Geophysical Research Abstracts
Vol. 12, EGU2010-10344-1, 2010
EGU General Assembly 2010

Functional Principal Components direction to cluster earthquake waveforms

Giada Adelfio, Marcello Chiodi, Antonino D'Alessandro, and Dario Luzio
University of Palermo, Dipartimento di Science Statistiche e Matematiche "S. Vianelli", Palermo, Italy (adelfio@unipa.it)

Looking for curves similarity could be a complex issue characterized by subjective choices related to continuous transformations of observed discrete data (Chiodi, 1989).

In this paper we combine the aim of finding clusters from a set of individual curves to the functional nature of data, applying a variant of a k-means algorithm based on the principal component rotation of data. We apply a classical clustering method to rotated data, according to the direction of maximum variance.

A k-means clustering algorithm based on PCA rotation of data is proposed, as an alternative to methods that require previous interpolation of data based on splines or linear fitting (García-Escudero and Gordaliza (2005), Tarpey (2007), Sangalli *et al.* (2008)).

1 Principal components analysis for functional data

When data are observed as function of time (such as financial time series, temperature recorded by some central source, etc.), we refer to as functional data. Since in many statistical applications realizations of continuous time series are available as observations of a process recorded in discrete time intervals, one crucial point is to convert discrete data to continuous functions, that is from vectors to curves or more generally functions in $\mathbb{R}^d, d \geq 1$.

Functional data are a very convenient approach to deal with data depending on time, providing theoretical tools indispensable to analyze observations of the process recorded in discrete time intervals and convert them to continuous functions.

When we talk about functional data, then we refer to n pairs of points (t_i, y_i) where y_i is the value of an observable variable x at time t_i, and we focus on a set of functions defined on $[0, T]$, such that:

$$\{y_i = x_i(t); i = 1, 2, \cdots, I; 0 \leq t \leq T\}$$

Therefore, assuming that a functional datum for replication i arrives as a set of discrete measured values $y_{i1}, y_{i2}, \cdots, y_{in}$ the first task is to convert these values to a function x_i with values $x_i(t)$ computable for any t, called functional objects. The conversion from discrete data to functions may involve smoothing (Ramsay and Silverman (2006)). One smoothing procedure often used is obtained representing each function as the linear combination of K base functions ϕ_k:

$$x_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t)$$

The conversion of functional data to functional objects requires to store the coefficients $\{c_{ik}\}$ in a $I \times K$ matrix; a main issue is the choice of the basis, such as Fourier basis useful for periodic data, B-splines, exponential basis.

Let $\{y_{ij}\}$ be the observed value of the i^{th} function at time t_j, the basis representation of $x(t)$ can be obtained by a least squares criterion, such that: $\min_{c_{ik}} \sum_{j=1}^{J} \left(y_{ij} - \sum_{k=1}^{K} c_{ik}\phi_k(t)\right)^2$ The smoothness degree depends on K, since small (large) values of K induce more (less) smoothed curves.

The best known basis expansion is obtained by the Fourier series, although it is useful when the observed functions are periodic and do not fluctuate in any particular interval more rapidly than the basis elsewhere. Therefore sometimes a roughness penalty approach can be used, that retains the advantages of the basis function and local expansion smoothing (like kernel and local polynomial fitting techniques), but overcome some of their limitations.

The spline smoothing method estimates a curve from observations with the aim to ensure both regularity and goodness of fit to the data, that is between variance and bias. For this purpose penalty terms can be added to the residual sum of squares.

Principal Components Analysis (PCA) is aimed to the reduction of the original set of variables $X_1, X_2, \cdots, X_q$ to a smaller set $f_1, f_2, \cdots, f_p (p < q)$ of linear orthogonal combinations

$$f_i = \sum_{j=1}^{q} \beta_{ij} x_j \ \ i = 1, 2, \cdots, p,$$

and able to display types of variation that are strongly represented in the data. The corresponding PCA for functional data (FPCA) is now reviewed to introduce some notation used throughout the paper. In the functional context the value of the j^{th} variable on the i^{th} unit is now a function of the time $x_i(t), i = 1, \cdots, p$; therefore the principal components are now function values and the discrete index j is now replaced by the continuous index s, such that:

$$f_i = \int \beta(s) x_i(s) ds \tag{1}$$

with $\beta(s)$ weight functions. Each functional principal component is obtained by maximizing:

$$\frac{1}{p} \sum_i f_i^2$$

and satisfying orthogonal constrains (Ramsay and Silverman (2006)).

2 Some analysis for waveforms data

The goal of this section is to use the functional analysis to highlight common characteristics of data and to summarize these characteristics by few components.

Waveforms correlation techniques have been introduced to characterize the degree of event similarity (Mezcua and Rueda (1994), Menke (1999)) and in facilitating more accurate relative locations within similar event clusters by providing more precise timing of P and S arrivals (Gillard and Okubom (1996), Phillips (1997)).

In this paper, FPCA together with a clustering approach are used to highlight common characteristics of data and to summarize these characteristics by few components. We analyze 32 earthquakes occurred in the South Tyrrhenian Sea in 2002 through their waveforms recorded by the station of Augusta (South East Sicily) and observed for 10 seconds, by intervals of 0.02 seconds. To construct functional data objects we specify a set of basis functions and a set of coefficients defining a linear combination of these basis functions. We use B-spline basis (Hastie, 1997), since they are often used for non-periodic functions. B-spline basis functions are polynomial segments jointed end-to-end at argument values called knots, breaks or join points; the segments have specifiable smoothness across these breaks. B-spline basis functions have the advantages of very fast computation and great flexibility.

Through this analysis we want to summarize the information given by all the 32 waveforms data in few curves, explaining a large part of the variability related to each event. For this purpose we use a variation of the trimmed k-means clustering algorithm proposed by García-Escudero and Gordaliza (2005) to find clusters of events according to the FPCA directions. Their algorithm is a kind of robust version of k-means methodology through a trimming procedure. In few words, given a q-variate data sample $X_1, \ldots, X_n$ with $X_i = X_{i1}, \ldots, X_{iq}$, and fixed the number of clusters k the trimmed k-means clustering algorithm looks for the k centers $C_1, \cdots, C_k$ that are solution of the minimization problem:

$$O_k(\alpha) = \min_{Y} \min_{C_1, \cdots, C_k} \frac{1}{[n(1-\alpha)]} \sum_{X_i \in Y} \inf_{1 \le j \le k} ||X_i - C_j||^2 \tag{2}$$

with α the trimming size and Y is the set of subsets of $X_1, \cdots, X_n$ containing $[n(1-\alpha)]$ data points, where $[x]$ is the integer part of x. This method allocates each non-trimmed observation to the cluster identified by its closest center C_j, dealing with possible outliers by the given proportion of observations to be discarded α. Their curve clustering procedure is based on a least-squares fit to cubic B-spline q-dimensional functions bases, applying the trimmed k-means clustering (2) on the resulting coefficients.

Here we use the PCA functions defined in (1) to get a linear approximation of each curve by a finite p dimensional vector of coefficients defined by the PCA scores. The number of starting clusters k is determined on the basis of a simple procedure based on the scores volume, such that we assign events to the clusters defined by events that have a distance less than a fixed threshold in the space of PCA scores. Once k is obtained we use the modified trimmed k-means algorithm, that consider the matrix of PCA scores instead of the coefficients of a linear fitting to B-spline bases. In other words we look for clusters in the direction of data with largest variance. In particular for a better fitting we consider up to the first four harmonics, that explain almost the 60% of the global variance of data, although from empirical results, we observed that the choice of the number of harmonics does not influence significantly neither the value of the object function nor in terms of clustering results.

Fixing $\alpha = 0.05$, our procedure defines six clusters of earthquakes with common features and two outlier curves (i.e. events), with a clear advantage related to an immediate use of PCA for functional data avoiding some objective choices related to splines fitting, since cluster results for the different fits based on B-splines, Fourier or power basis, can differ considerably (Tarpey, 2007).

References

[1] Chiodi, M. (1989). The clustering of longitudinal data when time series are short. *Multivariate data analysis*, pages 445–453.

[2] García-Escudero, L. A. and Gordaliza, A. (2005). A proposal for robust curve clustering. *Journal of classification*, **22**, 185–201.

[3] Gillard, D., R. A. and Okubom, P. (1996). Highly concentrated seismicity caused by deformation of kilauea's depp magma system. *Nature*, **384**, 343–346.

[4] Hastie, T. J. (1997). *Generalized additive models.* Chapman and Hall, London.

[5] Menke, W. (1999). Using waveform similarity to constrain earthquake locations. *Bull.Seismol. Soc. Am.*, **89**, 1143–1146.

[6] Mezcua, J. and Rueda, J. (1994). Earthquake relative location based on waveform similarity. *Tectonophysics*, **233**, 253–263.

[7] Phillips, W. S., H. L. F. J. (1997). Detailed joint structure in a geothermal reservoir from studies of induced microearthquake studies. *Journal of Geophysical Research*, **102**, 745–763.

[8] Ramsey, J. O. and Silverman, B. W. (2006). *Functional Data Analysis. Springer, New York.*

[9] Sangalli, L. M., Secchi, P., Vantini, S., and Vitelli, V. (2008). K-means alignment for curve clustering. *MOX (Modeling and Scientific Computing)-Report*, **13**.

[10] Tarpey, T. (2007). Linear transformations and the k-means clustering alghoritm: applications to clustering curves. *American statistician*, **61**(1), 34–40.