Notes on SSA in MLCore

Solidware

18 November 2019

1 SSA

Singular spectrum analysis (SSA) can be defined as principal component analysis (PCA) for univariate time series data. As a reminder, PCA consists of the decomposition of a covariance matrix $X^T X$, where $X \in \mathbb{R}^{n \times m}$ is a data matrix (n being the number of samples, m the number of features and the objective is dimension reduction).

On the other hand, SSA consists of the spectral decomposition of a lag-covariance matrix, also $X^T X$, though X (called the trajectory matrix) contains "the complete record of patterns that have occurred within a window of size L" (or embedding dimension). It is defined such that, given a centred time series $y = y_1, \ldots, y_N$ (with $N = L + K - 1$) of finite rank, one has:

$$X = \begin{bmatrix} y_1 & y_2 & y_3 & \cdots & y_L \\ y_2 & y_3 & y_4 & \cdots & y_{L+1} \\ y_3 & y_4 & y_5 & \cdots & y_{L+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_K & y_{K+1} & \cdots & \cdots & y_{L+K} \end{bmatrix}$$

Elser and Tsonis state that "by using lagged copies of a single time series, [we] can define the coordinates of the phase space that will approximate the dynamics of the system from which the time record was sampled".

1.1 Overview

SSA consists of 2 stages:

1. **decomposition**: (i) *embedding* (see trajectory matrix, X above) and (ii) *decomposition*: SVD is applied to X (equivalently, one may do an eigendecomposition of the covariance matrix) to obtain a decomposition into elementary rank-one matrix components.

2. **reconstruction**: (iii) *grouping*: grouped matrix components are created in a clever way and (iv) *diagonal averaging*: those are back-transformed to provide a decomposition of the time series.

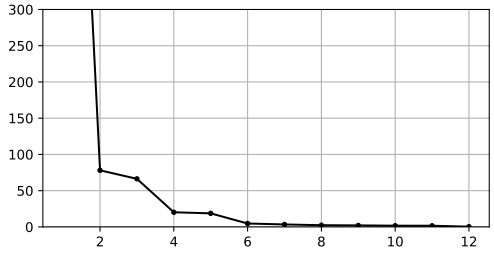

Figure 1: eigenvalues of lag-covariance matrix from passengers series. The first eigenvalue (related to the trend) is not shown here, because it is too large. In general, correlated components have similar eigenvalues (see (2,3) and (4,5), and noise can be related to small eigenvalues that decrease linearly (see 6 to 12). Here, k=6 meaning that one should use at most n_e=6 leading eigenvectors, and components (2,3) and (4,5) should be grouped (so $n_c = 3$).

1.2 Decomposition

The trajectory matrix is obtained via embedding, using `WindowTransform`, and it is decomposed using SVD: it provides the decomposition: $X = USV^T$, where the columns of U and V are bases for the row and column spaces of X.

$$\begin{aligned} X &= \sum_i \sigma_i U_i V_i^T \\ &= \sum_i \Pi_i X \end{aligned}$$

where (i) U, V are left singular and right singular vectors of X, or where (ii) $\Pi_i = U_i U_i^T$ is a projection matrix (with U_i, the eigenvectors of $X^T X$)[1].

One thereby has the approximation of $X = \sum_{i=1}^{r} X_i$, as a truncated sum of rank 1 quasi-Hankel matrices. This is the best approximation in terms of Euclidean distance[2], to any rectangular matrix. Note that by discarding the projections that correspond to the unwanted sources—such as the noise or artifact sources, usually associated with lower eigenvalues (Fig. 1)—and then inverting the transformation, we effectively perform a filtering of the signal (Fig. 2).

In MLCore, one can either:

- choose only a number of components, n_c (`n_components`), in which case, the number of leading eigenvectors used for matrix approximation is the

[1] since $V_i = \frac{1}{\sigma_i} X^T U_i$, one can rewrite the SVD decomposition of X as $US(X^T US^{-1})^T$

[2] the problem of low-rank approximation of a data matrix D is defined as the minimisation of the Frobenius norm $||D - \widetilde{D}||_F$ $\left(\text{def: } ||A||_F^2 = \text{Trace}(A^T A) = \sum \sigma_A^2\right)$ over $\widetilde{D}$ subject to $\text{rank}(\widetilde{D}) < r$, the desired rank. It has an analytic solution in terms of the SVD of D.

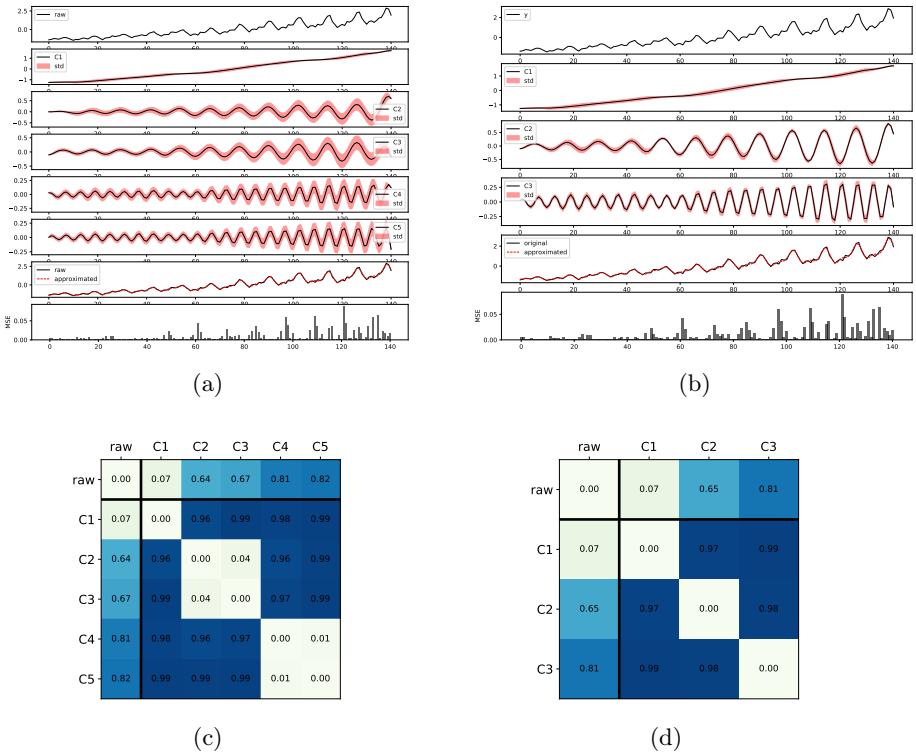

Figure 2: SSA decomposition without (2a) and with (2b) grouping (using weighted-correlations) of passengers series into respectively n_c=5 and n_c=3 (n_e=5); recall that grouping is "turned off" when $n_c = n_e$. The red shaded regions in the components plots correspond to standard deviations of antidiagonals during *diagonal averaging* (reconstruction). Resulting pairwise dependence is shown in the distance matrix (the larger, the more uncorrelated) (2c and 2d). Note that, (C1, C2) and (C3, C4) are clearly correlated; this is confirmed by Fig. 1 and those were (automatically) grouped to give the decomposition shown in 2b.

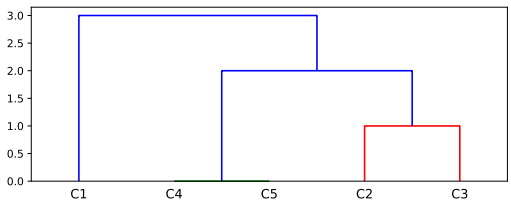

Figure 3: Hierarchical clustering (complete linkage) of elementary components C1, ..., C5 that come from the decomposition of passengers series ($n_c = n_e$=5). One can see that requesting n_c=3 gives {C1 (trend), (C2+C3), (C4+C5)}. Note that, should one ask for n_c=4, grouping would provide the set {C1, C2, C3, (C4+C5)}, which agrees with the distance matrix from Fig. 2c: $d(\text{C4}, \text{C5}) < d(\text{C2}, \text{C3})$; and similarly, n_c=2 would give {C2, (C2+C3+C4+C5)}.

smallest k such that $\sum_{i=1}^{k} \lambda_i / \sum_i \lambda_i > t$, where t is a variance threshold arbitrarily set to 99%, and $\lambda_1 > \lambda_2, \ldots$ are the eigenvalues of the lag-covariance matrix. Note that Donoho[3] proposed other thresholds (on $\sqrt{\lambda_i}$)—that depend on matrix dimensions and whether the noise level is known—which could be used in addition/replacement. From there, if $n_c > k$, the decomposition is elementary and one is likely to obtain noisy components (associated with small eigenvalues); otherwise, k elementary components will be grouped into n_c components.

- choose both a number of components, n_c and a number of eigenvectors (`n_leading_ev`), n_e with $n_c < n_e$. If $n_e > k$, then n_e first eigenvectors are used for elementary projections, but one must know that those associated with small eigenvalues are likely to be noise; the opposite case might result in a coarse decomposition (and therefore a rough approximation) as the proportion of variance explained is less that 99% (here). The elementary components are then grouped into n_c components, according to a grouping method.

1.3 Grouping

Since some of the elementary components may end up being correlated, we group them using hierarchical clustering (Fig 3). The distance matrix, D can either be $D_{ij} = 1 - \rho_{ij}$ where ρ is called the weighted-correlation or w-correlation:

$$\rho_{ij} = \frac{< C^{(i)}, C^{(j)} >_w}{||C^{(i)}||_w ||C^{(j)}||_w} \tag{1}$$

where $C^{(i)} = \{c_m^{(i)}\}_{m=1,..,N}$ are the reconstructed elementary components X_i (referred to as C1, C2, etc. in the plots) via diagonal averaging (see Section 1.4);

[3](Gavish and Donoho, 2014) The Optimal Hard Threshold for Singular Values is $4/\sqrt{3}$ (arXiv:1305.5870)

$< C^{(i)}, C^{(j)} >_w = \sum_m w_m c_m^{(i)} c_m^{(j)}$, with weights $w_m = \min(m, L, N-m)$ (i.e., the number of elements along the m-th antidiagonal of X_i or X_j, since they have the same dimensions); $||C^{(i)}||_w = \sqrt{< C^{(i)}, C^{(j)} >_w}$. Weighted-correlation quantifies how well different components can be separated from each other and can be seen as a measure of orthogonality (it relies on a dot product): the more orthogonal, the closer to 0 (and $d_{ij} \to 1$).

Else, the distance matrix can simply be the Euclidean distance between eigenvalues, though we observed more stable and interpretable results using the former.

1.4 Diagonal averaging

As the name suggests (actually it does not, strictly speaking), diagonal averaging consists of averaging elements of the antidiagonals of X_i. Because elementary components are only quasi-Hankel matrices, their antidiagonals are not constant, as opposed to the trajectory matrix defined in Section 1. Averaging is one way to reconstruct time components, although it is not clear why it should be preferred to e.g. median, min or max. From there we can also extract the standard deviation; it is worth mentioning that grouped components are closer to Hankel matrices, as illustrated in Fig. 2 (grouping actually converges to a Hankel structure, since one gets closer to the original series as components are summed).

1.5 Future work

1.5.1 "Reduced dimension" parameter

The fix would consist of using (i) the threshold introduced in (Gavish and Donoho, 2014) (on which the user has no control), which sets an upper bound on the number of eigenvectors to use (ii) and a threshold on the explained variance. The former would be overridden by the latter (if given by the user) and likely result in coarser decompositions.

1.5.2 Non-linear ssa

The idea would be to use principal geodesic analysis (PGA) in order to consider the geometric structure of the lag-covariance matrix (which is disregarded in the Euclidean formulation of SSA). One could also think of carrying out grouping in Riemannian geometry using components' covariance matrices—as opposed to, for now, grouping them in the time domain (which requires their reconstruction) based on a correlation criterion.

2 Missing value imputation

2.1 Principle

2.2 Results

2.3 Future work