Automatic extraction and forecast of time series cyclic components within the framework of SSA

Theodore Alexandrov, Nina Golyandina
theo@pdmi.ras.ru, nina@ng1174.spb.edu

"Caterpillar"-SSA: basic algorithm

- Decomposes time series into sum of additive components
  $F_N = (f_0, \ldots, f_{N-1}) \;\rightarrow\; F_N^{(1)} + \ldots + F_N^{(m)}$
- Provides the information about each component

1. Trajectory matrix construction:
$$F_N \rightarrow \mathbf{X} \in \mathbb{R}^{L\times K} \qquad \mathbf{X} = \begin{bmatrix} f_0 & f_1 & \cdots & f_{N-L} \\ f_1 & f_2 & \cdots & f_{N-L+1} \\ \vdots & \ddots & \ddots & \vdots \\ f_{L-1} & f_L & \cdots & f_{N-1} \end{bmatrix}$$
2. Singular Value Decomposition (SVD):
$$\mathbf{X} = \sum \mathbf{X}_j, \;\; j = 1\ldots d, \quad \mathbf{X}_j = \sqrt{\lambda_j} U_j V_j^{\mathrm{T}}$$
3. Grouping of SVD components:
$$\{1,\ldots,d\} = \bigoplus_{k=1}^{m} I_k, \quad I_k \leftrightarrow \mathbf{X}^{(k)} = \sum_{j\in I_k} \mathbf{X}_j$$
4. Reconstruction by diagonal averaging:
$$\mathbf{X}^{(k)} \rightarrow F_N^{(k)}, \quad F_N = F_N^{(1)} + \ldots + F_N^{(m)}$$

Origins of the approach

- Singular system approach to the method of delays
  Dynamic Systems – analysis of attractors [middle of 80's] (*Broomhead*)
- Singular Spectrum Analysis
  Geophysics/meteorology – signal/noise enhancing, distinguishing of a time series from the red noise realization [90's] (*Vautard, Ghil, Fraedrich, Allen, Smith*)
- "Caterpillar"
  Principal Component Analysis evolution [end of 60's]

Basic possibilities of the "Caterpillar"-SSA technique:

- Finding trends of different resolutions
- Smoothing
- Extraction of seasonality components. It allows to extract:
  - cycles with different periods (simultaneously)
  - periodicities with varying amplitudes
  - complex trends and periodicities (simultaneously)
- Forecast

Advantages

- Doesn't require the knowledge of parametric model of time series
- Works with a wide spectrum of real-life time series
- Matches up for non-stationary time series
- Allows to find structure in short time series

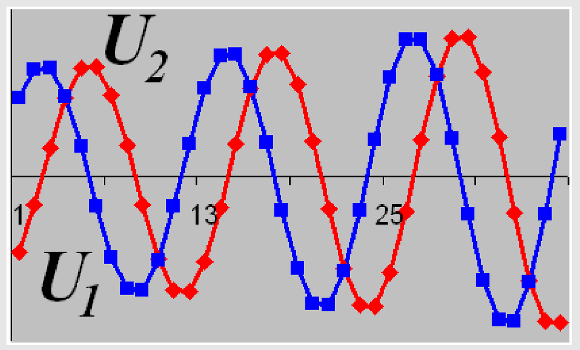
Identification

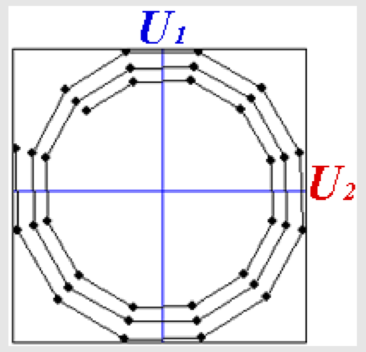
The algorithm of the "Caterpillar"-SSA is managed by
- setting of the L – *window length*,
- grouping components, ET (i.e. eigentriples)

For grouping necessary ETs we need to *identify* them.

An exponential-modulated (e-m) harmonic

$F_N = (f_0, \ldots, f_{N-1}), \quad f_n = Ae^{\alpha n}\cos(2\pi\omega n + \phi)$
- it generates two SVD components,
- eigenvectors:
  $U_1 = (u_1^{(1)}, \ldots, u_L^{(1)})^{\mathrm{T}}: \quad u_k^{(1)} = C_1 e^{\alpha k}\cos(2\pi\omega k + \phi)$
  $U_2 = (u_1^{(2)}, \ldots, u_L^{(2)})^{\mathrm{T}}: \quad u_k^{(2)} = C_2 e^{\alpha k}\sin(2\pi\omega k + \phi)$
  ("exponentially-modulated" form with the same α and ω)

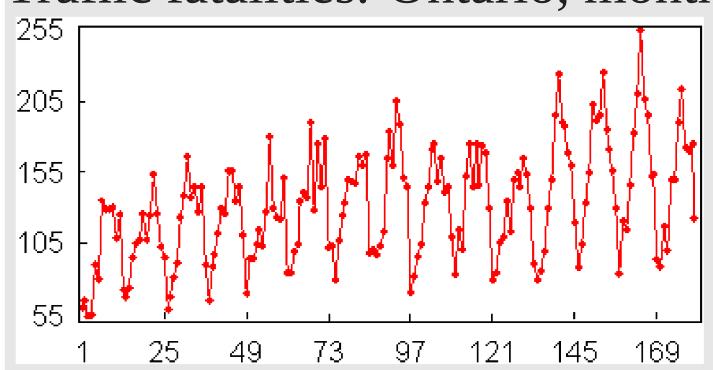




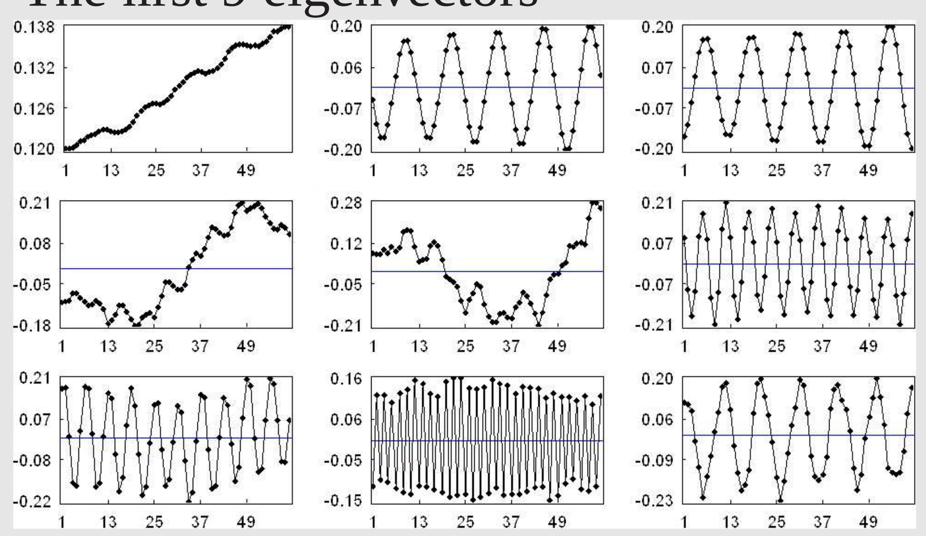
The sequences of elements of U_1, U_2 and the 2D connected scatterplot

Seasonality extraction

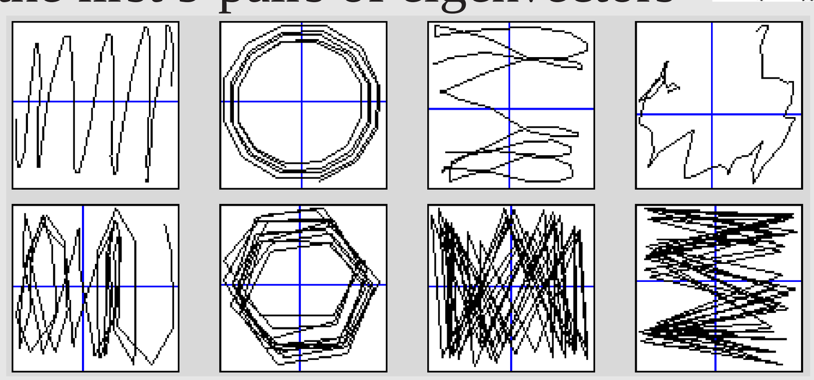
Traffic fatalities. Ontario, monthly, 1960-1974. N=180, L=60.



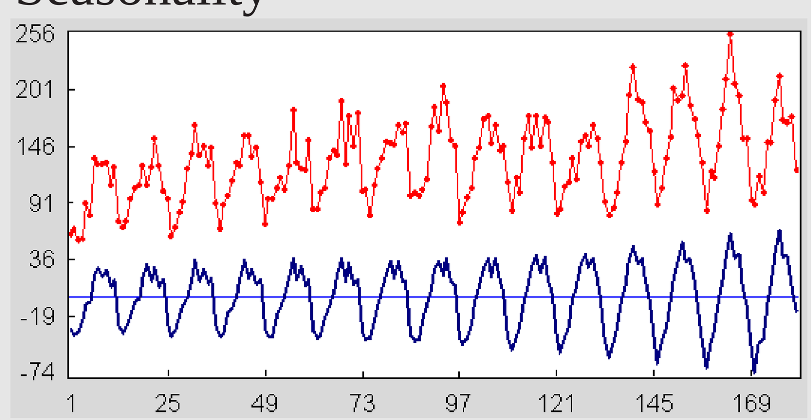
The first 9 eigenvectors



The scatterplots of the first 9 pairs of eigenvectors



Seasonality



SVD components forming seasonality (with periods estimations):
2-3 (12), 6-7(6), 8(2), 11-12(4), 13-14(2.4)

Fourier method: non-interactive identification

Examine eigenvectors U_j. For each $U = (u_1, \ldots, u_L)^{\mathrm{T}}$ consider Fourier decomposition of the sequence $u_1, \ldots, u_L$:
$$u_n = c_0 + \sum_{1\leqslant k\leqslant \frac{L-1}{2}} \big(c_k\cos(2\pi nk/L) + s_k\sin(2\pi nk/L)\big) + (-1)^n c_{L/2},$$
The periodogram of U: $\Pi_U^L(k/L) = \frac{L}{2}\begin{cases} 2{c_0}^2, & k = 0, \\ {c_k}^2 + {s_k}^2, & 1 \leqslant k \leqslant \frac{L-1}{2}, \\ 2{c_{L/2}}^2, & L \text{ is even}, k = L/2. \end{cases}$

$\Pi_U^L(\omega)$, $\omega \in \{k/L\}$, reflects the contribution of a harmonic with the frequency ω into the Fourier decomposition of U.

We define $\quad \rho_{j,j+1} = 0.5\max_{0\leqslant k\leqslant L/2}\left(\Pi_{U_j}^L(k/L) + \Pi_{U_{j+1}}^L(k/L)\right)$

We say that an eigenvectors pair U_j, U_{j+1} is identified as corresponding to some e-m harmonic if periodograms of U_j and U_{j+1} are peaked at the same frequency and $\rho_{j,j+1} \geqslant \rho_0$ for the given threshold $\rho_0 \in [0,1]$.

Facts for e-m harmonic in the lack of noise

For F_N e-m harmonic (on desired conditions):
1. $\alpha = 0 \;\;\Rightarrow\;\; \rho_{j,j+1} = 1$
2. $\alpha > 0 \;\;\Rightarrow\;\; \rho_{j,j+1} \approxeq \widetilde{\rho}(\gamma) = \frac{2}{\gamma}\frac{(e^\gamma - 1)}{(e^\gamma + 1)}$, where $\gamma = \alpha L$

Here we supposed that the value of α is small but L is large.

Noised e-m harmonic

$F_N = (f_0, \ldots, f_{N-1}): \quad f_n = s_n + \sigma e^{\alpha n}\varepsilon_n$, where $s_n = Ae^{\alpha n}\cos(2\pi\omega n)$ and ε_n is the normal white noise with zero mean and unit variance.

We investigated optimal threshold values and their properties by means of statistical simulation.

Results

1. If a time series fits the model described above and we know the values of α,ω,σ, then there is a way of choosing optimal threshold value ρ_0 so that **results of the method Fourier are similar to results of the ideal procedure (visual identification)**.
2. There is a strategy of threshold setting which makes possible **extraction and forecast of e-m harmonic under the condition of deficiency of information about parameters**.

   Example: we know only possible ranges of parameters α,ω,σ.

St.Petersburg State University, Russia

http://www.gistatgroup.com/cat/, http://www.pdmi.ras.ru/~theo/autossa/