Applied Mathematics and Computation 219 (2013) 7805–7819

Contents lists available at SciVerse ScienceDirect

Applied Mathematics and Computation

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

Penalized PCA approaches for B-spline expansions of smooth 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

Keywords:
Functional data
Principal component analysis
B-spline expansion
Roughness penalty
P-splines

ABSTRACT

Functional principal component analysis (FPCA) is a dimension reduction technique that explains the dependence structure of a functional data set in terms of uncorrelated variables. In many applications the data are a set of smooth functions observed with error. In these cases the principal components are difficult to interpret because the estimated weight functions have a lot of variability and lack of smoothness. The most common way to solve this problem is based on penalizing the roughness of a function by its integrated squared d-order derivative. Two alternative forms of penalized FPCA based on B-spline basis expansions of sample curves and a simpler discrete penalty that measures the roughness of a function by summing squared d-order differences between adjacent B-spline coefficients (P-spline penalty) are proposed in this paper. The main difference between both smoothed FPCA approaches is that the first uses the P-spline penalty in the least squares approximation of the sample curves in terms of a B-spline basis meanwhile the second introduces the P-spline penalty in the orthonormality constraint of the algorithm that computes the principal components. Leave-one-out cross-validation is adapted to select the smoothing parameter for these two smoothed FPCA approaches. A simulation study and an application with chemometric functional data are developed to test the performance of the proposed smoothed approaches and to compare the results with non penalized FPCA and regularized FPCA.

1. Introduction

Many application areas have functional data that come from the observation of a random function at a discrete set of sampling points. In the majority of cases the sample curves are functions of time but in many others the argument is a different magnitude. In spectroscopy, for example, the NIR spectrum is a functional variable whose observations are measured as functions of wavelengths. The potential of functional data analysis methodologies for the chemometric analysis of spectroscopic data was shown in [1]. A wide variety of applications with functional data in different fields were collected and analyzed by [2].

When analyzing a functional data set it is usual to have a large number of regularly spaced observations for each sample curve. Because of this a reduction dimension technique is necessary for explaining the main features of a set of sample curves in terms of a small set of uncorrelated variables. This problem was solved by generalizing principal component analysis to the case of a continuous-time stochastic process [3]. Asymptotic properties of the estimators of FPCA were deeply studied in the general context of functional variables [4]. Nonparametric methods were developed to perform FPCA for the case of a small number of irregularly spaced observations of each sample curve [5,6]. As in the multivariate case, the interpretation

* Corresponding author.
E-mail address: aaguiler@ugr.es (A.M. Aguilera).

0096-3003/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.amc.2013.02.009

of the principal component scores and loadings is a useful tool for discovering the relationships among the variables associated to a functional data set. To avoid misinterpretation of PCA, a new type of plots, named Structural and Variance Information plots, were recently introduced by [7].

FPCA is a flexible tool in functional data analysis that is successfully used to solve important problems as the estimation of the functional parameter in different functional regression models [8–14]. An alternative methodology for solving this estimation problem in the functional linear model is the functional version of partial least squares (PLS) regression. A estimation procedure based on basis expansions of sample curves was introduced by [15]. A Bayesian approach to FPCA based on a generative model for noisy and sparse observations of curves was developed in [16].

One usual form of estimating FPCA from discrete observations of the sample curves is based on basis expansion approximation. This way, FPCA of a set of curves is reduced to multivariate PCA of a transformation of the matrix of basis coefficients [17]. B-spline basis are appropriate to approximate smooth curves. Cubic spline interpolation with B-spline basis can be considered for approximating smooth sample curves observed without error [18]. On the other hand, least squares approximation with B-spline basis is appropriate for reconstructing the true functional form of noisy smooth curves. This type of approximation was performed to forecast lupus flares from time evolution of stress level [19]. The problem is that regression splines do not control the degree of smoothness and, consequently, the principal component curves show substantial variability and their interpretation is difficult. This problem must be solved by introducing some kind of smoothing in the estimation of principal component curves.

There are different ways of introducing smoothing in the estimation of FPCA. On the one hand, the data can be smooth first and then an unpenalized FPCA is carried out. A spline smoothing that penalizes the integrated squared second derivative of each sample path was considered by [20,21]. This approach was applied for smoothing and reconstructing a magnetic resonance imaging (FMRI) functional data in [22]. On the other hand, the smoothing can be introduced within the FPCA algorithm. Two different approaches for smoothing functional principal components analysis were proposed by [23] and [24]. Both approaches use a continuous penalty that measures the roughness of the principal component curves by their integrated squared d-order derivative but they differ in the way they incorporate the penalty. The Rice-Silverman approach introduces the roughness penalty in the definition of the sample variance of the principal component weight functions. The Silverman approach is known as regularized FPCA (RFPCA) and introduces the penalty in the orthonormality constraint between principal components. This FPCA approach was extended to the case of multivariate functional data sets by using Gaussian basis functions instead of B-splines [23]. An application of regularized FPCA with B-splines basis in actuarial science was performed to estimate the risk of occurrence of a claim in terms of the driver's age and others significative variables [24]. A third way of penalizing FPCA is based on smoothing not the data or the components, but the covariance operator, whose eigenfunctions are the principal component functions [6]. Penalized rank one approximation was recently proposed as an alternative approach to the estimation of FPCA [25].

Penalized spline regression [26] is an increasingly popular smoothing approach that was used to estimate the functional sample mean and to develop an iterative P-spline algorithm for estimating FPCA in [27]. The P-spline penalty measures the roughness of a function by summing squared d-order differences between adjacent basic coefficients. In this paper, two different versions of smoothed FPCA based on penalized splines (P-splines) with B-splines basis are introduced and compared. The first approach carries out an unpenalized FPCA on the P-spline smoothing of the sample curves. The second approximates the sample curves by unpenalized least squares and then incorporate the P-spline penalty in the orthonormality constraint within the FPCA algorithm. The accuracy of the estimates provided by both P-spline smoothed approaches is tested with simulated and real data, and the results compared with non penalized FPCA and regularized FPCA. In order to get an optimum estimation of the smoothing parameter, leave-one-out cross-validation is adapted to this context.

2. Functional principal component analysis

Let $\{x_i(t) : t \in T, i = 1, \ldots, n\}$ be a sample of functions that are the sample information related to a functional variable X. It will be supposed that they are observations of a second order stochastic process $X = \{X(t) : t \in T\}$, continuous in quadratic mean whose sample functions belong to the Hilbert space $L^2(T)$ of square integrable functions with the usual inner product $\langle f, g\rangle = \int_T f(t)g(t)dt, \forall f, g \in L^2(T)$. Multivariate PCA was extended to the functional case to reduce the infinite dimension of a functional predictor and to explain its dependence structure by a reduced set of uncorrelated variables [3]. In order to compute the functional principal components, let us assume without loss of generality that the observed curves are centered so that the sample mean $n^{-1}\sum_{i=1}^{n} x_i(t)$ is zero.

The principal components are obtained as uncorrelated generalized linear combinations with maximum variance (Var). In general, the j-th principal component scores are given by

$$\xi_{ij} = \int_T x_i(t) f_j(t) dt, \quad i = 1, \ldots, n, \tag{1}$$

where the weight function or loading f_j is obtained by maximizing the variance solving

$$\begin{cases} Max_f\ Var\left[\int_T x_i(t) f(t) dt\right], \\ s.t. \|f\|^2 = 1 \quad \text{and} \quad \int f_\ell(t) f(t) dt = 0,\ \ell = 1, \ldots, j-1. \end{cases}$$