Não preencher

Some contributions to PCA for time series

Isabel Silva

Fac. de Engenharia, Univ. do Porto e UIMA, Univ. de Aveiro, ims@fe.up.pt

Abstract

Multidimensional time series appear in many fields of application. Sometimes, it can be useful to use PCA to reach dimensionality reduction. However, formal inference procedures on PC rely on the independence of the variables. Therefore, several PC-like techniques, as Singular Spectrum Analysis, are used to attain this reduction by decomposing the original series into a sum of a small number of interpretable components. Here, SSA and its extension are described and applied to real datasets.

Keywords: Dimensionality reduction, Principal Component Analysis, Singular Spectrum Analysis, Time series.

1. Introduction

Multidimensional time series are common in many fields of application and are characterized by correlated data. In many situations the number of observations in each series outnumbers the number of series. Thus, it is of paramount importance to compress the series, extracting the most important information and discarding noise and redundant correlations. This is also useful for graphical representation and for future statistical analysis of the time series data. One very popular method for dimensionality reduction is Principal Component Analysis (PCA), which obtain a new set of variables, called Principal Components (PC), that are uncorrelated and that are ordered so that the first few retain most of the variation presented in the dataset (Jolliffe, 2002).

Pinto da Costa *et al* (2009) introduced a method of Weighted Principal Component Analysis specific for time series data which give different weights to the observation times, according to a certain goal. They have seen that large differences can occur between the usual PCA and weighted PCA; in particular, they have found that weighted PCA is capable of higher levels of compression of the data.

In some fields of application, PCA not only reduces the dimensionality of the dataset but also allows for reasonable interpretations of the retained PC. However, formal inference procedures based on PC rely on the independence (and multivariate normality) of the observations, a condition that is violated for time series data. Several techniques, like Singular Spectrum Analysis (SSA), that take in account the correlation in time (and space) have been developed. The main goal of SSA, and its extension to several time series called Multichannel SSA (MSSA), is to decompose the original series in a small number of independent and interpretable components that can be thought as trend, oscillatory components and a structureless noise.

In this work, the SSA is described and compared with PCA. In addition, the results of the application of these techniques to real datasets are exhibited.

2. Singular Spectrum Analysis

The central idea in SSA is to carry out a PCA on a suitable chosen lagged version of the original time series. More specifically, basic SSA consists in four steps: *embedding* and *Singular Value Decomposition* (SVD), for the **decomposition stage**, and *grouping* and *diagonal averaging*, for the **reconstruction stage** (Golyandina *et al*, 2001).

Given a time series of length N, $Y=\{y_1, \ldots y_N\}$, the embedding step consists in choosing an integer L ($1<L<N$), designated window length, in order to construct the so called trajectory matrix, **X,** which is the following LxK Hankel matrix ($K=N-L+1$):

$$\mathbf{X}=\begin{bmatrix}X_1 & X_2 & X_3 & \cdots & X_K\end{bmatrix}=\begin{bmatrix} y_1 & y_2 & y_3 & \cdots & y_K \\ y_2 & y_3 & y_4 & \cdots & y_{K+1} \\ y_3 & y_4 & y_5 & \cdots & y_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & y_{L+2} & \cdots & y_N \end{bmatrix}.$$

Next step is the SVD of **X**, which allow to represents **X** as a sum of rank-one bi-orthogonal elementary matrices. Denote by $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_L$ the eigenvalues of the LxL matrix $\mathbf{S}=\mathbf{X}\mathbf{X}^{\mathrm{T}}$ and by $U_1,\ldots,U_L$ the orthonormal system of the eigenvector of **S** corresponding to these eigenvalues. Let $d = rank(\mathbf{X}) = max\{i : \lambda_i > 0\} \leq L$. If $V_i = \mathbf{X}^T U_i / \sqrt{\lambda_i}$ $(i=1,..,d)$, then $\mathbf{X}=\mathbf{X}_1+\cdots+\mathbf{X}_d$, where $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^T$. The collection ($\lambda_i$, U_i, V_i) is called i-th eigentriple of **X**. Comparing with the PCA terminology, the vector U_i represents i-th (principal) eigenvector and the vector $Z_i = \sqrt{\lambda_i} V_i$ corresponds to the i-th principal component of **X**.

The reconstruction stage starts with the grouping step that consists of partitioning the set of indices $\{1,\ldots,d\}$ into m disjoint subsets $I_1,\ldots,I_m$ (the proposed number of PC) where each of these subset is of the form $I=\{i_1, \ldots, i_p\}$ and constructing the corresponding resultant matrix $\mathbf{X}_I$ which is defined as $\mathbf{X}_I = \mathbf{X}_{i_1} + \cdots + \mathbf{X}_{ip}$. Therefore, at the end of this step, it is found that $\mathbf{X} = \mathbf{X}_{I1} + \cdots + \mathbf{X}_{I_m}$. The contribution of the component $\mathbf{X}_I$ (for each $I \in \{I_1,\ldots,I_m\}$) is given by $\sum_{i\in I} \lambda_i \Big/ \sum_{i=1}^{d} \lambda_i$.

The last step of SSA, diagonal averaging, is a formal procedure that transforms each matrix of the grouped decomposition into a Hankel matrix and therefore into a new series of length *N* (for a detailed description see Golyandina *et al*, 2001). In this way, the initial series can be decomposed in a sum of *m* series (that represents the *m* PC) $Y \approx \tilde{\mathbf{X}}_{I1} + \cdots + \tilde{\mathbf{X}}_{I_m}$, where $\tilde{\mathrm{X}}_{I_k}$ is the result of applying diagonal averaging to X_{I_k}, with $k = 1,...,m$.

The application of SSA carries out some practical problems. The first practical issue is the choice of the dimension *L* of the embedding vector space: while larger values of *L* allow resolving longer-period oscillations, it also reduces the number of observations *K* from which to estimate the covariance matrix. In practice, *L* must be less than *N/2* and it depends on the maximum period of harmonics that can be detected. The second decision that the analyst has to make is the choice of the *m*, the number of eigentriples for the reconstruction stage. Rodrigues and de Carvalho (2008) have concluded that it is necessary a carefully choice of these parameters, since they can compromise the analysis of the datasets, specially the forecasting accuracy.

The extension to SSA to the analysis of several time series, called Multichannel SSA, is straightforward. The idea is to apply SSA to a large trajectory matrix constructed by concatenating the trajectory matrices corresponding to each of the *p* time series with length *N* that comprises the original dataset, given a certain window length *L*.

References:

GOLYANDINA, N., NEKRUTKIN, V. & ZHIGLJAVSKY, A. (2001) *Analysis of Time Series Structure: SSA and related techniques*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability.

JOLLIFFE, I. T. (2002) *Principal Component Analysis*. 2nd edition. Springer-Verlag.

PINTO DA COSTA, J., SILVA, I. & SILVA, M.E. (2009) Weighted principal components for time series. *Actas do XVII Congresso Anual da SPE* (submitted).

RODRIGUES, P. C. & DE CARVALHO, M. (2008) Monitoring Calibration of the Singular Spectrum Analysis Method. In *Proceedings of COMPSTAT'2008 (International Conference on Computational Statistics)*, Vol. 2, pp. 955–964, Physica-Verlag.