Singular Spectrum Analysis: a new decomposition technique applied to environmental systems

Girolamo Stea, Massimo Bilancia
Department of Statistical Sciences "Carlo Cecchi" – University of Bari
stea@dss.uniba.it, mabil@dss.uniba.it

Abstract: In the last few years Singular Spectrum Analysis (SSA), a powerful tool in time series analysis, has been developed and applied to many fields. In this paper we discuss the basics of SSA: reconstruction of components may based on the functional clustering algorithm introduced in Bilancia and Stea (2008). We report an example concerning an application in the environmental health field.

Keywords: Time series, Singular Spectrum Analysis (SSA).

1. Introduction

Environmental systems show a complex behavior evolving over many timescales: the problem of disentangling a mixture of several dynamical pattern may be based on Singular Spectrum Analysis (SSA, Golyandina et al., 2001), a model-free approach derived from dynamical system theory, and suitable for application in the environmental sciences area (see for example Bilancia and Stea, 2008, which propose an SSA-based approach to assess the association between airborne pollution and human health).

2. Singular Spectrum Analysis (SSA)

The SSA algorithm consists of four phases. In the first phase, given a one-dimensional time series $Y = (y_{0,\cdots,}y_{N-1})$ the *embedding* procedure constructs a sequence of $K = N - L + 1$ *lagged vectors* $X_i^{(L)} = (y_{i-1}, \cdots, y_{i+L-2})^T$, where L (the window length) is an integer such that $1 < L < N$. The *trajectory matrix* is given by

$$X = (x_{ij})_{i,j=1}^{L,K} = \left[X_1^{(L)}, X_2^{(L)}, \cdots, X_k^{(L)}\right] \quad (1)$$

It should be noted that the trajectory matrix X is an Hankel matrix, i.e. all the elements along the secondary diagonals such that $i + j = const$ are equal: the only free parameter is the *window length* L. In the second stage the Singular Value Decomposition (SVD) of the trajectory matrix X is computed. Given $S = XX^T$ and $d = \text{rank}(X)$, we denote by λ_i the eigenvalues of S ordered in the decreasing order, and by U_i the corresponding eigenvectors; in addition, let be and $V_i = X^T U_i / \sqrt{\lambda_i}$ (with $i = 1, \cdots, d$) . The SVD of the matrix X is given by

$$X = X_1 + \cdots + X_i + \cdots + X_d\,, \qquad X_i = \sqrt{\lambda_i} U_i V_i^T \tag{2}$$

The square roots $\sqrt{\lambda_i}$ are known as *singular values*, while U_i and V_i are called the *left* and *right singular vectors* of the matrix X: the collection $(\sqrt{\lambda_i}, U_i, V_i)$ is called *i-th eigentriple* of matrix X. Hence the trajectory matrix is decomposed into a sum of elementary rank-one, pairwise bi-orthogonal matrices. It can be proved that $\sum_{i=1}^{d} \lambda_i$ equals the squared Frobenius-Perron norm of the matrix X, as well as λ_i is the squared Frobenius-Perron norm of the matrix X_i $(i = 1, \cdots d)$. Thus, the ratio $\sum_{i=1}^{r} \lambda_i / \sum_{i=1}^{d} \lambda_i$ measures the degree of approximation of the trajectory matrix, when the approximation is made by means of the sum of the first r terms in (2). The third phase, the so called *eigentriple-grouping*, consists in dividing the index set $\{1, \cdots, d\}$ into m disjoint subsets $I_1, \cdots, I_m$ with $I_j = (j_1, \cdots, j_s)$, such that the decomposition (2) can be reformulated as

$$X = X_{I_1} + \cdots X_{I_j} + \cdots + X_{I_m} \tag{3}$$

with $X_{I_j} = X_{j_1} + \cdots + X_{j_s}$. Suppose now that the matrices on the right-side of (3) are Hankel: hence they are trajectory matrices from which component series on different timescales can be computed. Alternatively, *diagonal averaging* (or *hankelization*) consists in applying a suitable linear operator $\mathcal{H}$ to both sides of (3); after the hankelization we have that $\mathcal{H}X = X$ and $\mathcal{H}X_{I_j} = \tilde{X}_{I_j}$is Hankel, which is equivalent to decompose of the initial time series into a sum of m component series.

3. A functional clustering algorithm for component reconstruction

A suitable decomposition can be determined by a flexible use of the four-step algorithm described in the previous section (Bilancia and Stea, 2008). We suggest to apply Hankelization to each term in the full SVD decomposition (2): if $\tilde{X}_i = \mathcal{H}X_i$ then

$$X = \tilde{X}_1 + \cdots + \tilde{X}_i + \cdots + \tilde{X}_d \tag{4}$$

In general, elementary hankelized matrices on the right hand side of (4) are not pairwise row and column orthogonal, with the result that the sum of two of such matrices need not to be Hankel. It can be proved $\tilde{X}_l$ and $\tilde{X}_i$ are row and column orthogonal if and only if $\langle \tilde{X}_l, \tilde{X}_i \rangle_{\mathcal{M}} = 0$, where $\langle \tilde{X}_l, \tilde{X}_i \rangle_{\mathcal{M}}$ is the inner-product derived from the Frobenius-Perron norm; in addition (see Golyandina et al. 2001) from $\langle \tilde{X}_l, \tilde{X}_i \rangle_{\mathcal{M}} = 0$ it follows that $\tilde{X}_l + \tilde{X}_i$ is an Hankel matrix, hence it is the trajectory matrix of some component time series. This condition will be referred to as *weak L-separability*. By joining elementary hankelized components having minimum distance in terms of weak L-separability we often obtain a sensible grouping (3), whose component matrices are as close as possible to Hankel matrices, hence amenable to a suitable interpretation after the diagonal averaging step.

A sensible measure of weak L-separability between components l and i is the *w-correlation*

$$w_{li} = \langle \tilde{X}_l, \tilde{X}_i \rangle_{\mathcal{M}} / \|\tilde{X}_l\|_{\mathcal{M}} \|\tilde{X}_i\|_{\mathcal{M}} \tag{5}$$

where $l, i = 1, \cdots, d$, and the norm is the matrix Frobenius-Perron norm. If the absolute value of the w-correlation is small then the corresponding series are almost w-orthogonal, but if the value is large then the two series are far from being w-orthogonal and thus badly separable. Therefore, a functional clustering algorithm can be based on the dissimilarity matrix $1 - W = \{1 - |w_{li}|\}$ with the complete linkage method as the natural choice.

4. Estimating the number of components

Given the output of a hierarchical clustering routine, one needs to evaluate the number of clusters m: the reconstructed exposure variables should be easily interpretable, and they should vary at timescales of scientific interest (such as seasonality and trend). Each eigenvalue λ_i measures the degree of approximation of the i-th elementary component X_i to the trajectory matrix X: therefore, we can define (as a function of the number of groups m) an pseudo-R_m^2 index accounting for the total degree of approximation within each group. Let $T = \sum_{j=0}^{d} \left(\lambda_j - \bar{\lambda}\right)^2$ the total sum of squares with respect to the full eigenvalue spectrum; similarly, let $SSE_k = \sum_{i=1}^{n_k} \left(\lambda_{ik} - \bar{\lambda}_k\right)^2 n_k$ be the within cluster sum of squares (λ_{ik} is an eigenvalue belonging to the k-th group where $\bar{\lambda}_k$ is the average of such values, and n_k is corresponding number of eigenvalues). The pseudo-R_m^2 for a decomposition into m groups is defined as

$$R_m^2 = \frac{T - \sum_{k=1}^{m} SSE_k}{T} \tag{6}$$

A suitable decision criterion prescribes that a decomposition into m^* groups is chosen if $sup_m(R_m^2 - R_{m-1}^2)$ is reached for $m = m^*$, for m varying into a given range (typically $m = 3, \ldots, 8$).

We report an example based on a dataset concerning daily measurements of PM_{10}, obtained in Bari, Apulia, Italy, between June 1th 2000 and December 31th, 2001: the original dataset was collected by a monitoring network of Municipality of Bari (Department of Environmental Protection and Health), and a complete data-description and pre-processing is described in Bilancia and Stea (2008). SSA can be used to decompose air-pollution time series into a set of suitable exposure variables, each one representing a different timescale. We are interested to test the so called *mortality displacement* or *harvesting hypothesis*: a wealth of epidemiological studies based on time-series analysis has shown evidence for association between morbidity/mortality for cardiovascular adverse events and the exposure to airborne particles, but a considerable uncertainty remains to as whether these associations represents premature mortality within only few days among those already near to death. Result are reported in the Fig. 1; we set $L = 60$ for the window length, as adverse effect at timescales longer than two months are likely to be confounded with long-term effects due to other causes. The pseudo-R_m^2 criterion (6) yielded $m^* = 6$: the reconstructed exposure variables at diverse timescales are reported in bottom-right display in the Fig. 1: these variables can be used in a GAM model for estimating the net effect of past exposure on health.

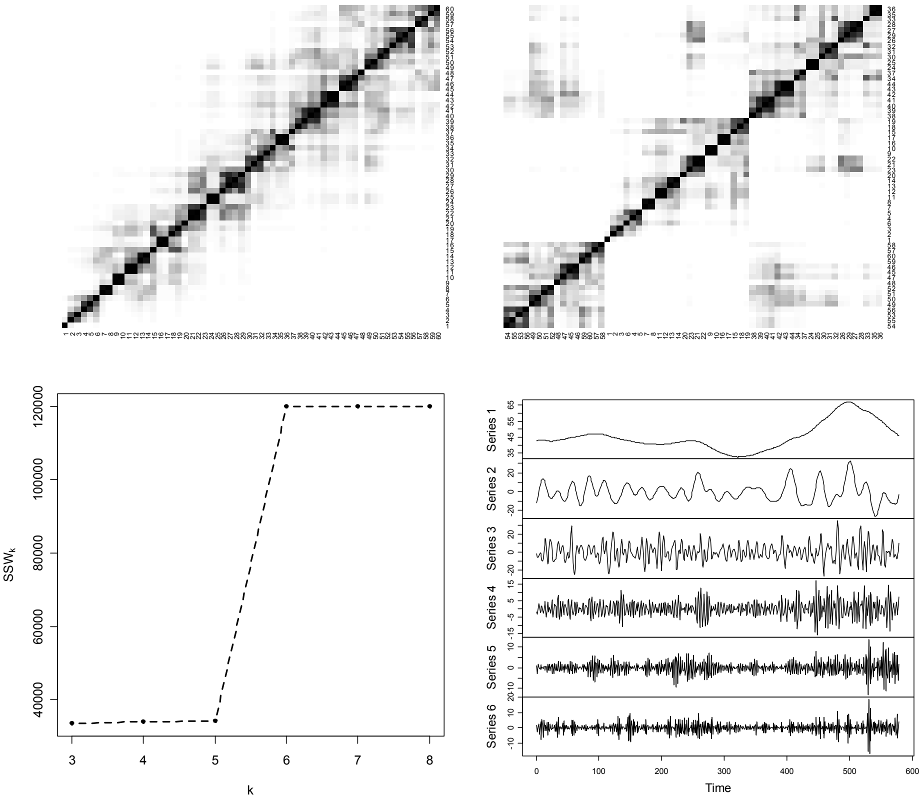

Figure 1: Top-left: w-correlation matrix of the original elementary components. Top-right: w-correlation matrix after row and column permutation according to the hierarchical clustering output for $m = 6$ groups. Bottom-left: the criterion for choosing the number of groups: it is apparent that $m^* = 6$. Bottom-right: the six reconstructed components (here $L = 60$).

Bibliography

Bilancia M., Stea G. (2008), Timescale effect estimation in time-series studies of air pollution and health: a singular spectrum analysis approach, Electronic Journal of Statistics, to appear.

Broomhead D.S., King G.P. (1986), Extracting qualitative dynamics from experimental data, Physica D 20, 217-236.

Elsner, J. B., Tsonis, A. A. (1996), Singular Spectral Analysis. A New Tool in Time Series Analysis, Plenum Press.

Golyandina N., Nekrutin A., Zhigljiavsky A. (2001), Analysis of Time Series Structure: SSA and related techniques, Chapman & Hall/CRC, New York.

Vautard R, Ghil M (1989) Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series, Physica D 35, 395-424.