Introduction to Singular Spectrum Analysis for series data

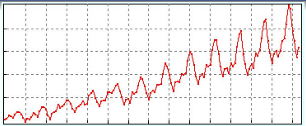
Andrey Pepelyshev

Sheffield
October 8, 2009

Examples of series data

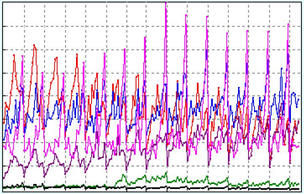
- time series

$$f_1 \quad f_2 \quad f_3 \quad \dots \quad f_n$$



- multivariate time series

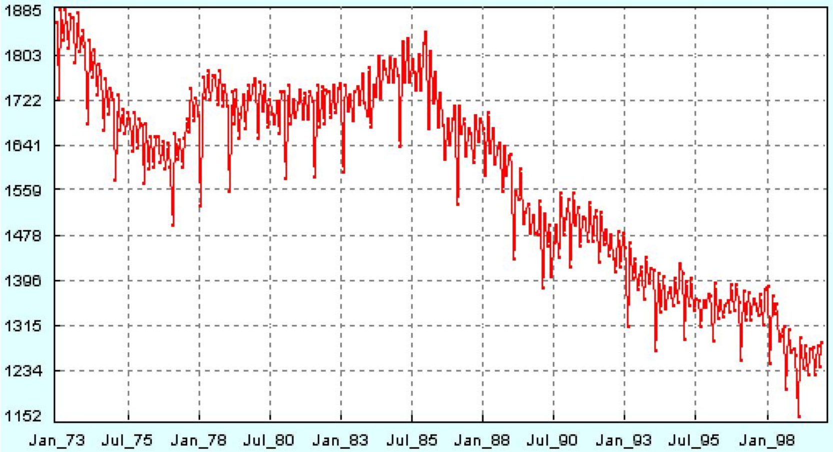
$$\begin{array}{ccccc} f_1 & f_2 & f_3 & \dots & f_n \\ g_1 & g_2 & g_3 & \dots & g_n \\ h_1 & h_2 & h_3 & \dots & h_n \end{array}$$



- images

$$\begin{array}{ccccc} f_{1,1} & f_{1,2} & f_{1,3} & \dots & f_{1,n_2} \\ f_{2,1} & f_{2,2} & f_{2,3} & \dots & f_{2,n_2} \\ \vdots & \vdots & \vdots & & \vdots \\ f_{n_1,1} & f_{n_1,2} & f_{n_1,3} & \dots & f_{n_1,n_2} \end{array}$$

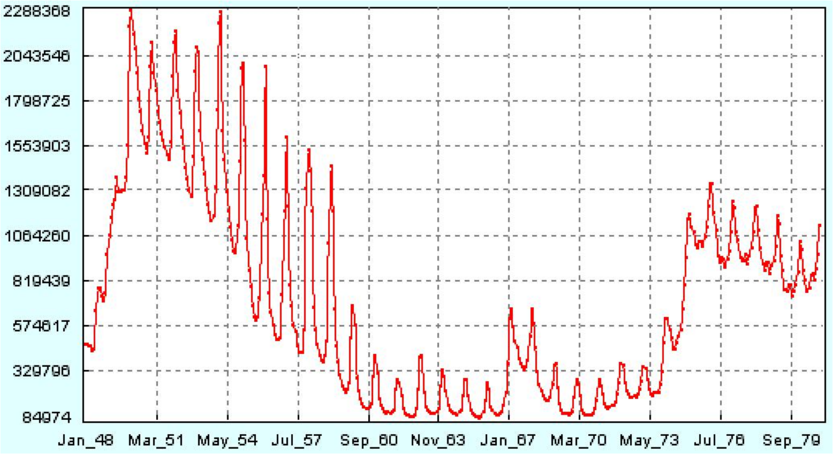
Example 1. Crude Oil Production



Monthly data, Y-unit is thousand barrels.

http://www.economagic.com/em-cgi/data.exe/frbg17/G211_ipsa

Example 2. Unemployment



Monthly data, West Germany, 1948-1980.

http://robjhyndman.com/TSDL/data/subbrao3.dat

Elements of structure to be discovered

- trend components
- oscilatory components
- noise components

Nonparametric approach for structure discovering

Nonparametric approach for structure discovering

Singular Spectrum Analysis

How it works

The vector of time series is not convenient to discover a structure.

f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9

The matrix is needed!

Transformation of data

f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9

f_1
f_2
f_3

Transformation of data

$$f_1 \quad \boxed{f_2 \mid f_3 \mid f_4} \quad f_5 \quad f_6 \quad f_7 \quad f_8 \quad f_9$$

f_1	f_2
f_2	f_3
f_3	f_4

Transformation of data

$f_1 \quad f_2 \quad \boxed{f_3}\boxed{f_4}\boxed{f_5} \quad f_6 \quad f_7 \quad f_8 \quad f_9$

f_1	f_2	f_3
f_2	f_3	f_4
f_3	f_4	f_5

Transformation of data

$$f_1 \quad f_2 \quad f_3 \quad \boxed{f_4 \mid f_5 \mid f_6} \quad f_7 \quad f_8 \quad f_9$$

f_1	f_2	f_3	f_4
f_2	f_3	f_4	f_5
f_3	f_4	f_5	f_6

Transformation of data

$$f_1 \quad f_2 \quad f_3 \quad f_4 \quad \boxed{f_5 \,|\, f_6 \,|\, f_7} \quad f_8 \quad f_9$$

f_1	f_2	f_3	f_4	f_5
f_2	f_3	f_4	f_5	f_6
f_3	f_4	f_5	f_6	f_7

Result of transformation

f_1	f_2	f_3	f_4	f_5	f_6	f_7	$\cdots$	f_n

$= F$

f_1	f_2	f_3	$\cdots$	f_{n-L+1}
f_2	f_3	f_4	$\cdots$	f_{n-L+2}
f_3	f_4	f_5	$\cdots$	f_{n-L+3}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
f_L	f_{L+1}	f_{L+2}	$\cdots$	f_n

$= \mathbf{X}$

$\mathbf{X}$ is called a trajectory matrix, $2 \leq L < n - L + 1$

Decomposition of trajectory matrix

$$\mathbf{X} = \sqrt{\lambda_1}U_1V_1^T + \sqrt{\lambda_2}U_2V_2^T + \sqrt{\lambda_3}U_3V_3^T + \dots$$

- (λ_i, U_i, V_i) is called an eigentriple
- λ_i is an eigenvalue of $\mathbf{X}\mathbf{X}^T$, $i = 1, \dots, L$
- U_i is called an eigenfunction
- V_i is called a factor vector
- $\dfrac{\lambda_i}{\sum_{j=1}^{L} \lambda_j}$ is a ratio of ith component

Decomposition of series

$$\mathbf{X} = \sqrt{\lambda_1}U_1V_1^T + \sqrt{\lambda_2}U_2V_2^T + \sqrt{\lambda_3}U_3V_3^T + \dots$$

$$\Uparrow \qquad\qquad \Downarrow \qquad\qquad \Downarrow \qquad\qquad \Downarrow$$

$$F \overset{\circlearrowright}{=} F_1 + F_2 + F_3 + \dots$$

- F_i is a trend component
- F_j is a harmonic component
- F_l is a noise component

Why it works

The equivalence of three definitions

- Series of finite rank
- Series of finite order
- Series of finite difference dimension

1. Series of finite rank

Definition

The series F has a rank d if the dimension of linear space spanned on columns of trajectory matrix is d for any $L > d$ and $N - L > d$, i.e. $\mathrm{rank}(F) := \mathrm{rank}(\mathbf{X})$.

$$\mathbf{X} = \begin{array}{|c|c|c|c|c|} \hline f_1 & f_2 & f_3 & \cdots & f_{n-L+1} \\ f_2 & f_3 & f_4 & \cdots & f_{n-L+2} \\ f_3 & f_4 & f_5 & \cdots & f_{n-L+3} \\ \vdots & \vdots & \vdots & & \vdots \\ f_L & f_{L+1} & f_{L+2} & \cdots & f_n \\ \hline \end{array}$$

2. Series of finite order

Definition

The series F has a order d if there exist two systems of functions

$$\begin{aligned} \phi_1, \ldots, \phi_d &: \{1, \ldots, L\} \to \mathbb{R}, \\ \psi_1, \ldots, \psi_d &: \{1, \ldots, N-L+1\} \to \mathbb{R} \end{aligned}$$

such that

$$f_{i+j} = \sum_{k=1}^{d} \phi_k(i)\psi_k(j)$$

for any $i = 1, \ldots, L,\ j = 1, \ldots, N-L+1$.

Example. Let $f_i = a^i$. Then $\mathrm{ord}(F) = 1$ since

$$f_{i+j} = a^{i+j} = \phi_1(i)\psi_1(i),$$

$\phi_1(i) = a^i$ and $\psi_1(j) = a^j$.

3. Series of finite difference dimension

Definition

The series F has a finite difference dimension d if there are numbers $\alpha_1, \ldots, \alpha_d$ such that $\alpha_d \neq 0$ and the series F satisfies by the linear recurrent formulae with coefficients $\alpha_1, \ldots, \alpha_d$, that is

$$f_{i+d} = \sum_{k=1}^{d} \alpha_k f_{i+d-k}$$

for any $i = 1, \ldots, n-d$.

Example. Let $f_i = a^i$. Then $\dim(F) = 1$ since $f_{i+1} = \alpha f_i$ for $\alpha = a$.

Results from theory of diff. equations

- Ordinary differential equation of order d w.r.t. $y = y(t)$

$$y^{(n)} + c_{n-1}y^{(n-1)} + \ldots + c_1 y' + c_0 y = 0$$

- The characteristic polynomial has n roots, $\gamma_k \in \mathbb{C}$

$$z^n + c_{n-1}z^{n-1} + \ldots + c_1 z + c_0 = \prod_{k=1}^{h}(z - \gamma_k)^{m_k}$$

and m_k is the corresponding multiplicities.

- Then the functions

$$y_{k,j}(t) = t^j e^{-\gamma_k t},\ j = 0, \ldots, m_k - 1,\ \kappa = 1, \ldots, h$$

are n linearly independent solutions of ODE.

Consequences

- The series given by

$$f_t = \sum_k P_k(t) e^{-\mu_k t} \sin(\alpha_k t + \beta_k), \; t = 1, \ldots, n$$

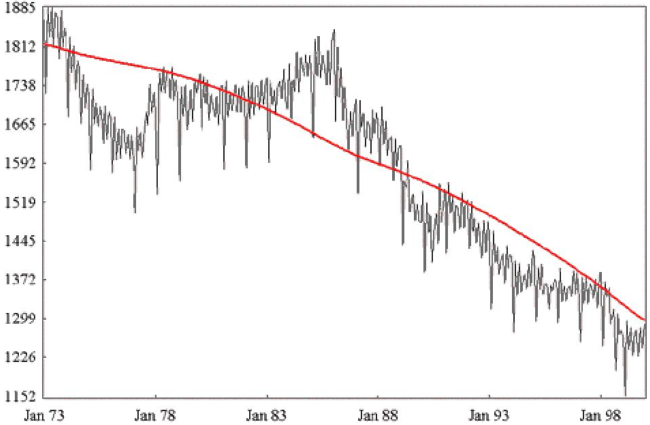
  is a series of finite rank (and of finite order...).
- SVD decomposition of $\mathbf{X}$ for the series of finite rank d has only d nonzero eigenvalues.

Consequences

- The series given by
$$f_t = \sum_k P_k(t) e^{-\mu_k t} \sin(\alpha_k t + \beta_k),\ t = 1, \ldots, n$$
is a series of finite rank (and of finite order...).
- SVD decomposition of $\mathbf{X}$ for the series of finite rank d has only d nonzero eigenvalues.

- Real time series is a sum of a series of finite rank and a disturbance series.
- SSA provide good results in more general cases.

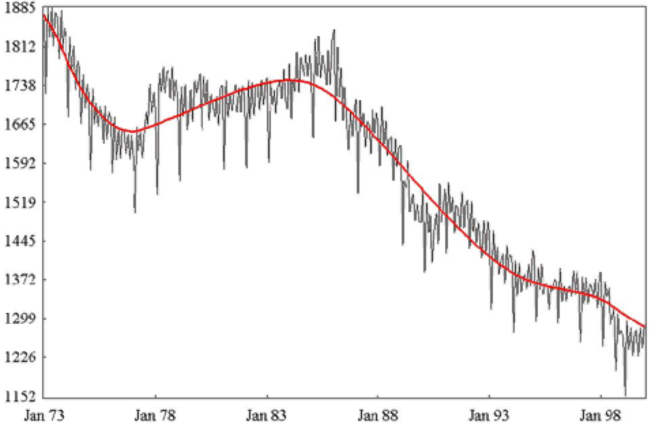
Rough trend extraction



Crude oil production, monthly data from Jan 1973 to Sep 1999

Reconstruction from the leading eigentriple, $L = 120$, $n = 324$

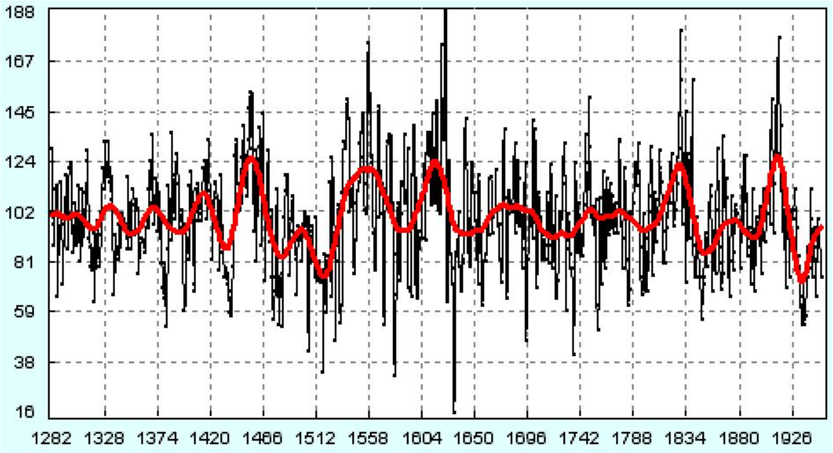
Fine trend extraction



Crude oil production, monthly data from Jan 1973 to Sep 1999

Reconstruction from 3 leading eigentriples, $L = 120$, $n = 324$

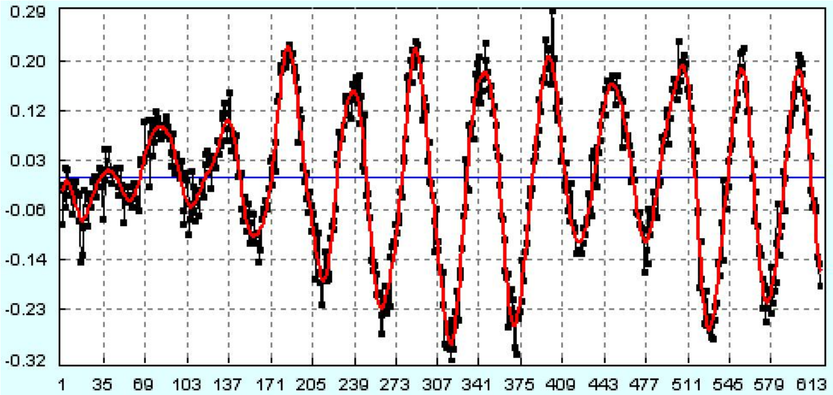
Smoothing



Tree ring indices, Douglas fir, annual, from 1282 to 1950

Reconstruction from 7 leading eigentriples, $L = 120$, $n = 669$

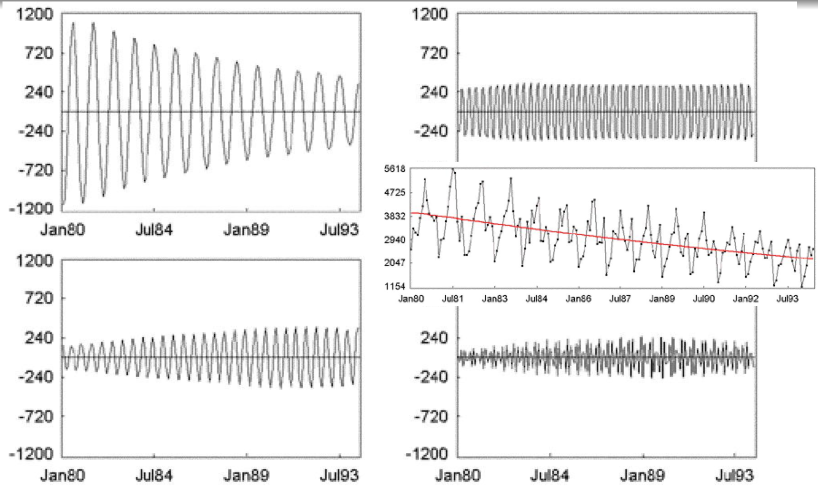
Noise reduction



Intensity of the white dwarf star PG1159-035 during March 1989

Reconstruction from 11 leading eigentriples, $L = 100$, $n = 618$

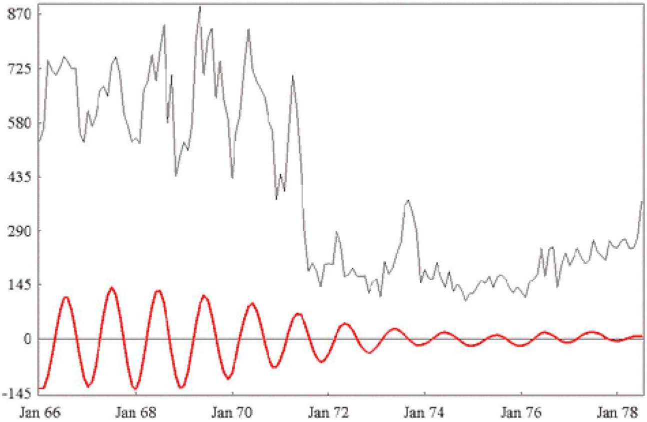
Extraction of seasonality components



Sales of fortified wines in Australia, monthly, Jan 1980-Jun 1994

Reconstruction from eigentriples 2-3, 4-5, 6-7, 8-9, $L = 84$, $n = 174$

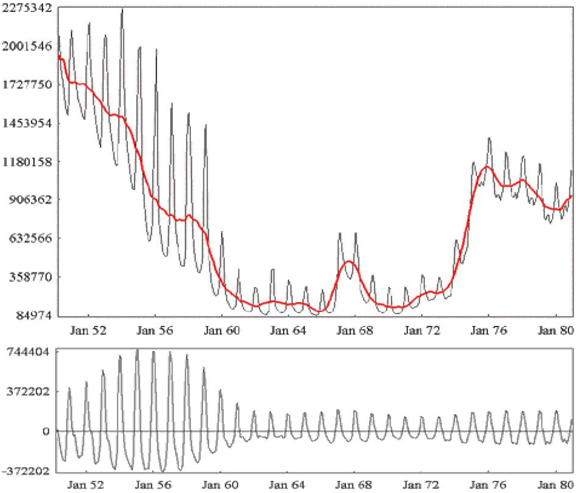
Extraction of periodicities with varying amplitudes



Monthly public drunkenness intakes, from Jan 1966 to Jul 1978

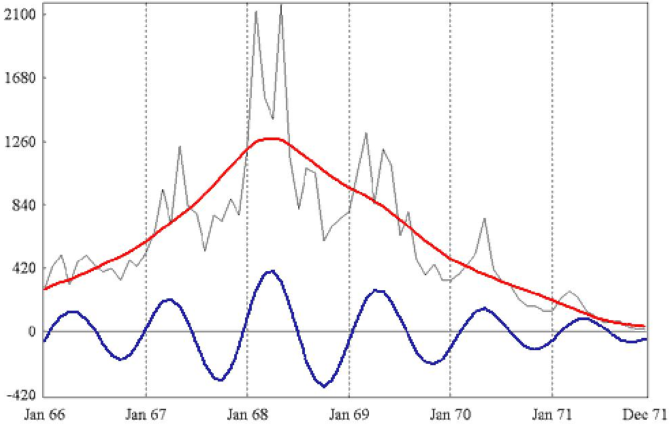
Reconstruction from the 4th-5th eigentriples, $L = 60$, $n = 151$

Complex trends and periodicities



Unemployment, West Germany, monthly, from Apr 1950 to Dec 1980

Finding structure in short time series

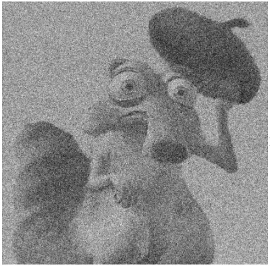


Deaths in the Indochina war, monthly, from 1966 to 1971

Reconstruction from eigentriples 1-2, 3-4, $L = 18$, $n = 72$

Noise reduction in images

⇒



⇒

Original Noised Denoised

Reconstruction from the leading 3 eigentriples,
$L = 16 \times 16$, $n = 310 \times 310$

Conclusions

SSA is a model-free methodology for the analysis of series data.

SSA is applied to a wide range of problems including

- Discovering of structure
- Signal/trend extraction
- Smoothing and noise reduction
- Extraction of harmonic/seasonal components
- Forecasting
- Change-point detection

Acknowledgement

- Centre for Optimisation and Its Applications
  http://www.cardiff.ac.uk/maths/subsites/coia/
- Prof Anatoly Zhigljavsky
- Dr Nina Golyandina
- Mr Konstantin Usevich

References

N. Golyandina, V. Nekrutkin, A. Zhigljavsky (2001) Analysis of Time Series Structure, SSA and Related Techniques. Taylor & Francis (Chapman & Hall/CRC).

SSA software http://www.gistatgroup.com/cat/