Monte Carlo SSA for extracting weak signals

Egor Poteshkin, Nina Golyandina

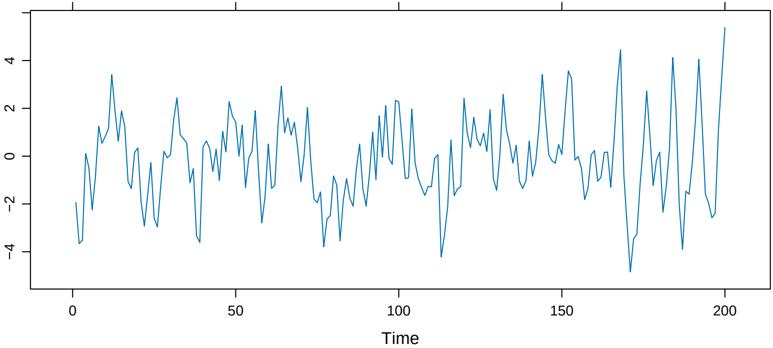
Saint Petersburg State University
Department of Statistical Modeling

CDAM'2025
September 24, 2025, Minsk, Belarus

Motivation

- There is a method for automatic trend and periodicity extraction [Dudnik, 2025]. However, it only works if they dominate over noise.
- The main aim of this work is 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 and Smith, 1996; Golyandina, 2023] — tests H_0 : S $= 0$.
2. Singular spectrum analysis (SSA) [Golyandina, Nekrutkin and Zhigljavsky, 2001].

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

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

Example: Applying SSA

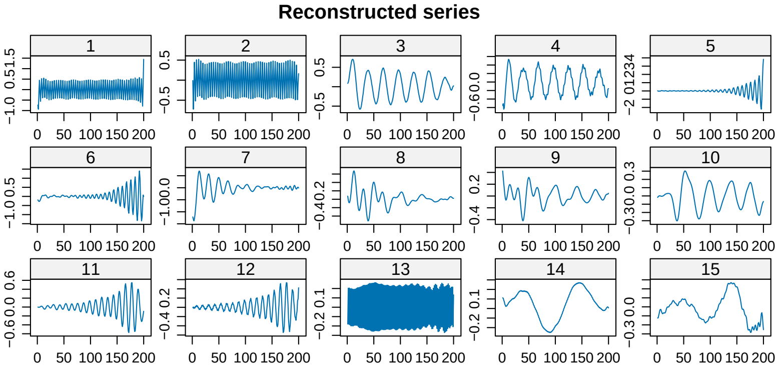


Figure: Elementary reconstructed components ($L = 100$)

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

Notations & Known Results: Monte Carlo SSA

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 generally unknown and 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]:

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

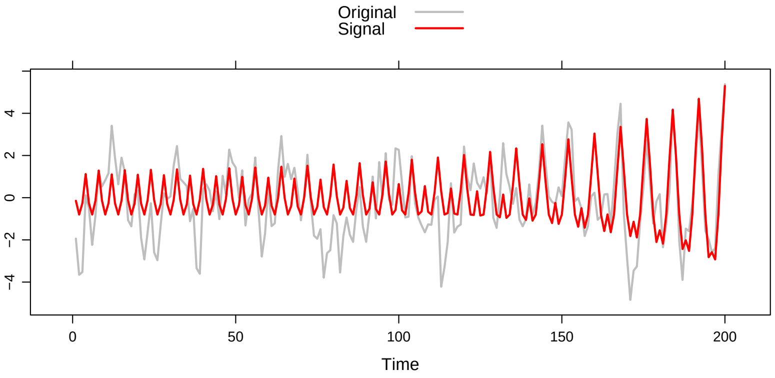
where $m = \lfloor (N-1)/2 \rfloor$, $f_{\boldsymbol{\theta}}$ is the spectral density of $\boldsymbol{\xi}$, I_N is the periodogram of the original series, and $\omega_j = j/N$.

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 $j \notin J$.

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

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



Example: Applying Monte Carlo SSA

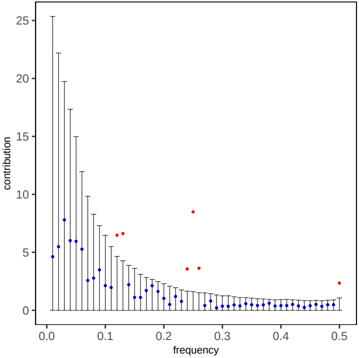


Figure: True noise model

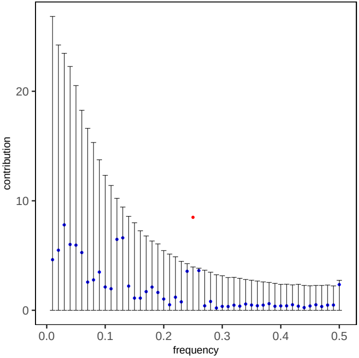


Figure: Estimated noise model

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.

Notations & Known Results: Automatic Grouping in SSA

If the frequencies corresponding to the signal are known, it can be extracted using Automatic Grouping in SSA [Golyandina, Zhigljavsky, 2013].

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

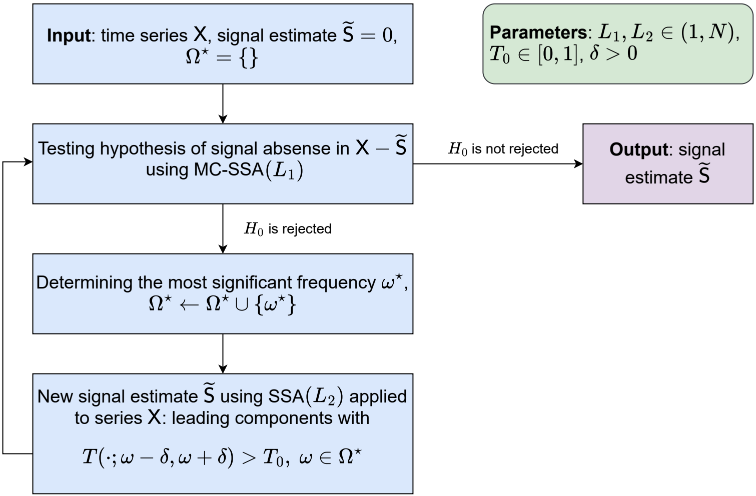
$$T(\mathsf{X}; \omega_1, \omega_2) = \frac{1}{\|\mathsf{X}\|^2} \sum_{k: \omega_1 \leqslant k/N \leqslant \omega_2} I_N(k/N),$$

where I_N is the periodogram of X. $T(\mathsf{X}; \omega_1, \omega_2)$ may be interpreted as the fraction of total contribution of frequencies within $[\omega_1, \omega_2]$.

Let T_0, $0 \leqslant T_0 \leqslant 1$, be a threshold. If $T(\cdot; \omega_1, \omega_2) > T_0$ for some elementary reconstructed component $\widetilde{\mathsf{X}}_i$, then this component can be considered as a part of the signal.

Let $\omega^\star$ be a significant frequency. Then we take $[\omega_1, \omega_2] = [\omega^\star - \delta, \omega^\star + \delta]$.

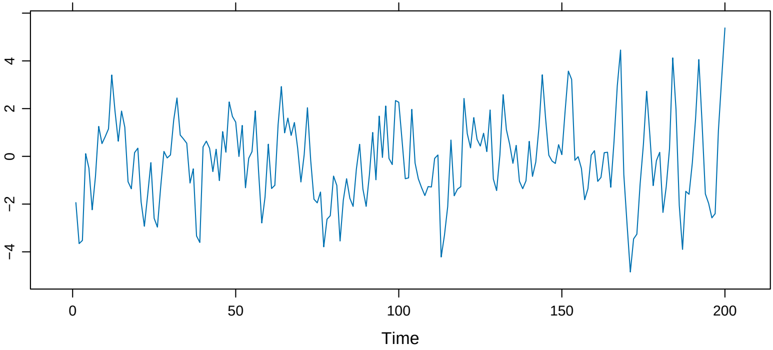
The autoMCSSA Algorithm



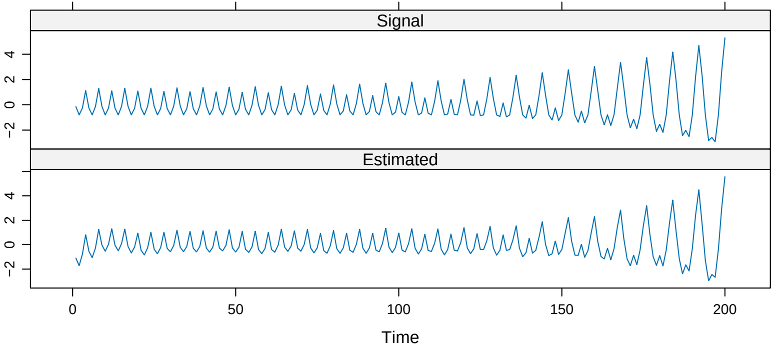
Example: Time series

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

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



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, we:

1. 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 autoSSA: autoMCSSA can extract signals whose SSA components are not necessarily dominant.

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