Computational Statistics and Data Analysis 54 (2010) 429–437

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

# Distance-based local linear regression for functional predictors

Eva Boj[a], Pedro Delicado[b,*], Josep Fortiana[c]

[a] Departament de Matemàtica Econòmica, Financera i Actuarial, Universitat de Barcelona, Barcelona, Spain
[b] Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, Barcelona, Spain
[c] Departament de Probabilitat, Lògica i Estadística, Universitat de Barcelona, Barcelona, Spain

ARTICLE INFO

*Article history:*
Received 26 February 2009
Received in revised form 7 September 2009
Accepted 7 September 2009
Available online 12 September 2009

ABSTRACT

The problem of nonparametrically predicting a scalar response variable from a functional predictor is considered. A sample of pairs (functional predictor and response) is observed. When predicting the response for a new functional predictor value, a semi-metric is used to compute the distances between the new and the previously observed functional predictors. Then each pair in the original sample is weighted according to a decreasing function of these distances. A Weighted (Linear) Distance-Based Regression is fitted, where the weights are as above and the distances are given by a possibly different semi-metric. This approach can be extended to nonparametric predictions from other kinds of explanatory variables (e.g., data of mixed type) in a natural way.



## 1. Introduction

Observing and saving complete functions as results of random experiments are nowadays possible by the development of real-time measurement instruments and data storage resources. For instance, continuous-time clinical monitoring is a common practice today. Functional Data Analysis (FDA) deals with the statistical description and modelization of samples of random functions. Functional versions for a wide range of statistical tools (ranging from exploratory and descriptive data analysis to linear models to multivariate techniques) have been recently developed. See Ramsay and Silverman (2005) for a general perspective on FDA and Ferraty and Vieu (2006) for a nonparametric approach. Special monographic issues recently dedicated to this topic by several journals (Davidian et al., 2004; González-Manteiga and Vieu, 2007; Valderrama, 2007) bear witness to the interest on this topic in the Statistics community. Other recent papers on FDA are Park et al. (2009), Ferraty and Vieu (2009), Aguilera et al. (2008) and Zheng (2008).

In this paper we consider the problem of predicting a scalar response using a functional predictor. Let us give an example: *Spectrometric Data* are described in Chapter 2 of Ferraty and Vieu (2006). This dataset includes information about 215 samples of chopped meat. For each of them, the function $\chi$, relating absorbance versus wavelength, has been recorded for 100 values of wavelength in the range 850–1050 nm. An additional response variable is observed: $y$, the sample fat content obtained by analytical chemical processing. Given that obtaining a spectrometric curve is less expensive than determining the fat content by chemical analysis, it is important to predict the fat content $y$ from the spectrometric curve $\chi$. In Section 4 the Spectrometric Data are used to illustrate the methods we propose in this work, jointly with another example on air pollution.

In technical terms, the problem is stated as follows: Let $(\boldsymbol{\chi}, Y)$ be a random element where the first component $\boldsymbol{\chi}$ is a random element of a functional space (typically a real function $\boldsymbol{\chi}$ from $[a, b] \subseteq \mathbb{R}$ to $\mathbb{R}$) and $Y$ is a real random variable.

* Corresponding address: Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, Edifici C5-214. C/Jordi Girona 1-3, 08034, Barcelona, Spain. Tel.: +34 934015698; fax: +34 934015855.
*E-mail address:* pedro.delicado@upc.edu (P. Delicado).

0167-9473/$ – see front matter 
doi:10.1016/j.csda.2009.09.010

We consider the problem of predicting the scalar response variable $y$ from the functional predictor $\chi$. We assume that we are given $n$ i.i.d. observations $(\chi_i, y_i)$, $i = 1, \ldots, n$, from $(\boldsymbol{\chi}, Y)$ as a training set. Let $m(\chi) = E(Y|\boldsymbol{\chi} = \chi)$ be the regression function. Then an estimate of $m(\chi)$ is a good prediction of $y$. The linear functional regression model, considered in Ramsay and Silverman (2005), assumes that

$$m(\chi) = \alpha + \int_a^b \chi(t)\beta(t)\mathrm{d}t, \quad \text{and} \quad y_i = m(\chi_i) + \varepsilon_i,$$

$\varepsilon_i$ having zero expectation. The parameter $\beta$ is a function and $\alpha \in \mathbb{R}$. These authors propose to estimate $\beta$ and $\alpha$ by penalized least squares:

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \left( y_i - \alpha - \int_a^b \chi_i(t)\beta(t)\mathrm{d}t \right)^2 + \lambda \int_a^b (L(\beta)(t))^2 \mathrm{d}t,$$

where $L(\beta)$ is a linear differential operator giving a penalty to avoid too much rough $\beta$ functions and $\lambda > 0$ acts as a smoothing parameter.

Ferraty and Vieu (2006) consider this linear regression as a parametric model because only a finite number of functional elements is required to describe it (in this case only one is needed: $\beta$). They consider a nonparametric functional regression model where few regularity assumptions are made on the regression function $m(\chi)$. They propose the following kernel estimator for $m(\chi)$:

$$\hat{m}_K(\chi) = \frac{\sum_{i=1}^{n} K(\delta(\chi, \chi_i)/h) y_i}{\sum_{i=1}^{n} K(\delta(\chi, \chi_i)/h)} = \sum_{i=1}^{n} w_i(\chi) y_i,$$

where $w_i(\chi) = K(\delta(\chi, \chi_i)/h) / \sum_{j=1}^{n} K(\delta(\chi, \chi_j)/h)$, $K$ is a kernel function with support $[0, 1]$, the bandwidth $h$ is the smoothing parameter (depending on $n$), and $\delta(\cdot, \cdot)$ is a semi-metric ($\delta(\chi, \chi) = 0, \delta(\chi, \gamma) = \delta(\gamma, \chi), \delta(\chi, \gamma) \le \delta(\chi, \psi) + \delta(\psi, \gamma)$) in the functional space $\mathcal{F} = \{\chi : [a, b] \to \mathbb{R}\}$ to which the data $\chi_i$ belong. Examples of semi-metrics in $\mathcal{F}$ are $L_2$ distances between derivatives,

$$d_r^{deriv}(\chi, \gamma) = \left( \int_a^b \left( \chi^{(r)}(t) - \gamma^{(r)}(t) \right)^2 \mathrm{d}t \right)^{1/2};$$

and the $L_2$ distance in the space of the first $q$ functional principal components of the functional dataset $\chi_i$, $i = 1, \ldots, n$: $d_q^{PCA}(\chi, \gamma) = (\sum_{k=1}^{q} (\psi_k^{\chi} - \psi_k^{\gamma})^2)^{1/2}$, where $\psi_k^{\chi}$ is the score of the function $\chi$ in the $k$th principal component. See Chapters 8 and 9 in Ramsay and Silverman (2005) or Chapter 3 in Ferraty and Vieu (2006) for more information about functional principal component analysis.

In Ferraty and Vieu (2006) it is proved that $\hat{m}_K(\chi)$ is a consistent estimator (in the sense of almost complete convergence) of $m(\chi)$ under regularity conditions on $m$, $\boldsymbol{\chi}$ (involving small balls probability), $Y$ and $K$. Moreover, Ferraty et al. (2007) prove the mean square convergence and find the asymptotic distribution of $\hat{m}_K(\chi)$.

The book of Ferraty and Vieu (2006) lists several interesting open problems concerning nonparametric functional regression. In particular, their *Open Question 5* addresses the transfer of local polynomial regression ideas to an infinite dimensional setting in order to extend the estimator $\hat{m}_K(\chi)$, that is a kind of Nadaraya–Watson regression estimator.

A first answer to this question is given in Baíllo and Grané (2009). They propose a natural extension of the finite dimensional local linear regression, by solving the problem

$$\min_{\alpha,\beta} \sum_{i=1}^{n} w_i(\chi) \left( y_i - \alpha - \int_a^b (\chi_i(t) - \chi(t))\beta(t)\mathrm{d}t \right)^2,$$

where local weights $w_i(\chi) = K(\|\chi - \chi_i\|/h) / \sum_{j=1}^{n} K(\|\chi - \chi_j\|/h)$ are defined by means of $L_2$ distances ($\|\chi\|^2 = \int_a^b \chi^2(t)\mathrm{d}t$; it is assumed that all the functions are in $L_2([a, b])$). Their estimator of $m(\chi)$ is $\hat{m}_{LL}(\chi) = \hat{\alpha}$. Closely related approaches can be seen in Berlinet et al. (2007) and Barrientos-Marin (2007).

In this work we give an alternative response to the same open question. Our proposal rests on Distance-Based Regression (DBR), a prediction tool based on inter-individual distances including both Ordinary and Weighted Least Squares (OLS, WLS) as particular cases. Section 2 presents the needed formulas. In Section 3 we introduce our proposal, Local Linear Distance-Based Regression and in Section 4 we apply it to studying two datasets: the Spectrometric Data mentioned above and another one arising from air pollution measures. Section 5 contains some concluding remarks.