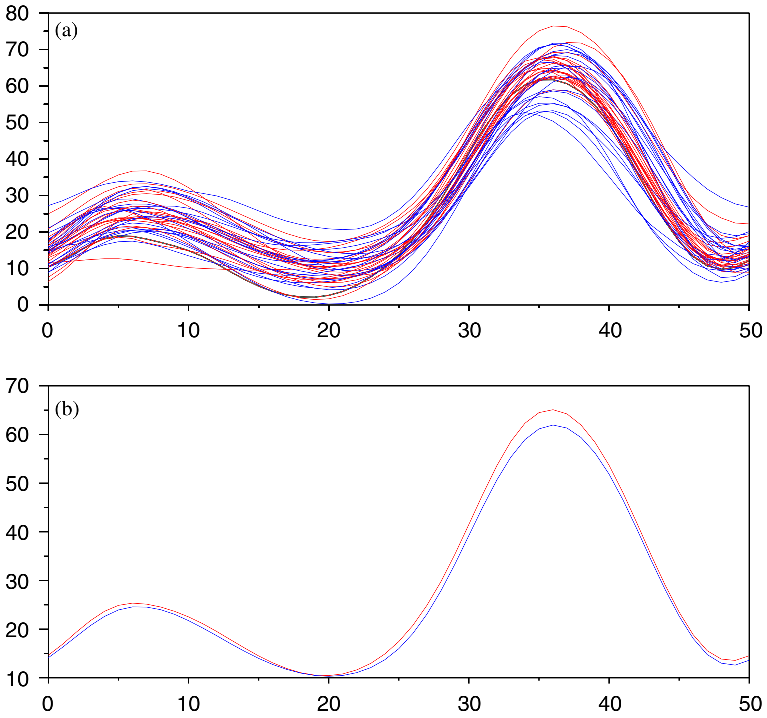

Fig. 4. Knee flexion angular data. (a) A sample of 40 cubic spline interpolated curves of the right knee angular rotation (20 for young subjects—in red, and 20 for senior subjects—in blue), (b) mean estimation of angular rotation of the right knee during a complete cycle for each group.

Table 4
Area under the ROC curve. Sample test estimation

Model	LDA_FPCR (4)	Log_FPCR (4)	Gaussian (1) $\lambda = 0.12$	IP (1, 2) $\lambda = 0.18$	IP (1, 3) $\lambda = 0.22$
Area	0.780	0.790	0.760	0.780	0.750

can be viewed as a sample of a stochastic process with continuous paths, a cubic spline interpolation is performed for each curve. For more details on the approximation of principal component analysis of cubic spline interpolation sample curves, see Aguilera et al. (1996).

Data is divided into two samples, a learning sample of 40 subjects (Fig. 4a) and a test sample of 20 patients. Each sample contains the same number of young and senior subjects.

We will use RKHS based regression models to estimate the probability $p(x(\omega,\cdot)) = P(\omega \in \text{young}|X = x(\omega,\cdot))$ using the logistic loss function given by $\mathscr{C}(X, Y, f(X)) = -Yf(X) + \log(1 + \mathrm{e}^{f(X)})$ where $f(x) = \log(p(x)/(1-p(x)))$ (Wahba, 2002). In order to estimate the coefficients $\{\alpha_i\}_{i=1,\dots,40}$, we use the iterative algorithm described in Remark 7, starting with $\alpha^{(0)} = \mathbf{1}$. The results are compared with the logistic regression (Log_FPCR(k)) and the linear discriminant analysis (LDA_FPCR(k)) using the first k principal components of the process $X = \{X_t\}_{t\in[0,50]}$ (the four first principal components explain 94.64% of the total inertia of X). The comparison criterion is the area under the receiver operating characteristic (ROC) curve (Table 4) estimated using the sample test.

The results obtained show that logistic and linear discriminant approaches based on principal components as well as the inhomogeneous polynomial RKHS regression model IP(1, 2) give the best performances. However, the smallest classification error rate for the Log_FPCR (4) model is about of 30% which means that the two groups are not well predicted from the knee angular curves.

6. Conclusion

Non-parametrical regression models based on reproducing kernel Hilbert spaces methods are adapted to functional data in a natural way. The main results established in finite dimension are preserved in the context of functional data and simpler conditions for consistency of obtained estimators are established for square (scalar response) and logistic (categorical response) loss function. A simulation study, as well as applications on real data, show the performances of this approach which generally improves those of linear models.

Acknowledgments

We are grateful to Groupe SBF (Bourse de Paris) and Department of Movement Disorders, Roger Salengro Hospital, Lille University Medical Center (France) for providing us with data for testing our methods.

References

Aguilera, A.M., Gutiérrez, R., Ocãna, F., Valderama, M.J., 1996. Approximation of estimators in the PCA of stochastic process using B-splines. Commun. Statist. Comput. Simulation 25 (3), 671–690.

Aguilera, A.M., Ocãna, F., Valderama, M.J., 1997. An approximated principal component prediction model for continuous-time stochastic process. Appl. Stochastic Models Data Anal. 13, 61–72.

Aguilera, A.M., Ocana, F., Valderama, M.J., 1999. Stochastic modeling for evolution of stock-prices by means of principal component analysis. Appl. Stochastic Models Business Ind. 15 (4), 227–234.

Aronszajn, N., 1950. Theory of reproducing kernels. Amer. Math. Soc. Trans. 63, 337–404.

Berlinet, A., Thomas-Agnan, C., 2004. Reproducing Kernel Hilbert spaces in Probability and Statistics. Kluwer Academic Publishers, Dordrecht.

Cardot, H., Sarda, P., 2005. Estimation in generalized linear models for functional data via penalized likelihood. J. Multivariate Anal. 92, 24–41.

Cardot, H., Ferraty, F., Sarda, P., 1999. Functional linear model. Statist. Probab. Lett. 45, 11–22.

Craven, P., Wahba, G., 1979. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. Numer. Math. 31, 377–403.

Cucker, F., Smale, S., 2001. On the mathematical foundations of learning. Bull. Amer. Math. Soc. 39 (1), 1–49.

Duhamel, A., Bourriez, J.L., Devos, P., Krystkowiak, P., Destee, A., Derambure, P., Defebvre, L., 2004. Statistical tools for clinical gait analysis. Gait Posture 20, 204–212.

Escabias, M., Aguilera, A.M., Valderama, M.J., 2004. Principal component estimation of functional logistic regression: discussion of two different approaches. J. Nonparametric Statist. 16 (3–4), 365–384.

Escabias, M., Aguilera, A.M., Valderama, M.J., 2005. Modelling environmental data by functional principal component logistic regression. Environmetrics 16 (1), 95–107.

Ferraty, F., Vieu, P., 2003. Curves discrimination: a nonparametric approach. Comput. Statist. Data Anal. 44, 161–173.

Ferraty, F., Vieu, P., 2004. Nonparametric models for functional data with application in regression, time series prediction and curve discrimination. J. Nonparametric Statist. 16 (1–2), 111–125.

Girard, D., 1989. A fast Monte-Carlo cross-validation procedure for large least squares problems with noisy data. Numer. Math. 56, 1–23.

Golub, G.H., Heath, M., Wahba, G., 1979. Generalized cross validation as a method for choosing a good ridge parameter. Technometrics 21, 215–224.

James, G.M., 2002. Generalized linear models with functional predictors. J. Roy. Statist. Soc. Ser. B 64 (3), 411–432.

Kimeldorf, G.S., Wahba, G., 1971. Some results on Tchebycheffian spline functions. J. Math. Anal. Appl. 33, 82–95.

Mendelson, S., 2002. Learnability in Hilbert spaces with reproducing Kernels. J. Complexity 18, 152–170.

Parzen, E., 1961. An approach to time series anlysis. Ann. Math. Statist. 32, 951–989.

Preda, C., Saporta, G., 2005. PLS regression on a sochastic process. Comput. Statist. Data Anal. 48 (1), 149–158.

Ramsay, J.O., Silverman, B.W., 1997. Functional data analysis. Springer Series in Statistics. Springer, New York.

Ramsay, J.O., Silverman, B.W., 2002. Applied Functional Data Analysis: Methods and Case Studies. Springer, Berlin.

Ratcliffe, S.J., Leader, L.R., Heller, G.Z., 2002a. Functional data analysis with application to periodically stimulated foetal heart rate data. I: functional regression. Statist. Med. 21, 1103–1114.

Ratcliffe, S.J., Leader, L.R., Heller, G.Z., 2002b. Functional data analysis with application to periodically stimulated foetal heart rate data II: functional logistic regression. Statist. Med. 21, 1115–1127.

Schölkopf, B., Smola, A., 2002. Learning with Kernes. MIT Press, Cambridge.

Vapnik, V., 1998. Statistical Learning Theory. Wiley, NY.

Wahba, G., 1990. Spline models for observational data. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, vol. 59.

Wahba, G., 2002. Soft and Hard Classification by Reproducing Kernel Hilbert Space Methods. Department of Statistics, University of Wisconsin. Technical Report 1067.

Wahba, G., Lin, X., Gao, F., Xiang, D., Klein, R., Klein, B., 1999. The biais-variance tradeoff and the randomized GACV. In: Kearns, M., Solla, S., Cohn, D. (Eds.), Advances in Information Processing Systems, vol. 11. MIT Press, Cambridge, pp. 620–626.

Xiang, D., Wahba, G., 1996. A generalized approximate cross validation for smoothing splines with non-Gaussian data. Statist. Sinica 6, 675–692.