The “Caterpillar”-SSA approach: automatic trend extraction and other applications

Theodore Alexandrov

theo@pdmi.ras.ru

St.Petersburg State University, Russia

Bremen University, 28 Feb 2006

History – Present

Origins of the "Caterpillar"-SSA approach

- Singular System Analysis *(Broomhead)*
  Dynamic Systems, method of delays for analysis of attractors [middle of 80's],
- Singular Spectrum Analysis *(Vautard, Ghil, Fraedrich)*
  Geophysics/meteorology – signal/noise enhancing, signal detection in red noise (Monte Carlo SSA) [90's],
- "Caterpillar" *(Danilov, Zhigljavsky, Solntsev, Nekrutkin, Golyandina)*
  Principal Component Analysis for time series [end of 90's],

Present

- Automation: papers are published, see http://www.pdmi.ras.ru/∼theo/autossa/ *(Alexandrov, Golyandina)*
- Change-point detection *(Golyandina, Nekrutkin, Zhigljavsky)*
- Missed observations: a paper is published, a software is on www.gistatgroup.com *(Golyandina, Osipov)*
- 2-channel SSA: a paper is published, see www.gistatgroup.com *(Golyandina, Stepanov)*
- Some generalizations

Future

- 2D, online "Caterpillar"-SSA...

Tasks and advantages

“Caterpillar”-SSA kernel

Theoretical framework, the most important concepts are

- Time series of finite rank (=order of Linear Recurrent Formula)
- Separability (possibility to separate/extract additive components)

General tasks

- Additive components extraction (for example trend, harmonics, exp.modulated harmonics)
- Smoothing (self-adaptive linear filter with small L)
- Automatic calculation of LRF for t.s. of finite rank => prolongation of an extracted additive component => forecast of an extracted additive component
- Change-point detection

Advantages

- Model-free
- Works with non-stationary time series (constrains will be described)
- Suits for short t.s., robust to noise model etc.

"Caterpillar"-SSA basic algorithm

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

Algorithm

1. Trajectory matrix construction:
   $F_N = (f_0, \ldots, f_{N-1})$, $F_N \to \mathbf{X} \in \mathbb{R}^{L \times K}$
   (L – window length, parameter)

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

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

$$\mathbf{X}_j = \sqrt{\lambda_j} U_j V_j^{\mathsf{T}}$$

λ_j – eigenvalue, U_j – e.vector of $\mathbf{X}\mathbf{X}^{\mathsf{T}}$,
V_j – e.vector of $\mathbf{X}^{\mathsf{T}}\mathbf{X}$, $V_j = \mathbf{X}^{\mathsf{T}} U_j / \sqrt{\lambda_j}$

3. Grouping of SVD components:
   $\{1, \ldots, d\} = \bigoplus I_k$,

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

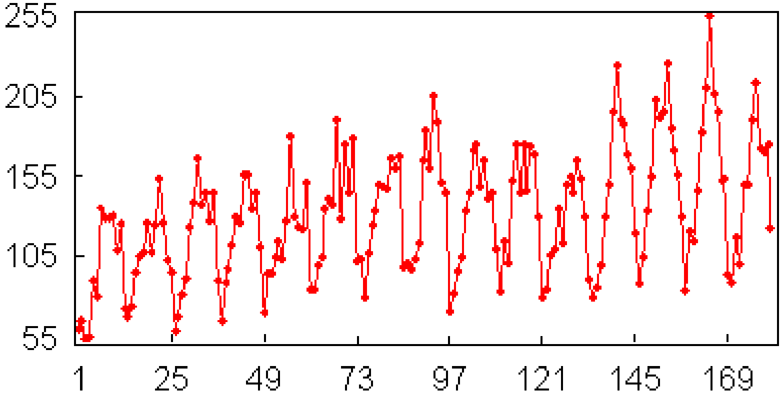
4. Reconstruction by diagonal averaging:
   $\mathbf{X}^{(k)} \to \widetilde{F}_N^{(k)}$

$$F_N = \widetilde{F}_N^{(1)} + \ldots + \widetilde{F}_N^{(m)}$$

Does exist an SVD such that it forms necessary additive component & how to group SVD components?

Example: trend and seasonality extraction

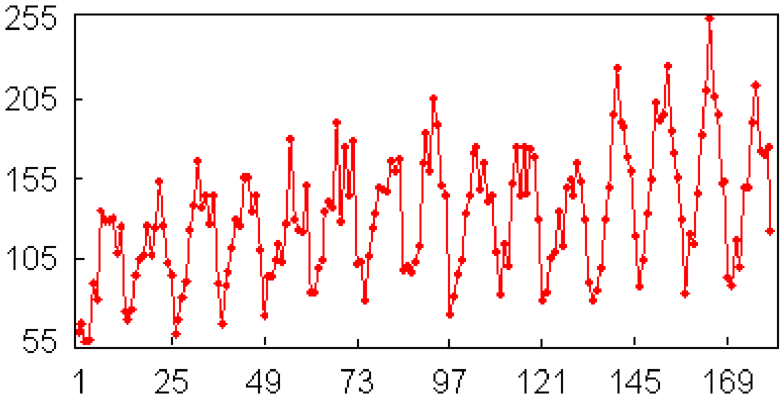
Traffic fatalities. Ontario, monthly, 1960-1974 (Abraham, Redolter. *Stat. Methods for Forecasting*, 1983)



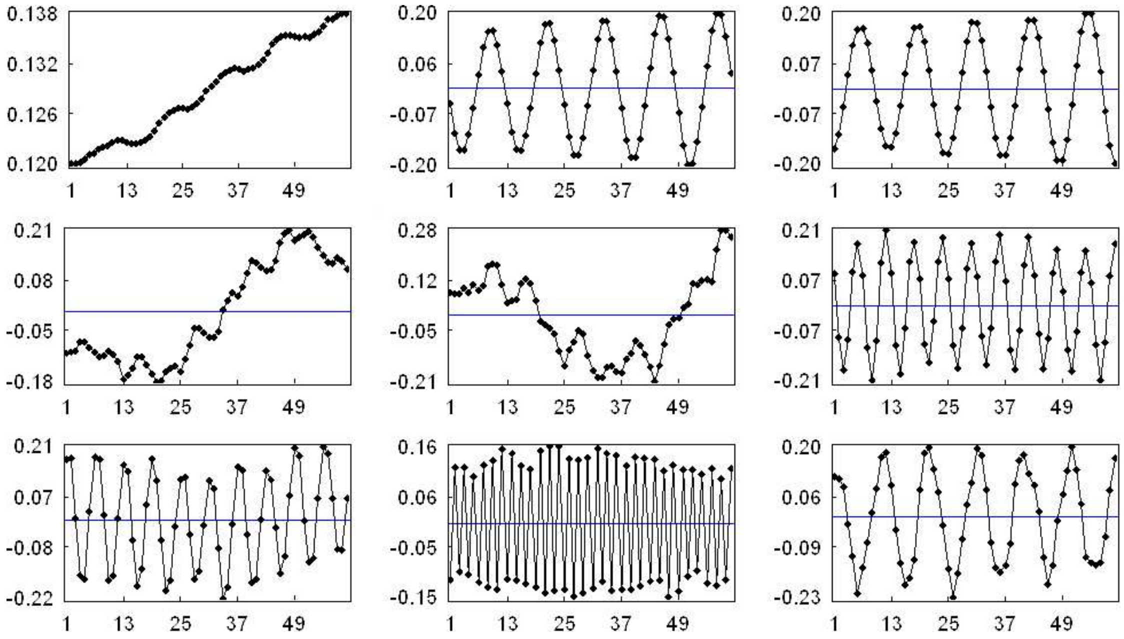
N=180, L=60

Example: trend and seasonality extraction

Traffic fatalities. Ontario, monthly, 1960-1974 (Abraham, Redolter. *Stat. Methods for Forecasting*, 1983)

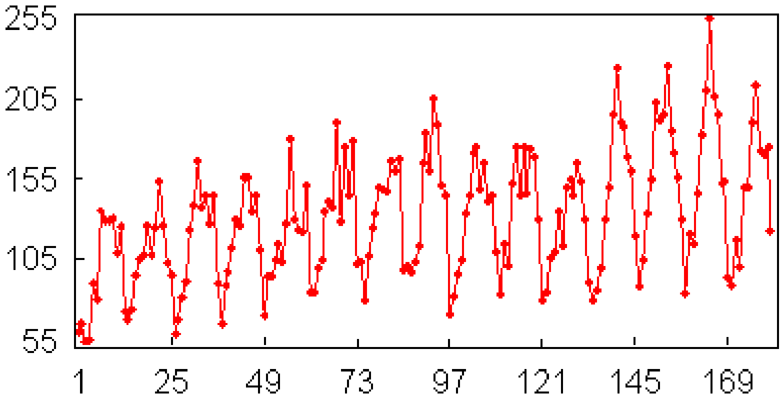


N=180, L=60

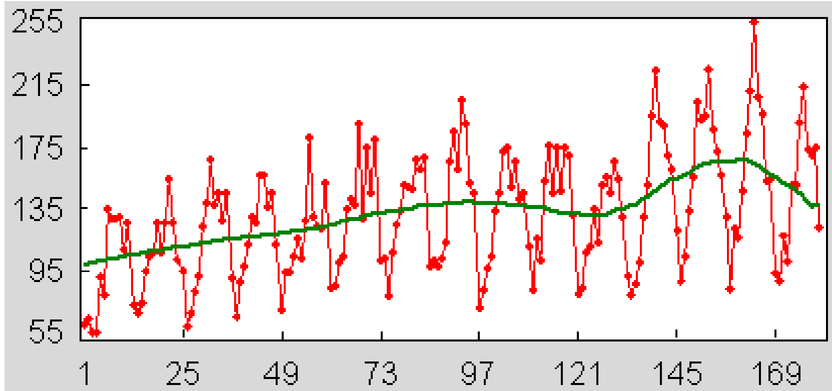


Example: trend and seasonality extraction

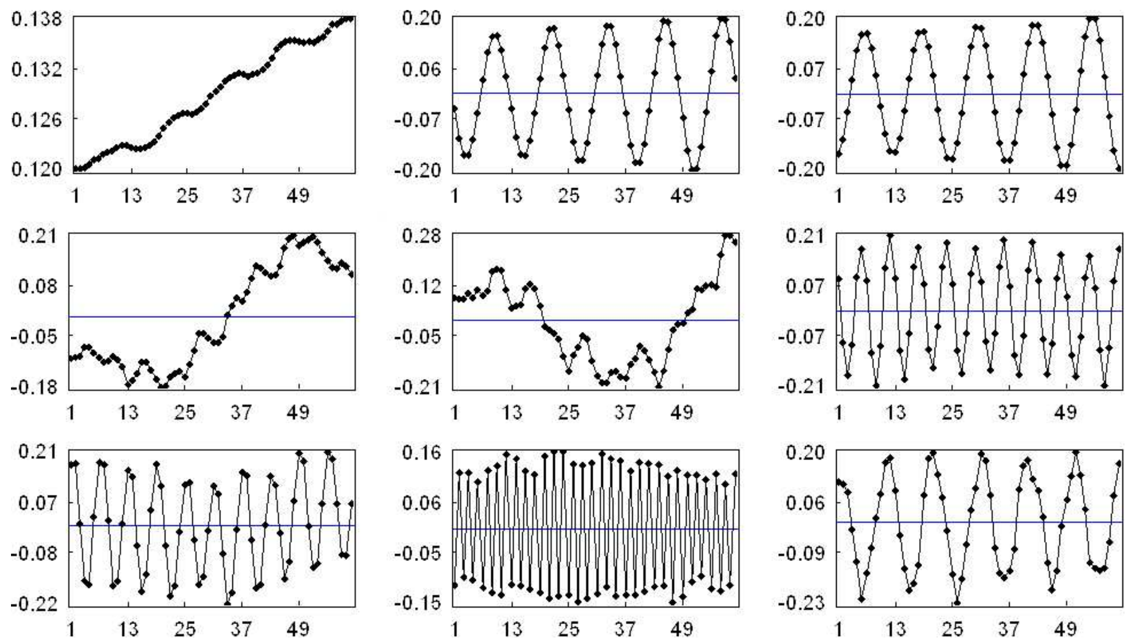
Traffic fatalities. Ontario, monthly, 1960-1974 (Abraham, Redolter. *Stat. Methods for Forecasting*, 1983)



N=180, L=60

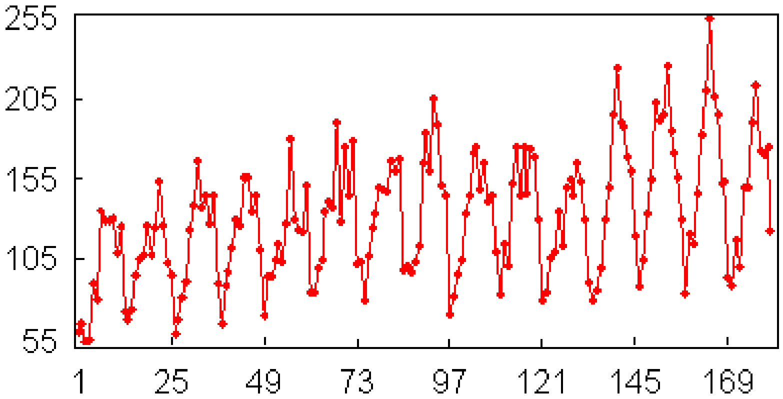


SVD components: 1, 4, 5

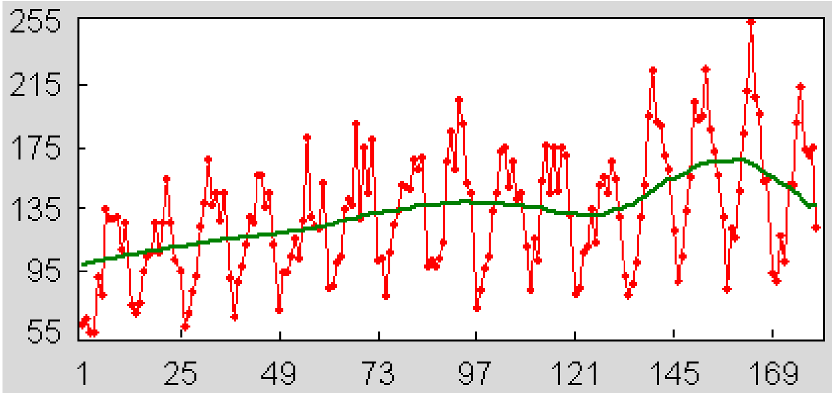


Example: trend and seasonality extraction

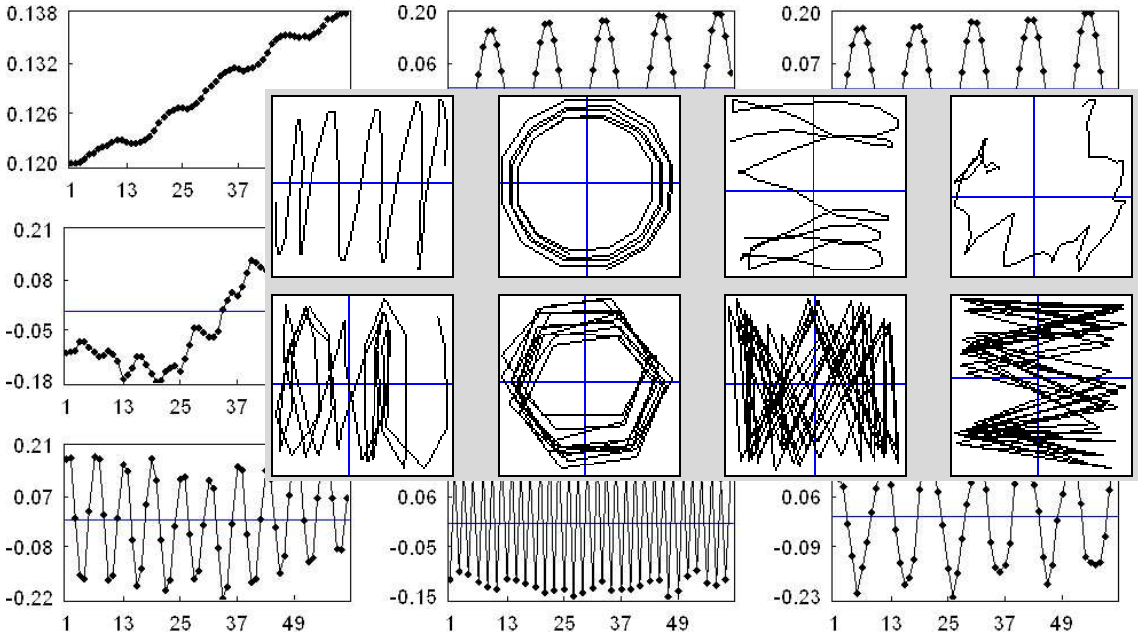
Traffic fatalities. Ontario, monthly, 1960-1974 (Abraham, Redolter. *Stat. Methods for Forecasting*, 1983)



N=180, L=60

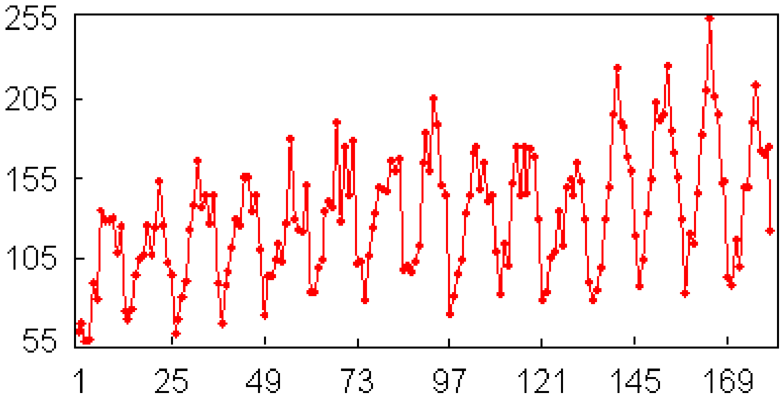


SVD components: 1, 4, 5

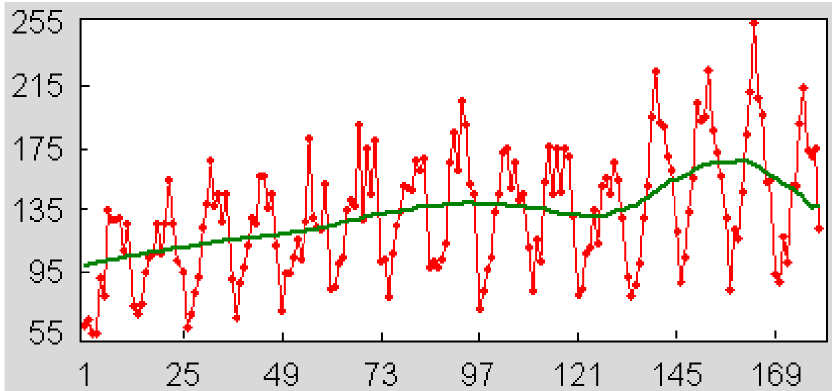


Example: trend and seasonality extraction

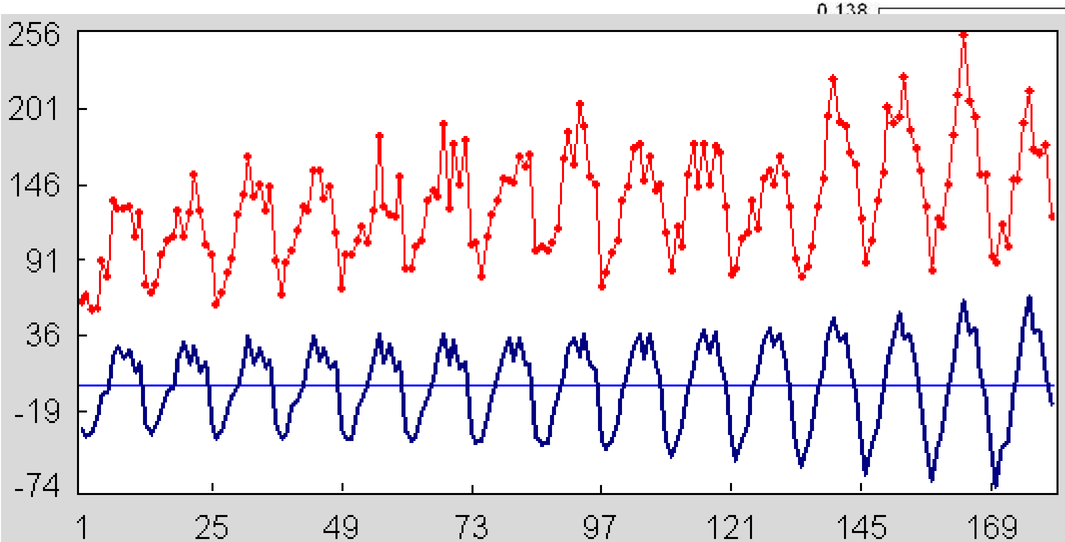
Traffic fatalities. Ontario, monthly, 1960-1974 (Abraham, Redolter. *Stat. Methods for Forecasting*, 1983)

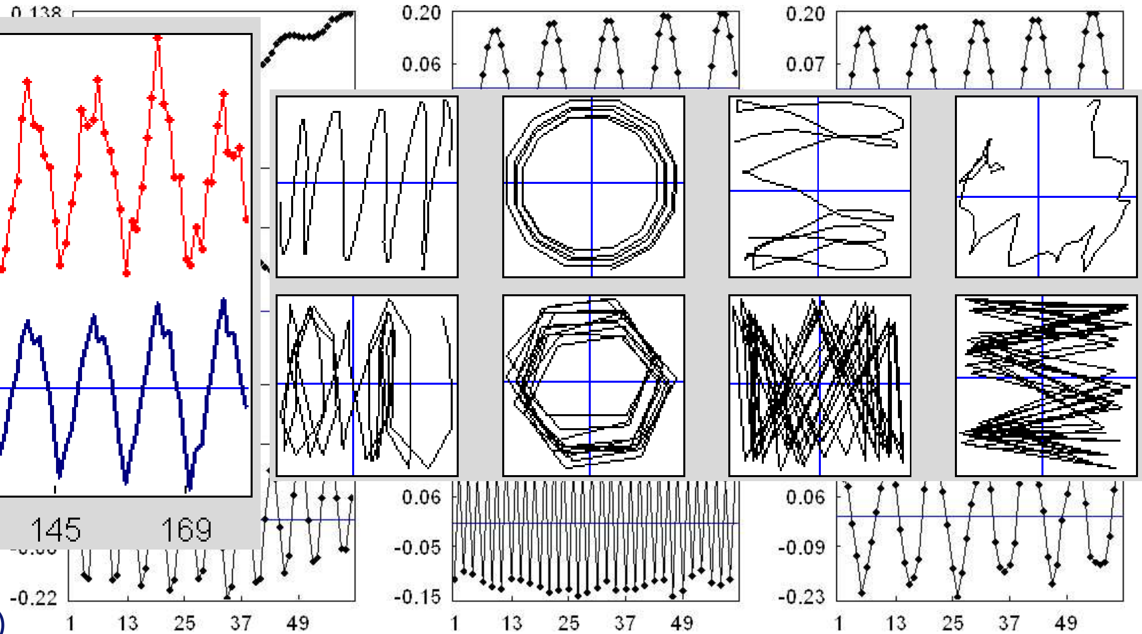


N=180, L=60



SVD components: 1, 4, 5





SVD components with estimated periods:
2-3 (T=12), 6-7(T=6), 8(T=2), 11-12(T=4), 13-14(T=2.4)

Finite rank time series

We said model-free, but the area of action is constrained to: span$(exp * cos * Pn)$.

Important concepts

- $\mathfrak{L}^{(L)} = \mathfrak{L}^{(L)} =$ span$(X_1, \ldots, X_K)$ – the trajectory space for F_N, $X_i = (f_{i-1}, \ldots, f_{i+L-2})^{\mathsf{T}}$.
- Time series F_N is a time series of (finite) rank d (rank$(F_N) = d$), if $\forall L$ dim $\mathfrak{L}^{(L)} = d$.

Rank $\leftrightarrow$ amount of SVD components $\leftrightarrow$ order of LRF

- rank$_L(F_N) =$ rank $\mathbf{X} \Rightarrow$ amount of SVD components with $\lambda_j \neq 0$ is equal to the rank.
- $F = (\ldots, f_{-1}, f_0, f_1, \ldots)$ – infinite time series, then

$$f_{i+d} = \sum_{k=1}^{d} a_k f_{i+d-k}, a_d \neq 0 \quad \Leftrightarrow \quad \text{rank}(F) = d.$$

Examples of finite rank time series

- Exponentially modulated (e-m) harmonic F_N: $f_n = Ae^{\alpha n}\cos(2\pi\omega n + \phi)$.
  - e-m harmonic ($0 < \omega < 1/2$): rank $= 2$
  - e-m saw ($\omega = 1/2$): rank $= 1$
  - exponential time series ($\omega = 0$): rank $= 1$
  - harmonic ($\alpha = 1$): rank $= 2$
- Polynomial F_N: $f_n = \sum_{k=0}^{m} a_k n^k$, $a_m \neq 0$: rank $= m + 1$

Separability

$F_N = F_N^{(1)} + F_N^{(2)}$, window length L, traj.matrices $\mathbf{X} = \mathbf{X}^{(1)} + \mathbf{X}^{(1)}$, traj.spaces $\mathfrak{L}^{(L,1)}$, $\mathfrak{L}^{(L,2)}$.

$F_N^{(1)}$ and $F_N^{(2)}$ are the L-separable if $\mathfrak{L}^{(L,1)} \perp \mathfrak{L}^{(L,2)}$ and $\mathfrak{L}^{(K,1)} \perp \mathfrak{L}^{(K,2)}$.
If $F_N^{(1)}$ and $F_N^{(2)}$ are separable then the SVD components of $\mathbf{X}$ can be grouped so that the first group corresponds to $\mathbf{X}^{(1)}$ and the second to $\mathbf{X}^{(2)}$.

i.e. separability (separation of trajectory spaces) $\Leftrightarrow$ separation of additive components

Reality:

- Approximate separability (approximate orthogonality of trajectory spaces)
- Asymptotic separability (with $L, N \to \infty$)

Examples

Separability (strict, asymptotic) on some conditions

	const	cos	exp	exp*cos	Pn
const	- -	+ +	- +	- +	- -
cos	+ +	+ +	- +	- +	- +
exp	- +	- +	- +	+ +	- +
exp*cos	- +	- +	+ +	+ +	- +
Pn	- -	- +	- +	- +	- -

- signal is asymptotically separated from noise
- periodicity is asymptotically separated from trend

Separability conditions (and the rate of convergence) $\Rightarrow$ rules for L setting (this problem had no solution before)

Trend – slow varying deterministic additive component.

Examples of parametric trends: exp, Pn, harmonic with large T ($T > N/2$).

How to identify trend SVD components

- Eigenvalues

  λ_j – contribution of $F^{(j)}$ to the form of F_N ($F^{(j)}$ is reconstructed by $\sqrt{\lambda_j} U_j V_j^{\mathsf{T}}$).

  Trend is large $\Rightarrow$ its SVD components are the first.

- Eigenvectors

$$U_j = (u_1^{(j)}, \ldots, u_L^{(j)})^{\mathsf{T}}$$

Form of eigenvectors for some slow-varying time series

f_n	$u_k^{(\cdot)}$
$e^{\alpha n}$	$e^{\alpha k}$
$\sum_m a_m n^m$	$\sum_m b_m k^m$
$e^{\alpha n} \cos(2\pi\omega n + \phi)$	$e^{\alpha k} \cos(2\pi\omega k + \psi)$

Trend SVD components have slow-varying eigenvectors.

Continuation and forecast

Continuation

$F_N = (f_0, \ldots, f_{N-1})$, rank$(F_N) = d < L$, then ,typically, F_N is governed by LRF of order d.

Main variant of the continuation: recurrent continuation using LRF.

- There are the unique minimal LRF (order d) and many LRFs of order $> d$
- The “Caterpillar”-SSA:
  - $\mathfrak{L}^{(L)}$: an orthogonal basis (e.g. eigenvectors) $\longrightarrow$ the LRF of order $L-1$ (automatically)
  - deflation of LRF (considering characteristic polynomial of LRF) (can be automated)

Continuation of an additive component

$F_N = F_N^{(1)} + F_N^{(2)}$, rank$(F_N^{(1)}) = d_1 < L$, rank$(F_N^{(1)}) = d_1 < L$.
$F_N^{(1)}$ and $F_N^{(2)}$ are separable $\Rightarrow$ we can continue them separately.

Forecast (approximate)

$F_N = F_N^{(1)} + F_N^{(2)}$, approximate separability (for example, signal+noise or slightly distorted time series) $\Rightarrow$ approximation of $F_N^{(1)}$ by d-dimension trajectory space $\Rightarrow$ approximation by LRF.

Change-point detection in brief

Problem statement

F_N is homogeneous if it is governed by some LRF with order $<< N$.

Assume that at one time it stops following the original LRF and after a certain time period it again becomes governed by an LRF.

We want to a posteriori detect such heterogeneities (change-points) and compare somehow homogeneous partes before and after change.

Solution

- F_N is governed by the LRF $\Rightarrow$ for suff. large N, L: $\mathfrak{L}^{(L)} = \mathrm{span}(X_1, \ldots, X_K)$ is independent of N
- Minimal LRF $\leftrightarrow \mathfrak{L}^{(L)}$

$\Rightarrow$ LRF heterogeneities can be described in terms of corresponding lagged vectors: the perturbations force the lagged vectors to leave the space $\mathfrak{L}^{(L)}$

We can measure this distance and detect a change-point

Automation of time series processing

Problem statement

Automation of manual processing of large set of similar time series (family)

Manual processing is the ideal | $\Rightarrow$ quality in comparison with manual processing
| $\Rightarrow$ we can use the stated theory

Why family?

Family processing $\Rightarrow$ several randomly taken time series can be used for:

- testing the auto-procedure, whether it works in general (necessary condition)
- finding proper parameters of the auto-procedure for the whole family (performance optimization)

Bootstrap test of the procedure

We must know noise (residual) model (or its approximation)

- Extract trend $\widetilde{F}_N^{(T)}$ manually
- Consider $F_N - \widetilde{F}_N^{(T)}$ and estimate parameters of noise
- Simulate noise using these parameters and generate surrogate data: $G_N = \widetilde{F}_N^{(T)} + noise$
- Extract trend from surrogate data $\widetilde{G}_N^{(T)}$
- MSE($\widetilde{F}_N^{(T)}, \widetilde{G}_N^{(T)}$) – a measure of procedure quality

Auto-method of trend extraction

Eigenvectors of trend SVD components have slow-varying form

Search all eigenvectors, let us assume we process an $U = (u_1, \ldots, u_L)^{\mathrm{T}}$.

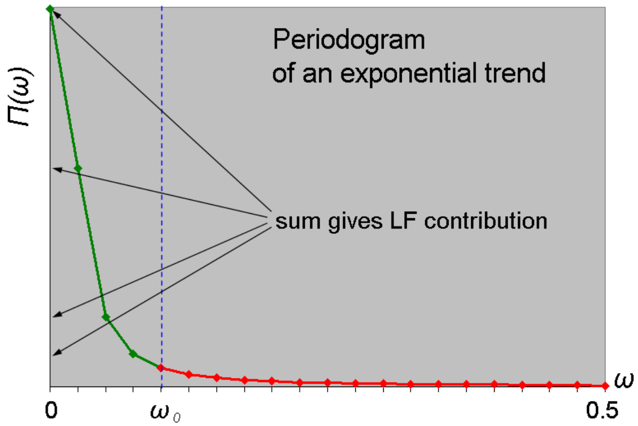
$$u_n = c_0 + \sum_{1 \leqslant k \leqslant \frac{L-1}{2}} \left(c_k \cos(2\pi nk/L) + s_k \sin(2\pi nk/L)\right) + (-1)^n c_{L/2},$$

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

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

Low Frequencies method

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



+ examples

Define $\mathcal{C}(U) = \dfrac{\sum_{\mathbf{0} \leqslant \omega \leqslant \omega_\mathbf{0}} \mathbf{\Pi}(\omega)}{\sum_{\mathbf{0} \leqslant \omega \leqslant \mathbf{0.5}} \mathbf{\Pi}(\omega)}$, $\omega \in k/L, k \in \mathbb{Z}$ – contribution of LF frequencies

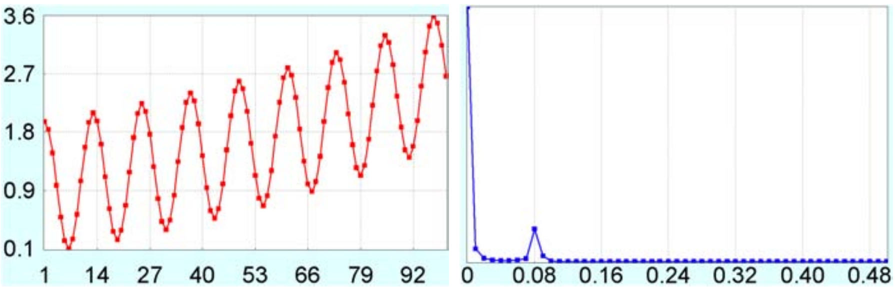
$\mathcal{C}(U) \geqslant \mathcal{C}_0 \Rightarrow$ eigenvector U corresponds to a trend, where $\mathcal{C}_0 \in (0, 1)$ – the threshold

Choice of ω_0

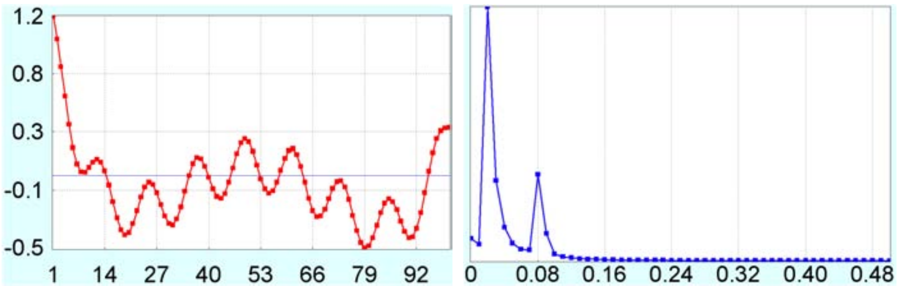
Examining the periodogram of an original time series

- Periodicity with period T exists $\Rightarrow \omega_0 < 1/T$

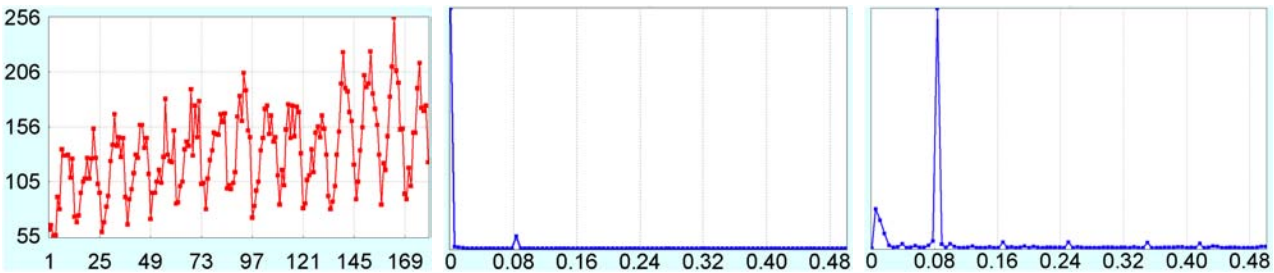
Examples



Exp+cos and its periodogram,
$f_n = e^{0.01*n} + \cos(2*\pi*n/12)$
$0.3 < \omega_0 < 0.8$



Pn+cos and its periodogram,
$f_n = (x-10)(x-40)(x-60)(x-95)$
$\cos(2*\pi*n/12)$
$\omega_0 \approxeq 0.7 < 0.8$



Traffat (left),
its periodogram (center) and
periodogram of normalized
time series (right)
$0.3 < \omega_0 < 0.8$

Choice of $\mathcal{C}_0$, measure of quality of trend extraction

If we have a measure $\mathcal{R}$ of quality of trend extraction $\Rightarrow$ $\mathcal{C}_{\text{opt}} = \text{argmin}_{\mathcal{C}_0 \in [0,1]} \mathcal{R}$

Measure

The natural measure of quality is MSE$(F^{(T)}, \widetilde{F}_0^{(T)})$, where $F^{(T)}$ is the real trend and $\widetilde{F}_0^{(T)}$ is the extracted trend (with $\mathcal{C}_0$), but it requires unknown $F^{(T)}$.

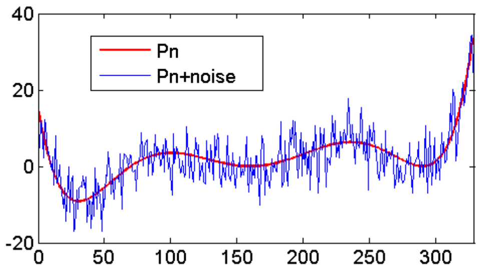
$$\text{We propose } \mathcal{R}(\mathcal{C}_0) = \frac{\mathcal{C}(F - \widetilde{F}_0^{(T)})}{\mathcal{C}(F)}, \quad \widetilde{F}_0^{(T)} \text{ is extracted with } \mathcal{C}_0$$

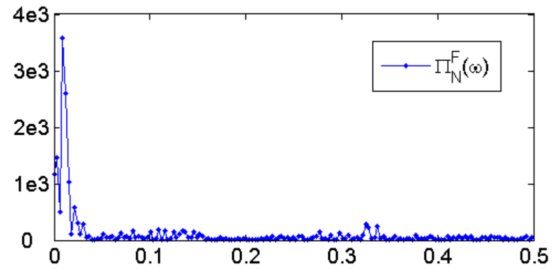
$\mathcal{R}(\mathcal{C}_0)$ is consistent with MSE$(F^{(T)}, \widetilde{F}_0^{(T)})$ in such a way:

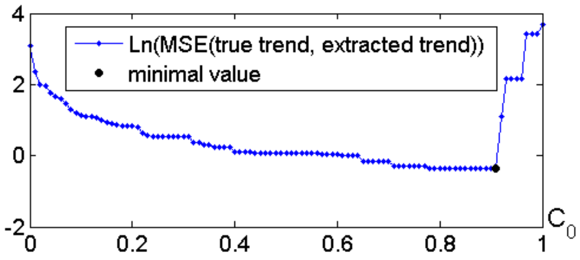
- it behaves like MSE
- by means of $\mathcal{R}(\mathcal{C}_0)$ we can define $\mathcal{C}_{\text{opt}} = \text{argmin}_{\mathcal{C}_0} \text{MSE}(F^{(T)}, \widetilde{F}_0^{(T)})$

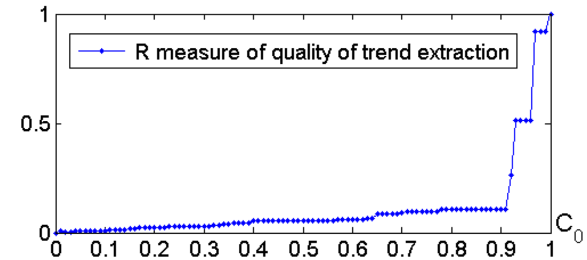
Examples of $\mathcal{C}_{\text{opt}}$ estimation

Model example, Pn+noise









$f_n = (n-10)(n-70)(n-160)^2 \cdot (n-290)^2/1e11 + N(0,25)$,

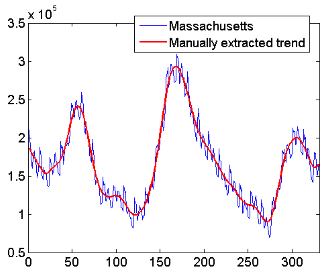
$N = 329$, $L = N/2 = 160$,

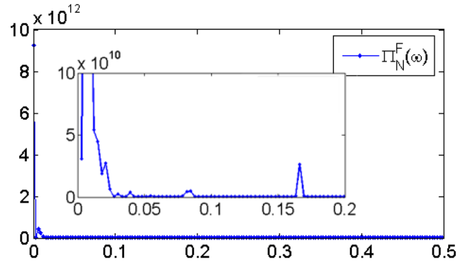
$\omega_0 = 0.07$

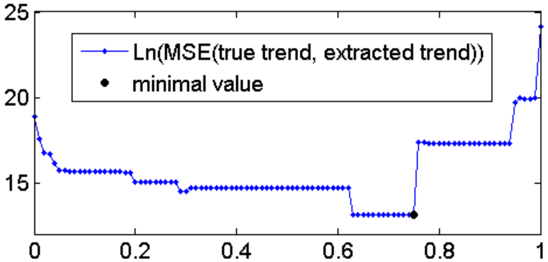
$\mathcal{C}_0 = 1 \ldots 0.9$: graphics reflect stepwise identification of trend SVD components $\Rightarrow$ considerable changes of MSE$(F^{(T)}, \widetilde{F}_0^{(T)})$

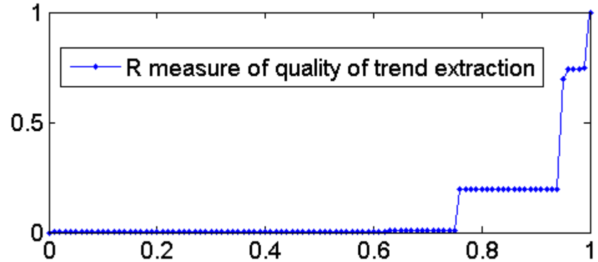
$\mathcal{C}_{\text{opt}} < 0.9 (\approx 0.9)$

Real-life example, Massachusetts unemployment









Massachusetts unemployment (thousands, monthly), from economagic.com

$N = 331$, $L = N/2 = 156$,

$\omega_0 = 0.05 < 1/12 = 0.08(3)$

$\mathcal{C}_0 = 1 \ldots 0.75$: graphics reflect stepwise identification of trend SVD components $\Rightarrow$ considerable changes of MSE$(F^{(T)}, \widetilde{F}_0^{(T)})$

$\mathcal{C}_{\text{opt}} < 0.75 (\approx 0.75)$

Final slide: to sum up

I. We have a family of similar time series $\mathfrak{F} = \{F_N\}$.

II. Take randomly (or somehow otherwise) a test subset $\mathfrak{T}$ of several characteristic time series.

On these time series perform:

1. Extract trends manually
2. Examine periodograms of the time series and choose ω_0
3. Check if the proposed auto-procedure works in general on such time series:
   - Bootstrap comparison: how trends automatically extracted from surrogate data are close to manually extracted trends
     If they are sufficiently close then auto-procedure is accepted
   - It requires knowledge of noise model but we can take a simple one as a first approximation
4. Estimate $\mathcal{C}_{\text{opt}}^{(F)}$ for each time series and take a minimum from them

$$\mathcal{C}_{\text{opt}}^{(\mathfrak{F})} = \min_{F \in \mathfrak{T}} \mathcal{C}_{\text{opt}}^{(F)}$$

   as the optimal $\mathcal{C}_0$ for the family $\mathfrak{F}$
   (this optimization step can be skipped, then during processing of $\mathfrak{F}$ we have to estimate $\mathcal{C}_{\text{opt}}$ for each time series)

III. Process all time series from the family $\mathfrak{F}$