SSA change-point detection for environmental data and monitoring the quality of photovoltaic modules

Andrey Pepelyshev

Institut für Statistik & Wirtschaftsmathematik
RWTH AACHEN

Rimini
May 22, 2013

Contents

- Algorithm of SSA change-point detection
- Possibilities with examples
- ARL in the presence of serial correlation
- Analysis of environmental data
- Application in photovoltaics

SSA for change-point detection

Singular Spectrum Analysis is a nonparametric method that decomposes a time series onto the sum of trend, periodics and noise.

SSA can be used to detect changes

- mean
- variance of noise
- amplitude of periodics
- frequency of periodics
- coefficients of a linear recurrent formulae

SSA change-point detection is proposed in Moskvina, Zhigljavsky (2003, 2007).

Informal description of SSA change-point

We say that there is a change-point in a series

$$\ldots, \underbrace{x_{n+1}, \ldots, x_{n+N}}_{\text{base series}}, \ldots, \underbrace{x_{n+p+1}, \ldots, x_{n+q+L-1}}_{\text{test series}}, \ldots$$

if the 'test' series $x_{n+p}, \ldots, x_{n+q+L-1}$ does not share the structure of the 'base' series $x_{n+1}, \ldots, x_{n+N}$.

SSA change-point algorithm

The main parameter is N, others are L, k, p, q.

Assumptions

- The distance between change-points is at least N.
- The first change-point occurs after N points.
- The parameter N is big enough to estimate a 'structure' of series.

Transformation of a series

Compound vectors $X_1, X_2, \ldots$ from a series $x_1, x_2, \ldots$ by applying the moving window of length L,

$$X_1 = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_L \end{pmatrix}, \quad X_2 = \begin{pmatrix} x_2 \\ x_3 \\ \vdots \\ x_{L+1} \end{pmatrix}, \quad X_3 = \begin{pmatrix} x_3 \\ x_4 \\ \vdots \\ x_{L+2} \end{pmatrix}, \ldots .$$

We say that a series has a structure of order k if $\dim \mathcal{L}(X_1, X_2, \ldots, X_n) = k$ for all $L, n > k$.

Examples of structured series

The series $x_j = A\sin(\gamma j + \omega)$ has a structure of order 2 since

$$X_{j+1} = \cos(\gamma j)\begin{pmatrix} \sin(\gamma+\omega) \\ \sin(2\gamma+\omega) \\ \vdots \\ \sin(L\gamma+\omega) \end{pmatrix} + \sin(\gamma j)\begin{pmatrix} \cos(\gamma+\omega) \\ \cos(2\gamma+\omega) \\ \vdots \\ \cos(L\gamma+\omega) \end{pmatrix}.$$

The series $x_j = \sum_{i=1}^{m} A_i e^{-\beta_i j}\sin(\gamma_i j + \omega_i)$ has a structure of order $2m$.

SSA estimation of a structure

1) For a noisy series $(x_{n+1}, \ldots, x_{n+N})$, we build vectors $X_{n+1}, \ldots, X_{n+N-L+1}$.

2) Make the SVD decomposition $\mathbb{X} = \sum_{i=1}^{L} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$, $\lambda_1 \geq \lambda_2 \geq \ldots$.

3) Select k principal components with largest λ_i. Components with small λ_i correspond to a noise.

4) Define the subspace as $\mathcal{L}(U_1, U_2, \ldots, U_k)$, which describes a structure of series.

Formal description of a change-point

Consider the statistic $D_{n,k,p,q}$ defined as a distance between vectors $X_{n+p+1}, \ldots, X_{n+q}$ and the subspace $\mathcal{L}(U_1, U_2, \ldots, U_k)$,

$$D_{n,k,p,q} = \frac{1}{L(q-p)} \sum_{i=n+p+1}^{n+q} [X_i^T X_i - X_i^T U U^T X_i]$$

where $U = (U_1, \ldots, U_k)$ is a 'structure' of the series $x_{n+1}, \ldots, x_{n+N}$.

Rule.
There is a change-point in a series if $D_{n,k,p,q} > h$.

Asymptotic behaviour

Theorem. [Moskvina, Zhigljavsky (2003)]
Under certain assumptions, we have

$$\frac{D_{n,k,p,q} - a}{s} \approx N(0,1)$$

where

$$a = \mathbf{E} D_{n,k,p,q} \approx \sigma^2 LQ, \quad Q = q - p,$$

and

$$s^2 = \mathrm{Var} D_{n,k,p,q} \approx \sigma^2 \frac{4}{3} Q(3LQ - Q^2 + 1).$$

Final step of algorithm

Define the normalized statistic

$$d_n = D_{n,k,p,q}/D_{n,k,0,N-L}.$$

Consider the process $W_1 = 0, W_2, W_3, \ldots$ defined as

$$W_{n+1} = \max\left\{W_n + d_{n+1} - d_n - 1/(3LQ),\ 0\right\}.$$

Rule.

The point $\tau = n + q + L - 1$ is a change-point if $W_n > h$,

$$h = \frac{2t_\alpha}{LQ}\sqrt{Q(3LQ - Q^2 + 1)/3}$$

and t_α is the $(1-\alpha)$-quantile of the standard normal d.

CUSUM and RS procedure

Define the score statistic $S_n = d_{n+1} - d_n - 1/(3LQ)$, where $d_n = D_{n,k,p,q}/D_{n,k,0,N-L}$.
The process $W_1 = 0, W_2, W_3, \ldots$ has the form of CUSUM,

$$W_{n+1} = \max\left\{W_n + S_n,\ 0\right\}.$$

The Shiryaev-Roberts procedure is

$$R_n = (1 + R_{n-1})e^{S_n},\ \ R_0 = 1.$$

Under traditional settings, the SR procedure is optimal in terms of the average detection delay.

Choice of parameters

1) N should be large enough to sufficiently well estimate a 'structure' of series.
2) Set $L = N/2$, $p = N$, $q = N + 1$.
3) Estimate k from all available data.

Thus, we have the situation where

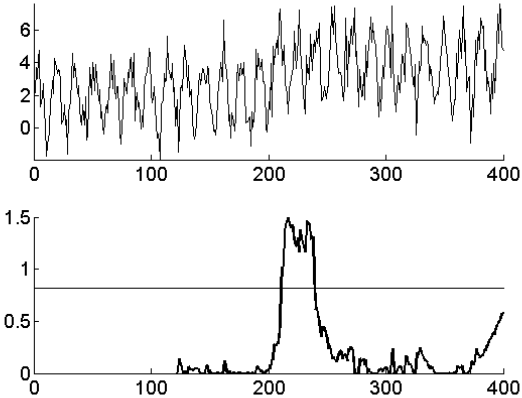
$$\ldots, \underbrace{x_{n+1}, \ldots, x_{n+N}}_{\text{base series}}, \underbrace{x_{n+N+1}, \ldots, x_{n+N+L}}_{\text{test series}}, \ldots$$

or, alternatively,

$$\underbrace{X_{n+1}, \ldots, X_{n+N-L+1}}_{\text{base vectors}}, \quad , \underbrace{X_{n+N+1}}_{\text{test vector}}$$

Change in mean

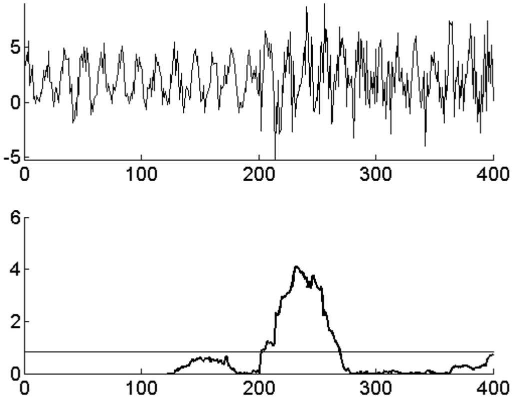
$$x_n = 2 + 2\sin(0.4n) + \varepsilon_n \text{ for } n \le 200 \text{ and}$$
$$x_n = 4 + 2\sin(0.4n) + \varepsilon_n \text{ for } n > 200,\ \varepsilon_n \sim N(0,1)$$



$N = 80,\ k = 3$

Change in variance

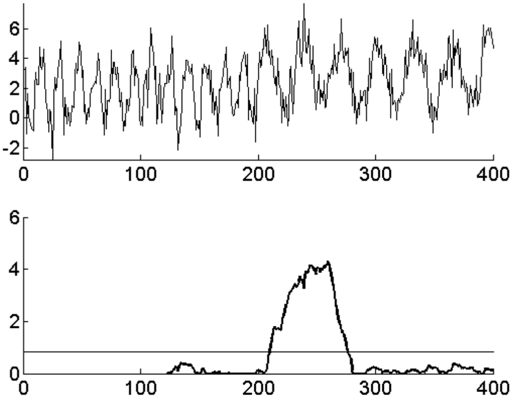
$x_n = 2 + 2\sin(0.4n) + \varepsilon_n$ for $n \le 200$ and
$x_n = 2 + 2\sin(0.4n) + 2\varepsilon_n$ for $n > 200$, $\varepsilon_n \sim N(0, 1)$



$N = 80$, $k = 3$

Change in frequency

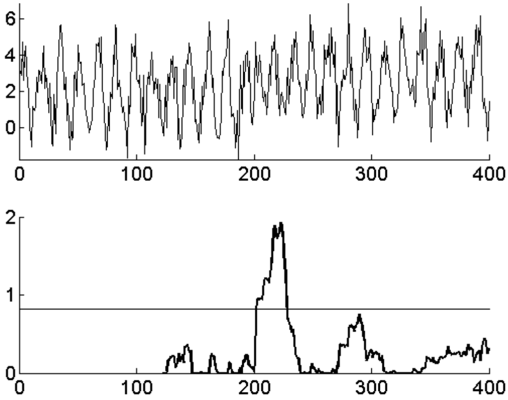
$x_n = 2 + 2\sin(0.4(n-200)) + \varepsilon_n$ for $n \le 200$ and
$x_n = 2 + 2\sin(0.2n) + \varepsilon_n$ for $n > 200$, $\varepsilon_n \sim N(0,1)$



$N = 80,\ k = 3$

Change in phase

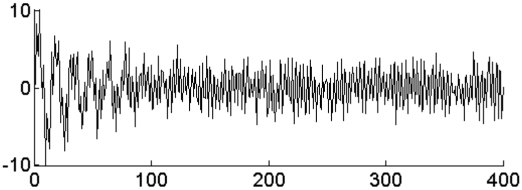
$x_n = 2 + 2\sin(0.4n) + \varepsilon_n$ for $n \le 200$ and
$x_n = 2+2\sin(0.4n+1)+\varepsilon_n$ for $n > 200$, $\varepsilon_n \sim N(0,1)$

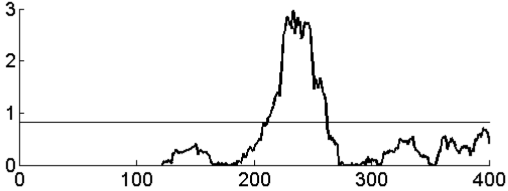


$N = 80,\ k = 3$

Change in AR model with iid noise

$$z_n = -0.96z_{n-4} + z_{n-3} - 0.5z_{n-2} + 0.97z_{n-1},\ n = 5, \ldots, 200,$$
$$z_n = -0.96z_{n-4} + z_{n-3} - 0.7z_{n-2} + 0.97z_{n-1},\ n = 201, \ldots, 400,$$
$$x_n = z_n + \varepsilon_n,\ z_1 = 0,\ z_2 = 8,\ z_3 = 6,\ z_4 = 4$$

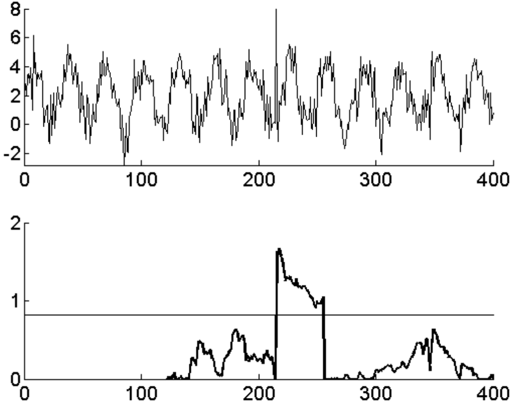




$$N = 80$$
$$k = 4$$

Detection of outliers

$x_n = 2 + 2\sin(0.2n) + \varepsilon_n$ for $n \neq 215$, $x_{215} = 8$, $\varepsilon_n \sim N(0,1)$



$N = 80,\ k = 3$

Average run length for AR(1) and MA(1)

SSA change-point detection with $k = 1$, $N = 80$

$x_n = \mu + \phi(x_{n-1} - \mu) + \varepsilon_n$, $\varepsilon_n \sim N(0, 1)$,
$\mu = 0.5$

ϕ	-0.5	-0.4	-0.2	0	0.2	0.4	0.5
ARL	308	391	498	524	460	224	192

$x_n = \mu + \varepsilon_n - \theta\varepsilon_{n-1}$, $\varepsilon_n \sim N(0, 1)$,
$\mu = 0.5$

θ	-0.5	-0.4	-0.2	0	0.2	0.4	0.5
ARL	272	302	441	524	512	503	430

Bagshaw M., Johnson R. A. (1974,1975)
The Effect of Serial Correlation on the Performance of CUSUM Tests

Average run length for AR(1) and MA(1)

SSA change-point detection with $k = 3$, $N = 80$

$$x_n = \mu + \phi(x_{n-1} - \mu) + \varepsilon_n,\ \varepsilon_n \sim N(0, 1),$$
$$\mu = 2 + 2\sin(0.4n)$$

ϕ	-0.5	-0.4	-0.2	0	0.2	0.4	0.5
ARL	209	333	412	450	355	162	115

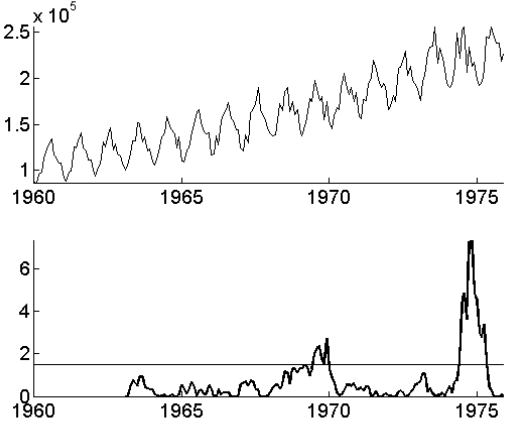
$$x_n = \mu + \varepsilon_n - \theta\varepsilon_{n-1},\ \varepsilon_n \sim N(0, 1),$$
$$\mu = 2 + 2\sin(0.4n)$$

θ	-0.5	-0.4	-0.2	0	0.2	0.4	0.5
ARL	190	243	337	450	423	415	356

Example: gasoline demand

Abraham, Redolter (1983) Statistical Methods for Forecasting, Wiley
`http://robjhyndman.com/TSDL/sales/`

Monthly gasoline demand in Ontario from 1960 to 1975



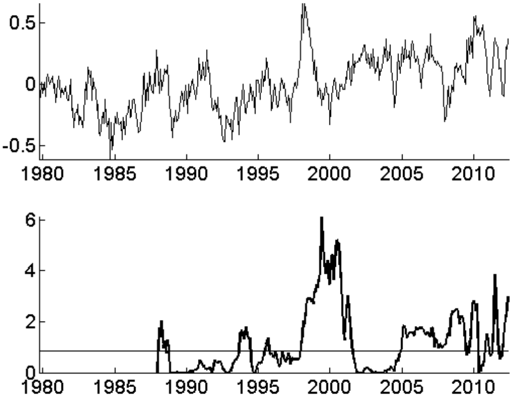
$$N = 24$$
$$k = 5$$

Example: global Earth temperature

National Space Science and Technology Center, USA, NASA

`http://vortex.nsstc.uah.edu/data/msu/t2lt/uahncdc.lt`

Monthly Earth temperatures from Dec 1978 to Jun 2012

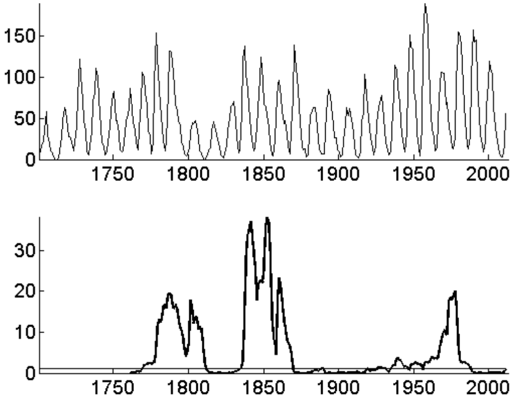


$$N = 72$$
$$k = 9$$

Example: yearly sunspot number

Solar Influences Data Analysis Center, Observatory of Belgium, `http://sidc.oma.be/sunspot-data/`

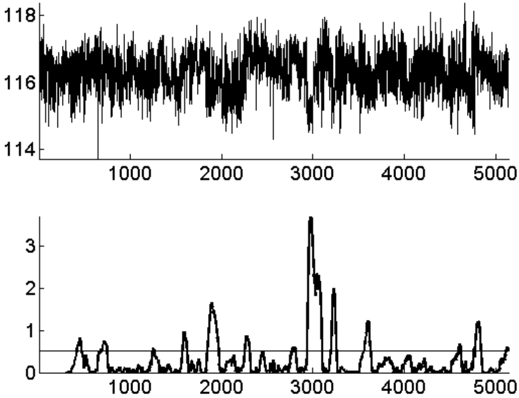
Yearly sunspot number from 1700 to 2011



$$N = 40$$
$$k = 4$$

Power measurements of PV modules

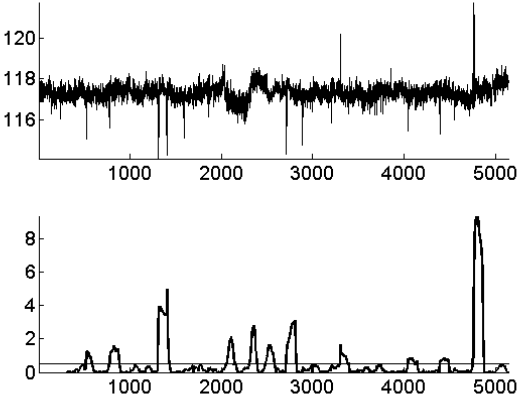
Power of PV modules obtained by using a flasher (sun simulator) from a production line 1



$$N = 200$$
$$k = 1$$

Power measurements of PV modules

Power of PV modules obtained by using a flasher (sun simulator) from a production line 2



$N = 200$
$k = 1$

Thank you for your attention!