Principal Component Analysis of Dependent Data

Paulo Canas Rodrigues
Department of Mathematics and CMA, Faculty of Science and Technology, Nova University of Lisbon, Monte de Caparica, 2829-516 Caparica, Portugal.

Keywords: Principal Component Analysis. Time Series. SSA. MSSA.
AMS: 62H25, 62M15

Abstract

When we perform a Principal Component Analysis (PCA) we assume, as we do in many other statistical methods, that the observations are independent. In this work we present an extension of PCA where the observations are not independent. To show how the method works in practice we present an application to a data set from a real time series.

1. Introduction

PCA is one of the most used methods in multivariate analysis. Its central idea is to reduce the dimensionality of a data set which consists of a larger number of interrelated variables and to identify underlying variables in the structure of the data (the principal components) that have physical meaning. Overall it allow us to see the structure from another point of view.

As it is known, the PCA operates on data sets where observations are independent. The aim of this work is to discuss the implications, in PCA, when the observations are not independent. In Jolliffe (2002) we can find some references to particular cases of this type of data as well as some procedures to its analysis using methods based on PCA. A possibility to analyze this type of data is to perform a PCA as usually but in such a way we lose precious information: the dependence of the observations. Then what should we do?

In this work we present the extension to time series, particularly the Singular Spectrum Analysis (SSA) for one-dimensional time series and the Multichannel Singular Spectrum Analysis (MSSA) for time series with more than one variable. In the next two sections we present the SSA together with one application. We also present a briefly discussion of the MSSA.

2. Singular Spectrum Analysis

This method consists in the decomposition of the time series within several components that usually can be identified as trends, oscillatory components or noise components.

Basic SSA performs four steps. At first step (called the *embedding step*), the one-dimensional series is represented as a multidimensional series whose dimension is called the *window length*. The multidimensional time series (which

is a sequence of vectors) forms the *trajectory matrix* (1). The sole parameter of this step is the window length and it should be proportional to the number of periods, for example if we have a time series measure monthly we must consider L proportional to 12, see Vautard *et al.* (1992).

If we consider $\mathbf{F} = (f_1, f_2, \ldots, f_n)$ the time series of length n, and L, $1 < L < n$, the window length, we obtain $K = n-L+1$ lagged vectors of length L, $\mathbf{X}_j = (f_j, f_{j+1}, \ldots, f_{j+L-1})'$, $j = 1, 2, \ldots, K$ and the trajectory matrix

$$\mathbf{X} = [\mathbf{X}_1 : \ldots : \mathbf{X}_K]' = \begin{bmatrix} f_1 & f_2 & \cdots & f_L \\ f_2 & f_3 & \cdots & f_{L+1} \\ \vdots & \vdots & \ddots & \vdots \\ f_K & f_{K+1} & \cdots & f_n \end{bmatrix}. \tag{1}$$

The second step, SVD step, is the singular value decomposition of the trajectory matrix (1) into a sum of rank-one bi-dimensional matrices as:

$$\mathbf{X} = \sum_{i=1}^{d} \mathbf{X}_i = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i'. \tag{2}$$

where λ_i, U_i and V_i, $i = 1, \ldots, L$, are the eigenvalues, the left and right singular vectors, respectively. The collection $(\sqrt{\lambda_i}, U_i, V_i)$ is called the ith *eigentriple*) of the SVD of matriz $\mathbf{S} = \mathbf{X}'\mathbf{X}$. The first two steps together are considered as the *decomposition stage* of Basic SSA.

The next two steps form the *reconstruction stage*. The point of this stage is the reconstruction of the original time series as the sum of principal components using the *diagonal averaging* procedure.

The *grouping step* corresponds to splitting the matrices, computed at the SVD step, into several groups (m - the intended number of principal components) and summing the matrices within each group. The result of the step is a representation of the trajectory matrix as a sum of several *resultant matrices*.

The last step transfers each resultant matrix into a time series, which is an additive component of the initial series. The corresponding operation is called *diagonal averaging*. It is a linear operation and maps the trajectory matrix of the initial series into the initial series itself. In this way we obtain a decomposition of the initial series into several addictive components.

If we apply this procedure to the matrix $\mathbf{X}_{I_k}$, where $I_k = \{i_{k1}, \ldots, i_{kp}\}$ is a set of indexes, we obtain, see Golyandina *et al.* (2001), the series $\widetilde{\mathbf{F}}^{(k)} = (\widetilde{f}_1^{(k)}, \ldots, \widetilde{f}_n^{(k)})$ and, therefore the initial series $\mathbf{F} = (f_1, \ldots, f_n)$ is decomposed into the sum of m series:

$$\mathbf{F} = \sum_{k=1}^{m} \widetilde{\mathbf{F}}^{(k)}, \qquad k = 1, \ldots, K. \tag{3}$$

This m series represent the first m principal components and the equality in (3) only happens if $m = L$, i.e., if we have the sum of all series/principal components.

To illustrate the course of the process followed by SSA we use a small data set constituted by a part of studied series, referring to the unemployment rate in Portugal (Section 3). In this exercise we use 8 values of this series, since the second trimester of 2004 to the first of 2006. The original time series, F, the trajectory matrix, X and the the principal components obtained using the covariance matrix, $\mathbf{S} = \frac{1}{K}\mathbf{X}'\mathbf{X}$, and a window length $L = 4$ ($K = n - L + 1 = 5$) are, respectively,

$$\mathbf{F} = \begin{bmatrix} 6.3 \\ 6.8 \\ 7.1 \\ 7.5 \\ 7.2 \\ 7.7 \\ 8.0 \\ 7.7 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 6.3 & 6.8 & 7.1 & 7.5 \\ 6.8 & 7.1 & 7.5 & 7.2 \\ 7.1 & 7.5 & 7.2 & 7.7 \\ 7.5 & 7.2 & 7.7 & 8.0 \\ 7.2 & 7.7 & 8.0 & 7.7 \end{bmatrix},$$

PC1	**PC2**	**PC3**	**PC4**
−0.331	0.047	−0.678	−0.025
−0.234	−0.063	−0.047	−0.143
−0.132	0.002	−0.060	0.003
−0.025	0.069	0.089	0.079
0.157	−0.117	−0.033	−0.094
0.262	0.021	0.100	0.029
0.354	0.202	0.055	0.101
0.384	−0.200	0.196	0.033

The sum of this components, each one with $K = 8$ values, should result in the original time series as in (3). However if we sum the $L = 4$ principal components in this example we obtain the series: -0.99; -0.49; -0.19; 0.21; -0.09; 0.41; 0.71 and 0.41 which is different of the original. If we look closer we find this series is the original minus its mean (7.3). This difference is due to the covariance matrix used in the routines implemented in MatLab by Eric Breitenberger and available in http://pangea.stanford.edu/research/Oceans/GES290/Breitenberger-SSAMatlab/ssa/ is not centered on the mean.

3. Application of SSA

In this section we present an application of SSA to the referred time series of the Portuguese unemployment rate, but from the first trimester of 1992 to the first trimester of 2006 (57 observations). This data set was analyzed using two softwares available in the web page of *GistaT Group* (http://www.gistatgroup.-com/cat/index.html) named "CatMV 1.00" and "CatSSA 1.00".

In this analysis we used different values for window length and we present the results for $L = 28$ in Figure 1. Through the figure analysis we can say we have a "good" adjustment. The analysis of the residuals confirm the "good" adjustment, with exception of two peaks more distant from the origin. Overall, the residuals are very close to the origin and its distribution throughout the time does not present any visible trend.

The proportions of variance explained by the first four principal components are 95.63%, 2.62%, 1.45% and 0.06%, respectively, and the last 24 only explain the remaining 0.34%. The plots of these four principal components are presented in Figure 2. In these plots we can see the three first principal components representing an evolutive trend with a great importance of the first (explains 96% of the variance). The fourth component represents the noise because of its oscillatory form.

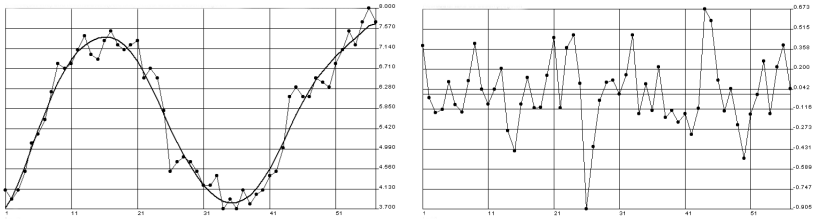

Figure 1: Time series of the unemployment rate and its reconstruction using the method of principal components (on the left) and the residuals of this reconstruction (on the right).

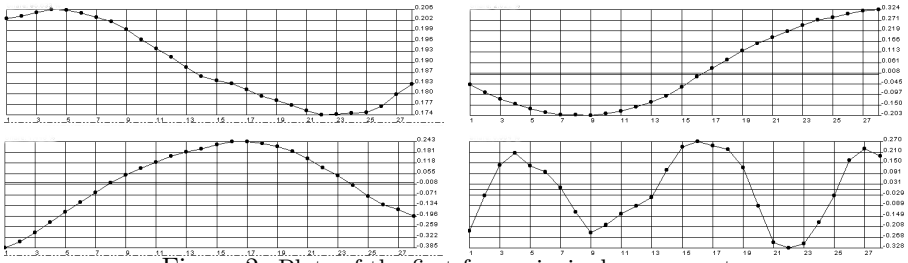

Figure 2: Plots of the first four principal components.

If we want to make forecasting based on the original time series of the unemployment rate in the next n trimesters, we can use the software "CatSSA 1.00".

4. Multichannel Singular Spectrum Analysis

MSSA is a generalization of SSA for p (>1) time series (or *channels*). MSSA is used in the same way as the SSA to extract oscillatory behavior within the multivariate time series using the information of correlation among observations and variables.

If we consider p time series, $\{\mathbf{x}_{i,t}\}_{t\in\{1,\ldots,n\}}$ for $i = 1,\ldots,p$ of length n, we can consider, as in SSA, the lagged matrix $\mathbf{X}_{K\times Lp}$:

$$\mathbf{X} = \begin{bmatrix} f_{1,1} & \cdots & f_{1,L} & \cdots & f_{p,1} & \cdots & f_{p,L} \\ f_{1,2} & \cdots & f_{1,L+1} & \cdots & f_{p,2} & \cdots & f_{p,L+1} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ f_{1,K} & \cdots & f_{1,n} & \cdots & f_{p,K} & \cdots & f_{p,n} \end{bmatrix}. \tag{4}$$

The covariance matrix obtained from the trajectory matrix (4) has information on interrelations between lagged versions of the original variables as well as different variables. MSSA uses all the information in the covariance matrix and instead of studying the usual $(n \times p)$ data matrix, it uses a $(n' \times p') = ((n - L + 1) \times (Lp))$ matrix with information from covariance

between variables at each lag (until lag L-1) with:

$$\mathbf{S}_X = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} & \cdots & \mathbf{S}_{1p} \\ \mathbf{S}_{21} & \mathbf{S}_{22} & \cdots & \mathbf{S}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_{p1} & \mathbf{S}_{p2} & \cdots & \mathbf{S}_{pp} \end{bmatrix} \tag{5}$$

where $\mathbf{S}_{kl}$ is the cross-covariance matrix between the k^{th} and l^{th} location until lag $L-1$. The rank of matrix $\mathbf{S}_X$ is pL.

To illustrate how MSSA works, a study was carried out involving five variables associated to the production and consumption of energy in Portugal. In this study we also compared the results from MSSA (which uses all correlations) with the results of PCA (uses only the information of the zero lagged correlations between variables). Due to space requirements it is not possible to include this study that can be found in Rodrigues (2007).

5. Comments and final remarks

The main concern resulting from this work debates on the necessity of a more complete and thorough exposition of the PCA, not limiting it to the case of data with independent observations, as is the case in most books of multivariate analysis. In most practical applications the assumption of independent observations is inadequate and therefore this case should be studied.

In the present text we chose to present a very usual case of dependence of observations, the time series, providing the use of procedures SSA and MSSA. The SSA was illustrated with an application to a time series, having been clearly useful in this type of analysis. However, an exercise was made previously with a small data set to perceive correctly the sequence of the steps covered by the method, which can have a pedagogical role in this contribution for the development of PCA.

6. Bibliography

[1] Golyandina, N., Nekrutkin, V. e Zhigljavsky, A. (2001). *Analysis of Time Séries Structure: SSA and Related Thechniques*. Chapman & Hall. New York.

[2] Jolliffe, I.T. (2002). *Principal Component Analysis*. Springer Verlag. New York, Inc.

[3] Rodrigues, P. C. (2007). *Componentes Principais: o método e suas generalizações*. MSc in Statistics - Instituto Superior Técnico, Universidade Técnica de Lisboa.

[4] Vautard, R., Yiou, P. e Ghil, M. (1992). Singular spectrum analysis: A toolkit for short noisy chaotic signals. *Physica D*, Vol.**58**, p.95-126.