Mathematical and Computer Modelling 58 (2013) 1568–1579

Contents lists available at SciVerse ScienceDirect

Mathematical and Computer Modelling

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

Comparative study of different *B*-spline approaches for functional data

A.M. Aguilera [a,*], M.C. Aguilera-Morillo [b]

[a] *Facultad de Ciencias. Campus de Fuentenueva s/n, 18071. Granada, Spain*

[b] *Facultad de Farmacia. Campus de Cartuja s/n, 18071. Granada, Spain*

ARTICLE INFO

Article history:
Received 19 October 2011
Received in revised form 7 April 2013
Accepted 30 April 2013

Keywords:
Functional data
B-spline expansions
Roughness penalty
P-splines

ABSTRACT

The sample observations of a functional variable are functions that come from the observation of a statistical variable in a continuous argument that in most cases is the time. But in practice, the sample functions are observed in a finite set of points. Then, the first step in functional data analysis is to reconstruct the functional form of sample curves from discrete observations. The sample curves are usually represented in terms of basis functions and the basis coefficients are fitted by interpolation, when data are observed without error, or by least squares approximation, in the other case. The main purpose of this paper is to compare three different approaches for estimating smooth sample curves observed with error in terms of *B*-spline basis: regression splines (non-penalized least squares approximation), smoothing splines (continuous roughness penalty) and *P*-splines (discrete roughness penalty). The performance of these spline smoothing approaches is studied via a simulation study and several applications with real data. Cross-validation and generalized cross-validation are adapted to select a common smoothing parameter for all sample curves with the roughness penalty approaches. From the results, it is concluded that both penalized approaches drastically reduced the mean squared errors with respect to the original smooth sample curves with *P*-splines giving the best approximations with less computational cost.

1. Introduction

Functional data analysis (FDA) is a topic of active statistical research devoted mainly to the extension of multivariate analysis techniques to the case where the data is a set of curves instead of vectors. In most of applications, sample curves come from the observation of a stochastic process in continuous time. There are many other applications, as for example chemometric [1], where the argument is not the time. An excellent collection of statistical methods to analyze functional data and a wide variety of applications of these methodologies in different fields can be found in [2,3].

Despite their continuous nature, sample curves are usually observed in a finite set of sampling points that could be unequally spaced and different among the sample units. Because of this it is necessary to reconstruct the true functional form of each sample curve from a finite set of discrete observations. Many approximation techniques such as interpolation or projection in a finite-dimensional space generated by basis functions were applied from the beginning to solve this problem. More recently, non-parametric techniques were used for approximating functional data [4].

In many applications the data are smooth functions observed with error. In this case least squares approximation can be used to estimate the basis coefficients of a basis expansion of the unobserved smooth sample functions. In this paper

* Corresponding author. Tel.: +34 958243267; fax: +34 958243267.
E-mail address: aaguiler@ugr.es (A.M. Aguilera).

0895-7177/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.mcm.2013.04.007

three different approaches for solving this problem in terms of B-spline basis functions are compared in the FDA context: regression splines, smoothing splines and penalized splines. The main purpose for this comparison is to provide guidelines about the best approach for estimating smooth sample curves from discrete noisy observations.

B-splines are constructed from polynomial pieces joined at a set of knots. Once the knots are given, B-splines can be evaluated recursively for any degree of the polynomial by using a numerically stable algorithm (see [5]). The choice of knots is an important problem when working with B-splines. If too many knots are selected you have an overfitting of the data. On the other hand too few knots provides an underfitting. This fact is specially significant in the case of non-penalized spline regression (regression splines). Some automatic numerical schemes for optimizing the number and the position of the knots were proposed to solve this problem (see for example [6]).

Soothing splines were first proposed by O'Sullivan [7] by introducing a penalty in the second derivative of the curve. This approach restricts the flexibility of the fitted curve and prevents the overfitting. This approximation was generalized later such that it could be applied in any context where regression on B-splines was useful [8]. This kind of penalized smoothers known as P-splines works with a relatively large number of equally spaced knots and a penalty based on differences between coefficients of adjacent B-splines.

The approximation of smooth functions with B-spline bases is used in the estimation of a wide variety of FDA methodologies as functional principal component analysis [9], functional linear regression models, functional generalized linear models and functional additive models, among others [10–15]. This justifies the importance of a comparison among the main smoothing approaches in terms of B-splines and to draw conclusions that allow the researchers and practitioners to use the most powerful tool in each case.

After this introduction section, a brief description of the different non-penalized and penalized spline smoothers with B-spline bases (regression splines, smoothing splines and P-splines) is presented in Section 2. The most used methods for choosing the smoothing parameter in the roughness penalty approaches (cross-validation and generalized cross-validation) are also adapted to select only one smoothing parameter for fitting all the sample curves in the FDA context. The comparison of the approximation results provided by the considered approaches is developed in Section 3 on a simulation study. Finally, the performance of these spline smoothers is also studied in two applications with real data.

2. Smoothing with *B*-spline bases

As indicated before, the firs step in FDA is to reconstruct the functional form of the sample curves from their discrete observations. The most usual way to solve this problem consists of assuming an expansion of each sample curve in terms of a basis of functions and to fit the basis coefficients using smoothing or interpolation.

2.1. Basis expansion of functional data

Let $\{x_i(t) : i = 1, \ldots, n\}$ be a sample of functions related to a functional variable X. The sample curves can be considered observations of a second order stochastic process $X = \{X(t) : t \in T\}$ whose sample functions belong to the Hilbert space $L^2(T)$ of square integrable functions with the usual inner product $\langle f, g\rangle = \int_T f(t)\, g(t)\, dt$, $\forall f, g \in L^2(T)$.

In practice, sample functions are observed in a finite set of time points $\{t_{i0}, t_{i1}, \ldots, t_{im_i} \in T\}$ $\forall i = 1, \ldots, n$. Then, the sample information is given by the vectors $x_i = \left(x_{i0}, \ldots, x_{im_i}\right)'$, with x_{ik} being the value of the ith sample path, $x_i(t)$, observed at the time t_{ik} $(k = 0, \ldots, m_i)$.

In this section, the sample paths are assumed to belong to a finite-dimension space generated by a basis $\left\{\phi_1(t), \ldots, \phi_p(t)\right\}$, so that

$$x_i(t) = \sum_{j=1}^{p} a_{ij}\phi_j(t), \quad i = 1, \ldots, n. \tag{1}$$

This equation can be expressed in matrix form as $x_i(t) = a_i'\phi(t)$, where $a_i = \left(a_{i1}, \ldots, a_{ip}\right)'$ and $\phi(t) = \left(\phi_1(t), \ldots, \phi_p(t)\right)'$.

There are different ways of obtaining the basis coefficients depending on the kind of observations we are working with.

If the sample curves are observed without error

$$x_{ik} = x_i(t_{ik}) \quad k = 0, \ldots, m_i, i = 1, \ldots, n,$$

some interpolation method, such as natural cubic spline interpolation, can be used [16]. Quasi-natural cubic spline interpolation with B-splines functions was used to reconstruct sample curves of temperatures from daily observations and to predict the annual risk of drought in terms of them [17].

If the functional predictor is observed with error

$$x_{ik} = x_i(t_{ik}) + \varepsilon_{ik} \quad k = 0, \ldots, m_i, i = 1, \ldots, n, \tag{2}$$

we can use a smooth approximation method as least squares after choosing an appropriate basis. An application of least squares smoothing with trigonometric and B-spline basis was developed for approximating the curves of stress of lupus patients from daily observations and determining the relationship between flares and stress level [18].

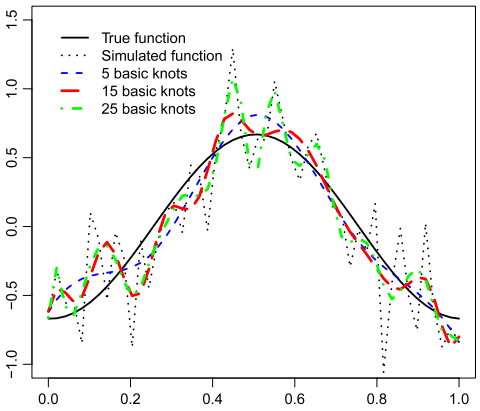


Fig. 1. Regression spline with different number of basis knots (5, 15 and 25).

With both methods, smoothing and interpolation, the functional form of sample paths is obtained by approximating the basis coefficients $\{a_{ij}\}$ from the observations of the sample curves at discrete points. In this work, smooth sample curves observed with error will be considered. Because of this, different types of least squares smoothing with spline functions are compared. *B*-spline basis functions that have excellent numerical properties are considered to span the spline smoothers.

The goal is fitting a function x_i from each vector $x_i = (x_{i1}, x_{i2}, \ldots, x_{im_i})'$ of discrete noisy observations by assuming model (2) and basis expansion (1) for each one of the n observed sample curves.

Choosing the ideal basis and its dimension p for approximating the functional form of a set of sample curves is very important and must be done according to the characteristics of the data. Useful basis systems are Fourier basis for periodic data, *B*-spline basis for non-periodic smooth data with continuous derivatives up to certain order, and wavelet basis for data with a strong local behavior whose derivatives are not required. In this paper, we consider the particular case in which we have smooth curves observed with noise. For this reason, *B*-spline bases are chosen.

The *B*-spline basis of order $q + 1$ (degree q) generates the space of the splines of the same degree, defined as curves consisting of piecewise polynomials of degree q that join up smoothly at a set of definition knots, denoted by $\tau_0 < \cdots < \tau_s$, with continuity in their derivatives up to order $q - 1$. The study of spline functions from an introductory level to a higher mathematical level can be followed in [19,5,20], respectively. In this paper, we will consider the iterative definition of *B*-splines introduced by De Boor [21]. In addition, cubic *B*-splines will be used in the simulations and applications to fit regular sample curves with first and second continuous derivatives.

2.2. Regression splines

The simplest linear smoother approximates the coefficients a_i by minimizing the least squares criterion

$$\mathrm{MSE}(a_i|x_i) = (x_i - \Phi_i a_i)'(x_i - \Phi_i a_i),$$

with $\Phi_i = \left(\phi_j(t_{ik})\right)_{m_i \times p}$. Thus, the estimate of a_i that minimizes this mean squared error is given by $\hat{a}_i = \left(\Phi_i'\Phi_i\right)^{-1}\Phi_i' x_i$. When a basis of *B*-splines is considered, these fitted curves are usually called regression splines.

This approximation is appropriate when the errors ε_{ik} are independently distributed with mean zero and constant variance $\forall k = 0, \ldots, m_i; i = 1, \ldots, n$. In many applications with functional data the errors could be non-stationary and/or autocorrelated so that this assumption is not realistic. In these cases weighted least squares regression can be used (see a detailed study in [2]).

The degree of smoothness of regression splines depends on the size of the *B*-spline basis which is a function of the number of knots and the degree of the spline. In Fig. 1 it can be seen how the largest number of knots is the worst fit to the underlying function because it does not filter out noise efficiently. The selection of the number and location of knots in regression splines is through quite complicated and non attractive algorithms (see for example [6,22,23]).

Localized smoothing methods such as kernel smoothing and local polynomial smoothing are an alternative class of weighted least squares smoothing with excellent computational properties but an important instability near the boundaries of the observational interval [4]. Continuous and discrete roughness penalty approaches are considered in this paper as a more flexible and powerful way of smoothing discrete data by a smooth function that solves the drawbacks of the ones mentioned before.

2.3. Smoothing splines

Let us remember that the goal is to estimate for each sample curve the coefficients of its basis expansion from a set of discrete noisy observations that verify Eq. (2). The curve fitted using roughness penalties provides a good fit to the data in terms of residual sum of squares and simultaneously controls the degree of smoothness.

The continuous penalty for smoothing splines measures the roughness of a function by means of the integrated squared second derivative and was first introduced by Reinsch [24]. If it is necessary, a higher order of derivative can be used to control the degree of smoothness of the true curve. The computation of this continuous penalty in terms of B-splines basis functions was considered in [7] to propose optimal algorithms for solving the inverse problem.

In order to quantify the roughness of each curve, $x_i(t)$, the integrated squared derivative of order d is considered

$$\int [D^d x_i(s)]^2 ds = a_i' R_d a_i,$$

where R_d is the matrix defined by $R_d = \int D^d\phi(s)\, D^d\phi(s)'\, ds$, with $D^d\phi(s) = \left(D^d\phi_1(s), \ldots, D^d\phi_p(s)\right)'$.

Then the basis coefficients of the smoother are obtained by minimizing the penalized least squares error given by

$$\mathrm{CPMSE}_d(a_i|x_i) = (x_i - \Phi_i a_i)'(x_i - \Phi_i a_i) + \lambda a_i' R_d a_i. \tag{3}$$

In practice, the most common penalty order is $d = 2$. The most usual computational approach for spline smoothing is to minimize penalized criterion (3) with respect to the coefficients of a basis expansion in terms of cubic B-splines functions with knots at the sampling points. In this case, the fitted function is called cubic spline smoother.

When a very large number of sampling points is involved, a lower number of appropriate knots can be sufficient to smooth the sample paths and capture their main features. In general, a smoothing spline is obtained assuming an expansion in terms of B-splines and minimizing (3). Then, the vector of estimated basis coefficients is $\hat{a}_i = \left(\Phi_i'\Phi_i + \lambda R_d\right)^{-1}\Phi_i' x_i$.

An interesting application of cubic smoothing splines for the implementation of the functional mixed effects models can be seen in [25].

2.4. P-splines

The roughness penalties considered for smoothing splines are defined in terms of integrated squared derivatives. The computational problem of this approach lies in the calculation of the matrix R_d whose elements are the integrals of products of d-order derivatives between B-spline basis functions. A simpler discrete penalty approach is based on defining the roughness of a function by summing squared d-order difference values. This kind of penalty depends on the considered basis and only works if the sampling points are equally spaced. A penalty based on differences of order d between coefficients of adjacent B-splines is used in [8]. Penalized splines can be also computed in terms of truncated power functions. A recent study has shown that penalized spline regression with B-splines with equally spaced knots and difference penalties outperforms the penalized spline approach based on truncated power functions with knots based on quantiles of the independent variable and a ridge penalty [26].

The basis coefficients of a penalized spline smoother in terms of B-spline basis can be computed by minimizing the penalized least squares error

$$\mathrm{DPMSE}_d(a_i|x_i) = (x_i - \Phi_i a_i)'(x_i - \Phi_i a_i) + \lambda a_i' P_d a_i, \tag{4}$$

where $P_d = \left(\Delta^d\right)'\Delta^d$ with Δ^d being the matrix representation of the d-order difference operator.

These smoothers are called penalized splines (P-splines) and their B-spline basis coefficients are estimated by $\hat{a}_i = \left(\Phi_i'\Phi_i + \lambda P_d\right)^{-1}\Phi_i' x_i$. The application of P-splines to different models with smooth components and a nonparametric strategy for the choice of the P-spline parameters has been performed by Currie and Durban [27], where mixed model (REML) methods were applied for smoothing parameter selection. Taking into account that the degree of smoothing is controlled by the smoothing parameter, the number and location of knots is not crucial for fitting a P-spline. Generally, the knots of a P-spline are equally spaced and the number of knots must be sufficiently large to fit the data and not so large that computation time is unnecessarily big. Two algorithms for automatic selection of the number of knots by using generalized cross validation were considered in [28].

Summarizing, we can say that P-splines combine the best of the regression and smoothing splines because they have less numerical complexity than smoothing splines and the selection of knots is not so determinant as in regression splines.

2.5. Choosing the smoothing parameter

The role of the smoothing parameter in penalized smoothing is to control the smoothness of the fitted curve. In order to compute the optimal value of the smoothing parameter λ, two selection criteria are considered and compared in this paper: leave-one-out cross validation (CV) and generalized cross validation (GCV). To select the same smoothing parameter for all the n sample paths we propose to minimize the mean of the cross-validation errors over all sample curves.

The CV (leave-one-out) method consists of selecting, for each curve, the smoothed parameter λ which minimizes the next expression

$$\mathrm{CV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{CV}_i(\lambda),$$

where

$$\mathrm{CV}_i(\lambda) = \sqrt{\sum_{k=0}^{m_i} \left(x_{ik} - \hat{x}_{ik}^{-k}\right)^2 / (m_i + 1)},$$

with $\hat{x}_{ik}^{-k}$ being the values of the ith sample path estimated at time t_{ik} avoiding the kth time point in the iterative estimation process. The CV approach has two main problems, in that it is very expensive from a computational point of view and can lead to undersmoothing the data.

The GCV method is computationally simpler and very well used in the literature about smoothing splines [29]. We consider two versions of GCV error, one for the smoothing splines and other for P-splines. The GCV method consists of selecting λ so that minimized

$$\mathrm{GCV}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{GCV}_i(\lambda),$$

where

$$\mathrm{GCV}_i(\lambda) = \left(\frac{(m_i + 1)}{(m_i + 1) - df(\lambda)}\right)\left(\frac{\mathrm{MSE}_i}{(m_i + 1) - df(\lambda)}\right),$$

with $\mathrm{MSE}_i = \sum_{k=0}^{m_i} \left(x_{ik} - \hat{x}_{ik}\right)^2$, $df(\lambda) = trace(H_i)$ and $H_i = \Phi_i \left(\Phi_i'\Phi_i + \lambda R_d\right)^{-1} \Phi_i'$ in the case of the smoothing splines. If we work with P-splines $H_i = \Phi_i \left(\Phi_i'\Phi_i + \lambda P_d\right)^{-1} \Phi_i'$.

3. Simulation study

The ability of P-splines, smoothing splines and regression splines to approximate smooth curves observed with noise is tested on simulated data. The simulated data set consists of 100 sample paths of the second order stochastic process with zero mean given by $X(t) = R\cos(2\pi t + \theta)$, where R and θ are i.r.v with distributions, Rayleigh (σ), with $\sigma = 0.3$, and Uniform $[0, 2\pi]$, respectively. Noisy observations of the sample paths were simulated at $m = 51$ equally spaced knots in the interval $T = [0, 1]$. That is, $x_{ik} = X(t_{ik}) + \epsilon_{ik}$ $(t_{ik} = k \times 0.02;\ k = 0, 1, \ldots, 50;\ i = 1, \ldots, 100)$, where the errors ϵ_{ik} were simulated from independent normal distributions, Normal $\left(0, \sigma^2\right)$ with $\sigma^2 = 0.07$. The error variances have been fixed to control the determination coefficient R^2 near 0.7.

The first step in this work was to select the smoothing parameter λ. In order to get the best smoothing parameter, we have compared the two different methods of selection of λ seen in 2.5. Fig. 2 (left) shows the box plot related to the mean squared error (MSE) of the approximation of curves provided by the smoothing splines and P-splines with λ selected by CV and GCV. We can see that with the smoothing spline approach the CV method minimizes the MSE regardless the number of basis knots. With the P-spline approach CV and GCV selection criteria provide similar approximation errors. In order to compare the three smoothing approaches with B-spline bases studied in this work, the smoothing parameter was selected by the CV method.

In Fig. 2 (right) and Fig. 3, the three different cubic spline approximations of different sample paths to the simulated discrete data with different number of basis knots (5, 15 and 25) are displayed. It can be observed that the three smoothers are good approximations to the true function for the case of a cubic B-spline basis with five knots. When the number of basis knots increases, regression splines and smoothing splines lose control of smoothness. However, P-splines maintain a good fit for any number of knots. The differences between smoothing splines and P-splines are not too big but P-splines are computationally easier to compute and the adjustment to the original function is not affected by the number of knots.

In order to obtain general conclusions, we have represented in Fig. 4 the mean curve and the MSE distribution provided by the three approximation approaches (regression splines, smoothing splines and P-splines) for the 100 simulated sample paths, considering different number of basis knots (5, 15 and 25). It can be seen that the P-spline approach provides the best fit to the true mean function and the smallest MSE in all considered cases. On the contrary, regression splines give the worst fit because they do not control the degree of smoothness.

4. Real data applications

Once the P-Splines have been chosen as the best smoothers to approximate noisy sample paths from discrete observations, their behaviors have been tested using two real functional data sets. Firstly, we approximate the pinch force data set by using P-splines and comparing the results with the other two methodologies summarized in this paper. In the second application, the P-splines approach is applied to smooth the spectrometric curves related to Flemish hog manures.

4.1. Pinch data

Pinch data, analyzed in [30], were collected at the Medical Research Council Applied Psychology Unit, Cambridge, and consist of records of the force exerted by pinching a force meter (width 6 cm) with the tips of the thumb and forefinger on opposite sides.

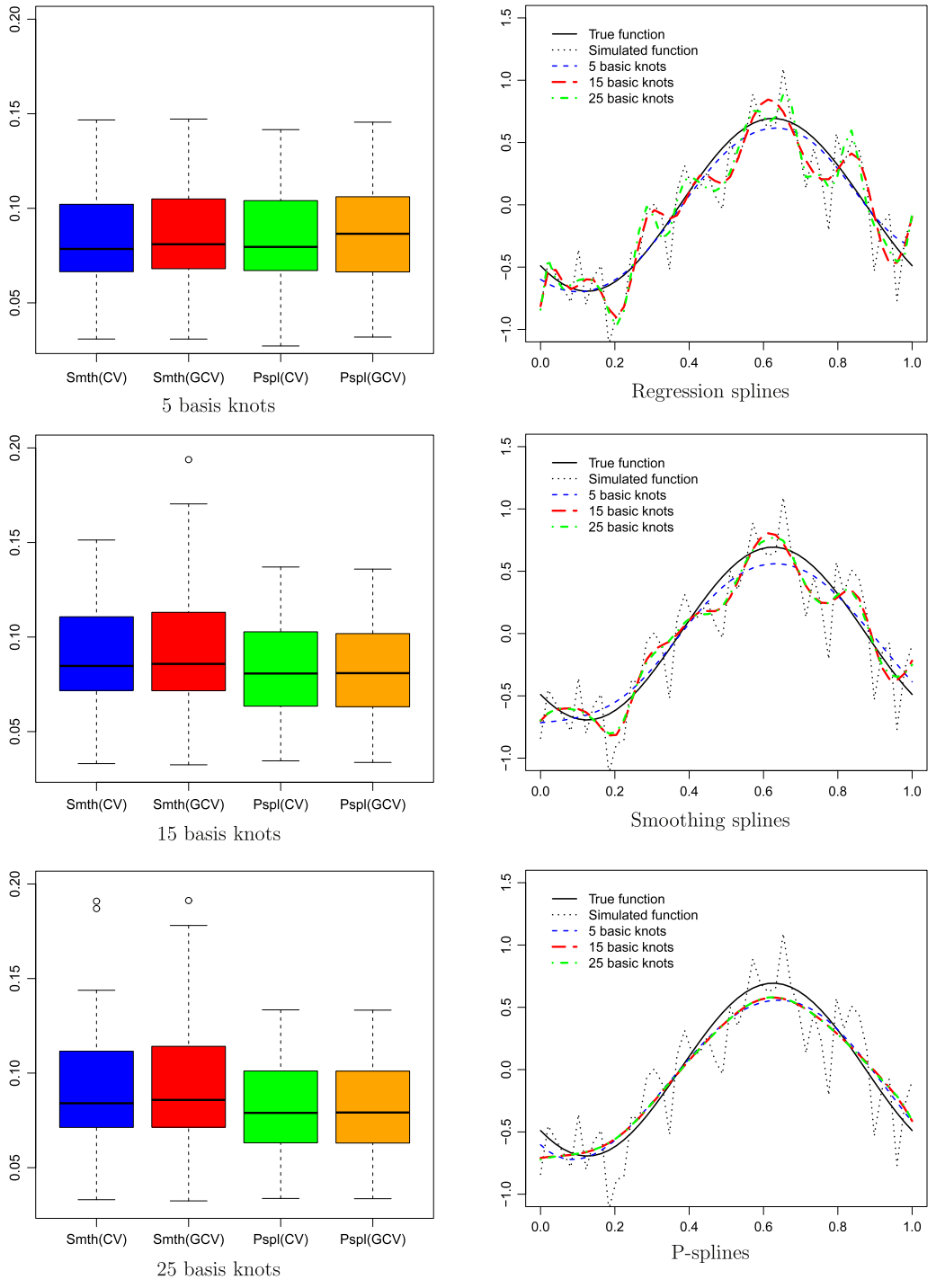


Fig. 2. Left: box plot related to the MSE of the approximation of curves by smoothing splines, with λ selected by CV (blue) and GCV (red), and *P*-splines, with λ selected by CV (green) and GCV (orange), by using different number of basis knots (5, 15 and 25). Right: regression splines, smoothing splines and *P*-splines approaches with different number of basis knots (5, 15 and 25) for one of the simulated sample curves. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The exerted force must be adapted to the characteristics of the gripped object (such as texture, weight, surface, acceleration, between others). Sometimes, the system is slow to the response speed required by the exterior world and in this case it is the brain who must exert the required force. So, the importance of studying this system is to make possible a better understanding of how the brain can control high performance motor systems.

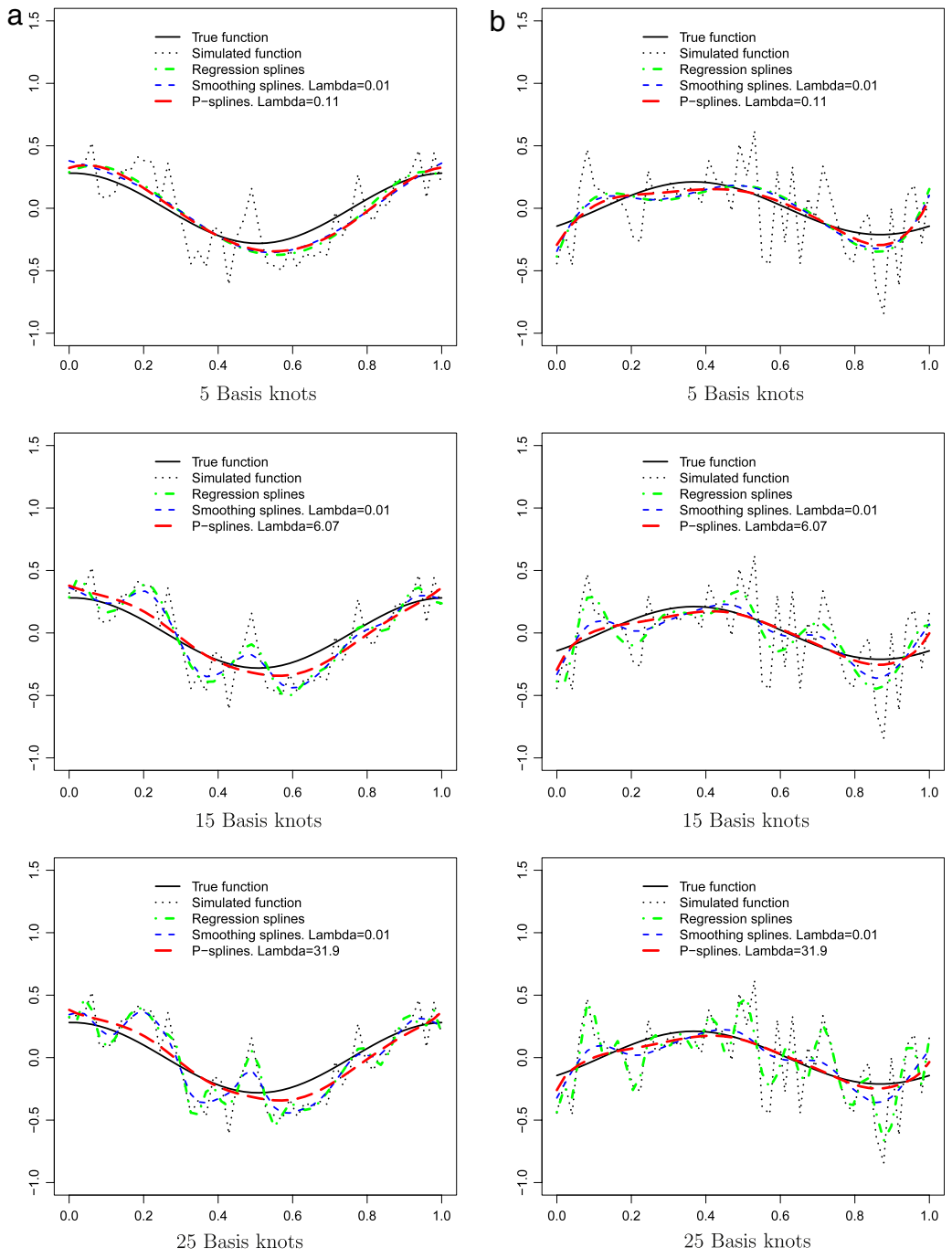


Fig. 3. Regression splines (green, dashed and dotted line), smoothing splines (blue and dashed line) and *P*-splines (red and long dashed line) approaches with 5, 15 and 25 basis knots, for two different sample paths (a) (left) and (b) (right). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The data set used in this paper consists of a sample of 20 records of the force exerted by the human thumb and forefinger during a brief squeeze. The force was sampled at 151 times (seconds). We have considered a cubic *B*-spline basis with 30 equally spaced knots to approximate the true sample paths. The smoothing parameter λ has been chosen by the CV method.

In Fig. 5(a) the necessity of smoothing the observed data is clear. The different spline approaches with *B*-spline basis studied in this paper have been applied and displayed in Fig. 5(b)–(d). Let us observe that regression splines cannot completely avoid the noise at the extremes. Between the two kinds of penalty applied (smoothing splines (c) and *P*-splines (d)), is the

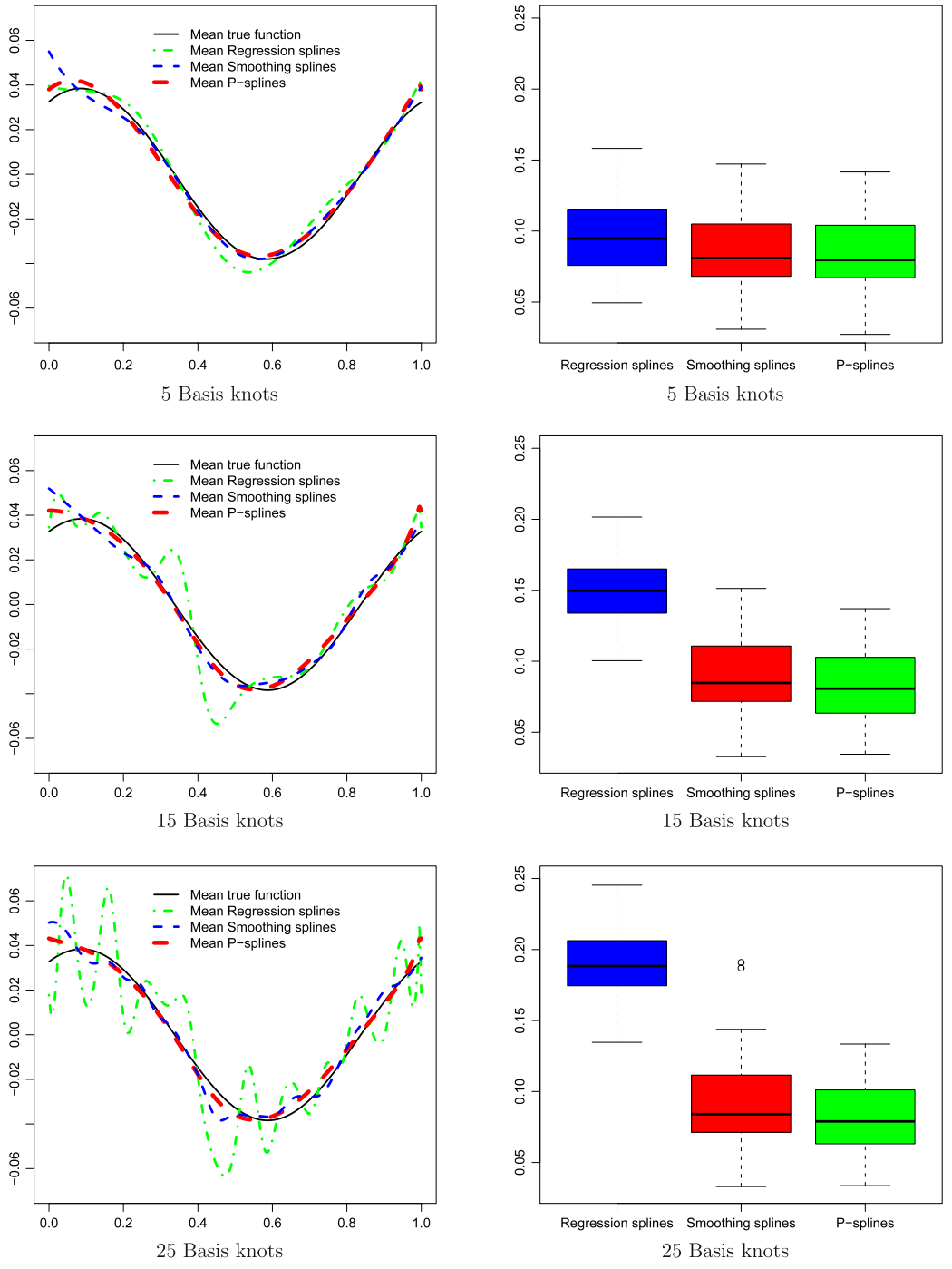
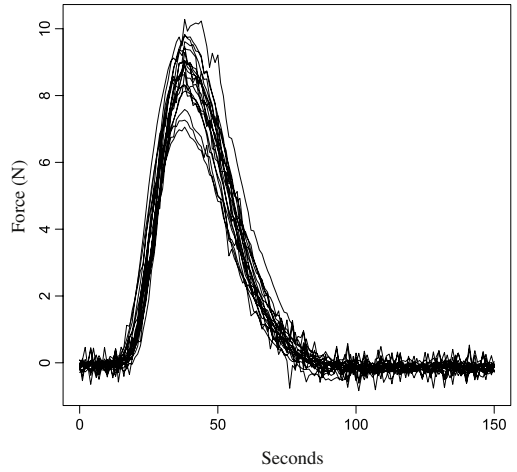


Fig. 4. Mean function (left) for 100 fitted curves through regression splines (green), smoothing splines (blue) and *P*-splines (red) approaches using 5, 15 and 25 basis knots. MSE (right) for 100 fitted curves through regression splines (blue), smoothing splines (red) and *P*-splines (green). λ has been chosen by CV. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

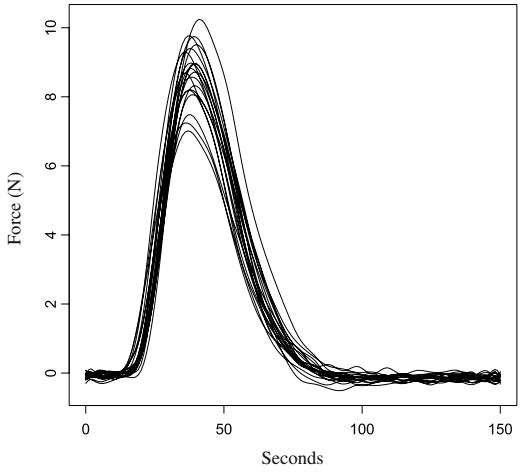
P-spline approach which provides the best smoothing of the sample paths. Two original sample paths and their *P*-spline approaches are shown in Fig. 6.

4.2. Manure data

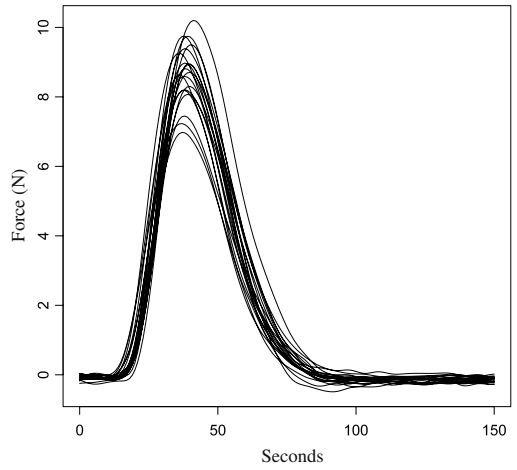
The manure data set was analyzed in [31] and consists of 138 sample paths about Flemish hog manures collected in the Spring of 2003 at almost as many different farms in Flanders by the Soil Service of Belgium. All samples were scanned in



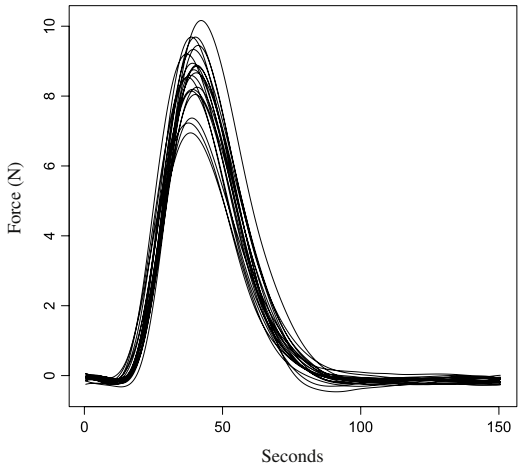
(a) Original sample paths.



(b) Regression splines.

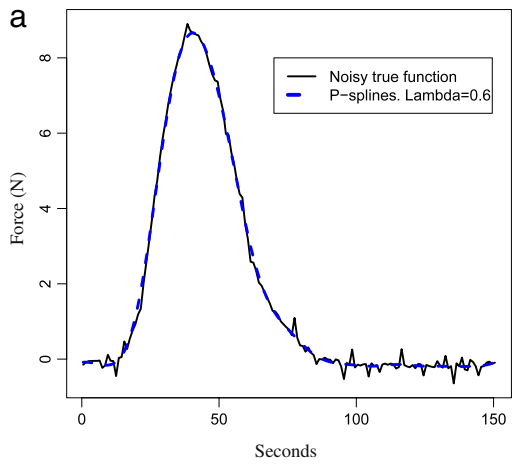


(c) Smoothing splines.



(d) *P*-splines.

Fig. 5. Original pinch data set (a) and its fit by regression splines (b), smoothing splines (c) and *P*-splines (d) using *B*-splines basis defined at 30 knots. The different values of λ have been chosen by CV.



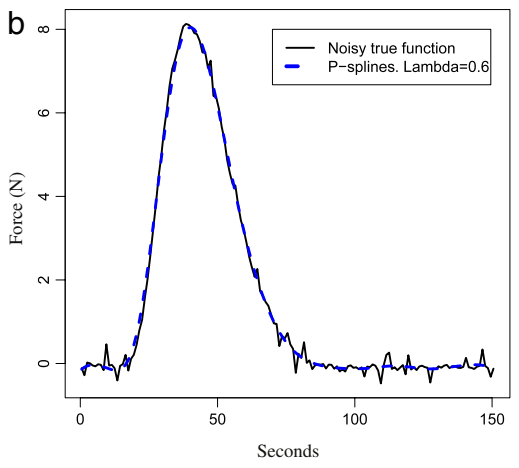
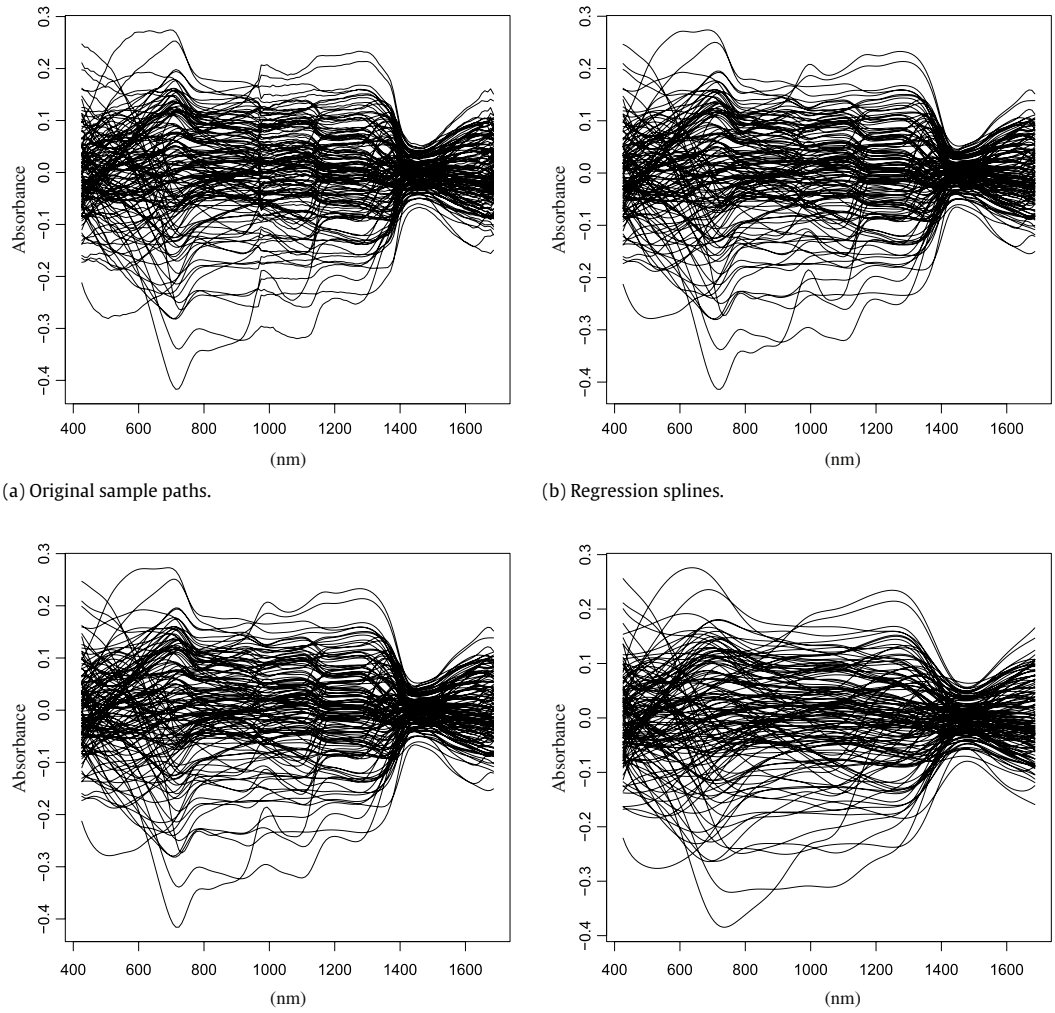


Fig. 6. Fitting two true functions observed with noise, (a) and (b) (black and solid line) by *P*-splines (blue and dashed line) using 30 basis knots and $\lambda = 0.6$ (chosen by CV). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(a) Original sample paths.

(b) Regression splines.

(c) Smoothing splines.

(d) *P*-splines.

Fig. 7. Original manure data set (a) and its fit by regression splines (b), smoothing splines (c) and *P*-splines (d) using *B*-splines basis defined at 30 knots. The different values of λ have been chosen by CV.

reflectance mode on a diode array Vis/NIR spectrophotometer. After that, data has been converted into absorbance units ranging from 426 to 1686 nm.

In order to compute the three different types of spline smoothers to the observed data we have considerate *B*-splines basis defined at 30 knots. The smoothing parameters λ have been chosen by the CV method. The original sample paths are represented in Fig. 7(a). The smoothing splines (c) and regression splines (b) are quite similar. However, *P*-splines lead to the smoothest approximation of the sample paths. Finally, two original sample paths and their *P*-spline approximations are shown in Fig. 8.

5. Software

All calculations in this research were performed on the 2.10.1 version of statistical software R project, making use of the spline package for the *B*-splines basis construction (included in the fda library). The *P*-splines approach was implemented through our own routine considering Eilers and Marx's definition. Smoothing parameters were chosen by our own routine based on the Cross Validation method (leave-one-out). Smoothing Splines approximation was made by means of a function available in the spline library (smooth.spline).

6. Conclusions

Non-penalized and penalized least squares smoothing in terms of *B*-spline bases have been compared in this paper to approximate a set of unobserved smooth curves from discrete noisy observations. A simulation study and two applications

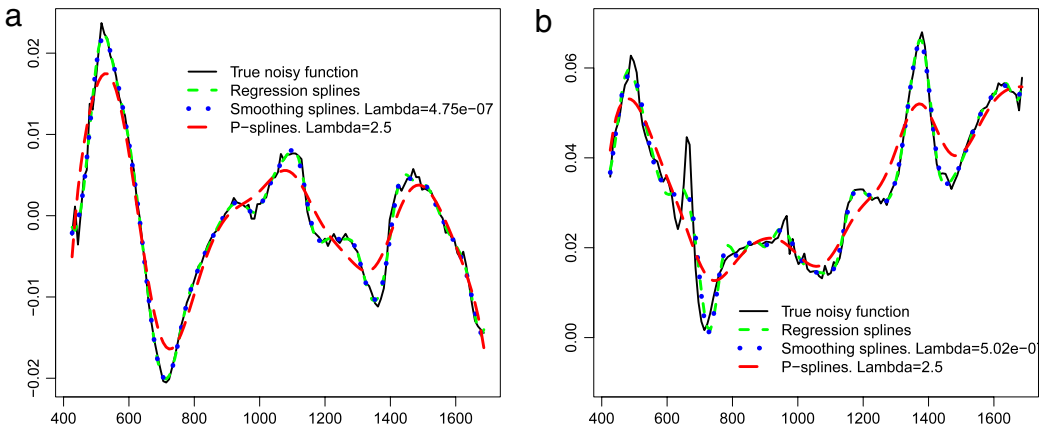


Fig. 8. Fitting two true functions observed with noise, (a) and (b) (black and solid line) by using *P*-splines (red and large dashed line), smoothing splines (blue and dotted line) and regression splines (green and dashed line), with 30 basis knots. λ chosen by CV. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

with real functional data have been developed to study and compare the performance of the three considered smoothers (regression splines, smoothing splines and *P*-splines) in the FDA context.

As a basis from these results we can conclude that regression splines and smoothing splines lose control of the smoothness when the number of knots increases. Both penalized approaches get to improve the fit providing mean squared errors with respect to the original smooth sample curves much smaller than the ones given by the non-penalized approach. On the other hand, *P*-splines provide the lowest approximation errors, have less numerical complexity making its computational implementation easier and are quite insensitive to the choice of knots, so it is sufficient to choose a relatively large number of equally spaced basis knots.

Acknowledgments

This research has been funded by project MTM2010-20502 from *Dirección General de Investigación, Ministerio de Educación y Ciencia Spain* and project P11-FQM-8068 from *Consejería de Innovación, Ciencia y Empresa. Junta de Andalucía, Spain.*

References

[1] W. Saeys, B. De Ketelaere, P. Darius, Potential applications of functional data analysis in chemometrics, Journal of Chemometrics 22 (5) (2008) 335–344.
[2] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, Springer-Verlag, 2005.
[3] J.O. Ramsay, B.W. Silverman, Applied Functional Data Analysis: Methods and Case Studies, Springer-Verlag, 2002.
[4] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis, Theory and Practice, Springer-Verlag, 2006.
[5] C. De Boor, A Practical Guide to Splines, Revised Edition, Springer, 2001.
[6] J. Friedman, B. Silverman, Flexible parsimonious smoothing and aditive modelling (with discussion and response), Technometrics 31 (1989) 1–39.
[7] F. O'Sullivan, A statistical perspective on ill-posed inverse problems, Statistical Science 1 (1986) 505–527.
[8] P. Eilers, B. Marx, Flexible smoothing with *b*-splines and penalties, Statistical Science 11 (2) (1996) 89–121.
[9] A. Aguilera, M. Aguilera-Morillo, Penalized pca approaches for *b*-spline expansions of smooth functional data, Applied Mathematics and Computation 219 (14) (2013) 7805–7819.
[10] B. Brumback, J. Rice, Smoothing spline models for the analysis of nested and crossed samples of curves, Journal of the American Statistical Association 93 (443) (1998) 961–976.
[11] B. Marx, P. Eilers, Generalized linear regression on sampled signals and curves: a *p*-spline approach, Technometrics 41 (1999) 1–13.
[12] H. Cardot, F. Ferraty, P. Sarda, Spline estimators for the functional linear model, Statistica Sinica 13 (2003) 571–591.
[13] C. Crambes, A.F. Kneip, P. Sarda, Smoothing splines estimators for functional linear regression, Annals of Statistics 37 (2009) 35–72.
[14] A. Aguilera, M. Escabias, C. Preda, G. Saporta, Using basis expansion for estimating functional pls regression applications with chemometric data, Chemometrics and Intelligent Laboratory Systems 104 (2010) 289–305.
[15] M.C. Aguilera-Morillo, A. Aguilera, M. Escabias, M.J. Valderrama, Penalized spline approaches for functional logit regression, Test 22 (2) (2013) 251–277.
[16] A. Aguilera, R. Gutiérrez, M. Valderrama, Approximation of estimators in the pca of a stochastic process using *b*-splines, Communications in Statistics Simulation and Computation 25 (3) (1996) 671–690.
[17] M. Escabias, A. Aguilera, M. Valderrama, Modeling environmental data by functional principal component logistic regression, Environmetrics 16 (1) (2005) 95–107.
[18] A. Aguilera, M. Escabias, M. Valderrama, Discussion of different logistic models with functional data, application to systemic lupus erythematosus, Computational Statistics and Data Analysis 53 (1) (2008) 151–163.
[19] P. Green, B. Silverman, Nonparametric Regression and Generalized Linear Models, in: Monographs on Statistics and Applied Probability, Chapman & Hall, 1994.
[20] G. Wahba, Spline Models for Observational Data, Society for Industrial and Applied Mathematics, 1990.
[21] C. De Boor, Package for calculating with *b*-splines, Journal of Numerical Analysis 14 (1977) 441–472.
[22] T. Lee, Regression spline smoothing using the minimum description length principle, Statistics & Probability Letters 48 (2000) 71–82.
[23] S. Zhou, X. Shen, Spatially adaptive regression splines and accurate knot selection schemes, Journal of the American Statistical Association 96 (453) (2001) 247–259.
[24] C. Reinsch, Smoothing by spline functions, Numerische Mathematik (10) (1967) 177–183.
[25] W. Guo, Functional data analysis in longitudinal settings using smoothing splines, Statistical Methods in Medical Research 13 (1) (2004) 49–62.

[26] P. Eilers, B. Marx, Splines, knots, and penalties, Wiley Interdisciplinary Reviews Computational Statistics 2 (2010) 637–653.
[27] I. Currie, M. Durban, Flexible smoothing with *p*-splines: a unified approach, Statistical Modelling (2) (2002) 333–349.
[28] D. Ruppert, Selecting the number of knots for penalized splines, Journal of Computational and Graphical Statistics 11 (2002) 735–757.
[29] P. Craven, G. Wahba, Smoothing noisy data with spline functions—estimating the correct degree of smoothing by the method of generalized cross-validation, Numerische Mathematik 31 (4) (1978) 377–403.
[30] J. Ramsay, X. Wang, A functional data analysis of the pinch force of human fingers, Applied Statistics 44 (1995) 17–30.
[31] W. Saeys, P. Darius, H. Ramon, Potential for on-site analysis of hog manure using a visual and near infrared diode array reflectance spectrometer, Journal of Near Infrared Spectroscopy 12 (2004) 299–309.