Insurance: Mathematics and Economics 45 (2009) 278–285

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

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

Explaining functional principal component analysis to actuarial science with an example on vehicle insurance

M.M. Segovia-Gonzalez *, F.M. Guerrero, P. Herranz

Universidad Pablo de Olavide, Department of Economics, Quantitative Methods and Economic History, Edificio Blanco White num. 3, Ctra. de Utrera km. 1, 41013 Sevilla, Spain

ARTICLE INFO

Article history:
Received October 2008
Received in revised form
May 2009
Accepted 12 July 2009

Keywords:
Vehicle insurance
Claims ratio
Functional principal component analysis

ABSTRACT

Given the high competitiveness in the vehicle insurance market, the need arises for an adequate pricing policy. To this end, insurance companies must select risks in a way that allows the expected claims ratio to come as close as possible to the real claims ratio. The use of new analytical tools which provide more information is of great interest. In this paper it is shown how functional principal component analysis can be useful in actuarial science. An empirical study is carried out with data from a Spanish insurance company to estimate the risk of occurrence of a claim in terms of the driver's age, whilst taking into account other relevant variables.

1. Introduction

In the motor vehicle insurance sector, a large number of insurance companies offer very similar products. The utilization of a set of criteria and actions to insure a vehicle is known as risk selection. This means that the conditions of the insurance policies allow the expected claims ratio to come as close as possible to the real claims ratio. The rules of contract hold great importance for proper conduct in the insurance business and a technical and commercial balance must be sought. Recent empirical work exists in the motor vehicle insurance market that shows no signs of adverse selection in these markets. Chiappori and Salanie (2000) found that there is no systematic relationship between risk and cover in the French motor vehicle insurance market. Dionne et al. (2001) found a similar result using data in Quebec. These authors all confirm that the socioeconomic, vehicle and policy characteristics play a very important role when establishing a suitable segmentation of risk. On the other hand, insurance companies are not only interested in keeping existing clients, but also in detecting potential policyholders with fraudulent behaviour. Along these lines, there are studies which show that when the policyholder is satisfied with the treatment and service given, there is increased susceptibility to policy renewal with the current insurance company (Pujol and Bolance, 2004). Moreover, Artis et al. (1999, 2002) explain how discrete-choice models may be useful for studying fraudulent behaviour.

Insurance rates in the motor vehicle sector are structured according to a number of variables. Here we highlight the application of multivariate and actuarial calculation methods in risk selection for pricing (Boj et al., 2004; Guillen et al., 2005). Insurance companies are able to estimate 'a priori' the claims ratio of homogeneous groups of policyholders. To this end, the characteristics that are most relevant when it comes to having a major or minor number of accidents are taken into account. These categories are: the pricing group of the vehicle, the area of circulation, the use of the vehicle, the driver's personal circumstances and any no-claims bonus. It is very important to know the personal circumstances of the insured: gender, age, years of holding a driving licence, marital status, … . All these characteristics are taken into account in the various kinds of motor vehicle insurance policies.

In the case of a car insurance policy, a large amount of information is made available and this can lead to difficulties in the analysis. Traditionally this difficulty is overcome by means of principal component analysis. This multivariate analysis technique enables such information to be summarized into a smaller number of variables, thereby making the workload lighter. In the last decade, however, techniques, which we label functional analysis techniques, have been developed which provide appropriate processing when dealing with observations from different periods. In this paper we show that the analysis in functional principal components provides stabler estimations than those of conventional principal component analysis (conventional PCA). This is especially relevant to actuarial sciences, since the information commonly available is of this type (analyzing the volume of premiums, the market quota, the cost of claims, … at a regional level in order to analyze the situation of the principal entities in each of the sectors in a determined time period). As an illustration, an empirical application is carried out using these techniques with real car insurance data.

* Corresponding author. Tel.: +34 954349741; fax: +34 954349339.
E-mail addresses: mmseggon@upo.es (M.M. Segovia-Gonzalez), fguecas@upo.es (F.M. Guerrero), pherpei@upo.es (P. Herranz).

0167-6687/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.insmatheco.2009.07.003

In order to develop this idea, the work has been structured as follows.

Section 2 refers to the statistical methodology used in the paper and describes a series of considerations concerning each of the different techniques.

In Section 3, a specific implementation of the techniques is given. We propose a study on how the age variable influences the behaviour of drivers. The function of the risk of occurrence of a claim during an insured driver's lifetime is defined, which is the natural way to deal with a continuous process. To this end, conventional PCA and the functional techniques are applied to data supplied by the vehicle branch of a Spanish insurance company, and more specifically, to those data relating to the study of claims ratios, and the results obtained by these techniques are compared. Once a thorough filtering of this data was carried out and the information was prepared for treatment, a series of algorithms developed by Ramsay and Silverman had to be adapted in order to implement the technique described in Section 2 and to interpret and evaluate the benefits of that technique. The mathematical program MATLAB was used, which has its own programming language that allows us to work with oriented objects and thus implement the techniques of functional data analysis (Ramsay, 2003).

In Section 4, the main conclusions drawn from this work are given, and then we list the bibliographic references used.

2. Theoretical framework

The theoretical framework on which our estimates are based is set out. Two ways of carrying out the functional principal component analysis are given. One consists of smoothing the data first, and then carrying out an unsmoothed functional principal component analysis (functional PCA). The second involves carrying out the smoothing step within the functional principal component analysis (regularized FPCA). A method for the interpretation of these analyses is expounded and it is shown how the estimates are made. And finally, a discussion on the utilization of classic and functional analysis is included.

2.1. Functional PCA

Let a second-order stochastic process $\{X(t) : t \in [t_1, t_2]\}$ be defined on the probabilistic space (Ω, A, P), continuous in quadratic mean and whose sample functions have squares integrable on $[t_1, t_2]$, i.e., belong to the Hilbert space $L^2[t_1, t_2]$. The inner product of $L^2[t_1, t_2]$ is defined as

$$\langle X, Y\rangle = \int_{t_1}^{t_2} X(t)Y(t)\mathrm{d}t \quad \text{for all } X, Y \in L^2[t_1, t_2].$$

The covariance function of the process is defined as

$$C : [t_1, t_2]x[t_1, t_2] \longrightarrow \mathcal{R}$$
$$(t, s) \longrightarrow C(t, s) = E[\{X(t) - \mu(t)\}\{X(s) - \mu(s)\}],$$

where $\mu(t) = E[X(t)]$.

The covariance operator associated with the process is defined as the integral operator in $L^2[t_1, t_2]$ whose kernel is the covariance function, i.e.,

$$\mathcal{C} : L^2[t_1, t_2] \longrightarrow L^2[t_1, t_2]$$
$$f \longrightarrow \mathcal{C}(f)$$

where $\mathcal{C}(f)(t) = \int_{t_1}^{t_2} C(t, s)f(s)\mathrm{d}s$.

The hth principal component associated with the process $\{X(t)\}$ is defined as $\xi_h = \int_{t_1}^{t_2}(X(t) - \mu(t))f_h(t)\mathrm{d}t$, where f_h is the normalized eigenfunction corresponding to the hth-largest eigenvalue ρ_h of the covariance operator defined above and $\mu(t)$ is the mean function of the process. These principal components have the same optimal properties as in the finite case. Furthermore, ξ_h is the normalized generalized linear combination of the process which has the maximum variance ρ_h out of all generalized linear combinations which are uncorrelated with $\xi_1, \ldots, \xi_{h-1}$ (Deville, 1974). Moreover, the variance explained by the hth principal component is ρ_h/V, where V is the total variance of the process $\{X(t)\}$ given by $V = E(\|X\|^2_{L^2[t_1,t_2]}) = \sum_{h=1}^{\infty}\rho_h$. Therefore, the spectral representation of $\mathcal{C}$ provides the following principal component decomposition for the process—this is known as the Karhunem–Love orthogonal expansion: (Todorovic, 1992)

$$X(t) = \mu(t) + \sum_{h=1}^{\infty} f_h(t)\xi_h, \tag{1}$$

where the infinite series converges in quadratic mean to $X(t)$ uniformly in t. Furthermore, the orthogonal representation for the process in terms of its principal components is optimal, i.e., the best linear approximation of the process $X(t)-\mu(t)$ in the least squares sense is the series (1) truncated at the mth term (Fukunaga, 1990). Hence the variance explained by the model is $\sum_{h=1}^{m}\rho_h/V$ and the minimum mean square error is $\sum_{h=m+1}^{\infty}\rho_h$.

2.2. Regularized FPCA

In the application developed in this section, smoothing is incorporated into the principal component analysis itself (Ramsay and Silverman, 1997). Without any loss of generality, it will be assumed that the mean function is identically null. The challenge is to find functions $g \in L^2[t_1, t_2]$ that provide the maximum variance for $\xi = \int_{t_1}^{t_2} X(t)g(t)\mathrm{d}t$ and that are smoothed. This raises the optimization problem

$$\mathrm{Max}_g \mathrm{var}(\xi) = \mathrm{Max}_g \frac{\langle \mathcal{C}(g), g\rangle}{\|g\|^2 + \lambda\, PEN_2(g)}, \tag{2}$$

where λ is the smoothing parameter (Green and Silverman, 1994), t_1 and t_2 are any two ages, and its integrated squared second derivative is considered as the measure of roughness of a function

$$PEN_2(g) = \|D^2 g\|^2 = \langle g, D^4 g\rangle. \tag{3}$$

By virtue of expression (3), the optimization problem (2) can be expressed as

$$\mathrm{Max}_g \mathrm{var}(\xi) = \mathrm{Max}_g \frac{\langle \mathcal{C}(g), g\rangle}{\langle g, (I + \lambda D^4)g\rangle},$$

whereby the problem is equivalent to that of Ramsay and Silverman (1997)

$$\begin{cases} \mathrm{Max}_g \langle \mathcal{C}(g), g\rangle \\ \text{s.t.} \langle g, (I + \lambda D^4)g\rangle = 1. \end{cases} \tag{4}$$

Hence, in order to obtain the first principal component, $g \in L^2[t_1, t_2]$ must be found, which provides the maximum variance for $\xi_1 = \int_{t_1}^{t_2} X(t)g(t)\mathrm{d}t$ and is smoothed under the terms imposed previously, that is, (4) must be solved. In the case of the hth principal component, which is given in the form $\xi_h = \int_{t_1}^{t_2} X(t)g(t)\mathrm{d}t$, the same problem of maximization must be solved by imposing a further restriction. In other words,

$$\begin{cases} \mathrm{Max}_g \langle \mathcal{C}(g), g\rangle \\ \text{s.t.} \langle g, (I + \lambda D^4)g\rangle = 1 \\ \quad \langle g, f_k\rangle + \lambda\langle D^2 g, D^2 f_k\rangle = 0 \quad \text{para } k = 1, \ldots, h-1. \end{cases} \tag{5}$$

In general, the solution to (4) and (5) is given by solving the equation

$$\mathcal{C}(g) = \rho(I + \lambda D^4)g. \tag{6}$$

Therefore, the problem now is to obtain the eigenvalues ρ and eigenfunctions g of Eq. (6), where the autofunction associated with the largest eigenvalue must be the solution to (4), and is denoted as f_1. The first component is $\xi_1 = \int_{t_1}^{t_2} X(t)f_1(t)\mathrm{d}t$. The solution of (5) for the hth component involves calculating f_h, the eigenfunction corresponding to the hth-largest eigenvalue, by imposing that $< f_h, (I + \lambda D^4)f_h >= 1$ and also that $< f_h, f_k > +\lambda\langle D^2 f_h, D^2 f_k\rangle = 0$ for $k = 1, \ldots, h-1$, the orthogonality condition. Hence, the hth component would be given as $\xi_h = \int_{t_1}^{t_2} X(t)f_h(t)\mathrm{d}t$.

Besse and Ramsay (1986) carried out the estimation of a smoothed eigenfunction from a rather different point of view. In Silverman (1996), a theoretical discussion of the advantages of this approach to the problem can be found. The procedure proposed by Rice and Silverman (1991) is much more complex. To incorporate the penalization of roughness, a step-by-step procedure is considered in the analysis and the use of a different smoothed parameter for each component is proposed. Hence, the more components we want to calculate, the greater the number of problems that remain to be solved.

2.3. Interpretation of functional techniques by means of the correlation function

In this section, a method is described for carrying out the interpretation of the aforementioned functional techniques. To this end, use is made of the correlation function of the stochastic process and of each of the principal components obtained.

In Valderrama et al. (2000) it can be seen that the correlation function between the process and the hth component ($h = 1, 2, 3, 4, \ldots$) is given by

$$r_h(t) = r(X(t), \xi_h) = \frac{\sqrt{\rho_h}}{\sigma(t)} f_h(t),$$

where $X(t)$ is the stochastic process, ρ_h and $f_h(t)$ the eigenvalue and eigenfunction for the hth component respectively, and where $\sigma(t)$ denotes the standard deviation of the original process.

Therefore, in an analogous way to that for the multivariate case, correlations between each principal component and the functions of the original process can be obtained. In addition, we know from the way in which the optimization problem is set out in order to obtain each of the principal components that it is the first principal component which explains a higher percentage of the variability of the process, followed by the second, the third and so on. The percentage of variability explained by the hth component is given by $\frac{\rho_h}{\sum_{h=1}^{\infty}\rho_h}$.

Like for the classic case, the first few eigenfunctions and eigenvalues can be used for data reduction, as part of feature extraction (Rice and Silverman, 1991). The scores of individual curves on the leading eigenfunctions can be used for description, clustering, classification and prediction (Aubrey et al., 1980). As in the classic case, where there is a strong positive correlation between two variables, when one grows then the other also increases, and when one diminishes the other also decreases. Therefore, if in a segment $t \in [t_i, t_j]$ the correlation between the process and the hth component is strongly positive, then what happens on this segment may be explained by the hth principal component. As a result, an increase in the score of the hth component will suppose an increase in $X(t)$, $t \in [t_i, t_j]$. Otherwise, if the correlation is strongly negative, when one variable grows then the other diminishes, and vice versa. In this way we can arrange individuals depending on the value that they take in each of the components obtained. Hence, using the correlation function, we can interpret each of the components, by individually associating them with the behaviour of $X(t)$ in a segment given in $t \in [t_i, t_j] \subset [t_0, t_1]$.

2.4. Estimation of the functional techniques

In order to obtain estimators of eigenvalues and eigenfunctions associated with the sample covariance operator, approximation methods have to be employed (Aguilera, 1993).

In the case of functional PCA, it is necessary to solve

$$\hat{\mathcal{C}}\hat{f}(t) = \int_{t_1}^{t_2} \hat{C}(t,s)\hat{f}(s)\mathrm{d}s = \hat{\rho}\hat{f}(t), \quad t \in [t_1, t_2]. \tag{7}$$

In the same way, in the case of regularized FPCA, we want to solve

$$\hat{\mathcal{C}}\hat{f}(t) = \int_{t_1}^{t_2} \hat{C}(t,s)\hat{f}(s)\mathrm{d}s = \hat{\rho}(I + D^4)\hat{f}(t), \quad t \in [t_1, t_2]. \tag{8}$$

In Ramsay and Silverman (1997) the problems above are solved for both the periodic and the non-periodic case. In the application carried out below, the algorithm developed by the authors for the non-periodic case is used, where the functions are base B-splines (Aguilera et al., 1996a). The B-splines have compact support, that is, they are zero everywhere except over a finite interval. In our case, each B-spline is a cubic spline with support on the interval $[t_{k-2}, t_{k+2}]$, and with shorter support at the ends (Ramsay and Silverman, 1997). Every spline function of a given degree, smoothness, and domain partition can be represented as a linear combination of B-splines of that same degree and smoothness, and over that same partition.

2.5. Considerations of these methods

Over the last decade, studies have focused on the problem that arises when data analysis is carried out from a functional perspective. Some papers (Ramsay and Dalzell, 1991; Rice and Silverman, 1991; Silverman, 1995) have considered versions of principal component analysis for data that may be considered as curves rather than the vectors of classic multivariate analysis. There is a close relationship between these two methods. In the same way as in the classic multivariate case with matrices, the variance–covariance and correlation functions can be difficult to interpret, and do not always give a fully comprehensible presentation of the structure of variability in the observed data directly. Both of the analyses provide a way of looking at covariance structure that can be much more informative and can complement, or even replace altogether, a direct examination of the variance–covariance function or variance–covariance matrix. The objective of these techniques is the same: to summarize the most important information in the data by representing the variables in a limited number of components that explain a maximal amount of variance. That is to say, both methodologies aim not only to reduce dimension but also to reconstruct the complete process between the sampling time points.

A significant, intrinsic difference between the two methods lies in the perception that functional data are observed in the continuum, without noise, whereas traditional data are observed at distributed time points and are often subject to experimental error (Hall et al., 2006). When the data are only observed a finite set of times, as usually happens in practice, and the sampling time points are not evenly spaced, the functional PCA produces stable and consistent estimates of the lead terms of the Karhunem–Love expansion, while the conventional PCA of the observed data gives erroneous results and unstable estimates of process variability (Castro et al., 1986). However, when the temporary observations are equally spaced, Aguilera et al. (1996a) propose applying functional PCA to the natural cubic spline interpolation of the sample paths since it corrects the failings of conventional PCA. Yet another reason for applying the functional procedure instead of the classic version is that the first methodology considers a very special

Table 1
Profiles considered in the study of claims ratio.

Profiles	Gender	Type of car	Geographical area	Claims	Total number of policyholders
1	Female	Intermediate-premium	South	1553	6,687
2	Female	Intermediate-premium	Centre–North	728	3,574
3	Female	Intermediate-premium	Mediterranean	443	2,604
4	Female	Economy	South	1810	9,330
5	Female	Economy	Centre–North	865	5,134
6	Female	Economy	Mediterranean	516	3,658
7	Male	Intermediate-premium	South	8799	43,709
8	Male	Intermediate-premium	Centre–North	3985	22,286
9	Male	Intermediate-premium	Mediterranean	2239	13,649
10	Male	Economy	South	5646	35,794
11	Male	Economy	Centre–North	2506	18,157
12	Male	Economy	Mediterranean	1393	10,609

covariance structure for holding sampled values unrecognized by the conventional methodology. The closest items distributed over time are expected to have greater covariance than those which are more separated. For functional PCA, the order in which the variables are given is crucial, while conventional PCA produces results independent from the order of variables (Ramsay et al., 1994).

On the other hand, it is known that smoothing methods are useful in functional data analysis when preprocessing the data to obtain functional observations. In the case of regularized FPCA, instead of carrying out the smoothing within the functional PCA, the data are smoothed first, and then an unsmoothed functional PCA is carried out. Studies exist which consider that carrying out the smoothing within the functional PCA is more advantageous. Along these lines, in Pezzulli and Silverman (1993), it can be observed that under the hypothesis of normality and very mild conditions on the smoothing parameters, estimated eigenvalues and eigenfunctions are consistent. These authors asymptotically study whether smoothing is advantageous in terms of the mean squared error of the smoothed eigenfunctions and eigenvalues. Smoothing improves the quality of the empirical analysis if the eigenfunctions that correspond to higher variability also have relatively low roughness.

In the application developed in this paper, only the regularized FPCA is shown in detail, due to the similarity of results on applying the other functional procedure.

3. Application to the Spanish automobile insurance market

Premiums, or insurance prices, which the insurers apply, depend directly on the number of accidents and their average cost. For some time, insurance companies have been very interested in a much more thorough analysis in order to increase their margins and competitiveness. These analyses are carried out on the frequency of claims, or on the costs derived from these claims, or a combination of the two. It is on the study of the frequency that we base our analysis.

In the section above, two functional techniques have been expounded, although on occasions the difference between them can be too small to affect any interpretation (Ramsay and Silverman, 1997, 2002). In this work the two aforementioned functional techniques are carried out although only one set of results, those of the regularized FPCA, is shown for the sake of simplicity since the application of the other technique produces similar results. In Section 3.5, the explained variability and the error committed in the reconstruction of the original process are shown when the three procedures, two functional and one traditional, are compared.

3.1. The data

The database (2001–2003) has been obtained from a Spanish insurance company. We restrict ourselves to those policyholders in the said period who drive a private car and who are directly or partly to blame for any accident in the period under study. If other vehicles such as trucks, motorcycles, taxis were considered then the results could be distorted. In addition, we are interested in the characteristics of individuals responsible for an accident. Once the above limitations are applied and the data filtered, 175,191 policyholders and 30,483 claims are found to be the object of our study. A stratification of the sample is carried out as follows: gender, geographical location and type of car. The gender variable is justified by the predictable difference in behaviour between men and women when driving. The geographical area is chosen, since the weather, population density and infrastructure serve as important components in the results. Three different areas are considered: 'Centre–North' (Castilla-Leon, Castilla-La Mancha, Madrid, Aragon, La Rioja, Galicia, Asturias, Cantabria, Navarra and the Basque Country), the Mediterranean area (Catalonia, Valencia, Murcia and Baleares) and the southern zone (Andalucia, Extremadura and Canary Islands). Lastly, the type of car is also studied, by distinguishing between economy and intermediate-premium cars. A profile is designated for any group of individuals who have a number of common characteristics. In our case, the profiles are determined by the variables: gender, geographic location and type of car. As a result of this stratification, 12 groups of individuals were employed.

In Table 1 the different profiles that we are going to work with are shown, including the number of accidents and the number of policyholders in each profile. Of the individuals involved, 17.69% are women, 52.8% correspond to those owning intermediate-premium cars, and as regards the distribution by geographical areas under consideration, in the South 54.5% are insured, in the Centre–North 28.1%, and in the Mediterranean area 17.4%.

Our main intention is to see what claims are made according to the different ages of the driver. Therefore, the risk of occurrence of a claim is studied within each of the profiles and within each of the age segments. This risk is defined as the ratio of the number of accidents in a certain profile for a certain age to the total number of policyholders who are of the same age and profile. Therefore, in our study, the main variable is the risk of occurrence of an accident of the ith profile at the age t, with $i = 1, \ldots, 12$, and $t = \langle 25, 25, 26, \ldots, 70, 71, \rangle 71$.

3.2. Approximation of functional data

The first step in our study is to convert data into a functional form. We are going to work with 12 functions $x_i(t)$, with $i = 1, \ldots, 12$, and $t \in [t_0, t_1]$, where t_0 represents all those insured of under 25 years of age and t_1 those older than 71. We use the basis function methods (Ramsay and Silverman, 1997, 2002), and want to represent the function by a linear combination of K known basis functions (ϕ_k), i.e., $x_i(t) = \sum_{k=1}^{K} c_{ik}\phi_{ik}(t)$, for all $i = 1, \ldots, 12$.

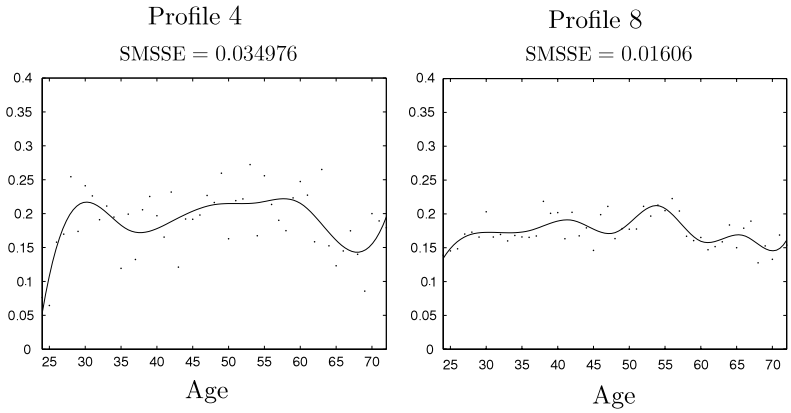

Fig. 1. Risk of occurrence of a claim.

The Fourier series is especially useful for extremely stable functions but this is not our case. We consider cubic B-splines since these combine the computational ease of polynomials with great flexibility (Aguilera et al., 1996b). Therefore, an attempt is made to minimize the ith-profile value

$$\text{SMSSE} = \sum_{j=t_0}^{t_1}\left[y_{ij} - \sum_{k=1}^{K} c_{ik}\phi_{ik}(t_j)\right]^2.$$

The number of base functions, K, must be equal to the number of nodes plus the order of polynomials which intervene minus 2 (Ramsay and Silverman, 1997). Nodes are taken as those drivers 25 years old or younger, of between 26 to 30 years old and the rest of the ages are taken as equally spaced. We therefore work with nine nodes. The order of polynomials is 4 since cubic B-splines are considered so that the curvature of the function under study can be controlled. Hence 11 base functions are used.

Graphs of the approximations of each profile used in the study are presented. We show the original data and curves that approximate such data (Fig. 1). These curves feature the characteristics described above (Table 1). In this way the error committed in each curve is given by applying the criterion of least squares.

3.3. Application of functional PCA to claims

The regularized FPCA is carried out using the 12 curves previously estimated, and thus the size of the problem is reduced. To this end, the algorithm developed by Ramsay (2003) is used.

As a result of the analysis, a series of eigenfunctions is obtained that allows the principal components to be calculated. In practice, estimates are calculated and the hth principal component is given by $\hat{\xi}_h = \int_{t_0}^{t_1} \hat{X}(t)\hat{f}_h(t)\mathrm{d}t$, $h = 1, 2, 3, 4\ldots$, where $\hat{X}(t)$ and $\hat{f}_h$ are estimates of the curves of the original process and of the hth eigenfunction, respectively. In addition, for each profile, the score is found for each of the components. For the hth component, the scores in each of the individuals concerned are $\hat{\xi}_h = (\hat{\xi}_{1h}, \hat{\xi}_{2h}, \ldots, \hat{\xi}_{12h})$ with $\hat{\xi}_{ih} = \int_{t_0}^{t_1} \hat{x}_i(t)\hat{f}_h(t)\mathrm{d}t$ for all $i = 1, \ldots, 12$, where $\hat{f}_h(t)$ and $\hat{x}_i(t)$ are estimates of the hth eigenfunction and of the function corresponding to the hth profile, respectively.

From the way in which the optimization problem is set out, it is known that the first principal component must be that which accumulates the greatest variability of the original process; followed by the second principal component, the third and so on. Information is given about the corresponding scores for each of the principal components for the different profiles (Table 2). Only the first four principal components are shown since with the first three and four components explain 91.99% and 96.07%, of the total variability of the process, respectively, which is a fairly high percentage.

To measure the reconstruction error at each $t \in [t_0, t_1]$, the mean square error is used, given by $\text{MSE}^{(m)}(t) = \frac{1}{12}\sum_{i=1}^{12}[\hat{x}_i(t) - \hat{x}_i{}^{(m)}(t)]^2$, where the estimation of the reconstructed process, using m components, is calculated as $\hat{x}_i{}^{(m)}(t) = \sum_{h=1}^{m}\hat{f}_h(t)\hat{\xi}_{ih}$, where $\hat{f}_h(t)$ is the hth estimated eigenfunction and $\hat{\xi}_{ih}$ is the value taken by the ith observation in the hth component.

3.4. Interpretation of the results

The estimation of the correlation function between the process and the hth component ($h = 1, 2, 3, 4$) is given by $\hat{r}_h(t) = r(\hat{X}(t), \hat{\xi}_h) = \frac{\sqrt{\hat{\rho}_h}}{\hat{\sigma}(t)}\hat{f}_h(t)$, where $\hat{X}(t)$ is the estimate of the curves obtained for different profiles, $\hat{\rho}_h$ and $\hat{f}_h(t)$ are the estimated eigenvalue and eigenfunction for the hth component, respectively, and where $\hat{\sigma}(t)$ denotes the estimation of the standard deviation of the original process. It is generally considered to be a strong correlation if this is greater than or equal to 0.7 in absolute terms (Hair, 2006). In Fig. 2, the correlation functions between the estimated process and the first four components are presented.

One can observe that the first component is highly positively correlated with the segment that runs from 31 to 67 year olds inclusively, and with drivers over 71 years of age. The information obtained for those older than 71 years of age is not very reliable, since these were grouped together in our study. Therefore, the statistical significance detected for the first component in that age bracket is not taken into account. The second and third components are very positively correlated with the age groups ranging from 68 to 70 year olds and from 27 to 30 year olds, respectively. And finally, the fourth component is not strongly correlated with any age group. However, with the first three principal components, 91.99% of the total variability of the process is explained.

Furthermore, the twelve ratings corresponding to each of the observations under consideration are known. Thus, it can be deduced when, in the age bracket of 31–67 years old, the risk of occurrence of an accident is higher or lower within the different profiles considered. However, if the correlation were negative, the profiles that take the highest scores would behave in such a way that the risk of occurrence of an incident in this age range would be lower with respect to the profiles that take lower values in this variable. Scores of each profile in the second and third component indicate the behaviour in the case of a claim in the age groups ranging from 68 to 70 years old and from 27 to 30 years old, respectively.

Table 2
Scores of different profiles in each of the components.

Profiles component	1st component	2nd component	3rd component	4th component
1	0.454173	−0.123593	0.027439	0.008693
2	0.222051	0.169771	−0.127397	−0.032286
3	0.003438	−0.122665	−0.146860	0.038278
4	0.131653	−0.011524	0.101664	−0.018921
5	−0.004001	−0.109032	−0.013333	−0.098486
6	−0.231211	0.072026	−0.025004	−0.080370
7	0.167503	0.107711	0.098156	0.026214
8	0.006047	0.073371	0.022335	0.042549
9	−0.094135	0.016874	−0.053302	0.089978
10	−0.110139	−0.001048	0.041606	−0.015850
11	−0.256171	−0.043447	0.041342	0.026575
12	−0.289207	−0.028445	0.033354	0.013626

Table 3
Description of the profiles ranked by scores in the first three components.

Profiles 1st CP	Characteristics	Profiles 2nd CP	Characteristics	Profiles 3rd CP	Characteristics
12	(M, E, M)	1	(F, I-P, S)	3	(F, I-P, M)
11	(M, E, C–N)	3	(F, I-P, M)	2	(F, I-P, C–N)
6	(F, E, M)	5	(F, E, C–N)	9	(M, I-P, M)
10	(M, E, S)	11	(M, E, C–N)	6	(F, E, M)
9	(M, I-P, M)	12	(M, E, M)	5	(F, E, C–N)
5	(F, E, C–N)	4	(F, E, S)	8	(M, I-P, S)
3	(F, I-P, M)	10	(M, E, S)	1	(F, I-P, S)
8	(M, I-P, C–N)	9	(M, I-P, M)	12	(M, E, M)
4	(F,E, S)	6	(F, E, M)	11	(M, E, C–N)
7	(M, I-P, S)	8	(M, I-P, C–N)	10	(M, E, S)
2	(F, I-P, C–N)	7	(M, I-P, S)	7	(M, I-P, S)
1	(F, I-P, S)	2	(F, I-P, C–N)	4	(F, E, S)

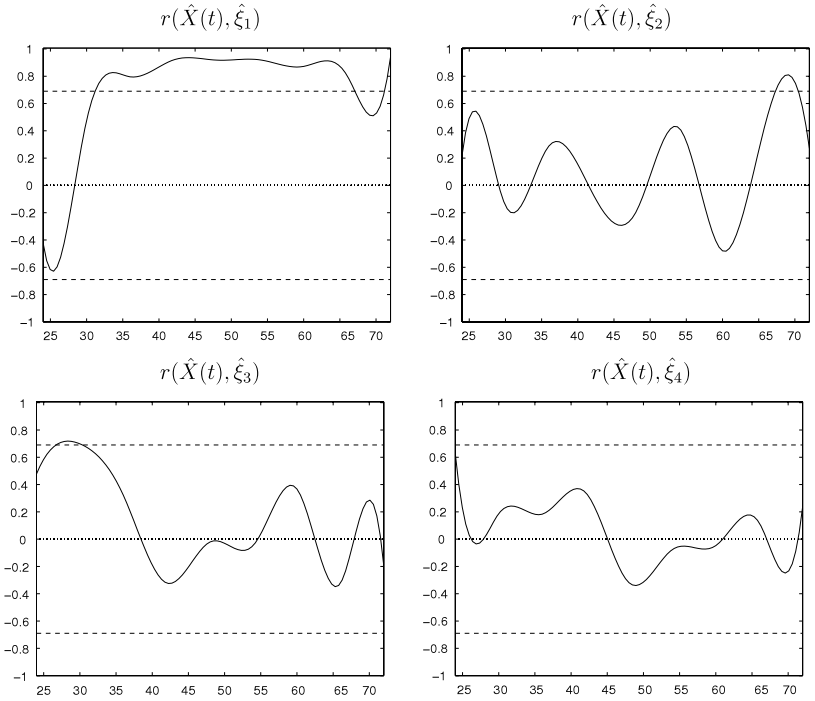

Fig. 2. Correlation functions.

The profiles are presented in Table 3 arranged from the lowest to the highest scores in the first three components. As for the first component, it can be observed that those who drive an economy car generally have a lower score in this component. The two lowest ratings correspond to men who drive economy cars and whose place of residence is the Mediterranean or the Centre–North. By contrast, the two highest values in this variable correspond to women who drive intermediate-premium cars and reside in the South or the Centre–North. When considering the second component, which describes the behaviour of 68 to 70 year olds, the best-behaved individuals are women with intermediate-premium cars who reside in the South and the Mediterranean. As for the third component, the behaviour in the age group ranging from 27 to 30 years old is described, and those individuals who are best behaved are women with intermediate-premium cars who are residents in the Mediterranean or in the Centre–North.

In Figs. 3 and 4 graphic representations of the scores are presented of the first component together with the second or third

Table 4
Variance explained by the principal components.

PCA			Functional PCA		Regularized FPCA	
No. p.c.	var. expl. (%)	Acum. var. expl. (%)	var. expl. (%)	Acum. var. expl. (%)	var. expl. (%)	Acum. var. expl. (%)
1	34.14	34.14	70.13	70.13	70.19	70.19
2	16.90	51.04	12.86	82.99	12.86	83.05
3	15.74	66.79	8.96	91.95	8.94	91.99
4	9.22	76.01	4.09	96.04	4.08	96.07
5	7.77	83.78	1.56	97.6	1.59	97.61
6	5.27	89.05	1.30	98.9	1.29	98.90
7	4.69	93.75	0.46	99.36	0.46	99.36
8	2.56	96.32	0.32	99.68	0.32	99.67
9	1.99	98.31	0.19	99.87	0.19	99.87

Table 5
Mean square error ($MSE^{(m)}(t)$) using m components under different techniques.

	25 years old	36 years old	41 years old	52 years old	61 years old	68 years old
PCA (1)	0.28773	0.20174	0.24561	0.31188	0.28319	0.28689
PCA (2)	0.22086	0.18882	0.24234	0.30105	0.21744	0.26604
PCA (3)	0.17500	0.18044	0.23928	0.29069	0.21702	0.24331
PCA (4)	0.15238	0.17978	0.23725	0.29029	0.21672	0.23909
Functional PCA (1)	0.01412	0.02647	0.03278	0.03806	0.02851	0.01984
Functional PCA (2)	0.01393	0.02642	0.03275	0.03797	0.02827	0.01915
Functional PCA (3)	0.01357	0.02636	0.03266	0.03798	0.02810	0.01908
Functional PCA (4)	0.01335	0.02632	0.03257	0.03792	0.02810	0.01903
Regularized FPCA (1)	0.01412	0.02647	0.03278	0.03806	0.02851	0.01984
Regularized FPCA (2)	0.01393	0.02642	0.03275	0.03797	0.02827	0.01915
Regularized FPCA (3)	0.01357	0.02636	0.03266	0.03798	0.02810	0.01908
Regularized FPCA (4)	0.01335	0.02632	0.03257	0.03792	0.02810	0.01903

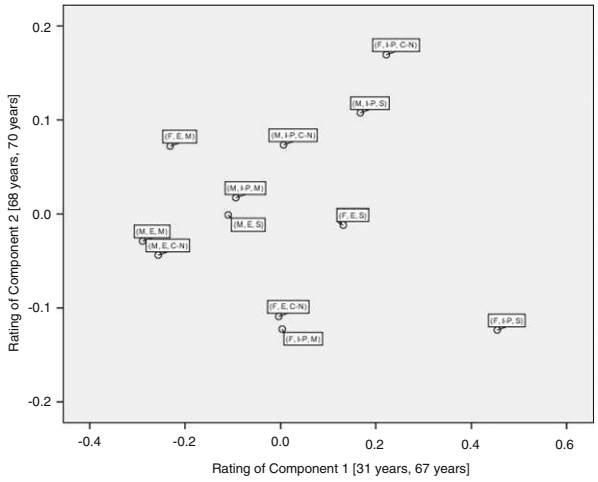

Fig. 3. Ratings of Component 1 versus Component 2.

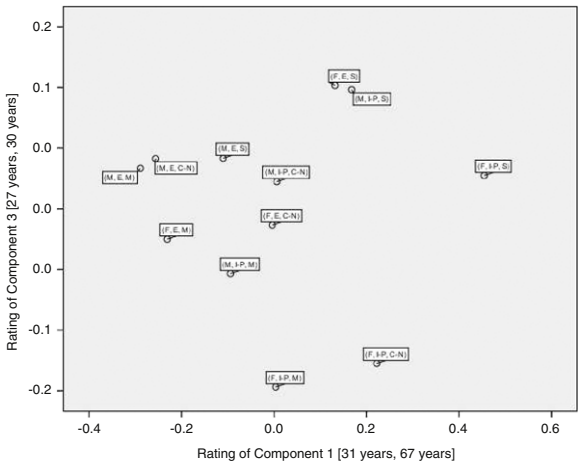

Fig. 4. Ratings of Component 1 versus Component 3.

components respectively. In this way the behaviour of individual profiles can be observed in unison.

3.5. Comparison of the results obtained

In the comparison between the traditional method (conventional PCA) and the two functional methods (functional PCA and regularized FPCA) in our empirical study, it was found that the traditional technique needed a larger number of components than those functional procedures in order to explain the same amount of variability. In the functional case, with only three or four components, more than 91% or 96% could be explained, while in the traditional case at least seven or eight components were needed for this to occur. It was then observed that in the functional case, with a lesser number of components, much more information was gathered. In Table 4, the results for the first nine components are given.

Furthermore, when the mean square error ($MSE^{(m)}(t)$) using m components was calculated in order to evaluate the reliability of the process reconstruction, it was observed that a major error was always committed in the discrete case. In Table 5, some values are given of this measure for each of the techniques.

However, in the classic case, since the chronological order of the variables has been disregarded, the interpretation of the data is not as useful for the company as that from the functional case.

As regards the explained variability and the reconstruction of the process by means of the principal components, very similar results were obtained with the two functional procedures.

4. Conclusion

Our orientation in this paper has been methodological rather than theoretical. We have shown how the functional techniques in the principal component analysis may be useful in the study of

insurance claims ratios. We have shown both the analogies and the differences between conventional PCA and functional procedures (functional PCA and regularized FPCA). The application of these functional techniques to a Spanish sample reveals a number of interesting findings. A series of profiles is considered on the basis of gender, vehicle type and geographic location. The function that is studied is the risk of occurrence of a claim in the light of the variable age of the driver. The regularized FPCA technique is applied, which is widely used in other areas when the study of the phenomenon is a function rather than a p-dimensional vector. And thus we have been able to rank the various profiles according to their greater or lower score in the components obtained by applying the technique. Making use of the correlation function, we can identify the information that each principal component yields, in the same way as in the classic case. In this way the behaviour of individual policyholders in each age group identified by the technique is given. With this information in our empirical work, results are obtained such as that in the age bracket of 27–30 years old, the best performance in terms of claims ratio is held by women with intermediate-premium cars in the Mediterranean area. For the age bracket of 31–67 years of age, the best result is given by men with economy cars in the Mediterranean area. And finally, the best performance in the segment of 68 to 70 year olds is again held by women with intermediate-premium cars in the South.

Furthermore, we carry out a comparison of the three procedures and deduce that, with the functional methods, much more information on the process under study is gathered in a smaller number of components, and therefore, a lesser error is committed on reconstructing the original process.

We believe that this kind of analysis could be of considerable interest for insurance companies, since it allows them to offer a product at a better-adjusted price in accordance with the profile of the insured.

References

Aguilera, A.M., 1993. Metodos de aproximacion de estimadores en el ACP de un proceso estocastico. Tesis Doctoral, Universidad de Granada.

Aguilera, A.M., Gutirrez, R., Valderrama, M.J., 1996a. Approximations of estimators in the PCA of a stochastic process using B-splines. Communational Statistics (simulation) 25 (3), 671–690.

Aguilera, A.M., Ocaa, F.A., Valderrama, M.J., 1996b. Analisis en componentes principales de un proceso estocastico con funciones muestrales escalonadas. Qestiio 20 (1), 7–28.

Artis, M., Ayuso, M., Guillen, M., 1999. Modeling different types of automobile insurance fraud behaviour in the Spanish market. Mathematics and Economics 24, 67–81.

Artis, M., Ayuso, M., Guillen, M., 2002. Detection of automobile insurance fraud with discrete choice models and misclassified claims. Journal of Risk and Insurance 69 (3), 325–340.

Aubrey, D.G., Inman, D.L., Winant, C.D., 1980. The statistical prediction of beach changes in Southern California. Journal of Geophysical Research 85, 3264–3276.

Besse, P., Ramsay, J.O., 1986. Principal components analysis of sampled functions. Psychometrika 51 (2), 285–311.

Boj, E., Claramunt, M.M., Fortiana, J., 2004. Analisis multivariante aplicado a la seleccion de factores de riesgo en la tarificacion. Cuadernos de la Fundacion Mapfre estudios. Instituto de Ciencias del Seguro, Madrid.

Castro, P.E., Lawton, W.H., Sylvestre, E.A., 1986. Principal modes of variation for processes with continuous sample curves. Technometrics 86, 329–337.

Chiappori, P.A., Salanie, B., 2000. Testing for asymmetric information in insurance markets. Journal of Political Economics 108 (1), 56–78.

Deville, J.C., 1974. Methodes statistiques et numeriques de l'analyse harmonique. Annales de l'INSEE 15, 3–101.

Dionne, G.C., Gourieroux, Vanasse, C., 2001. Testing for evidence of adverse selection in the automobile insurance market: A comment. Journal of Political Economy 109 (2), 444–453.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition. Academic Press.

Green, P.J., Silverman, B.W., 1994. Nonparametric regression and generalized linear models. A roughness penalty approach. Chapman and Hall/CRC.

Guillen, M., Ayuso, M., Bermudez, L., Morillo, I., 2005. El Seguro de automoviles: estado actual y perspectiva de la tecnica actuarial. Fundacion Mapfre estudios. Instituto de Ciencias del Seguro, Madrid.

Hair, J., 2006. Multivariate Data Analysis. Prentice Hall.

Hall, P., Miller, H.G., Wang, J.L., 2006. Properties of principal component methods for functional and longitudinal data analysis. The Annals of Statistics 34 (3), 1493–1517.

Pezzulli, S., Silverman, B.W., 1993. Some properties of smoothed principal components analysis for functional data. Computational Statistics 8, 1–16.

Pujol, M., Bolance, C., 2004. La matriz valor fidelidad en el analisis de los asegurados en el ramo del automovil. Fundacin Mapfre estudios. Instituto de Ciencias del Seguro.

Ramsay, J.O., 2003. R and S-PLUS Functions for Functional Data Analysis. McGill University.

Ramsay, J.O., Altman, N., Bock, R.D., 1994. Variation in height acceleration in the Fels growth data. The Canadian Journal of Statistics 22 (1), 89–102.

Ramsay, J.O., Dalzell, C.J., 1991. Some tools for functional data analysis (with discussion). Journal of the Royal Statistical Society Series C 44, 17–30.

Ramsay, J.O., Silverman, B.W., 1997. Functional Data Analysis. In: Springer Series in Statistics.

Ramsay, J.O., Silverman, B.W., 2002. Applied Functional Data Analysis. In: Springer Series in Statistics.

Rice, J.A., Silverman, B.W., 1991. Estimating the mean and covariance structure nonparametrically when the data are curves. Journal of the Royal Statistical Society Series B 53, 233–243.

Silverman, B.W., 1995. Incorporating parametric effects into functional principal components analysis. Journal of the Royal Statistical Society Series B 57 (4), 673–689.

Silverman, B.W., 1996. Smoothed functional principal components analysis by choice of norm. Annals of Statistics 24 (1), 1–24.

Todorovic, P., 1992. An Introduction to Stochastic Processes and Their Applications. Springer-Verlag, New York.

Valderrama, M.J., Aguilera, A.M., Ocana, F.A., 2000. Prediccion dinamica mediante analisis de datos funcionales. La Murralla-Hesperides.