2009 International Symposium on Nonlinear Theory and its Applications
NOLTA'09, Sapporo, Japan, October 18-21, 2009

Singular Spectrum Analysis of Equatorial Precipitation Data

Naoki Itoh† and Jürgen Kurths

†Interdisciplinary Center for Dynamics of Complex Systems
University of Potsdam, D-14476 Potsdam, Germany
Email: naoki@agnld.uni-potsdam.de

Abstract—Singular spectrum analysis (SSA) is adopted to the time series of the monthly equatorial precipitation data observed at three stations, Nakuru [1904-1991], Naivasha [1950-1985], and Narok [1913-1991], in Kenya, to explain how climate works in tropical East Africa. By singular value decomposition (SVD) method in the SSA technique, basically, the bivariate precipitation data is mathematically decomposed and then a similarity between the two kinds of time series is discussed. A comparison of the data is performed by a heterogeneous correlation and an expansion coefficient. Annual structures obtained by estimates of this correlation show a similar form in each pair of the stations. 1-st expansion coefficient which explains the most dominant feature in the precipitation, shows relatively high values in rainy season of spring. In the next study some sub time series with background features can be represented by applying caterpillar-SSA to each single time series. A harmonic curve of 12 month cycle resulted in the 1-st mode by using this method can be interpreted as the most dominant characteristics in the time series. Such a result can be obtained at all the stations, i.e. there exists such a common seasonal cycle in Kenya. And results of the shorter cycles in a seasonal sense obtained from the other sub time series, can be interpreted as cycles of a rainy season. On the other hand, 15 month cycle is shown in the higher modes of the sub time series in Nakuru, which may be interpreted as an irregular period in the sense of annual cycle.

1. Introduction

One of the goals of time series analysis is an extraction of properties from time series by using the singular spectrum analysis (SSA) based on the principal component analysis (PCA). In many previous studies of the SSA, successful findings have been provided [1–5, 7]. A basic advantage of the SSA is possible to apply to both of a square matrix and a rectangular matrix [1]. Besides, some spots which should be notable in all field are as follows: 1.finding trends of different resolution; 2.smoothing; 3.extraction of seasonality components; 4.simultaneous extraction of cycles with small and large periods; 5.extraction of periodicities with varying amplitudes; 6.simultaneous extraction of complex trends and periodicities; 7.finding structure in short time series; and 8.change-point detection. The above points will lead well into interpretation of an objective time series [5].

In section 2, procedures of the SSA are briefly explained: i) comparative analysis of bivariate data and ii) analysis for a single time series. Then section 3 describes their results. Especially, in the first approach the similarity between two different data is discussed and the results of the next approach are focused on the 3. and 4. of the above 8 problems. Finally, some conclusions are described in the section 4.

2. Methods

SSA is essentially a model-free technique. In order to find out background properties in observed data series, this method will decompose the original time series into the sub time series. These decomposition time series are shown, in principle, in descending order of strong correlation for the original time series, because it follows a similar process to the eigenvalue decomposition.

2.1. Singular Value Decomposition (SVD)

If there are bivariate or multivariate data, SVD in the SSA can be applied to each pair of data.

Assume that there are two different kinds of standardized data matrices, X : $(n \times m)$ and Y : $(n \times l)$. The covariance/correlation matrices of them will be calculated as follows:

$$C = \mathrm{X}^T \mathrm{Y}. \tag{1}$$

By using the SVD method the matrix can be formed

$$C = USV^T, \tag{2}$$

where the columns of matrix U are the singular vectors for X, the columns of matrix V the singular vectors for Y, and then the diagonal S contain the singular values($\sqrt{\lambda_1} \geq \cdots \geq \sqrt{\lambda_d} > 0$, $d = min(m, l)$). The k-th singular vector means in general a representation of the k-th dominant spatial patterns for the original time series. In order to find time series describing how each mode of variability oscillates in time, the expansion coefficients are defined as follows [1–3]:

$$E_X = \mathrm{X}U, \quad E_Y = \mathrm{Y}V. \tag{3}$$

2.2. Caterpillar-SSA Technique

Consider that there is a single time series of length N, $Y = (y_1, \cdots, y_N)$. The process of caterpillar-SSA is, mainly,

constructed by a decomposition of the single time series and a reconstruction of the elements.

In the decomposition process, firstly, the single time series will be transformed into the multidimensional one with the following lagged vectors,

$$X_k = (y_k, \cdots, y_{k+L-1})^T, \quad 1 \le k \le K, \tag{4}$$

i.e., it is to create a Hankel matrix form [4–6], which is a trajectory matrix with $(L \times K)$,

$$\mathrm{X} = [X_1, \cdots, X_K] = (x_{ij})_{i,j=1}^{L,K}. \tag{5}$$

The lag number L called a window length, is an integer such that $2 < L < N$. Then let us define $K = N - L + 1$ as a parameter. The relationship between i and j in the Eq. (5) is diagonally constant.

In the next step, the SVD will be applied to the trajectory matrix. This method estimates two kinds of eigenvalues from matrices such as $M = \mathrm{X}\mathrm{X}^T$ and $N = \mathrm{X}^T\mathrm{X}$. Sets of the eigenvalues, Λ_M and Λ_N are sorted in descending order of magnitude. Note that, however, the intersection of them consists of positive elements ($\sqrt{\lambda_1} \ge \cdots \ge \sqrt{\lambda_d} > 0$, where $d = min(L, K)$). If $V_i = \mathrm{X}^T U_i / \sqrt{\lambda_i}$ is defined, then the SVD of the X can be written as

$$\mathrm{X} = \mathrm{X}_1 + \cdots + \mathrm{X}_d, \tag{6}$$

where the matrix X_i can be reformed as $\mathrm{X}_i = \sqrt{\lambda_i} U_i V_i^T$, $(i = 1, \cdots, d)$, having rank 1. The collection $(\sqrt{\lambda_i}, U_i, V_i)$ is called the i-th eigentriple of the matrix X, where $\sqrt{\lambda_i}$ is the singular value, U_i empirical orthogonal functions (EOFs) and V_i principal components (PC) [4, 5].

In order to reconstruct sub time series from the matrices X_i the grouping of them and the summation of the grouped one will be performed firstly. If a group of indices is assumed as $I = \{i_1, \cdots, i_p\}$, the matrix X_I can be defined as $\mathrm{X}_I = \mathrm{X}_{i_1} + \cdots + \mathrm{X}_{i_p}$. Therefore, the trajectory matrix X can be represented as $\mathrm{X} = \mathrm{X}_{I_1} + \cdots + \mathrm{X}_{I_M}$. The above procedure is called the eingentriple grouping.

In the last step of the reconstruction the procedure called the diagonal averaging transfer [4,5] will be applied to each matrix, X_I, $(I = 1, \cdots, M)$.

Assume that there is an $(L \times K)$ matrix, X_I with elements $x_{i,j}^{(I)}$, then the matrix has to be transformed to a standardized matrix by

$$\mathrm{X}_I^* = (x_{ij}^{*(I)})_{i,j=1}^{L,K}, \quad x_{ij}^* = \frac{(x_{ij} - \bar{x_j})}{s_j}, \tag{7}$$

where $\bar{x_j}$ and s_j are a mean and a standard deviation respectively [7]. Elements of reconstructed time series, $\tilde{y}_n^{(I)}$ can be estimated by the diagonal averaging as a following formula [4, 5]:

$$\tilde{y}_n^{(I)} = \begin{cases} \frac{1}{n} \sum_{m=1}^{n} x_{m,(n-m+1)}^{*(I)} & (1 \le n < L) \\ \frac{1}{L} \sum_{m=1}^{L} x_{m,(n-m+1)}^{*(I)} & (L \le n < K) \\ \frac{1}{N-n} \sum_{m=n-K+2}^{N-K+1} x_{m,(n-m+1)}^{*(I)} & (K \le n < N) \end{cases}, \tag{8}$$

whose estimates lead to the original time series as follows:

$$Y \approx (\hat{y}_1, \cdots, \hat{y}_M) + \text{"}trend\ term\text{"}, \ \hat{y}_t = \sum_{s=1}^{M} \tilde{y}_t^{(s)}. \tag{9}$$

3. Applications

In this study, the SSA is applied to precipitation data time series at three stations, Nakuru ([1904-1991], 1056 points), Naivasha ([1950-1985], 432 points) and Narok ([1913-1991], 948 points), in Kenya (see the fig.1) from GHCN v2 database (`http://www.ncdc.noaa.gov/`). In subsection 3.1, these data will be used just for an overlapping time range ([1950-1985]) of all the stations. Next approach in subsection 3.2 will adopt the whole data.

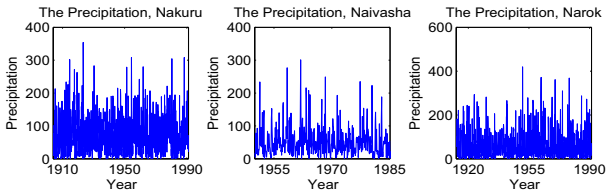

Figure 1: Monthly precipitation in Nakuru, [1904-1991] (left panel), in Naivasha, [1950-1985] (middle panel), and in Narok, [1913-1991] (right panel) from GHCN v2 database.

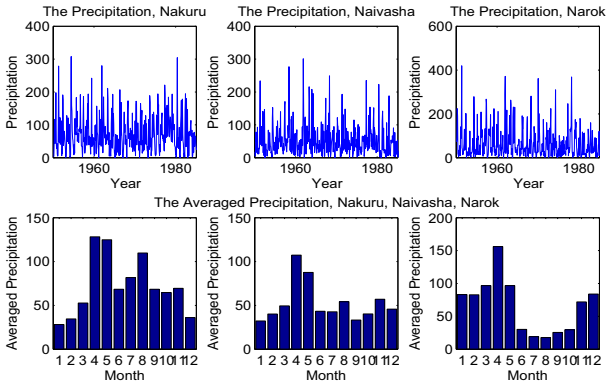

Figure 2: Monthly precipitation [1950-1985], (top panels) and the averages [Jan.-Dec.], (bottom panels): Nakuru (left panels); Naivasha (middle panels); Narok (right panels).

3.1. Correlation Coefficient of Bivariate Data

First approach is to investigate some similarities of bivariate precipitation data by the SVD. These data matrices, however, consist of two kinds of elements that the columns are monthly and the rows are yearly, i.e. the time scales are shown in the both of elements instead of time and space scales. Here, the number of years in the rows is conveniently edited to the overlapping time range, which is, in this case, for 36 years ([1950-1985]), as shown in the figure

2. As a comparison method of the bivariate data, correlation coefficients between expansion coefficients defined in Eq. (3) and the observed data will be estimated.

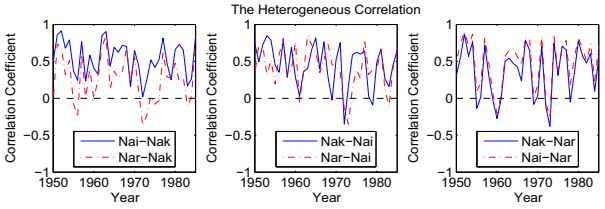

Figure 3: The heterogeneous correlation coefficients: Nakuru-Naivasha (solid line) and -Narok (dashdot-), (left panel); Naivasha-Nakuru (solid-) and -Narok (dashdot-), (middle panel); Narok-Nakuru (solid-) and -Naivasha (dashdot-), (right panel).

The figure 3 shows the correlation coefficients between 1-st expansion coefficient at one station and the observed data at the other station, which is called a heterogeneous correlation. The results represent that these years structures are similar with each other in all the cases of bivariate data. However, the left panel of figure 3 showing the heterogeneous correlation coefficients between Naivasha and Nakuru and that between Narok and Nakuru results that the first pair correlates slightly stronger than the second pair.

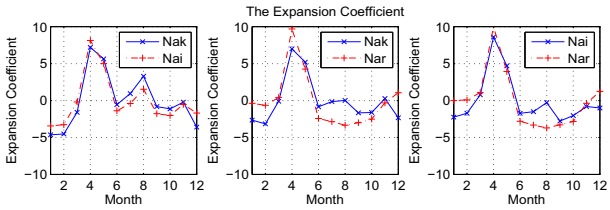

Figure 4: The time series of monthly expansion coefficients at the 1-st mode for the precipitation data: Nakuru (solid line) and Naivasha (dash-), (left panel); Nakuru (solid-) and Narok (dash-), (middle panel); Naivasha (solid-) and Narok (dash-), (right panel).

Furthermore, the time series of the monthly expansion coefficients for each of them are shown in the figure 4. In April and May there are relatively strong coefficients. The result corresponds to a long rainy season as shown in the bottom panels of figure 2, which can be, thus, regarded as one of the dominant properties of the precipitation in Kenya.

How many modes should be taken as the useful components by the SVD can be estimated by the Frobenius norm [4] which is a sum of squared singular values, S. The contribution ratio can be quantified as the squared covariance fraction (SCF) and the cumulative SCF ($CSCF$) [1]:

$$SCF(m) = \frac{l_m^2}{\sum_{r=1}^{R} l_r^2}, \; CSCF(m) = \frac{\sum_{m=1}^{R} l_m^2}{\sum_{r=1}^{R} l_r^2}, \qquad (10)$$

where $l = \sqrt{\lambda}$. The results are shown in the figure 5. The valid modes should be, at the longest, chosen until the 8-th component in all the cases, because the ratio already reaches mostly 100%. It means that the original data may be mostly explained by the information of such modes.

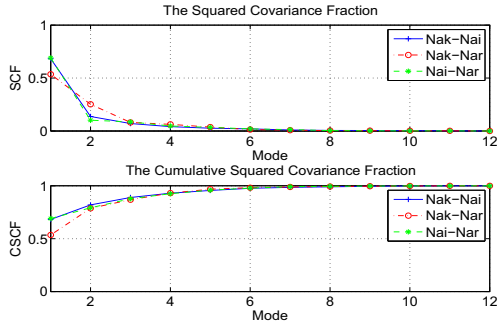

Figure 5: The SCF (top panel) and the CSCF (bottom panels): Nakuru-Naivasha (solid line); Nakuru-Narok (dashdot-); Naivasha-Narok (dash-).

3.2. Caterpillar SSA

The caterpillar SSA [4,5] will be applied to a single time series of precipitation so that the background properties can be shown from the precipitation. Firstly, it has to be transformed to a matrix form by using the Hankelization technique [4–6]. An expansion parameter K for creating the matrix is, in principle, assumed as $K < N/2$, here, as a tenth of the data length, ($K = N/10$), which is then called caterpillar length [4,5]. From the results of the singular values spectrum and the SCF in the figure 6, it is possible to separate the time series reconstructed by the diagonal averaging defined in the Eq. (8) into several groups: the group 1. (1-st,2-nd) mode, -2. (3-rd,4-th), -3. (5-th,6-th), -4. (7-th,8-th), and so on [4, 5]. The results from the group 1. to -3. in the figure 7 are actually not surprising because they are typical seasonal cycles (corresponding to 12-, 6-, and 4 months). Note that there is, however, an untypical period in the seasonal sense in the group 4., which is 15 month cycle. Indeed the contribution of these eigentriples for the original data is low as shown in the figure 6, but this finding may admit of an interpretation that there is some irregular cycle in the precipitation separately from seasonal cycles.

4. Conclusions

In this paper, the two different kinds of the SSA techniques are applied to the precipitation data in Kenya, to investigate climate features of a tropical East Africa. The approach to comparing the precipitation at different locations provides that there are a similarity of years structures and relatively high correlations in the spring rainy season, and

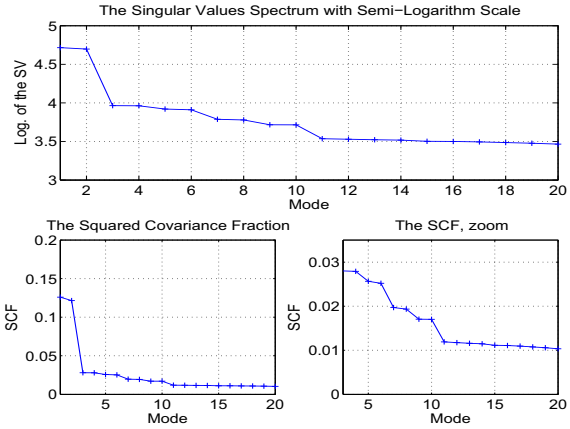

Figure 6: The singular value spectrum with semilogarithm scale (top panel), the SCF (bottom left panel), the SCF with a zoom, 3rd-20th (bottom right panel) of Nakuru.

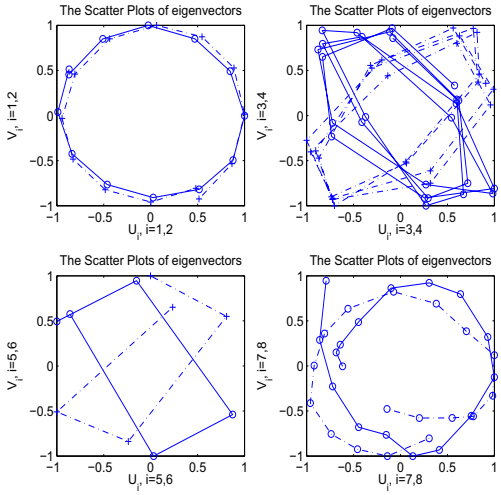

Figure 7: the scatter plots between two kinds of eigenvectors, U and V, in Eq. (6): 12-month period (1st (solid line) and 2nd (dashdot-)), (top left panel); 6-month period (3rd (solid-) and 4th (dashdot-)), (top right panel); 4-month period (5th (solid-) and 6th (dashdot-)), (bottom left panel); 15-month period (7th (solid line) and 8th (dashdot-)), (bottom right panel); in Nakuru.

all the data can be explained by 8 components by the estimates of the contribution ratio. By the caterpillar SSA not only some properties of seasonal harmonics are extracted, but also obviously the anomaly period in the seasonal sense can be found from the time series reconstructed by the diagonal averaging. Since this anomaly period is, however, not yet considered enough from a climate point of view in this paper, it is necessary to associate results with more detail climate background in the future work.

Acknowledgements

The precipitation data from GHCN v2 database (`http://www.ncdc.noaa.gov/`) used in this study was provided by Norbert Marwan (PIK Potsdam), he and Martin Trauth (*Institut für Geowissenschaften*, University of Potsdam) gave me some quite valuable comments and suggestions. And Udo Schwarz (Center for Dynamics of Complex Systems, University of Potsdam) took time to discuss this study. We would like to acknowledge the help of them.

References

[1] H. Björnsson, S.A. Venegas, "A Manual for EOF and SVD analyses of Climate Data," *Centre for Climate and Global Change Research*, Report No. 97–1, pp.52, 1997.

[2] J.M. Wallace, C. Smith, and C.S. Bretherton, "Singular Value Decomposition of Wintertime Sea Surface Temperature and 500-mb Height Anomalies," *Journal of Climate*,vol.5, pp.561–577, Jun. 1992.

[3] S. Minobe, and N. Mantua, "Interdecadal modulation of interannual atmospheric and oceanic variability over the North Pacific," *Progress in Oceanography*,vol.43, pp.163–192, 1999.

[4] N. Golyandina, V. Nekrutkin and A. Zhigljavsky, "Analysis of Time Series Structure," *SSA and related techniques. Chapman & Hall/CRC.*, pp.303, 2001.

[5] H. Hassani, "Singular Spectrum Analysis: Methodology and Comparison," *Journal of Data Science.*,vol.5, pp.239–257, 2007.

[6] J.L. Phillips, "The Triangular Decomposition of Hankel Matrices," *MATHEMATICS OF COMPUTATION*,vol.5, no.115, Jul. 1971.

[7] A. Serita, K. Hattori, C. Yoshino, M. Hayakawa, and N. Isezaki "Principal component analysis and singular spectrum analysis of ULF geomagnetic data associated with earthquakes," *Natural Hazards and Earth System Sciences*,vol.5, pp.685–689, 2005.