Monte Carlo SSA for extracting weak signals

Egor Poteshkin, Nina Golyandina

Saint Petersburg State University
Department of Statistical Modeling

CDAM'2025
September 24, 2025, Minsk, Belarus

Motivation

Let $\mathsf{X} = (x_1, \ldots, x_N)$, $x_i \in \mathbb{R}$, be a time series.

Observed: $\mathsf{X} = \mathsf{T} + \mathsf{H} + \mathsf{R}$, where T is a trend, H is a periodic component, and R is a noise.

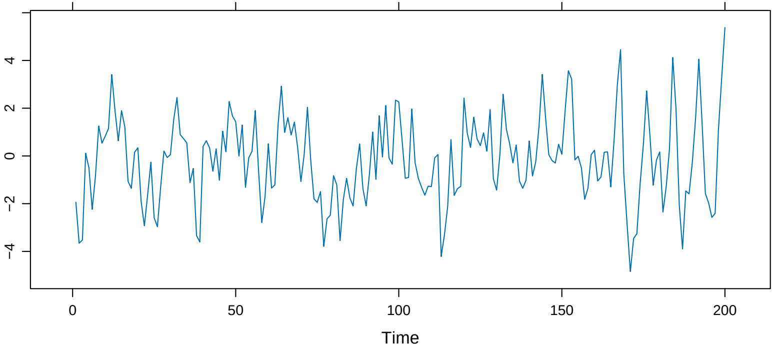
Problem: How to extract the signal $\mathsf{S} = \mathsf{T} + \mathsf{H}$, if it is present?

There is a method for automatic trend and periodicity extraction [Golyandina, Dudnik and Shlemov, 2023].

Disadvantage: method only works if components dominate over noise.

Aim of this work: to develop a method for automatic signal extraction that does not necessarily dominate.

Is There a Signal?



Question: is it pure noise, or is there a signal, and if so, how to extract it?

Problem Statement

Let X $= (x_1, \ldots, x_N)$, $x_i \in \mathbb{R}$, be a time series.

Observed: X $=$ T $+$ H $+$ R, where T is a trend, H is a periodic component, and R is a noise.

Problems:

1. How to test for the presence of a signal S $=$ T $+$ H?
2. How to extract the signal S, if it is present?

Methods:

1. Monte Carlo SSA (MC-SSA) [Allen & Smith, 1996; Golyandina, 2023] — tests H_0 : S $= 0$.
2. Singular spectrum analysis (SSA) [Broomhead & King, 1985; Golyandina, Nekrutkin and Zhigljavsky, 2001].

Aim: to develop an algorithm for automatic signal extraction based on MC-SSA

Notations & Known Results: Embedding and Hankelization

For $\mathsf{X} = (x_1, \ldots, x_N)$, fix L $(1 < L < N)$.

Embedding operator $\mathcal{T}_{\mathsf{SSA}}$:

$$\mathcal{T}_{\mathsf{SSA}}(\mathsf{X}) = \mathbf{X} = \begin{pmatrix} x_1 & x_2 & \cdots & x_K \\ x_2 & x_3 & \cdots & x_{K+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & \cdots & x_N \end{pmatrix},$$

where $K = N - L + 1$.

Hankelization operator $\mathcal{H}$ — averaging the matrix over its anti-diagonals.

Notations & Known Results: The SSA Algorithm

Input: time series $\mathsf{X} = (x_1, \ldots, x_N)$.
Parameters: window length L, index set $I \subset \{1, \ldots, d\}$.
Output: signal estimate $\widetilde{\mathsf{S}}$.

Input: X — time series

1. Embedding

Trajectory matrix
$$\mathbf{X} = \mathcal{T}_{\mathsf{SSA}}(\mathsf{X})$$

2. Decomposition

Sum of rank-one matrices
$$\mathbf{X} = \sum_{j=1}^{d} \mathbf{X}_j$$

3. Reconstruction

Elementary reconstructed components
$$\widetilde{\mathsf{X}}_j = \mathcal{T}_{\mathsf{SSA}}^{-1} \circ \mathcal{H}(\mathbf{X}_j)$$

4. Grouping

Grouping components, corresponding to the signal

Result: SSA decomposition
$$\mathsf{X} = \widetilde{\mathsf{S}} + \widetilde{\mathsf{R}},\ \widetilde{\mathsf{S}} = \sum_{i \in I} \widetilde{\mathsf{X}}_i$$

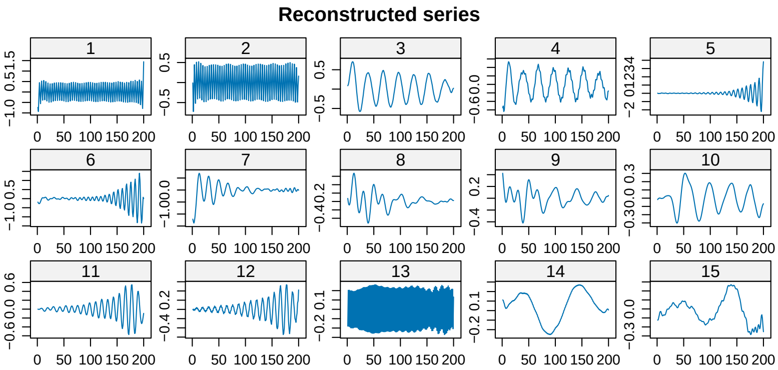


Figure: Elementary reconstructed components $(L = 100)$

Inside information: the components corresponding to the signal are 1, 2, 5, 6 and 13.

Input: $\mathsf{X} = \mathsf{S} + \mathsf{R}$, where S is the signal and R is a realization of a zero-mean stationary process $\boldsymbol{\xi}$ with spectral density $f_{\boldsymbol{\theta}}$.

Parameter: window length L.

Test statistics:

$$\widehat{p}_k = \left\| \mathbf{X}^{\mathrm{T}} W_k \right\|^2, \quad k = 1, \ldots, L,$$

where $W_1, \ldots, W_L$ are normalized sine waves with equidistant frequencies $\omega_k = k/(2L)$: $V_k = \{\cos(2\pi\omega_k j)\}_{j=1}^{L}$, $W_k = V_k / \|V_k\|$.

Distribution of $\widehat{p}_k$ under H_0 is estimated via Monte Carlo by modeling $\boldsymbol{\xi}$ with density $f_{\boldsymbol{\theta}}$ (what is $\boldsymbol{\theta}$ equal to?).

Result: $(1-\alpha)$-confidence intervals for each $\widehat{p}_k$ under H_0.

Problem: intervals are liberal due to uncontrolled FWER.

Multiple MC-SSA [Golyandina, 2023]: modification with multiple comparisons correction.

Noise Parameters Estimation

Noise parameters $\boldsymbol{\theta}$ are generally unknown and must be estimated.

Parameter estimates can be obtained by maximizing the Whittle's likelihood [Whittle, 1953]:

$$\ell_W(\boldsymbol{\theta}) = -\frac{1}{m}\sum_{j=1}^{m}\left(\ln f_{\boldsymbol{\theta}}(\omega_j) + \frac{I_N(\omega_j)}{f_{\boldsymbol{\theta}}(\omega_j)}\right),$$

where $m = \lfloor (N-1)/2 \rfloor$, $f_{\boldsymbol{\theta}}$ is the spectral density of $\boldsymbol{\xi}$, I_N is the periodogram of the original series, and $\omega_j = j/N$.

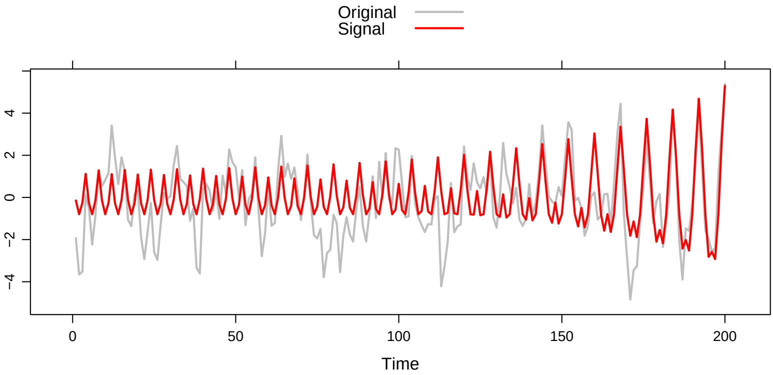
Problem: after detrending a time series, the periodogram values at very low frequencies are unreliable.

Solution: estimate parameters on part of the spectrum.
Let $J = \{j_1, \ldots, j_p\}$ be frequency indices we want to exclude when estimating parameters. Then $\ell_W(\boldsymbol{\theta})$ is computed only over indices $j \notin J$.

Example: Time series with signal

$\mathsf{X} = \mathsf{S} + \boldsymbol{\xi}$, where $\boldsymbol{\xi}$ is AR(1) with $\phi = 0.7$ and $\sigma^2 = 1$ (red noise), $N = 200$,

$$s_n = 0.075\, e^{0.02n} \cos(2\pi n/8) + 2\cos(2\pi n/4) + 0.2 \cdot (-1)^n.$$



Example: Applying Monte Carlo SSA

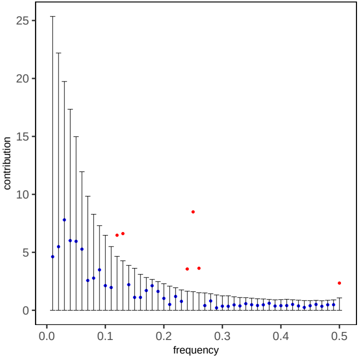


Figure: True noise model

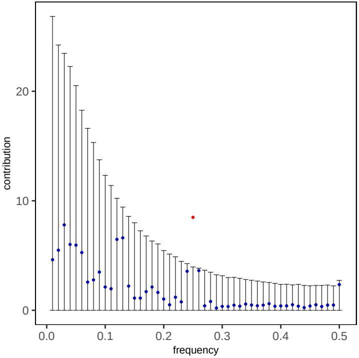


Figure: Estimated noise model

Problem: not all frequencies are detected with estimated noise model.

Solution: apply the test iteratively, extracting one harmonic at a time until $H_0 : \mathsf{S} = 0$ is no longer rejected.

Problem: how to automate signal extraction if the frequency range is known?

Solution: Automatic Grouping in SSA [Golyandina and Zhigljavsky, 2013].

For a series X of length N and $0 \leqslant \omega_1 \leqslant \omega_2 \leqslant 0.5$, define

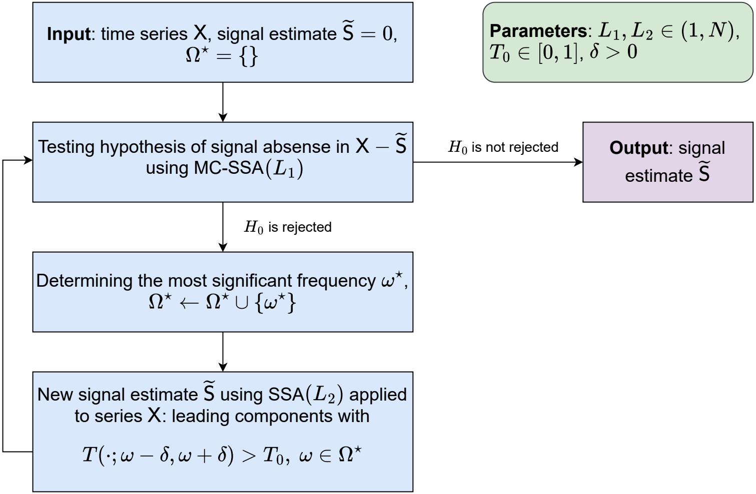
$$T(\mathsf{X}; \omega_1, \omega_2) = \frac{1}{\|\mathsf{X}\|^2} \sum_{k: \omega_1 \leqslant k/N \leqslant \omega_2} I_N(k/N),$$

where I_N is the periodogram of X.

Let T_0, $0 \leqslant T_0 \leqslant 1$, be a threshold, and $\widetilde{\mathsf{X}}_i$ be i-th elementary reconstructed component. Then:

$$T\left(\widetilde{\mathsf{X}}_i; \omega_1, \omega_2\right) > T_0 \implies \widetilde{\mathsf{X}}_i \text{ corresponds to the signal.}$$

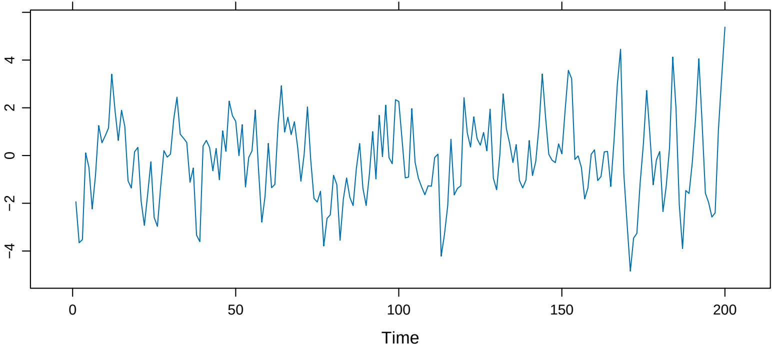
Idea: for every significant frequency $\omega^\star$ take $\omega_{1,2} = \omega^\star \mp \delta$.


Input: time series X, signal estimate $\widetilde{\mathsf{S}} = 0$, $\Omega^\star = \{\}$
Parameters: $L_1, L_2 \in (1, N)$, $T_0 \in [0, 1]$, $\delta > 0$
Testing hypothesis of signal absense in $\mathsf{X} - \widetilde{\mathsf{S}}$ using MC-SSA(L_1)
H_0 is not rejected
Output: signal estimate $\widetilde{\mathsf{S}}$
H_0 is rejected
Determining the most significant frequency $\omega^\star$, $\Omega^\star \leftarrow \Omega^\star \cup \{\omega^\star\}$
New signal estimate $\widetilde{\mathsf{S}}$ using SSA(L_2) applied to series X: leading components with
$T(\cdot; \omega - \delta, \omega + \delta) > T_0,\ \omega \in \Omega^\star$

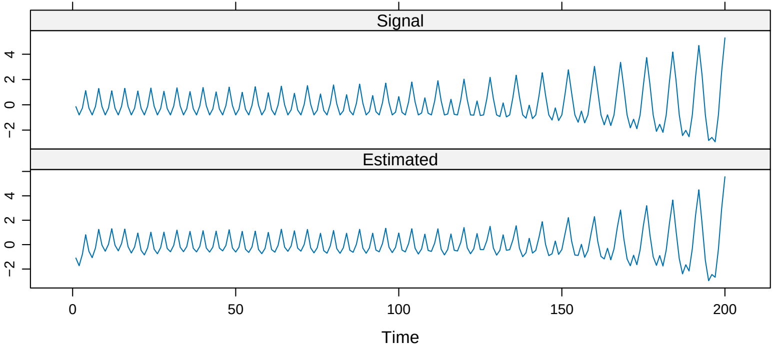
Example: Time series

$\mathsf{X} = \mathsf{S} + \boldsymbol{\xi}$, where $\boldsymbol{\xi}$ is AR(1) with $\phi = 0.7$ and $\sigma^2 = 1$ (red noise), $N = 200$,

$$s_n = 0.075\, e^{0.02n} \cos(2\pi n/8) + 2\cos(2\pi n/4) + 0.2 \cdot (-1)^n.$$



Example: Applying autoMCSSA



Parameters: $L_1 = 50$, $L_2 = 100$, $\delta = 1/80$, $T_0 = 0.5$.

Result: autoMCSSA correctly identified significant components (1, 2, 5, 6 and 13).

Conclusions

To sum up:

1. Main results: we developed and implemented autoMCSSA, which automatically extracts a significant signal, plus a modification of the Whittle approach using part of the spectrum.

2. Advantage over previous method: autoMCSSA can extract signals whose SSA components are not necessarily dominant.

In the future, we plan to formulate a strategy for selecting autoMCSSA parameters.