

Functional principal components analysis to study environmental data

Analisi in componenti principali funzionali per lo studio di dati ambientali

Miriam Daniele

Dipartimento di Scienze Statistiche e Matematiche “S. Vianelli”, Università di Palermo

Riassunto: Nel presente lavoro vengono analizzati i dati di concentrazione di alcuni inquinanti dell'aria e variabili meteorologiche forniti dalla rete di rilevamento per il monitoraggio dell'inquinamento atmosferico nel comune di Palermo. Dopo uno studio preliminare ed alcune trasformazioni elementari, viene proposta una sintesi multivariata: la tecnica utilizzata è una analisi in componenti principali per dati funzionali. L'analisi iniziale effettuata su 12 serie temporali, mostra che le prime due componenti principali spiegano il 73% della variabilità complessiva e sono facilmente interpretabili.

Keywords: principal component, functional data analysis, air pollution.

1. Introduction

In the last decades the interest in environmental protection is widely increased. Many researches dealt with the investigation and identification of the main air pollutants and several monitoring programme have been carried out to prevent and reduce the air pollution. A monitoring network usually consists in stations, located in the urban area, in which the concentrations of the main pollutants and of some weather variables are instantaneously recorded; such data are essentially continuous. Ramsay and Dalzell (1991) called *functional data analysis* the analysis of such data. The term functional refers to the intrinsic structure of the data rather than to their explicit form, see Ramsay and Silverman (1997, 2002) for a broad perspective of the available methods and case studies in functional data.

In practice, these data are usually observed and recorded discretely and the aim of functional data analysis is to gain additional intuition by considering functions as single observations in function space rather than as high-dimensional vectors (Silverman, 1996). Statistical techniques for functional data analysis work on a vector space of real-valued functions defined on a closed interval for which the integral of their squares is finite; in the case of functions having finite norm, the resulting space is a Hilbert space.

The first step, in the functional data analysis, is to convert the discrete values of a data matrix to a function $x_i(t)$ ($\forall i = 1, \dots, n$) computable for any desired argument value $t \in [0, T]$; this occurs converting suitably the data through smoothing procedures, for example representing them as linear combinations of K known basis functions. The resulting set $\mathcal{X}_t = \{x_1(t), \dots, x_n(t)\}_{t \in T}$ is called *functional dataset*. In particular, the basis selected here is the *Fourier basis*, which is the most appropriate when we work with temporal series, enough long and showing particular periodicity.

The goal of this paper is to explore the air pollution in Palermo, highlighting modes of variation and common trends among the pollutants and the weather variables, in order

to suitably reduce the phenomenon dimensionality and to evaluate the possibility to construct indices of environmental quality.

2. Functional Principal Components Analysis

The functional principal components analysis have been investigated by many authors: Ocaña *et al.* (1998) focused on the importance of the norm choice, James *et al* (2000) dealt with sparse data with measurements for individuals occurring at possibly differing time points, Locantore and Marron *et al.* (1999) discussed the robustness of this method.

Principal component analysis (PCA) is a standard approach to the exploration of variability in multivariate data. PCA uses an eigenvalue decomposition of the variance matrix of the data to find directions in the observations space along which the data have the highest variability. For each principal component, the analysis yields a loading vector or weight vector which gives the direction of variability corresponding to that component. In the functional context, each principal component is specified by a principal component weight function or *eigenfunction* $\xi(t)$ defined of the same range of t as the functional data.

The principal component scores of the individuals in the sample are the values z_i given by

$$z_i = \int \xi(t)x_i(t)dt. \quad (1)$$

The aim of PCA is to find the weight function $\xi_j(t)$ that maximizes the *variance function* of the principal component scores z_i

$$v(s, t) = \frac{1}{n-1} \sum_{i=1}^n \{z_i(s) - \bar{z}(s)\} \{z_i(t) - \bar{z}(t)\}, \quad (2)$$

subject to the constraints

$$\int \xi_j^2(t)dt = 1 \quad (3)$$

$$\int \xi_j(t)\xi_i(t)dt = 0 \quad \forall i, j \text{ con } i \leq j. \quad (4)$$

The condition (3) is necessary because the variance is superiorly unlimited, the (4) ensure that all the estimated principal components are mutually orthogonal. The principal components analysis of functional data is often enhanced by the use of smoothing (Silverman, 1996). To obtain a smoothed functional PCA, we have to control the size of ξ , but also its roughness. In practice, the constraints (3) and (4) are replaced by the following ones

$$\int \xi^2(t)dt + \lambda \int \{\xi''(t)\}^2 dt = 1 \quad (5)$$

$$\int \xi_j(t)\xi_i(t)dt + \lambda \int \xi_i''(t)\xi_j''(t)dt = 0 \quad \forall i, j \text{ con } i \neq j. \quad (6)$$

The smoothing parameter $\lambda \geq 0$ controls the amount of smoothing inherent in the procedure. The smoothing parameter choice is usually a consequence of empirical subjective considerations together with a cross-validation criterion.

3. Data description and analysis

Since 1996 a network of monitoring stations for air pollution was established in Palermo; the network includes ten stations managed by AMIA, the Palermo municipal environmental agency. The stations are located in the urban area except one, situated in a suburban zone, acting a control station. After a preliminary analysis of all stations we decided to work with the only station, located in the center of the town, that also register the weather variables.

The dataset considered consists of a (2191×12) matrix, constituted by the series of daily maxima of the pollutants (CO, NO₂, NO, NO_x, SO₂, PM₁₀, O₃) and weather variables (temperature, humidity, pressure, wind, solar radiation) in the period January 1998-December 2003. The choice to analyze the maxima is motivated by different legislative and statistical reasons: from a statistical point of view the distributional properties of the maxima are described by the distribution tails and not much by the distribution body; this allows us to underline the common trends in the variables rather than the variables different peculiarities. Some preliminary manipulation of the dataset were necessary: in particular, the data were standardized to eliminate the effect of the different size and the unit of measurement. The functional PCA was obtained by the following steps:

1. the series were converted in the functional form using a Fourier basis with 314 expansion coefficients (one for each week);
2. we performed functional PCA on the functional dataset.

Figure 1: *First two Harmonics simultaneous variations*

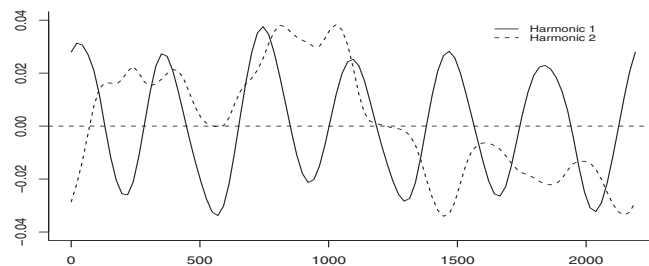
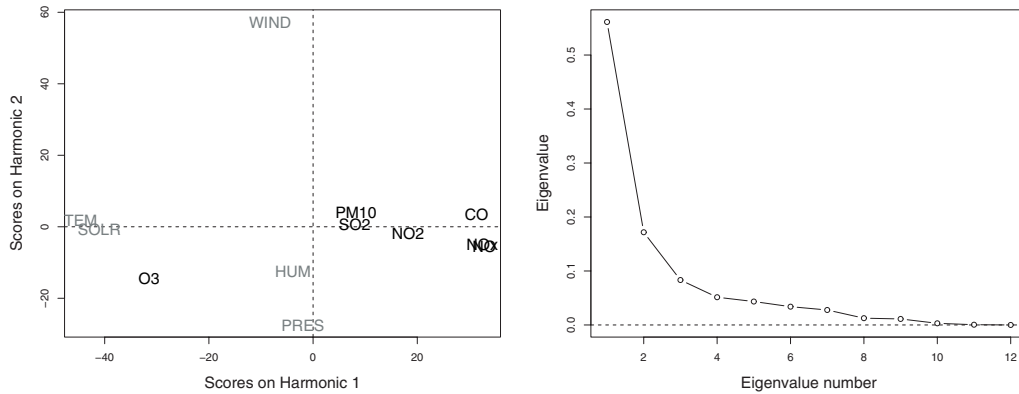


Table 1: *Explained variability by the first four Principal Components.*

	<i>PC1</i>	<i>PC2</i>	<i>PC3</i>	<i>PC4</i>
Variance proportion (%)	56	17	8	5
Cumulative variance proportion (%)	56	73	81	86

The graphics (1) above shows the simultaneous variation of the first two harmonics, that together accounts for the 73% of the whole variability (see the table 1).

Figure 2: Variables scores (black labels for the pollutants and gray labels for the weather variables) and screeplot on the the first two principal components



We note (see Fig. 2) that the first harmonic expresses a contrast between the weather variables and the pollutants, except for the ozone, that is strongly positively correlated with weather variables and negatively correlated with the pollutants. The second harmonic variation depends mainly on the wind velocity variable.

4. Conclusions

The FPCA carried out allowed us to appreciate the main relations (linear and non linear) among the pollutants and weather variables and to suitably summarize the available information about the air pollution. The detailed interpretation of the obtained results will be discussed.

References

- James G. M., Hastie T. J. and Sugar C. A. (2000) Principal component models for sparse functional data, *Biometrika*, 87, 587–602.
- Locantore N. and Marron J. S., Simpson D. G., Tripoli N., Zhang J. T. and Cohen K. L. (1999) Robust principal component analysis for functional data, *Test*, Vol. 8, 1.
- Ocaña F. A., Aguilera A. M. and Valderrama M. J. (1999) Functional principal components analysis by choice of norm, *Journal of Multivariate Analysis*, 71, 262–276.
- Ramsay J. O. and Dalzell C. J. (1991) Some tools for functional data analysis, *J. R. Statist. Soc., B*, 53, 539–572.
- Ramsay J. O. and Silverman B. W. (1997) *Functional Data Analysis*, Springer-Verlag, New York.
- Ramsay J. O. and Silverman B. W. (2002) *Applied Functional Data Analysis*, Springer-Verlag, New York.
- Silverman B. W. (1996) Smoothed functional principal components analysis by choice of norm, *The Annals of Statistics*, 24, 1, 1–24.