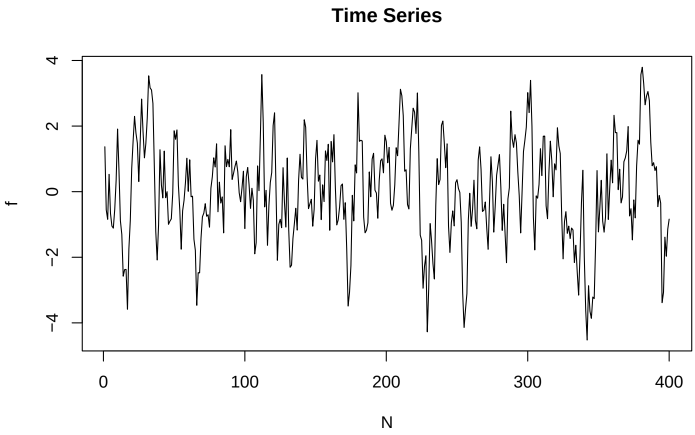
Выбор параметров в методе Monte Carlo SSA

Потешкин Егор Павлович

Санкт-Петербургский государственный университет
Кафедра статистического моделирования

Процессы управления и устойчивость
2 апреля 2024, Санкт-Петербург

Введение



Вопрос: это чистый шум или там есть сигнал?

Постановка задачи

$\mathsf{X} = (x_1, \ldots, x_N)$, $x_i \in \mathbb{R}$ — временной ряд.

Дано: $\mathsf{X} = \mathsf{T} + \mathsf{H} + \mathsf{R}$, где T — тренд, H — сезонность и R — шум.

Проблемы:

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

Методы:

1. Singular spectrum analysis (SSA) [Broomhead and King, 1986].
2. Monte-Carlo SSA (MC-SSA) [Allen and Smith, 1996] проверяет $H_0 : \mathsf{S} = 0$.

Задача: исследовать зависимость радикальности и мощности MC-SSA от параметра L.

Обозначения и известные результаты: оператор вложения и ганкелизации

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

Оператор вложения $\mathcal{T}_{\mathsf{SSA}}$:

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

где $K = N - L + 1$.

Оператор ганкелизации $\mathcal{H}$ — усреднение матрицы по побочным диагоналям.

Обозначения и известные результаты: SSA

Входные данные: временной ряд $\mathsf{X} = (x_1, \ldots, x_N)$.
Параметр: длина окна L.
Результат: m восстановленных составляющих временного ряда.

Рис.: Алгоритм SSA

Обозначения и известные результаты: Toeplitz SSA

Модификации SSA отличаются только шагом разложения.

Basic SSA: сингулярное разложение траекторной матрицы, универсальный метод.

Toeplitz SSA: теплицево разложение траекторной матрицы, имеет преимущество для стационарных рядов:

$$\mathbf{X} = \sum_{i=1}^{L} \sigma_i P_i Q_i^{\mathrm{T}} = \mathbf{X}_1 + \ldots + \mathbf{X}_L,$$

где $Q_i = \mathbf{X}^{\mathrm{T}} P_i / \sigma_i$, $\sigma_i = \|\mathbf{X}^{\mathrm{T}} P_i\|$, $\{P_i\}_{i=1}^{L}$ — собственные векторы матрицы $\mathbf{C}$ с элементами

$$c_{ij} = \frac{1}{N - |i-j|} \sum_{m=1}^{N-|i-j|} x_m x_{m+|i-j|}, 1 \leqslant i, j \leqslant L.$$

Обозначения и известные результаты: Monte-Carlo SSA

Рассмотрим задачу поиска сигнала (неслучайной составляющей) во временном ряде.

Модель: $\mathsf{X} = \mathsf{S} + \boldsymbol{\xi}$, где S — сигнал, $\boldsymbol{\xi}$ — стационарный процесс с нулевым средним.

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

Метод: Monte-Carlo SSA.

Определение

Случайный вектор $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)$ *называют красным шумом с параметрами* φ *и* δ, *если* $\xi_n = \varphi \xi_{n-1} + \delta \varepsilon_n$, *где* $0 < \varphi < 1$, $\varepsilon_n \sim N(0,1)$ *и* $\xi_1 \sim N(0, \delta^2/(1-\varphi^2))$.

Далее предполагаем, что $\boldsymbol{\xi}$ — красный шум.
Для исследования предполагаем, что параметры известны.

Обозначения и известные результаты: Monte-Carlo SSA. Алгоритм

Дано: $\mathsf{X} = \mathsf{S} + \mathsf{R}$, где S — сигнал, R — реализация $\boldsymbol{\xi}$.

Параметры: длина окна L, $W \in \mathbb{R}^L$ — вектор с какой-то частотой.

Результат: решение, отвергать H_0 или нет.

1. Построить статистику критерия $\widehat{p} = \|\mathbf{X}^\mathrm{T} W\|^2$.
2. Построить доверительную область случайной величины $p = \|\mathbf{\Xi}^\mathrm{T} W\|^2$: распределение p оценивается методом Монте-Карло.
3. Если $\widehat{p}$ не попадает в построенный интервал, то H_0 отвергается.

Обозначения и известные результаты: множественное тестирование

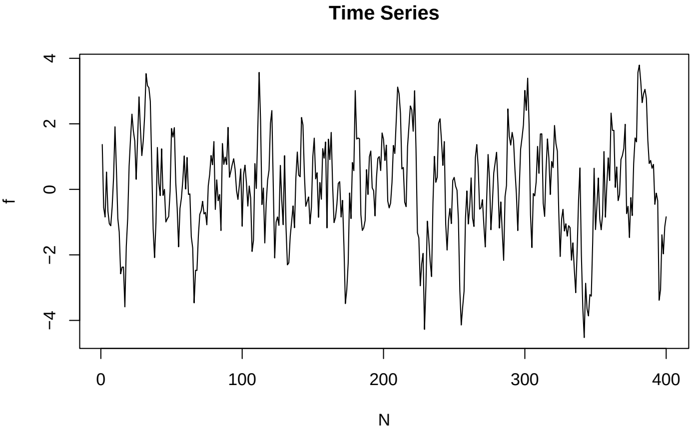
Если частота ω сигнала S известна, то в качестве W можно взять синусоиду с частотой ω. Но на практике ω редко бывает известна, поэтому необходимо рассматривать несколько векторов W_k, $k = 1, \ldots, H$.

Проблему множественного тестирования решает метод Multiple MC-SSA [Golyandina, 2023].

Гипотеза об отсутствии сигнала отвергается, если хотя бы для одного вектора $W = W_k$ значение $\widehat{p}$ оказывается значимым.

Monte Carlo SSA: пример

$\mathsf{X} = \mathsf{S} + \boldsymbol{\xi}$, где $\mathsf{S} = \{A\cos(2\pi\omega n)\}_{n=1}^{N}$, $A = 1$, $\omega = 0.1$, $N = 400$,
$\boldsymbol{\xi}$ – красный шум с параметрами $\varphi = 0.7$ и $\delta = 1$.



Monte Carlo SSA: пример

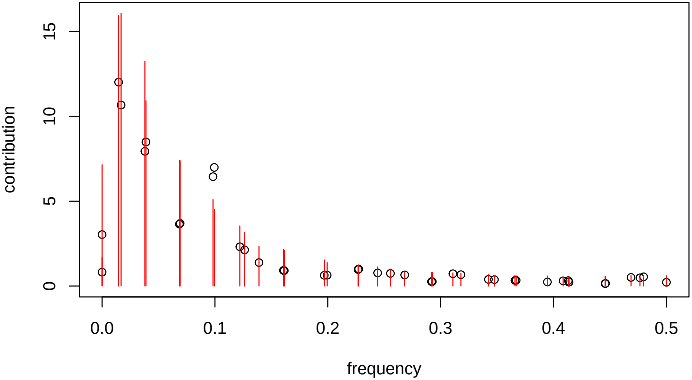


Рис.: Результат работы метода MC-SSA

MC-SSA: выбор векторов для проекции

В качестве векторов для проекции будем брать собственные векторы матрицы $\mathbf{C}$ теплицева разложения $\mathbf{X}$.

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

Минус: этот вариант дает радикальный критерий, поскольку W_k зависят от ряда X, в котором ищется сигнал.
Решение: использовать метод эмпирической поправки критерия, чтобы сделать его точным.

Поправка неточных критериев

Зафиксируем H_0, уровень значимости α^*, количество выборок M_1 для оценки $\alpha_I(\alpha)$ и их объем N:

1. Моделируется M_1 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 $M_1 \to \infty$.

Заметим, что если критерий сильно радикальный, то функция $\alpha_I(\alpha)$ имеет большую производную в нуле, что существенно затрудняет оценку $\alpha_I^{-1}(\alpha^*)$.

ROC-кривая

Определение

ROC-кривая — это кривая, задаваемая параметрически

$$\begin{cases} x = \alpha_I(\alpha) \\ y = \beta(\alpha) \end{cases}, \quad \alpha \in [0, 1],$$

где $\alpha_I(\alpha)$ — функция зависимости ошибки первого рода α_I от уровня значимости α, $\beta(\alpha)$ — функция зависимости мощности β от уровня значимости α.

С помощью ROC-кривых можно сравнивать по мощности неточные (в частности, радикальные) критерии после того, как к ним применена поправка.

Зависимость радикальности и мощности MC-SSA от параметра L

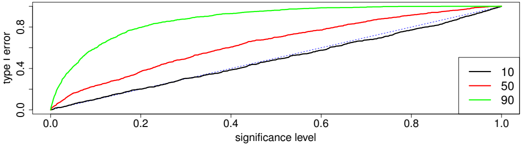
Задача: выбрать такую длину окна L, которая дает максимально мощный критерий, но при этом не слишком радикальный, чтобы можно было применить поправку.

В следующих примерах рассматриваем следующую модель:

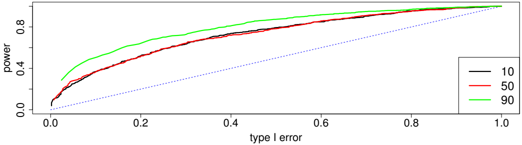
$$\mathsf{X} = \mathsf{S} + \boldsymbol{\xi},$$

где $\mathsf{S} = \{A\cos(2\pi\omega n\}_{n=1}^{N}$ — сигнал, $\boldsymbol{\xi}$ — красный шум с параметрами φ и $\delta = 1$. Тогда $H_0 : A = 0$, $H_1 : A \neq 0$.

Пример 1. $\varphi = 0.7$, $N = 100$



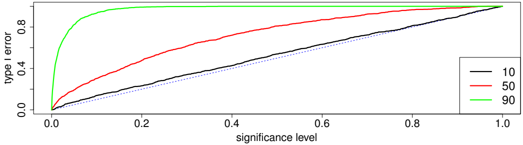
(a) Ошибка первого рода



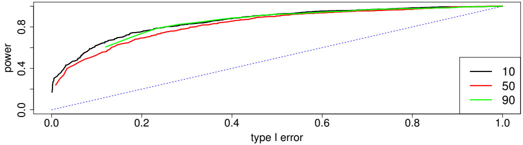
(b) ROC-кривая ($\omega = 0.075$)

Рис.: Пример 1. $\varphi = 0.7$, $N = 100$

Пример 2. $\varphi = 0.3,\ N = 100$



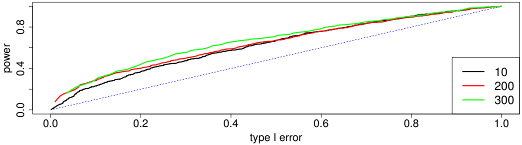
(a) Ошибка первого рода



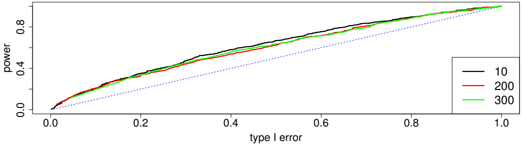
(b) ROC-кривая ($\omega = 0.075$)

Рис.: Пример 2. $\varphi = 0.3,\ N = 100$

Пример 3. $N = 400$



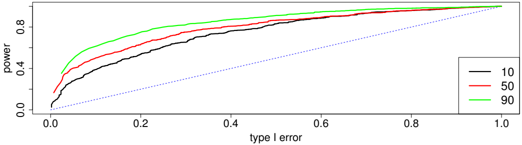
(a) ROC-кривая ($\varphi = 0.7$, $\omega = 0.075$)



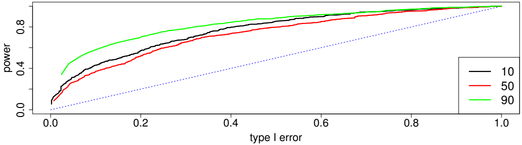
(b) ROC-кривая ($\varphi = 0.3$, $\omega = 0.075$)

Рис.: Пример 3. $N = 400$

Пример 4. Зависимость от параметров сигнала



(a) ROC-кривая ($\omega = 0.175$)



(b) ROC-кривая ($\omega = 0.025$)

Рис.: Пример 4. $\varphi = 0.7$, $N = 100$

Заключение

Численные эксперименты показали, что длина окна L, дающая наибольшую мощность, зависит от параметров шума, длины ряда и частоты сигнала в H_1.

На их основе были выработаны следующие рекомендации:

1. Первый вариант — использовать поправленный критерий MC-SSA с $L = 10$. Плюсы: нетрудозатратно, а также критерий не сильно радикальный. Минус: такой выбор L может являться неоптимальным, т.е. возможна некоторая потеря в мощности.
2. Второй вариант — построить зависимость оптимальной длины окна от параметров ряда с помощью численного моделирования. Это возможно, если есть дополнительная информация о диапазоне возможных частот в ряде.