8

Principal components analysis for functional data

8.1 Introduction

For many reasons, principal components analysis (PCA) of functional data is a key technique to consider. First, our own experience is that, after the preliminary steps of registering and displaying the data, the user wants to explore that data to see the features characterizing typical functions. Some of these features are expected to be there, for example the sinusoidal nature of temperature curves, but other aspects may be surprising. Some indication of the complexity of the data is also required, in the sense of how many types of curves and characteristics are to be found. Principal components analysis serves these ends admirably, and it is perhaps also for these reasons that it was the first method to be considered in the early literature on FDA.

Just as for the corresponding matrices in the classical multivariate case, the variance-covariance and correlation functions can be difficult to interpret, and do not always give a fully comprehensible presentation of the structure of the variability in the observed data directly. The same is true, of course, for variance-covariance and correlation matrices in classical multivariate analysis. A principal components analysis provides 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.

PCA also offers an opportunity to consider some issues that reappear in subsequent chapters. For example, we consider immediately how PCA is

defined by the notion of a linear combination of function values, and why this notion, at the heart of most of multivariate data analysis, requires some care in a functional context. A second issue is that of *regularization*; for many data sets, PCA of functional data is more revealing if some type of smoothness is required of the principal components themselves. We consider this topic in detail in Chapter 9.

8.2 Defining functional PCA

8.2.1 PCA for multivariate data

The central concept exploited over and over again in multivariate statistics is that of taking a linear combination of variable values,

$$f_i = \sum_{j=1}^{p} \beta_j x_{ij},\ i = 1, \ldots, N, \tag{8.1}$$

where β_j is a weighting coefficient applied to the observed values x_{ij} of the jth variable. We can express (8.1) as

$$f_i = \beta' x_i,\ i = 1, \ldots, N, \tag{8.2}$$

where β is the vector $(\beta_1, \ldots, \beta_p)'$ and x_i is the vector $(x_{i1}, \ldots, x_{ip})'$.

In the multivariate situation, we choose the weights so as to highlight or display types of variation that are very strongly represented in the data. Principal components analysis can be defined in terms of the following stepwise procedure, which defines sets of normalized weights that maximize variation in the f_i's:

1. Find the weight vector $\boldsymbol{\xi}_1 = (\xi_{11}, \ldots, \xi_{p1})'$ for which the linear combination values
$$f_{i1} = \sum_j \xi_{j1} x_{ij} = \boldsymbol{\xi}_1' x_i$$
have the largest possible mean square $N^{-1} \sum_i f_{i1}^2$ subject to the constraint
$$\sum_j \xi_{j1}^2 = \|\boldsymbol{\xi}_1\|^2 = 1.$$

2. Carry out second and subsequent steps, possibly up to a limit of the number of variables p. On the mth step, compute a new weight vector $\boldsymbol{\xi}_m$ with components ξ_{jm} and new values $f_{im} = \boldsymbol{\xi}_m' x_i$. Thus, the values f_{im} have maximum mean square, subject to the constraint $\|\boldsymbol{\xi}_m\|^2 = 1$ and the $m - 1$ additional constraint(s)
$$\sum_j \xi_{jk} \xi_{jm} = \boldsymbol{\xi}_k' \boldsymbol{\xi}_m = 0,\ k < m.$$

The motivation for the first step is that by maximizing the mean square, we are identifying the strongest and most important mode of variation in the variables. The unit sum of squares constraint on the weights is essential to make the problem well defined; without it, the mean squares of the linear combination values could be made arbitrarily large. On second and subsequent steps, we seek the most important modes of variation again, but require the weights defining them to be orthogonal to those identified previously, so that they are indicating something new. Of course, the amount of variation measured in terms of $N^{-1}\sum_i f_{im}^2$ will decline on each step. At some point, usually well short of the maximum index p, we expect to lose interest in modes of variation thus defined.

The definition of principal components analysis does not actually specify the weights uniquely; for example, it is always possible to change the signs of all the values in any vector $\boldsymbol{\xi}_m$ without changing the value of the variance that it defines.

The values of the linear combinations f_{im} are called *principal component scores* and are often of great help in describing what these important components of variation mean in terms of the characteristics of specific cases or replicates.

To be sure, the mean is a very important aspect of the data, but we already have an easy technique for identifying it. Therefore, we usually subtract the mean for each variable from corresponding variable values before doing PCA. When this is done, maximizing the mean square of the principal component scores corresponds to maximizing their sample variance.

8.2.2 Defining PCA for functional data

How does PCA work in the functional context? The counterparts of variable values are function values $x_i(s)$, so that the discrete index j in the multivariate context has been replaced by the continuous index s. When we were considering vectors, the appropriate way of combining a weight vector β with a data vector x was to calculate the inner product

$$\beta' x = \sum_j \beta_j x_j.$$

When β and x are functions $\beta(s)$ and $x(s)$, summations over j are replaced by integrations over s to define the inner product

$$\int \beta x = \int \beta(s)x(s)\,ds. \tag{8.3}$$

Within the principal components analysis, the weights β_j now become functions with values $\beta_j(s)$. Using the notation (8.3), the principal

component scores corresponding to weight β are now

$$f_i = \int \beta x_i = \int \beta(s) x_i(s)\, ds. \tag{8.4}$$

For the rest of our discussion, we will often use the short form $\int \beta x_i$ for integrals in order to minimize notational clutter.

In the first functional PCA step, the weight function $\xi_1(s)$ is chosen to maximize $N^{-1}\sum_i f_{i1}^2 = N^{-1}\sum_i (\int \xi_1 x_i)^2$ subject to the continuous analogue $\int \xi_1(s)^2 ds = 1$ of the unit sum of squares constraint. This time, the notation $\|\xi_1\|^2$ is used to mean the squared norm $\int \xi_1(s)^2 ds = \int \xi_1^2$ of the *function* ξ_1.

Postponing computational details until Section 8.4, now consider as an illustration in the upper left panel in Figure 8.1. This displays the weight function ξ_1 for the Canadian temperature data after the mean across all 35 weather stations has been removed from each station's monthly temperature record. Although ξ_1 is positive throughout the year, the weight placed on the winter temperatures is about four times that placed on summer temperatures. This means that the greatest variability between weather stations will be found by heavily weighting winter temperatures, with only a light contribution from the summer months; Canadian weather is most variable in the wintertime, in short. Moreover, the percentage 89.3% at the top of the panel indicates that this type of variation strongly dominates all other types of variation. Weather stations for which the score f_{i1} is high will have much warmer than average winters combined with warm summers, and the two highest scores are in fact assigned to Vancouver and Victoria on the Pacific Coast. To no one's surprise, the largest negative score goes to Resolute in the High Arctic.

As for multivariate PCA, the weight function ξ_m is also required to satisfy the orthogonality constraint(s) $\int \xi_k \xi_m = 0,\ k < m$ on subsequent steps. Each weight function has the task of defining the most important mode of variation in the curves subject to each mode being orthogonal to all modes defined on previous steps. Note again that the weight functions are defined only to within a sign change.

The weight function ξ_2 for the temperature data is displayed in the upper right panel of Figure 8.1. Because it must be orthogonal to ξ_1, we cannot expect that it will define a mode of variation in the temperature functions that will be as important as the first. In fact, this second mode accounts for only 8.3% of the total variation, and consists of a positive contribution for the winter months and a negative contribution for the summer months, therefore corresponding to a measure of uniformity of temperature through the year. On this component, one of the highest scores f_{i2} goes to Prince Rupert, also on the Pacific coast, for which there is comparatively low discrepancy between winter and summer. Prairie stations such as Winnipeg, on the other hand, have hot summers and very cold winters, and receive large negative second component scores.

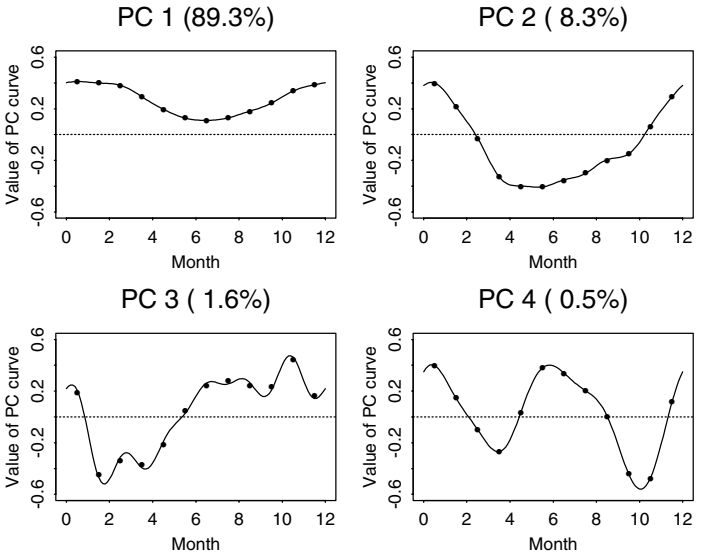

Figure 8.1. The first four principal component curves of the Canadian temperature data estimated by two techniques. The points are the estimates from the discretization approach, and the curves are the estimates from the expansion of the data in terms of a 12-term Fourier series. The percentages indicate the amount of total variation accounted for by each principal component.

The third and fourth components account for small proportions of the variation, since they are required to be orthogonal to the first two as well as to each other. At this point they are difficult to interpret, but we look at techniques for understanding them in Section 8.3.

Displays such as Figure 8.1 can remind one of the diagrams of modes of vibration in a string fixed at both ends always found in introductory physics texts. The first and dominant type is simple in structure and resembles a single cycle of a sine wave. Subdominant or higher order components are also roughly sinusoidal, but with more and more cycles. With this analogy in mind, we find the term *harmonics* evocative in referring to principal components of variation in curves in general.

8.2.3 Defining an optimal empirical orthonormal basis

There are several other ways to motivate PCA, and one is to define the following problem: We want to find a set of exactly K orthonormal functions ξ_m so that the expansion of each curve in terms of these basis functions approximates the curve as closely as possible. Since these basis functions

are orthonormal, it follows that the expansion will be of the form

$$\hat{x}_i(t) = \sum_{k=1}^{K} f_{ik}\xi_k(t),$$

where f_{ik} is the principal component value $\int x_i\xi_k$. As a fitting criterion for an individual curve, consider the integrated squared error

$$\|x_i - \hat{x}_i\|^2 = \int [x(s) - \hat{x}(s)]^2\, ds$$

and as a global measure of approximation,

$$\texttt{PCASSE} = \sum_{i=1}^{N} \|x_i - \hat{x}_i\|^2. \tag{8.5}$$

The problem is then, more precisely, what choice of basis will minimize the error criterion (8.5)?

The answer, it turns out, is precisely the same set of principal component weight functions that maximize variance components as defined above. For this reason, these functions ξ_m are referred to in some fields as *empirical orthonormal functions*, because they are determined by the data they are used to expand.

8.2.4 PCA and eigenanalysis

In this section, we investigate another characterization of PCA, in terms of the eigenanalysis of the variance-covariance function or operator.

Assume for this section that our observed values, x_{ij} in the multivariate context and $x_i(t)$ in the functional situation, result from subtracting the mean variable or function values, so that their sample means $N^{-1}\sum_i x_{ij}$, or cross-sectional means $N^{-1}\sum_i x_i(t)$, respectively, are zero.

Texts on multivariate data analysis tend to define principal components analysis as the task of finding the eigenvalues and eigenvectors of the covariance or correlation matrix. The logic for this is as follows. Let the $N \times p$ matrix $\mathbf{X}$ contain the values x_{ij} and the vector $\boldsymbol{\xi}$ of length p contain the weights for a linear combination. Then the mean square criterion for finding the first principal component weight vector can be written as

$$\max_{\boldsymbol{\xi}'\boldsymbol{\xi}=1} N^{-1}\boldsymbol{\xi}'\mathbf{X}'\mathbf{X}\boldsymbol{\xi} \tag{8.6}$$

since the vector of principal component scores f_i can be written as $\mathbf{X}\boldsymbol{\xi}$.

Use the $p \times p$ matrix $\mathbf{V}$ to indicate the sample variance-covariance matrix $\mathbf{V} = N^{-1}\mathbf{X}'\mathbf{X}$. (One may prefer to use a divisor of $N-1$ to N since the means have been estimated, but it makes no essential difference to the principal components analysis.) The criterion (8.6) can now be expressed