Метод Монте-Карло SSA для одномерных и многомерных временных рядов

Потешкин Егор Павлович, гр.20.Б04-мм

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Кафедра статистического моделирования

Научный руководитель — д. ф.-м. н. Н. Э. Голяндина
Рецензент — программист А. Ю. Шлемов, Майкрософт

Санкт-Петербург, 2024

Введение и постановка задачи

$\mathsf{X} = (x_1, \ldots, x_N)$ — временной ряд длины N.

Модель: $\mathsf{X} = \mathsf{T} + \mathsf{H} + \mathsf{R}$, где T — тренд, H — периодическая компонента и R — шум, случайная составляющая.

Проблемы:

1. Как выделить неслучайные компоненты T и H?
2. Как проверить наличие сигнала $\mathsf{S} = \mathsf{T} + \mathsf{H}$?

Методы:

1. Singular spectrum analysis (SSA) [Broomhead and King, 1986], [Golyandina, Nekrutkin and Zhigljavsky, 2001].
2. Monte Carlo SSA (MC-SSA) [Allen and Smith, 1996].

MSSA и MC-MSSA — обобщение SSA и MC-SSA на многомерный случай.

Задачи:

1. Реализовать Toeplitz MSSA и сравнить с обычным MSSA.
2. Сравнить модификации MC-MSSA.
3. Рассмотреть MC-SSA в условиях реальных задач.

Глава 1. Метод MSSA

Метод SSA и его многомерное обобщение

Входные данные: временной ряд $\mathsf{X} = (x_1, \ldots, x_N)$.

Параметр: длина окна L, $1 < L < N$.

Результат: m восстановленных составляющих временного ряда.

1. **Вложение**: $\mathbf{X} = \mathcal{T}(\mathsf{X}) = [X_1 : \ldots : X_K]$, где $X_i = (x_1, \ldots, x_{i+L-1})^{\mathrm{T}}$, $K = N - L + 1$.
2. **Разложение**: $\mathbf{X} = \sum_{i=1}^{r} \sigma_i P_i Q_i^{\mathrm{T}} = \mathbf{X}_1 + \ldots + \mathbf{X}_r$, $\operatorname{rank} \mathbf{X}_i = 1$.
3. **Группировка**: $\mathbf{X} = \mathbf{X}_{I_1} + \ldots + \mathbf{X}_{I_m}$, где $\mathbf{X}_{I_k} = \sum_{i \in I_k} \mathbf{X}_i$.
4. **Восстановление**: $\mathsf{X} = \widetilde{\mathsf{X}}_{I_1} + \ldots + \widetilde{\mathsf{X}}_{I_m}$, где $\widetilde{\mathsf{X}}_{I_k} = \mathcal{T}^{-1} \circ \mathcal{H}(\mathbf{X}_{I_k})$.

MSSA: $\mathsf{X}^{(1)}, \ldots, \mathsf{X}^{(D)}$ — временные ряды.

Составим $\mathsf{X} = \{\mathsf{X}^{(d)}\}_{d=1}^{D}$ — D-канальный временной ряд.

Тогда на шаге вложения $\mathbf{X} = [\mathbf{X}^{(1)} : \ldots : \mathbf{X}^{(D)}]$, где $\mathbf{X}^{(i)} = \mathcal{T}(\mathsf{X}^{(i)})$.

Модификации MSSA

Модификации MSSA отличаются только шагом разложения.

Basic MSSA: сингулярное разложение (SVD) $\mathbf{X}$.

Toeplitz Block MSSA [Plaut and Vautard, 1994]: Q_i — ортонормированные собственные векторы матрицы

$$\mathbf{T}_{\text{Block}} = \begin{pmatrix} \mathbf{T}^{(K)}_{1,1} & \mathbf{T}^{(K)}_{1,2} & \dots & \mathbf{T}^{(K)}_{1,D} \\ \mathbf{T}^{(K)}_{2,1} & \mathbf{T}^{(K)}_{2,2} & \dots & \mathbf{T}^{(K)}_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{T}^{(K)}_{D,1} & \mathbf{T}^{(K)}_{D,D} & \dots & \mathbf{T}^{(K)}_{D,D} \end{pmatrix} \in \mathbb{R}^{DK \times DK},$$

где $\mathbf{T}^{(K)}_{l,k}$ — матрица с элементами

$$\left(\mathbf{T}^{(K)}_{l,k}\right)_{ij} = \frac{1}{N - |i-j|} \sum_{n=1}^{N-|i-j|} x^{(l)}_n x^{(k)}_{n+|i-j|}, \ 1 \leqslant i, j \leqslant K.$$

Toeplitz Sum MSSA: P_i — ортонормированные собственные векторы матрицы $\mathbf{T}_{\text{Sum}} = \sum_{i=1}^{D} \mathbf{T}^{(L)}_{i,i} \in \mathbb{R}^{L \times L}$, предлагается в этой работе.

Численное исследование

Дано: $\{\mathsf{F}^{(1)}, \mathsf{F}^{(2)}\} = \{\mathsf{S}^{(1)}, \mathsf{S}^{(2)}\} + \{\mathsf{R}^{(1)}, \mathsf{R}^{(2)}\}$, где R — независимые реализации белого гауссовского шума с $\sigma^2 = 25$, $N = 71$.

Задача: проверить точность базового и модифицированных методов для разных значений параметра L.

Рассмотрим 3 случая:

1. Косинусы с одинаковыми частотами:

$$s_n^{(1)} = 30\cos(2\pi n/12), \quad s_n^{(2)} = 20\cos(2\pi n/12), \quad n = 1, \ldots, N.$$

2. Косинусы с разными частотами:

$$s_n^{(1)} = 30\cos(2\pi n/12), \quad s_n^{(2)} = 20\cos(2\pi n/8), \quad n = 1, \ldots, N.$$

3. Полиномы первой степени (нестационарные ряды):

$$s_n^{(1)} = 1.2n, \quad s_n^{(2)} = 0.8n, \quad n = 1, \ldots, N.$$

Численное исследование. Результаты

Таблица: MSE восстановления сигнала.

	$L = 12$ $DK = 120$	$L = 24$ $DK = 96$	$L = 36$ $DK = 72$	$L = 48$ $DK = 48$	$L = 60$ $DK = 24$
Случай 1 ($\omega_1 = \omega_2$)					
MSSA	3.18	1.83	1.59	**1.47**	2.00
Toeplitz Sum MSSA	3.17	1.75	1.44	**1.32**	**1.33**
Toeplitz Block MSSA	1.39	**1.26**	**1.25**	1.33	1.97
Случай 2 ($\omega_1 \neq \omega_2$)					
MSSA	6.91	3.77	3.07	**2.88**	3.84
Toeplitz Sum MSSA	6.88	3.65	2.64	2.37	**2.27**
Toeplitz Block MSSA	4.47	3.67	**3.22**	**3.23**	3.8
Случай 3 (тренд)					
MSSA	3.42	1.94	1.63	**1.57**	2.27
Toeplitz Sum MSSA	3.32	**2.24**	3.04	5.91	11.95
Toeplitz Block MSSA	12.55	6.18	2.97	**1.78**	1.97

Трудоемкость: нахождение собственных векторов. Toeplitz Block MSSA зависит от DK, а Toeplitz Sum MSSA — от L.

Глава 2. Метод Monte Carlo (M)SSA

Статистический критерий с суррогатными выборками

Если распределение статистики T при верной H_0 неизвестно, оно оценивается с помощью G суррогатных выборок:

1. По выборке $X = (X_1, \ldots, X_n)$ строится статистика критерия $t = T(X)$.
2. Моделируется G выборок $R_1, \ldots, R_G$ объема n при верной H_0, и строятся статистики $t_i = T(R_i)$, $i = 1, \ldots, G$.
3. Находится оценка t_α критического значения $\hat{t}_\alpha$ как выборочный $(1-\alpha)$-квантиль $(t_1, \ldots, t_G)$.
4. Если $t > \hat{t}_\alpha$, то H_0 отвергается.

Замечание

Оценка критического значения t_α по квантилям выборки объема G имеет смысл при $\alpha > 1/G$, так как минимальное и максимальное значения выборки соответствуют α- и $(1-\alpha)$-квантилям при $\alpha = 1/G$. Если $\alpha < 1/G$, критерий с суррогатными выборками радикальнее исходного критерия.

Поправка неточных критериев

Приведем алгоритм, преобразовывающий неточный критерий (консервативный или радикальный) в асимптотически точный.

Зафиксируем H_0, уровень значимости α^*, количество выборок M для оценки $\alpha_I(\alpha)$ и их объем N:

1. Моделируется M выборок объема N при верной H_0.
2. По моделированным данным строится зависимость ошибки первого рода от уровня значимости $\alpha_I(\alpha)$.
3. Рассчитывается формальный уровень значимости: $\widetilde{\alpha}^* = \alpha_I^{-1}(\alpha^*)$. Критерий с таким уровнем значимости является асимптотически точным при $M \to \infty$.

Замечание

Критерий с суррогатными выборками, к которому применяется поправка, не должен быть слишком радикальным. Для сильно радикальных критериев с очень маленьким $\widetilde{\alpha}^*$ условие $G > 1/\widetilde{\alpha}^*$ приводит к очень большим вычислительным затратам и, тем самым, к практической нереализуемости.

Multiple Monte Carlo SSA

Модель: $\mathsf{X} = \mathsf{S} + \boldsymbol{\xi}$, где S — сигнал (колебания с какой-то частотой), $\boldsymbol{\xi}$ — стационарный процесс с нулевым средним.

Задача: проверить $H_0 : \mathsf{S} = 0$ — отсутствие сигнала.

Метод: Multiple Monte Carlo SSA [Golyandina, 2023]:

Входные данные: временной ряд X.
Параметры: длина окна L, выбор векторов $W_1, \ldots, W_H \in \mathbb{R}^L$, количество суррогатных реализаций G.

1. Строятся статистики критерия $\widehat{p}_k = \|\mathbf{X}^{\mathrm{T}} W_k\|^2$, $k = 1, \ldots, H$.
2. Оцениваются их распределения на основе G суррогатных реализаций $\boldsymbol{\xi}$.
3. Строятся предсказательные интервалы с поправкой на множественные сравнения с уровнем доверия $(1 - \alpha)$.

Далее предполагаем, что $\boldsymbol{\xi}$ — красный шум, причем с известными параметрами.

Используемый варинт MC-SSA

Определение

Пусть $\mathbf{X} = \sum_i \sigma_i P_i Q_i^T$ *— любое разложение* $\mathbf{X}$ *в сумму матриц единичного ранга. Будем называть* P_i *левыми, а* Q_i *— правыми векторами матрицы* $\mathbf{X}$.

В качестве векторов для проекции будем брать левые векторы матрицы $\mathbf{X}$. Варианты разложения траекторной матрицы: сингулярное, теплицево.

Плюс: если H_0 отверглась, можно восстановить сигнал с помощью SSA на основе значимых W_i.

Минус: этот вариант дает радикальный критерий, но Toeplitz MC-SSA менее радикален, что дает возможность использовать поправку неточных критериев [Ларин, 2022].

Численное сравнение MC-SSA с другими критериями

Подход: отбелить красный шум, т. е. сделать его белым, и затем применить статистический критерий, проверяющий гипотезу, что ряд является реализацией белого шума.

Для проверки исходного ряда на белый шум возьмем Q-тест Бокса-Пирса [Box and Pierce, 1970] и тест с использованием вейвлетов [Nason and Savchev, 2014].

Особенность второго теста: длина ряда должна быть степенью двойки.

Пусть $N = 128$. Сравним такой метод с MC-SSA при разных частотах в альтернативе.

Модель: $\mathsf{X} = \mathsf{S} + \boldsymbol{\xi}$, где $\mathsf{S} = \{A\cos(2\pi\omega n)\}_{n=1}^{N}$, $\boldsymbol{\xi}$ — красный шум с параметрами $\varphi = 0.7$ и $\delta = 1$.

$H_0 : A = 0$, $H_1 : A \neq 0$.

Результат численного сравнения MC-SSA с другими критериями

Таблица: Результаты численного сравнения MC-SSA с другими критериями ($\alpha^* = 0.1$)

Метод	$\alpha_I(\alpha^*)$	$\beta(\widetilde{\alpha}^*)$ $A = 1.5$ $\omega = 0.025$	$\beta(\widetilde{\alpha}^*)$ $A = 0.8$ $\omega = 0.125$	$\beta(\widetilde{\alpha}^*)$ $A = 0.5$ $\omega = 0.225$
MC-SSA ($L = 10$)	0.101	0.57	0.51	0.465
MC-SSA ($L = 32$)	0.163	0.566	0.678	0.668
MC-SSA ($L = 64$)	0.25	0.556	0.684	0.665
MC-SSA ($L = 96$)	0.593	0.599	0.734	0.709
MC-SSA ($L = 115$)	0.668	**0.668**	**0.791**	**0.753**
box	0.103	0.289	0.269	0.064
wavelet	0.091	0.354	0.414	0.57

Вывод: MC-SSA намного мощнее box и wavelet, особенно при малых частотах сигнала.

MC-SSA: обобщение на многомерный случай

MC-MSSA: SSA заменяется на MSSA и красный шум генерируется с тем же количеством каналов, что и у исходного временного ряда.

Замечание

*В одномерном случае левые векторы матрицы $\mathbf{X}$ становятся правыми заменой $L \mapsto N - L + 1$, поэтому по-отдельности рассматривать в качестве векторов для проекции правые векторы не нужно. В многомерном случае левые и правые векторы дают **разные** критерии.*

Замечание

Для рассмотрения правых векторов матрицы $\mathbf{X}$ в качестве W_k требуется заменить в алгоритме MC-SSA $\mathbf{X}$ на $\mathbf{X}^{\mathrm{T}}$ и $\boldsymbol{\Xi}$ на $\boldsymbol{\Xi}^{\mathrm{T}}$.

Численное сравнение модификаций MC-MSSA

Модель: $\mathsf{X} = \mathsf{S} + \boldsymbol{\xi}$, где $\boldsymbol{\xi}$ — красный шум с параметрами φ и $\delta = 1$, а S — сигнал с

$$s_n^{(1)} = s_n^{(2)} = A\cos(2\pi n\omega), \quad n = 1, \ldots, N, \quad N = 100.$$

H_0: $A = 0$, $H_1 : A \neq 0$.

Задача: сравнить критерии Basic MC-MSSA и Toeplitz MC-MSSA в двух вариациях.

Рассмотрим следующие примеры:

1. $\varphi = 0.7$, $\omega = 0.075$;
2. $\varphi = 0.3$, $\omega = 0.075$;
3. $\varphi = 0.7$, $\omega = 0.225$.

Результат численного сравнения модификаций MC-MSSA

Таблица: Мощность методов для оптимальных L при $\alpha^* = 0.1$

Метод	левые/правые векторы	$\beta(\widetilde{\alpha}^*)$ (пример 1)	$\beta(\widetilde{\alpha}^*)$ (пример 2)	$\beta(\widetilde{\alpha}^*)$ (пример 3)
SVD	левые	0.754	**0.399**	0.573
SVD	правые	0.754	0.382	0.442
Block	левые	**0.796**	**0.398**	**0.597**
Block	правые	0.717	0.389	0.473
Sum	левые	**0.806**	**0.421**	**0.625**
Sum	правые	0.748	**0.412**	**0.613**

Таблица: Размеры матриц методов

Метод	левые/правые векторы	Размер матрицы (пример 1)	Размер матрицы (пример 2)	Размер матрицы (пример 3)
SVD	левые	50	10	20
SVD	правые	80	80	80
Block	левые	**102**	**162**	**162**
Block	правые	42	22	42
Sum	левые	**80**	**20**	**80**
Sum	правые	80	80	80

Глава 3. Применение метода Monte Carlo SSA на практике

Зависимость радикальности и мощности MC-SSA от параметра L

Было произведено исследование зависимости мощности MC-SSA от длины окна L.

Численные эксперименты показали, что длина окна L, дающая наибольшую мощность, зависит от параметров шума $\boldsymbol{\xi}$, длины ряда N и **частоты сигнала** ω в H_1. Также с ростом N радикальность критерия при фиксированном L едва заметно уменьшается.

На их основе были выработаны следующие рекомендации:

1. Без поправки использовать MC-SSA можно только с $L \approx 10$.
2. Можно построить зависимость оптимальной длины окна от параметров ряда с помощью численного моделирования. Но это возможно, если есть дополнительная информация о диапазоне возможных частот в ряде.

MC-SSA с мешающим сигналом

Мешающий сигнал — это сигнал, который не нужно обнаруживать при проверки гипотезы о наличии сигнала (например, тренд, сезонность).

Подход: устранить влияние мешающего сигнала на векторы, на которые делается проекция.

Модель: $\mathsf{X} = \mathsf{F} + \mathsf{S} + \boldsymbol{\xi}$, где F — мешающий сигнал, S — неизвестный сигнал и $\boldsymbol{\xi}$ — красный шум.

Тогда $H_0 : \mathsf{S} = 0$ и $H_1 : \mathsf{S} \neq 0$.

Получено, что мощность критерия понижается, но больше от оценки параметров шума, чем от мешающего сигнала, причем чем меньше параметр φ, тем больше потеря в мощности.

Анализ реального примера

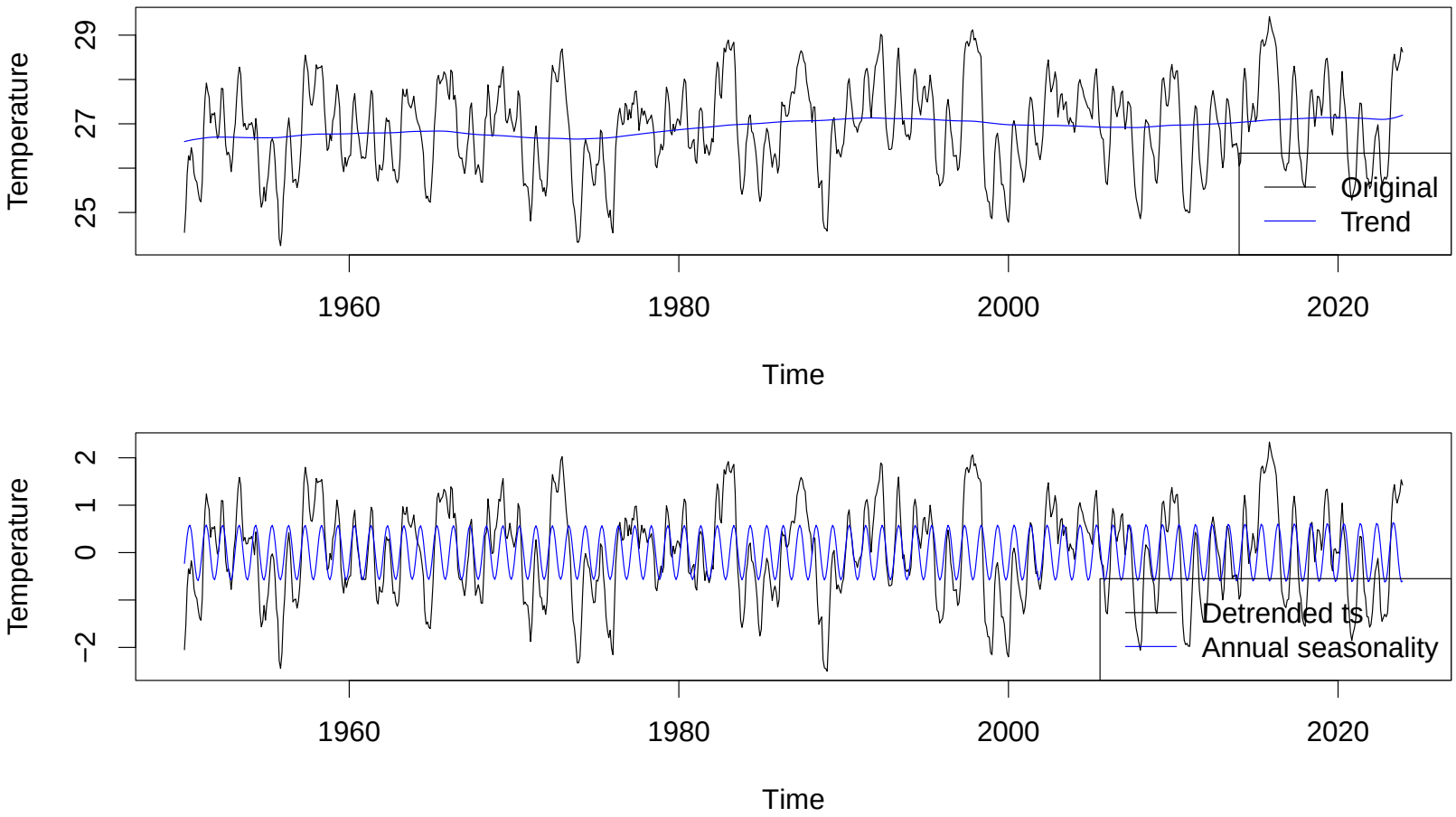

Рис.: Выделенный тренд ($L = 120$) и годовая периодичность ($L = 444$) с помощью SSA (данные по месяцам)

Анализ реального примера

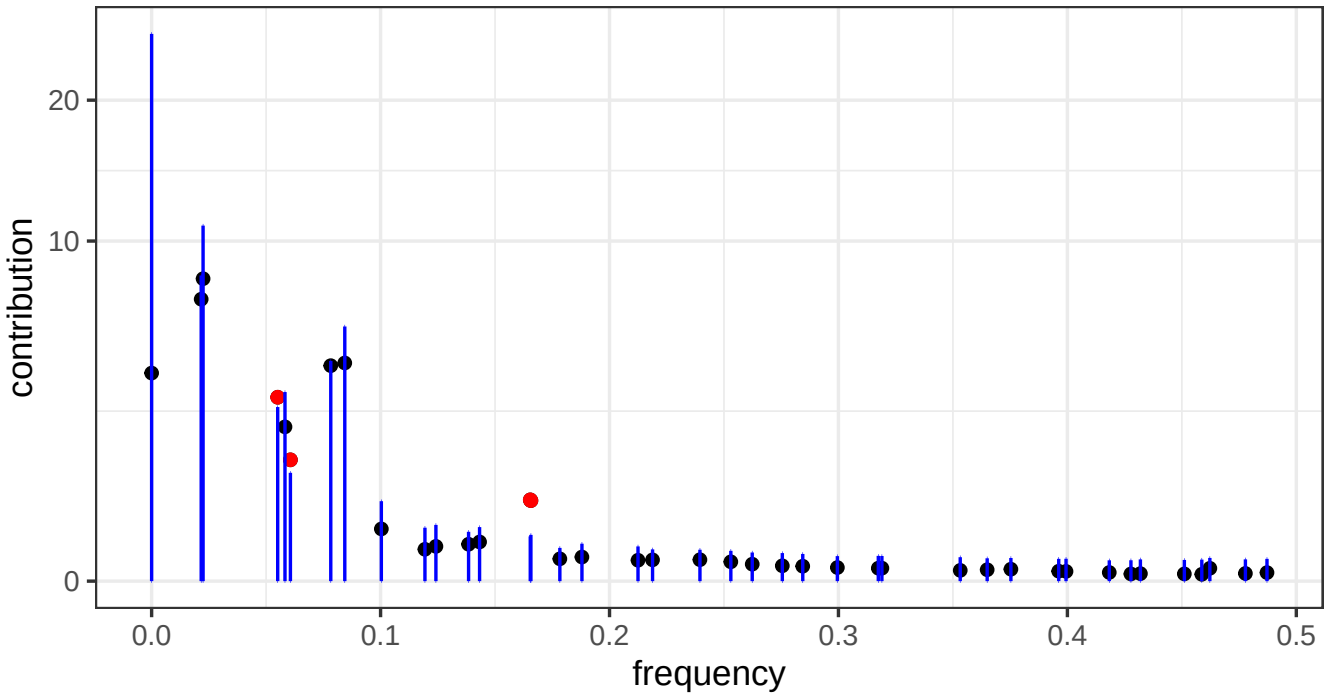

Рис.: Результат работы MC-SSA ($\alpha = 0.05$)

Значимыми являются четыре компоненты, две компоненты, имеющие **период** приблизительно $\mathbf{6}$, легко интерпретируются — это замеченная **полугодовая периодичность**.

Заключение

Мои результаты:

1. Был реализован метод Toeplitz MSSA в вариантах Block и Sum (на языке R). На основе численных исследований рекомендуется использовать вариант Sum в виду численной эффективности и реализации, подходящей под структуру пакета Rssa.
2. Был выработан подход к сравнению критериев, построенных на основе радикальных и использующих суррогатные выборки: выбирать наиболее мощный критерий среди не слишком радикальных.
3. На основе этого подхода по результатам численных исследований рекомендуется использовать модификацию Toeplitz Sum MC-MSSA с проекцией на левые векторы.
4. Проведены численные исследования по выбору оптимальной длины окна, степени искажения критерия при оценке параметров красного шума, а также случая мешающего сигнала.
5. Показано, что метод MC-SSA намного мощнее двух рассмотренных критериев в задаче обнаружения сигнала в красном шуме, особенно при малых частотах сигнала в альтернативе.