Thresholds for methods of automatic extraction of time series trend and periodical components with the help of the “Caterpillar”-SSA approach

Th.Alexandrov, N.Golyandina

theo@pdmi.ras.ru, nina@ng1174.spb.edu

St.Petersburg State University

Signal approximation

$F_N = (f_0, \dots, f_{N-1}) : \quad f_n = s_n + \varepsilon_n,$

$S_N = (s_0, \dots, s_{N-1})$ – determinate signal,
$(\varepsilon_0, \varepsilon_1, \varepsilon_2, \dots, \varepsilon_{N-1})$ – residual (noise).

Signal approximation – in mean-square terms.

We want to approximation such signals:

- non-stationary,
- without information about its parametric model,
- and more, without knowledge of its structure.

“Caterpillar”-SSA approach

The method accomplishes such tasks:

- finding trend of different resolution,
- smoothing,
- seasonality extraction,
- extraction periodicities with changing amplitudes,
- forecast,
- change-point detection.

History:

- USA, UK – SSA (Singular Spectrum Analysis),
- Russia – “Caterpillar”-SSA.

Advantages:

- doesn’t require the knowledge of parametric model of time series,
- processes wide spectrum of real-life time series,
- match up for non-stationary time series,
- work with such natural components as modulated harmonics.

"Caterpillar"-SSA: base algorithm

- Decomposition into sum of components: $F_N = F_N^{(1)} + \ldots + F_N^{(m)}$.
- Gives the information about each component.

Algorithm:

1. Trajectory matrix construction: $F_N \to \mathbf{X} \in \mathbb{R}^{L \times K}$ (L – window length, parameter)

$$\mathbf{X} = \begin{bmatrix} f_0 & f_1 & \ldots & f_{N-L} \\ f_1 & f_2 & \ldots & f_{N-L+1} \\ \vdots & \ddots & \ddots & \vdots \\ f_{L-1} & f_L & \ldots & f_{N-1} \end{bmatrix}.$$

2. Singular Value Decomposition (SVD): $\mathbf{X} = \sum \mathbf{X}_j$,

$\mathbf{X}_j = \sqrt{\lambda_j} U_j V_j^{\mathrm{T}}$,
λ_j – e.val. $\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$, U_j – e.v-r $\mathbf{S}$,
V_j – e.v-r $\mathbf{S}^{\mathrm{T}}$, $V_j = \mathbf{X}^{\mathrm{T}} U_j \sqrt{\lambda_j}$.

3. Components grouping SVD: $\{1, \ldots, d\} = \bigoplus I_k$,

$\mathbf{X}^{(k)} = \sum_{j \in I_k} \mathbf{X}_j$.

4. Reconstruction by diagonal averaging: $\mathbf{X}^{(k)} \to \widetilde{F_N}^{(k)}$.

Grouping

Common case: $F_N = F_N^{(1)} + F_N^{(2)}$ $\qquad I_1 : \mathbf{X}^{(1)} \leftrightarrow \widetilde{F_N}^{(1)}$.

Grouping is possible, if:

1. $F_N^{(1)}$ – has finite amount of components,
2. $F_N^{(1)}$ is separable from a residual.

Approximation case:

$$F_N = F_N^{(1)} + F_N^{(2)}$$

signal, noise

$I_1 : \mathbf{X}^{(1)} \leftrightarrow \widetilde{F_N}^{(1)}$ – approximation of a signal.

1. Every linear combination of multiplication of **exponents**, **e-m harmonics** and **polynomials** has finite amount of components.

2. Asymptotic separability examples:
   - A determinate signal is asympt. separable from a white noise.
   - A periodicity is asympt. separable from a trend.

Identification

Identification – choosing of components during grouping.

Exponential trend: $f_n = Ae^{\alpha n}$.

- it generates one SVD component,
- eigenvector:
  $U = (u_1, \ldots, u_L)^{\mathrm{T}}: \quad u_k = Ce^{\alpha k}$.
  ("exponential" form with the same α)

Exponentially-modulated harmonic: $f_n = Ae^{\alpha n}\cos(2\pi\omega n)$.

- it generates two SVD components,
- eigenvectors:

$U_1 = (u_1^{(1)}, \ldots, u_L^{(1)})^{\mathrm{T}}: \quad u_k^{(1)} = C_1 e^{\alpha k}\cos(2\pi\omega k)$.

$U_2 = (u_1^{(2)}, \ldots, u_L^{(2)})^{\mathrm{T}}: \quad u_k^{(2)} = C_2 e^{\alpha k}\sin(2\pi\omega k)$.

("exponentially-modulated" form with the same α и ω)

Identification

Identification – choosing of components during grouping.

Exponential trend: $f_n = Ae^{\alpha n}$.

- it generates one SVD component,
- eigenvector:
  $U = (u_1, \ldots, u_L)^{\mathrm{T}}: \quad u_k = Ce^{\alpha k}$.
  ("exponential" form with the same α)

Exponentially-modulated harmonic: $f_n = Ae^{\alpha n}\cos(2\pi\omega n)$.

- it generates two SVD components,
- eigenvectors:

$$U_1 = (u_1^{(1)}, \ldots, u_L^{(1)})^{\mathrm{T}}: \quad u_k^{(1)} = C_1 e^{\alpha k}\cos(2\pi\omega k).$$

$$U_2 = (u_1^{(2)}, \ldots, u_L^{(2)})^{\mathrm{T}}: \quad u_k^{(2)} = C_2 e^{\alpha k}\sin(2\pi\omega k).$$

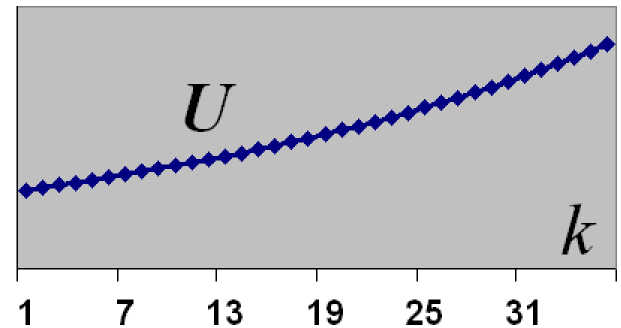
("exponentially-modulated" form with the same α и ω)

Identification

Identification – choosing of components during grouping.

Exponential trend: $f_n = Ae^{\alpha n}$.

- it generates one SVD component,
- eigenvector:
  $U = (u_1, \ldots, u_L)^{\mathrm{T}}: \quad u_k = Ce^{\alpha k}$.
  ("exponential" form with the same α)

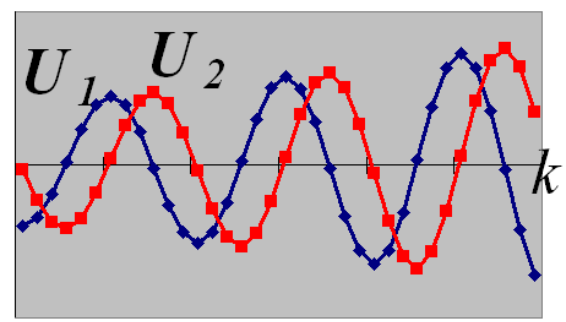


Exponentially-modulated harmonic: $f_n = Ae^{\alpha n}\cos(2\pi\omega n)$.

- it generates two SVD components,
- eigenvectors:

$$U_1 = (u_1^{(1)}, \ldots, u_L^{(1)})^{\mathrm{T}}: \quad u_k^{(1)} = C_1 e^{\alpha k}\cos(2\pi\omega k).$$
$$U_2 = (u_1^{(2)}, \ldots, u_L^{(2)})^{\mathrm{T}}: \quad u_k^{(2)} = C_2 e^{\alpha k}\sin(2\pi\omega k).$$

("exponentially-modulated" form with the same α и ω)



Trend: low frequencies method

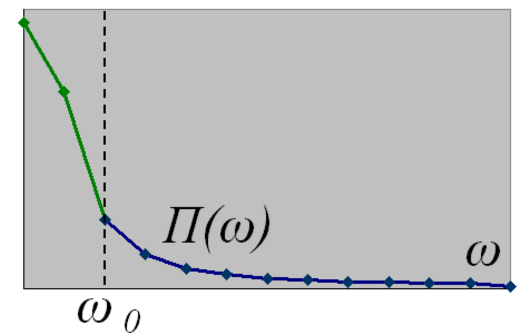
Investigate every eigenvector U_j. Let us take $U = (u_1, \ldots, u_L)^{\mathrm{T}}$.

LOW FREQUENCIES METHOD

- $u_n = c_0 + \sum\limits_{1 \leqslant k \leqslant \frac{L-1}{2}} \big(c_k \cos(2\pi nk/L) + s_k \sin(2\pi nk/L)\big) + (-1)^n c_{L/2},$

- Periodogram:

$$\Pi_U^L(k/L) = \frac{L}{4} \begin{cases} 2{c_0}^2, & k = 0, \\ {c_k}^2 + {s_k}^2, & 1 \leqslant k \leqslant \frac{L-1}{2}, \\ 2{c_{L/2}}^2, & L \text{ – even and } k = L/2. \end{cases}$$



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

- Parameter: ω_0 – upper boundary for the “low frequencies” interval

$$\mathcal{C}(U) = \frac{\sum_{\mathbf{0 \leqslant k \leqslant L\omega_0}} \mathbf{\Pi_U^L(k/L)}}{\sum_{\mathbf{0 \leqslant k \leqslant L/2}} \mathbf{\Pi_U^L(k/L)}}$$ – contribution of LF frequencies.

$$\boxed{\mathcal{C}(U) \geqslant \mathcal{C}_0 \Rightarrow \textbf{e. v-r } U \textbf{ corresponds to a trend.}}$$

($\mathcal{C}_0 \in (0,1)$ – threshold)

LF method: optimal thresholds values

This slide isn't translated and omitted due to its obsoleteness.

Periodicity: Fourier method

Let us investigate sequences of eigenvectors elements U_j,U_{j+1} for all pairs of neighbor components.

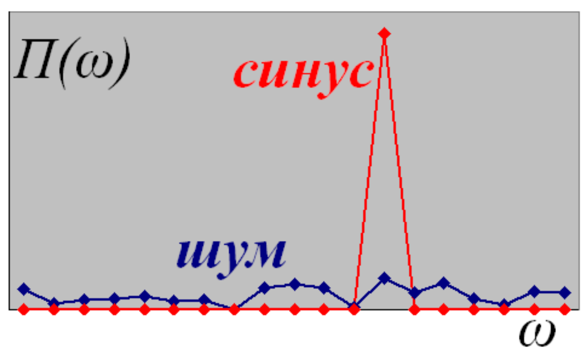
FOURIER METHOD

- **Stage 1.** Check “maximal” frequencies: $\theta_j = \arg\min_k \Pi^M_{U_j}(k/M)$,

  $M|\theta_j - \theta_{j+1}| \leqslant s_0 \Rightarrow$ the pair $(j, j+1)$ is a “harmonical” pair.

- **Stage 2.** Check the form of periodogram:

  $\rho_{(j,j+1)} = \frac{1}{2}\max_k \left(\Pi^M_{U_j}(k/M) + \Pi^M_{U_{j+1}}(k/M)\right)$, for a harm. pair $\rho_{(j,j+1)} = 1$.



$\rho_{(j,j+1)} \geqslant \rho_0 \Rightarrow$ **the pair** $(j, j+1)$ **corresponds to a harmonic.**

($\rho_0 \in (0, 1)$ is the threshold.)

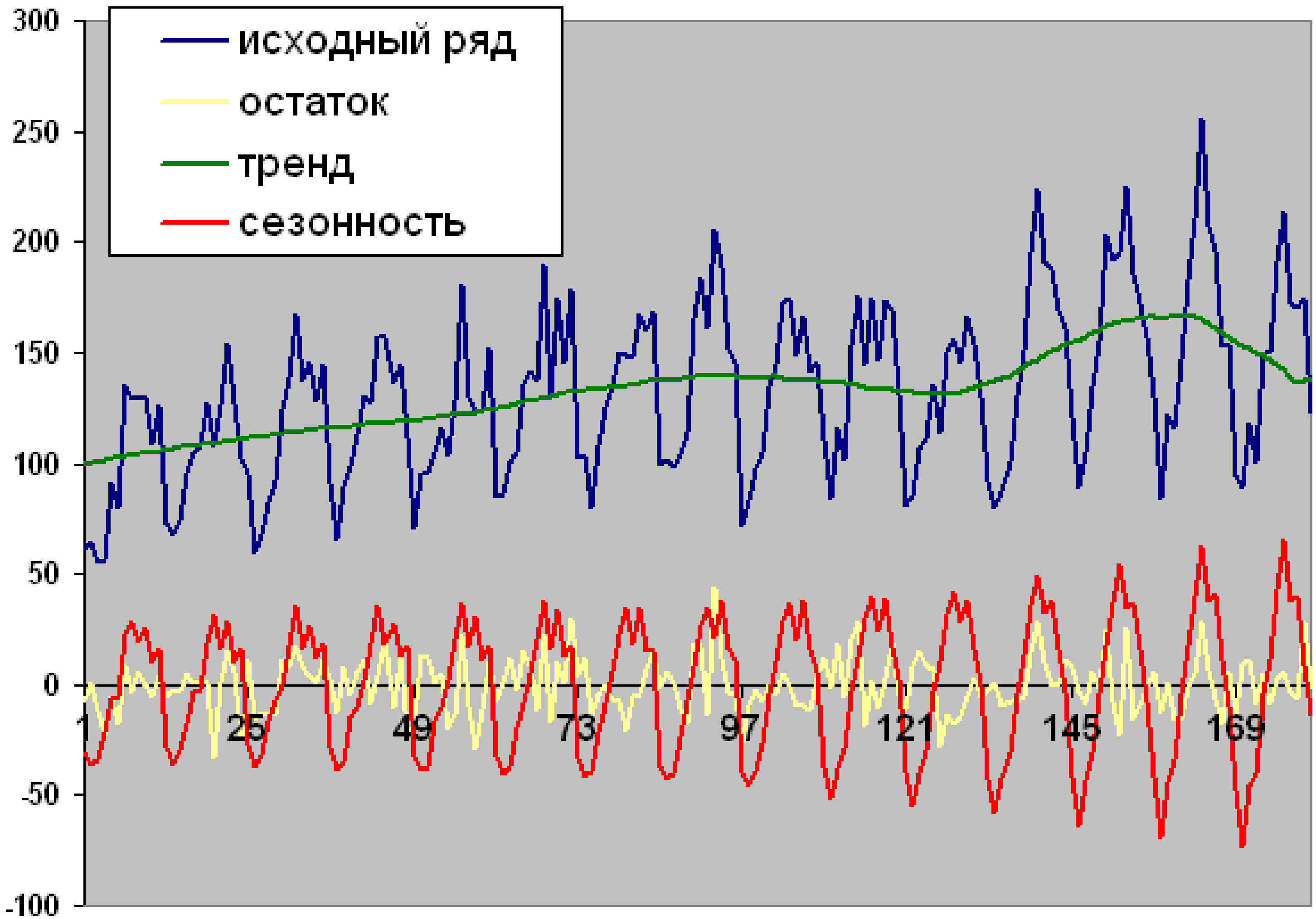
Fourier method: optimal thresholds values

This slide isn't translated and omitted due to its obsoleteness.

Real-life situation

This slide isn't translated and omitted due to its obsoleteness.

Conclusion



Monthly data: traffic fatalities, 1960-1974, Ontario.

Trend components numbers: 1, 4, 5.

Seasonality components numbers: 2, 3, 6-8, 11-14.