Санкт-Петербургский государственный университет

Прикладная математика и информатика

Отчет по учебной практике 4 (научно-исследовательской работе) (семестр 8)

Метод Монте-Карло SSA для одномерных и многомерных временных рядов

Выполнил:

Потешкин Егор Павлович

группа 20.Б04-мм

Научный руководитель:

д. ф.-м. н., доцент

Голяндина Нина Эдуардовна

Кафедра Статистического Моделирования

Санкт-Петербург

2024

Оглавление

Введение

Метод Singular Spectrum Analysis (SSA) является мощным инструментом для анализа временных рядов. Он позволяет разложить ряд на интерпретируемые компоненты, такие как тренд, периодические колебания и шум, что значительно упрощает процесс анализа. Метод Monte-Carlo SSA, в свою очередь, решает задачу обнаружения сигнала в шуме [1].

Однако, вариант Monte-Carlo SSA для анализа многомерных временных рядов мало исследован. В работе [2] рассматривается применение метода Monte Carlo SSA для анализа многомерных временных рядов, и авторы сталкиваются с проблемой отсутствия реализации Тёплицева варианта MSSA в пакете Rssa [3].

В этой работе была поставлена задача реализовать двумя способами метод Toeplitz MSSA, сравнить их между собой и с обычным MSSA как для методов оценки сигнала, так и для использования в Monte-Carlo MSSA, а также рассмотреть Monte-Carlo SSA в условиях реальных задач.

В главе 1 приведено описание метода MSSA и двух его модификаций, и их численное сравнение. В главе 2 представлен метод Monte-Carlo SSA, и приведено численное сравнение метода с разными параметрами. В ходе работы в этом семестре была добавлена глава 3, в которой рассмотрено два способа оценки неизвестных параметров красного шума и их сравнение, а также разобран случай Monte-Carlo SSA, когда во временном ряде присутствует мешающий сигнал.

Впервые Monte-Carlo SSA упоминается в [4], алгоритм которого совпадает одиночным тестом MC-SSA, представленным в этой работе. В [5] метод был применен на данных о зимнем индексе Североатлантического колебания (NAO index), и были обнаружены значимые колебания с периодом 7.7 лет. В [6] повторили результат работы [5] уже на ежемесячном индексе Североатлантического колебания: с помощью усовершенствованного Monte-Carlo SSA (enhanced MC-SSA) был обнаружен сигнал с периодом в 7.8 лет. В [7] представлен нелинейный вариант Monte-Carlo SSA, использующий ядерный анализ главных компонент (kernel PCA). В работе [8] метод применяется для извлечения сезонных колебаний из непрерывных GPS-наблюдений, однако отмечается, что модель AR(1) может быть неподходящей для этих данных.

Глава 1

Метод MSSA и его модификации

В этой главе рассматривается метод Multivariate Singular Spectrum Analysis [9] (сокращенно MSSA) и его модификации. В разделе 1.1 представлены вспомогательные определения, нужные в дальнейшем. В разделе 1.2 представлен алгоритм метода MSSA, а в разделах 1.3.1 и 1.3.2 — его базовый и стандартный теплицев [10] варианты. В разделе 1.3.3 предлагается другая теплицева модификация MSSA, и в разделе 1.4 происходит сравнение методов MSSA: как теоретическое, так и численное.

1.1. Вспомогательные определения

Определение 1. Пусть X — одномерный временной ряд длины N. Выберем параметр L, называемый *длиной окна*, $1 < L < N$. Рассмотрим $K = N - L + 1$ векторов вложения $X_i = (x_i, \ldots, x_{i+L-1})^{\mathrm{T}}$, $1 \leqslant j \leqslant K$. Определим оператор вложения $\mathcal{T}$ следующим образом:

$$\mathcal{T}(\mathsf{X}) = \mathbf{X} = [X_1 : \ldots : X_K] = \begin{pmatrix} x_1 & x_2 & \cdots & x_K \\ x_2 & x_3 & \cdots & x_{K+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & \cdots & x_N \end{pmatrix}. \tag{1.1}$$

Определение 2. Матрицу $\mathbf{X}$ из (1.1) называют траекторной матрицей.

Заметим, что матрица $\mathbf{X}$ является *ганкелевой*, т.е на всех ее побочных диагоналях стоят одинаковые элементы, а оператор $\mathcal{T}$ задает взаимно-однозначное соответствие между множеством временных рядов длины N и множеством ганкелевых матриц $L \times K$.

Определение 3. Пусть $\mathbf{Y} = \{y_{ij}\}_{i,j=1}^{L,K}$ — некоторая матрица. Определим оператор ганкелизации $\mathcal{H}$:

$$(\mathcal{H}(\mathbf{Y}))_{ij} = \sum_{(l,k)\in A_s} y_{lk}/w_s, \tag{1.2}$$

где $s = i+j-1$, $A_s = \{(l,k) : l+k = s+1, 1 \leqslant l \leqslant L, 1 \leqslant k \leqslant K\}$ и $w_s = |A_s|$ — количество элементов в множестве A_s. Это соответствует усреднению элементов матрицы $\mathbf{Y}$ по побочным диагоналям.

Определение 4. Случайный процесс $\xi = (\xi_1, \ldots, \xi_n, \ldots)$ называется стационарным, если $\forall k \geqslant 1$ $\mathsf{E}\xi_k = \text{const}$ и $\forall k, l \geqslant 1$

$$K(k,l) \stackrel{\text{def}}{=} \text{cov}(\xi_k, \xi_l) = \widetilde{K}(k-l).$$

Определение 5. Белый гауссовский шум $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n, \ldots)$ — стационарный случайный процес с $\varepsilon_n \sim N(0, \sigma^2)$ $\forall n$ и $\widetilde{K}(n) = 0$, при $n > 0$.

Определение 6. Детерминированный временной ряд $\mathsf{X} = (x_1, \ldots, x_n, \ldots)$ называют стационарным, если существует функция $R_x(k)$ $(-\infty < k < +\infty)$ такая, что $\forall k, l \geqslant 0$

$$R_x^{(N)}(k,l) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{m=1}^{N} x_{k+m} x_{l+m} \underset{N\to\infty}{\longrightarrow} R_x(k-l).$$

Определение 7. Рассмотрим вещественнозначные одномерные временные ряды $\mathsf{X}^{(d)}$ длины N_d, $d = 1, \ldots, D$. Тогда составленный из этих рядов $\mathsf{X} = \{\mathsf{X}^{(d)}\}_{d=1}^{D}$ — D-канальный временной ряд с длинами N_d.

Определение 8. Пусть $\mathsf{X} = \{\mathsf{X}^{(d)}\}_{d=1}^{D}$ — D-канальный временной ряд. Ряд X называют стационарным, если каждый канал $\mathsf{X}^{(d)}$ — стационарный.

1.2. Метод MSSA

Метод MSSA состоит из четырех этапов: *вложения*, *разложения*, *группировки* и *диагонального усреднения*.

1.2.1. Вложение

Зафиксируем L, $1 < L < \min(N_1, \ldots, N_D)$. Для каждого ряда $\mathsf{X}^{(d)}$ составим траекторную матрицу $\mathbf{X}^{(d)}$. Обозначим $K = \sum_{d=1}^{D} K_d$. Результатом этапа вложения является траекторная матрица многоканального временного ряда

$$\mathbf{X} = \mathcal{T}_{\text{MSSA}}(\mathsf{X}) = [\mathcal{T}(\mathsf{X}^{(1)}) : \ldots : \mathcal{T}(\mathsf{X}^{(D)})] = [\mathbf{X}^{(1)} : \ldots : \mathbf{X}^{(D)}]. \tag{1.3}$$

1.2.2. Разложение

Задача этапа разложения — разбить траекторную матрицу $\mathbf{X}$ в сумму матриц единичного ранга: $\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_p$.

Определение 9. Пусть $\mathbf{X} = \sum_i \sigma_i P_i Q_i^T$ — любое разложение $\mathbf{X}$ в сумму матриц ранга 1. Будем называть P_i *левыми*, а Q_i — *правыми векторами* матрицы $\mathbf{X}$.

1.2.3. Группировка

На этом шаге множество индексов $I = \{1, \ldots, p\}$ разбивается на m непересекающихся множеств $I_1, \ldots, I_m$ и матрица $\mathbf{X}$ представляется в виде суммы

$$\mathbf{X} = \sum_{k=1}^{m} \mathbf{X}_{I_k},$$

где $\mathbf{X}_{I_k} = \sum_{i \in I_k} \mathbf{X}_i$.

1.2.4. Диагональное усреднение

Пусть $\mathbf{Y} = [\mathbf{Y}^{(1)} : \ldots : \mathbf{Y}^{(M)}]$ — некоторая составная матрица, тогда оператор ганкелизации для составной матрицы

$$\mathcal{H}_{\text{stacked}}(\mathbf{Y}) = [\mathcal{H}(\mathbf{Y}^{(1)}) : \ldots : \mathcal{H}(\mathbf{Y}^{(M)})]. \tag{1.4}$$

Финальным шагом MSSA является преобразование каждой матрицы $\mathbf{X}_{I_k}$, составленной в разделе 1.2.3, в D-канальный временной ряд:

$$\widetilde{\mathsf{X}}_{I_k} = \mathcal{T}^{-1}_{\text{MSSA}} \circ \mathcal{H}_{\text{stacked}}\left(\mathbf{X}_{I_k}\right), \tag{1.5}$$

где $\mathcal{T}_{\text{MSSA}}$ — оператор вложения (1.3), $\mathcal{H}_{\text{stacked}}$ — оператор ганкелизации (1.4).

Замечание 1. При $D = 1$ X — одномерный временной ряд, и приведенный выше алгоритм совпадает с алгоритмом Basic SSA, описанный в [11].

1.3. Этап разложения

Модификации MSSA отличаются только этапом разложения, остальные этапы остаются неизменными.

1.3.1. Basic MSSA

Базовый вариант MSSA использует сингулярное разложение (SVD) матрицы $\mathbf{X}$. Положим $\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$. Пусть λ_i — собственные числа, а U_i — ортонормированная система векторов матрицы $\mathbf{S}$. Упорядочим λ_i по убыванию и найдем p такое, что $\lambda_p > 0$, а $\lambda_{p+1} = 0$. Тогда

$$\mathbf{X} = \sum_{i=1}^{p} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}} = \sum_{i=1}^{p} \mathbf{X}_i, \tag{1.6}$$

где $V_i = \mathbf{X}^{\mathrm{T}} U_i / \sqrt{\lambda_i}$. Тройку $(\sqrt{\lambda_i}, U_i, V_i)$ принято называть i-й собственной тройкой сингулярного разложения, $\sqrt{\lambda_i}$ — сингулярным числом, U_i — левым сингулярным вектором, а V_i — правым сингулярным вектором.

Замечание 2. Левые сингулярные векторы U_i являются собственными векторами матрицы $\mathbf{X}\mathbf{X}^{\mathrm{T}}$, а правые V_i в свою очередь — матрицы $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. В одномерном случае U_i и V_i равносильны с точностью до замены L на $N - L + 1$. Но при $D > 1$ это не так по построению матрицы $\mathbf{X}$. Для наглядности рассмотрим случай $D = 2$, тогда $\mathbf{X}\mathbf{X}^{\mathrm{T}} = \mathbf{X}^{(1)}(\mathbf{X}^{(1)})^{\mathrm{T}} + \mathbf{X}^{(2)}(\mathbf{X}^{(2)})^{\mathrm{T}}$, а

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{pmatrix} (\mathbf{X}^{(1)})^{\mathrm{T}}\mathbf{X}^{(1)} & (\mathbf{X}^{(1)})^{\mathrm{T}}\mathbf{X}^{(2)} \\ (\mathbf{X}^{(2)})^{\mathrm{T}}\mathbf{X}^{(1)} & (\mathbf{X}^{(2)})^{\mathrm{T}}\mathbf{X}^{(2)} \end{pmatrix}.$$

1.3.2. Toeplitz Block MSSA

Если предполагается, что ряд X стационарен, то можно заменить матрицу $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ (или $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ в силу замечания 2) на некоторую другую, учитывая стационарность исходного ряда. Для этого введем следующее обозначение.

Определение 10. Пусть $\mathsf{X} = \{\mathsf{X}^{(d)}\}_{d=1}^{D}$ — D-канальный временной ряд с $N_d = N$. Зафиксируем $1 < M < N$. Обозначим за $\mathbf{T}_{l,k}^{(M)} \in \mathbb{R}^{M \times M}$ матрицу с элементами

$$\left(\mathbf{T}_{l,k}^{(M)}\right)_{ij} = \frac{1}{N - |i-j|} \sum_{n=1}^{N-|i-j|} x_n^{(l)} x_{n+|i-j|}^{(k)}, \; 1 \leqslant i, j \leqslant M,$$

Замечание 3. Если ряд X стационарен, матрица $\mathbf{T}_{l,k}^{(M)}$ является оценкой кросс-ковариационной матрицы l и k-го каналов.

В работе [10] предложен способ разложения $\mathbf{X}$, который мы назовем Toeplitz Block MSSA. Этот способ использует вместо матрицы $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ матрицу

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

где $K = N - L + 1$. Найдя ортонормированные собственные векторы $Q_1, \ldots, Q_{DK}$ матрицы $\mathbf{T}_{\mathrm{Block}}$, получаем разложение траекторной матрицы $\mathbf{X}$:

$$\mathbf{X} = \sum_{i=1}^{DK} \sigma_i P_i Q_i^{\mathrm{T}} = \mathbf{X}_1 + \ldots + \mathbf{X}_{DK}, \tag{1.7}$$

где $Z_i = \mathbf{X}Q_i$, $P_i = Z_i / \|Z_i\|$, $\sigma_i = \|Z_i\|$.

1.3.3. Toeplitz Sum MSSA

Вместе с методом Block рассмотрим другой вариант разложения, который назовем Sum. Рассмотрим вместо матрицы $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ матрицу $\mathbf{T}_{\mathrm{Sum}} = \sum_{d=1}^{D} \mathbf{T}_{d,d}^{(L)} \in \mathbb{R}^{L\times L}$. Найдем ортонормированные собственные векторы $P_1, \ldots, P_L$ матрицы $\mathbf{T}_{\mathrm{Sum}}$ и разложим траекторную матрицу $\mathbf{X}$ следующим образом:

$$\mathbf{X} = \sum_{i=1}^{L} \sigma_i P_i Q_i^{\mathrm{T}} = \mathbf{X}_1 + \ldots + \mathbf{X}_L, \tag{1.8}$$

где $S_i = \mathbf{X}^{\mathrm{T}} P_i$, $Q_i = S_i/\|S_i\|$, $\sigma_i = \|S_i\|$.

1.4. Сравнение методов MSSA

1.4.1. Теоретическое сравнение методов

Базовый вариант MSSA можно использовать для временных рядов с разными длинами каналов. Toeplitz Sum MSSA также позволяет это делать, в отличие от Toeplitz Block MSSA. Связано это с вычислением матриц $\mathbf{T}_{l,k}^{(K)}$, для которых при $l \neq k$ обязательно условие $N_l = N_k$. В методе Sum такой проблемы не возникает, поскольку $l = k$ всегда.

Также имеет смысл сравнить обе модификации по трудоемкости построения матриц и нахождения их собственных векторов. В методе Block нужно построить $D(D{+}1)/2$ теплицевых матриц размера $K\times K$, в методе Sum — D теплицевых матриц размера $L\times L$. Тогда, учитывая, что для матрица $\mathbf{T}_{l,k}^{(M)}$ задается M элементами, метод Sum численно эффективнее метода Block в построении матрицы, когда выполняется неравенство

$$\frac{D(D+1)K}{2} > DL \iff (D+1)N - (D+1)L + (D+1) > 2L \iff L < \frac{(D+1)(N+1)}{D+3}.$$

Что касается нахождения собственных векторов, если использовать спектральное разложение матрицы, то трудоемкость составляет $\mathcal{O}(D^3K^3)$ для метода Block и $\mathcal{O}(L^3)$ для метода Sum.

1.4.2. Численное сравнение методов

Посмотрим на точность базового и модифицированных методов MSSA для разных значений параметра L. Рассмотрим следующий двухканальный временной ряд длины

$N = 71$:

$$\{\mathsf{F}^{(1)}, \mathsf{F}^{(2)}\} = \{\mathsf{S}^{(1)}, \mathsf{S}^{(2)}\} + \{\mathsf{N}^{(1)}, \mathsf{N}^{(2)}\},$$

где $\mathsf{S}^{(1)}$, $\mathsf{S}^{(2)}$ — некоторые сигналы, а $\mathsf{N}^{(1)}$, $\mathsf{N}^{(2)}$ — независимые реализации гауссовского белого шума с $\sigma^2 = 25$.

Рассмотрим 3 случая, первые два из которых рассматривались ранее в [9]:

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):

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

В качестве оценки точности восстановления сигнала было взято среднеквадратичное отклонение от истинного значения. В таблице 1.1 представлены результаты на основе 10000 реализаций шума. Наиболее точные результаты для каждого метода были выделены жирным шрифтом. Лучший результат для каждого случая выделен отдельно синим.

Как видно из таблицы 1.1, в первом случае метод Block лучше всего выделял сигнал. В случае разных частот каналы имеют разную структуру, поэтому наиболее оптимальным является использовать Toeplitz SSA для каждого канала по отдельности. В третьем случае мы имеем дело с нестационарными рядами одинаковой структуры, поэтому стандартный MSSA справляется лучше всего.

Заметим, что преимущество Block перед Sum в первом случае не очень больше. Также, если сравнивать методы по трудоемкости, для оптимальной длины окна метод Sum численно эффективнее Block: в случае Sum для оптимального $L \approx 2N/3$ количество параметров равно $2 \cdot 2N/3$, в случае Block $L \approx N/2$ и количество параметров равно $3 \cdot N/2$. И еще раз отметим, что Sum можно применять к временным рядам разной длины, поэтому рекомендуется использовать именно Toeplitz Sum MSSA.

Таблица 1.1. MSE восстановления сигнала.

Случай 1 ($\omega_1 = \omega_2$)	$L = 12$	$L = 24$	$L = 36$	$L = 48$	$L = 60$
SSA	3.25	**2.01**	**2.00**	**2.01**	3.25
Toeplitz SSA	3.2	1.87	1.63	**1.59**	1.67
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 ($\omega_1 \neq \omega_2$)	$L = 12$	$L = 24$	$L = 36$	$L = 48$	$L = 60$
SSA	3.25	**2.01**	**2.00**	**2.01**	3.25
Toeplitz SSA	3.2	1.87	1.63	**1.59**	1.67
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)	$L = 12$	$L = 24$	$L = 36$	$L = 48$	$L = 60$
SSA	3.65	2.08	**1.96**	2.08	3.65
Toeplitz SSA	3.33	**2.43**	3.74	7.84	16.29
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

Глава 2

Метод Monte-Carlo (M)SSA

Это глава делится на две части. В разделе 2.1 рассматривается метод Monte-Carlo SSA (сокращенно MC-SSA), а в разделе 2.2 его многомерное обобщение (MC-MSSA). В названии методов присутсвует (M)SSA, поскольку при построении критериев используется траекторная матрица временного ряда, а самый распространенный вариант использует разложение этой траекторной матрицы. В связи с тем, что получить разложение траекторной матрицы можно разными способами, возникают разные варианты MC-(M)SSA.

В разделах 2.1.2 и 2.1.3 приведены алгоритмы метода MC-SSA и его модификации с поправкой на множественые тестирования [1]. В разделе 2.1.4 рассмотрен выбор параметров метода, который используется в дальнейших численных исследованиях. Известное решение проблемы радикальности критерия при таком выборе параметров приведено в разделе 2.1.5. В разделе 2.1.6 проведено численное сравнение MC-SSA с другими статистическими критериями. Завершает эту главу раздел 2.2.2, где происходит численное сравнение модификаций метода MC-MSSA на одном примере.

2.1. Monte Carlo SSA

2.1.1. Постановка задачи

Рассмотрим задачу поиска сигнала (неслучайной составляющей) во временном ряде. Модель выглядит следующим образом:

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

где S — сигнал, $\boldsymbol{\xi}$ — стационарный процесс с нулевым средним. Тогда нулевая гипотеза $H_0 : \mathsf{S} = 0$ (отсутствие сигнала, ряд состоит из чистого шума) и альтернатива $H_1 : \mathsf{S} \neq 0$ (ряд содержит сигнал, например, периодическую составляющую).

Определение 11. Случайный процесс $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n, \ldots)$ называют красным шумом с параметрами φ и δ, если $\xi_n = \varphi\xi_{n-1} + \delta\varepsilon_n$, где $0 < \varphi < 1$, ε_n — белый гауссовский шум с дисперсией 1 и ξ_1 имеет нормальное распределение с нулевым средним и дисперсией $\delta^2/(1-\varphi^2)$.

В этой и следующей главах под шумом будем подразумевать именно красный, причем в данной главе с известными параметрами. Также будем рассматривать только односторонние критерии.

2.1.2. Одиночный тест

Пусть $\boldsymbol{\xi}$ — красный шум. Зафиксируем длину окна L и обозначим траекторную матрицу ряда $\boldsymbol{\xi}$ как $\boldsymbol{\Xi}$. Рассмотрим вектор $W \in \mathbb{R}^L$ такой, что $\|W\| = 1$. Введем величину

$$p = \|\boldsymbol{\Xi}^{\mathrm{T}} W\|^2.$$

Статистикой критерия является величина

$$\widehat{p} = \|\mathbf{X}^{\mathrm{T}} W\|^2.$$

Если вектор W — синусоида с частотой ω, то $\widehat{p}$ отражает вклад частоты ω в исходный ряд.

Рассмотрим алгоритм статистического критерия проверки наличия сигнала в ряде с проекцией на один вектор W, описанный в работе [1].

Алгоритм 1. Одиночный тест [1]

1. Построить статистику критерия $\widehat{p}$.
2. Построить доверительную область случайной величины p: интервал от нуля до $(1-\alpha)$-квантиля, где α — уровень значимости.
3. Если $\widehat{p}$ не попадает в построенный интервал — H_0 отвергается.

Построенная доверительная область называется *прогнозируемым интервалом* с уровнем доверия $1-\alpha$.

Замечание 4. В большинстве случаев, распределение p неизвестно. Поэтому оно оценивается методом Монте-Карло: берется G реализаций случайной величины $\boldsymbol{\xi}$, для каждой вычисляется p и строится эмпирическое распределение. В связи с этим описанный выше алгоритм называют методом Monte-Carlo SSA.

Замечание 5. Если частота ω сигнала S известна, то в качестве W можно взять синусоиду с частотой ω. Но на практике ω редко бывает известна, что делает данный критерий несостоятельным против H_1.

2.1.3. Множественный тест

Пусть теперь частоты периодических компонент неизвестны, что не редкость на практике. Тогда подобно одиночному тесту рассмотрим набор $W_1, \ldots, W_H$ векторов для проекции, и для каждого $k = 1, \ldots, H$ построим статистику критерия $\widehat{p}_k$:

$$\widehat{p}_k = \|\mathbf{X}^{\mathrm{T}} W_k\|^2, \quad k = 1, \ldots, H. \tag{2.1}$$

В таком случае нужно построить H предсказательных интервалов для каждого W_k по выборкам $P_k = \{p_{ki}\}_{i=1}^{G}$ с элементами

$$p_{ki} = \|\boldsymbol{\Xi}_i^{\mathrm{T}} W_k\|^2, \quad i = 1, \ldots, G;\ k = 1, \ldots, H, \tag{2.2}$$

где G — количество суррогатных реализаций $\boldsymbol{\xi}$, $\boldsymbol{\Xi}_i$ — траекторная матрица i-й реализации $\boldsymbol{\xi}$.

В работе [1] подробно описана проблема множественного тестирования, когда вероятность ложного обнаружения периодической составляющей для одной из рассматриваемых частот (групповая ошибка I рода) неизвестна и значительно превышает заданный уровень значимости (частота ошибок одиночного теста), и ее решение. Приведем модифицированный алгоритм построения критерия в случае множественного тестирования, который будем использовать в дальнейшем.

Алгоритм 2. Multiple MC-SSA [1]

1. Для $k = 1, \ldots, H$ вычисляется статистика $\widehat{p}_k$, выборка $P_k = \{p_{ki}\}_{i=1}^{G}$, ее среднее μ_k и стандартное отклонение σ_k.

2. Вычисляется $\eta = (\eta_1, \ldots, \eta_G)$, где

$$\eta_i = \max_{1 \leqslant k \leqslant H} (p_{ki} - \mu_k)/\sigma_k, \quad i = 1, \ldots, G.$$

3. Находится q как выборочный $(1 - \alpha)$-квантиль η, где α — уровень значимости.

4. Нулевая гипотеза не отвергается, если

$$\max_{1 \leqslant k \leqslant H} (\widehat{p}_k - \mu_k)/\sigma_k < q.$$

5. Если H_0 отвергнута, вклад W_k (и соответствующей частоты) значим, если $\widehat{p}_k$ превосходит $\mu_k + q\sigma_k$. Таким образом, $[0, \mu_k + q\sigma_k]$ считаются скорректированными интервалами прогнозирования.

2.1.4. Выбор векторов для проекции

Отметим, что в SSA правые векторы матрицы $\mathbf{X}$ становятся левыми заменой L на $N-L+1$, поэтому рассматривать по-отдельности левые и правые не нужно. Это не так в случае MSSA, который рассмотрен ниже.

В данной работе в качестве $W_1, \ldots, W_H$ берутся левые векторы матрицы $\mathbf{X}$. Такой способ выбора векторов для проекции самый распространенный, поскольку, если есть значимые векторы, можно восстановить сигнал с помощью SSA на их основе. Но этот вариант, вообще говоря, дает радикальный критерий. Бороться с этой проблемой позволяет метод эмпирической поправки критерия, основанный на оцененных ошибках первого рода в зависимости от уровня значимости.

2.1.5. Поправка неточных критериев

Приведем алгоритм поправки, преобразовывающий радикальные и консервативные критерии в точные. Зафиксируем уровень значимости α^*, количество выборок M для оценки $\alpha_I(\alpha)$ и их объем N.

Алгоритм 3. Поправка уровня значимости по зависимости $\alpha_I(\alpha)$ [2]

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 $M \to \infty$.

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

Определение 12. 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. Отметим, что для точного критерия ROC-кривая совпадает с графиком мощности, так как $\alpha_I(\alpha) = \alpha$.

2.1.6. Численное сравнение MC-SSA с другими критериями

Рассотрим другой подход проверки ряда на наличие сигнала. Пусть дан временной ряд $\mathsf{X} = \mathsf{S} + \boldsymbol{\xi}$, где S — сигнал, $\boldsymbol{\xi}$ — красный шум с параметрами φ и δ. Подход заключается в следующем:

1. Провести отбеливание: вычислить $\mathsf{Y} = \Sigma^{-1/2}\mathsf{X}$, где Σ — теоретическая автокорреляционная матрица красного шума с элементами $\varphi^{|i-j|}$, которая при верной H_0 совпадает с автокорреляционной матрицей ряда X.

2. Проверить гипотезу о том, что Y является реализацией белого шума.

В качестве тестов на белый шум был взят Q-тест Бокса-Пирса [12] и тест с использованием вейвлетов [13] (будем далее называть их box и wavelet). Отметим, что второй тест применим только к рядам длины $N = 2^k$, где $k \in \mathbb{N}$. В связи с этим положим $N = 128$, параметры шума возьмем $\varphi = 0.7$, $\delta = 1$. За альтернативу возьмем $\mathsf{S} = \{A\cos(2\pi n\omega)\}_{n=1}^{N}$ и сравним мощность методов с отбеливанием и MC-SSA при помощи ROC-кривых для разных ω.

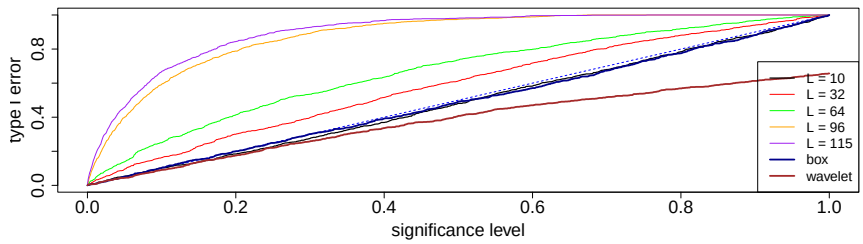
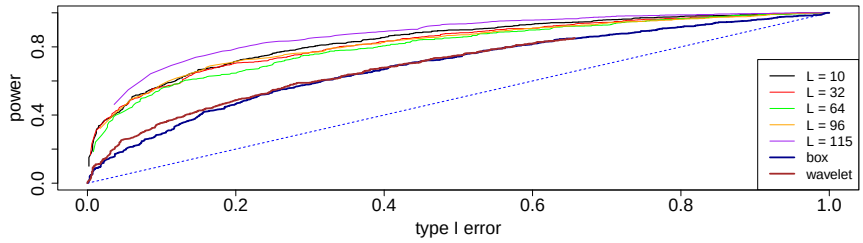
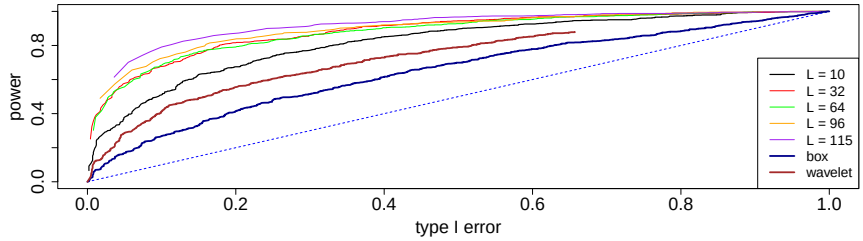


Рис. 2.1. Сравнение ошибки I рода с другими методами

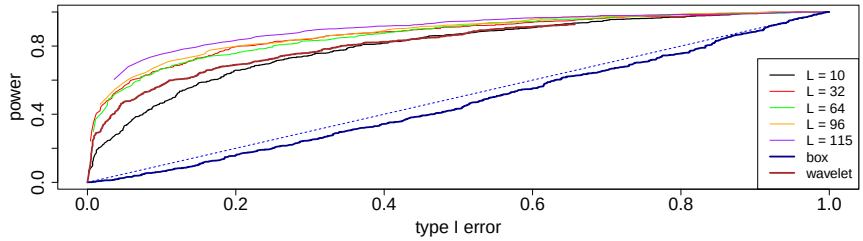
На рис. 2.2, *а*, 2.2, *б*, 2.2, *в* изображены ROC-кривые методов при разных ω. Отметим, что для wavelet построить ROC-кривые для больших ошибок I рода не удалось, поскольку на рис. 2.1 видно, что $\alpha_I(\alpha) < 1\ \forall\alpha$. Для всех рассмотренных ω MC-SSA при всех длинах окна мощнее, чем box и wavelet, кроме случая с высокой частотой ($\omega = 0.225$), где wavelet оказался немного мощнее MC-SSA с $L = 10$. Отметим также,



($а$) ROC-кривая ($\omega = 0.025$)



($б$) ROC-кривая ($\omega = 0.125$)



($в$) ROC-кривая ($\omega = 0.225$)

Рис. 2.2. Сравнение с другими методами

Таблица 2.1. Результаты численного сравнения MC-SSA с другими критериями ($\alpha^* = 0.1$)

Метод	$\alpha_I(\alpha^*)$	$\beta(\widetilde{\alpha}^*)$ ($\omega = 0.025$)	$\beta(\widetilde{\alpha}^*)$ ($\omega = 0.125$)	$\beta(\widetilde{\alpha}^*)$ ($\omega = 0.225$)
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

что на рис. 2.2, *в* ROC-кривая метода box лежит ниже прямой $y = x$, поэтому этот тест не имеет смысл применять для выявления высоких частот. Для удобства сравнения в таблице 2.1 для каждого критерия указана ошибка первого рода и мощность поправленного критерия для каждой рассмотренной альтернативы при уровне значимости $\alpha^* = 0.1$.

2.2. Monte-Carlo MSSA

2.2.1. Отличие от одномерного случая

MC-SSA легко обобщается на многомерный случай: нужно просто заменить SSA на MSSA и генерировать красный шум с тем же количеством каналов, что и у исходного ряда [14].

Стоит отметить, что, в отличие от одномерного случая, левые и правые векторы матрицы отличаются по построению $\mathbf{X}$ (1.3), поэтому в MC-MSSA в качестве векторов для проекции рассмотрены и левые, и правые векторы. Если $W_1, \ldots, W_H$ — левые векторы матрицы $\mathbf{X}$, метод совпадает с алгоритмом 2. Если рассматривать в качестве векторов для проекции правые векторы, то в формулах (2.1) и (2.2) нужно заменить $\mathbf{X}$ на $\mathbf{X}^{\mathrm{T}}$ и $\mathbf{\Xi}_i$ на $\mathbf{\Xi}_i^{\mathrm{T}}$ соответственно.

2.2.2. Численное сравнение модификаций MC-MSSA

Алгоритм 3 поправки радикальных критериев плохо работает для сильно радикальных критериев. Как было показано в [2, Приложение Б.2.4], метод MC-SSA с проекцией на левые (правые) векторы SVD разложения матрицы $\mathbf{X}$ (1.6) дает очень радикальный критерий для больших (малых) значений длины окна L, что делает невозможным построение поправки.

Однако, в одномерном случае было установлено [2], что если вместо SVD разложения матрицы $\mathbf{X}$ использовать тёплицево, то радикальность критерия уменьшается, и уже можно применить поправку. Установим, что будет в многомерном случае, если использовать модификации, описанные в разделе 1.3.

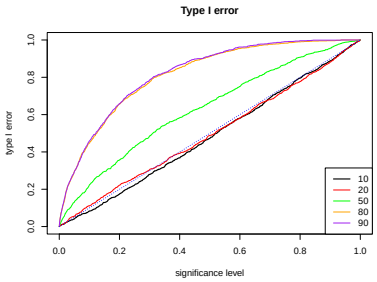
Пусть количество каналов равно двум, количество суррогатных реализаций красного шума $G = 1000$. Для оценки ошибки первого рода, будем рассматривать красный шум $\boldsymbol{\xi}$ с параметрами $\varphi = 0.7$ и $\delta = 1$, а для оценки мощности будет рассматривать

временной ряд $\mathsf{X} = \mathsf{S} + \boldsymbol{\xi}$, где S — сигнал с элементами

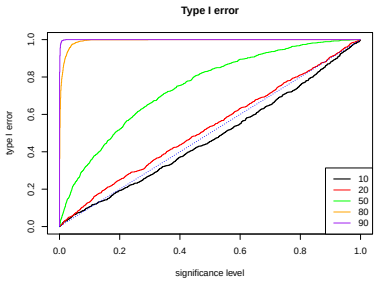
$$s_n^{(1)} = s_n^{(2)} = \cos(2\pi\omega n), \quad n = 1, \ldots, N,$$

где $\omega = 0.075$, $N = 100$.

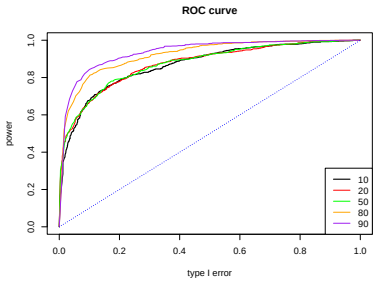
Построим графики ошибки первого рода и ROC-кривые для каждой длины окна $L = 10, 20, 50, 80, 90$. Будем воспринимать ROC-кривую как график мощности критерия, к которому был применен алгоритм 3.



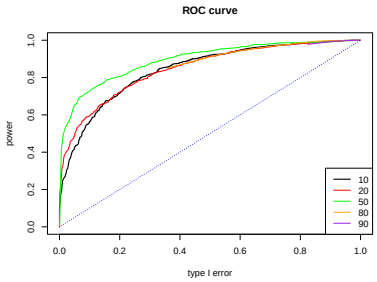
($а$) Ошибка первого рода (Sum).



($б$) Ошибка первого рода (базовый MSSA).



($в$) ROC-кривая (Sum).

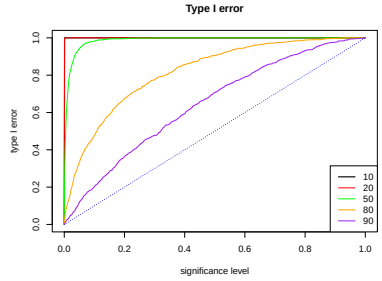


($г$) ROC-кривая (базовый MSSA).

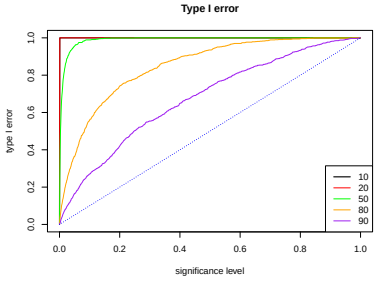
Рис. 2.3. Сравнение методов Sum и базового MSSA (проекция на левые векторы).

На рис. 2.3 и 2.4 векторы для проекции были взяты из разложения (1.8). На рис. 2.3, $а$ видно, что при $L > 20$ метод радикальный, а наибольшая мощность достигается при $L = 90$. На рис. 2.4, $а$ отчетливо заметно, что метод радикальный для всех L. Наибольшая мощность наблюдается при $L = 90$, но отметим, что из-за слишком большой ошибки первого рода построить ROC-кривую на промежутке [0,3) для $L = 50$ и на всем промежутке для $L = 10$ и $L = 20$ не получилось.

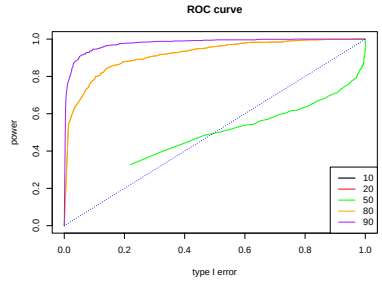
На рис. 2.5 и 2.6 векторы для проекции были взяты из разложения (1.7). Если рассматривать проекцию на левые векторы, то на рис. 2.5, $а$ видно, что метод радикальный, а наибольшая мощность достигается при $L = 20$. Проекция на правые векторы также



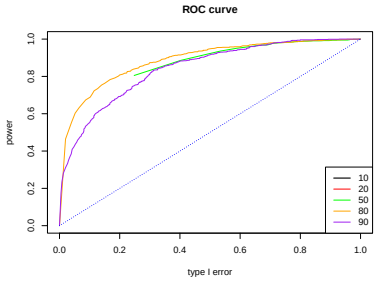
(*а*) Ошибка первого рода (Sum).



(*б*) Ошибка первого рода (базовый MSSA).

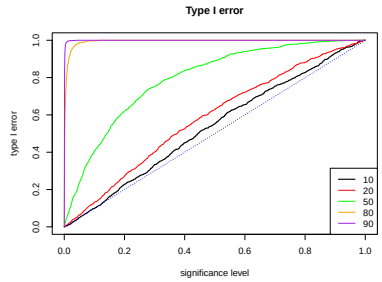


(*в*) ROC-кривая (Sum).

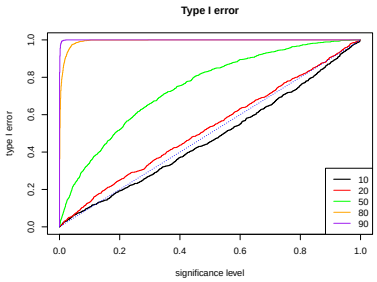


(*г*) ROC-кривая (базовый MSSA).

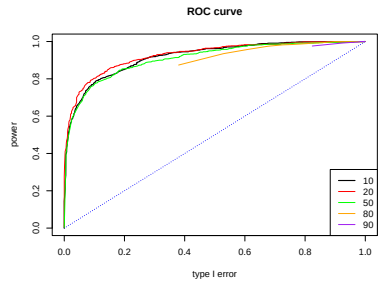
Рис. 2.4. Сравнение методов Sum и базового MSSA (проекция на правые векторы).



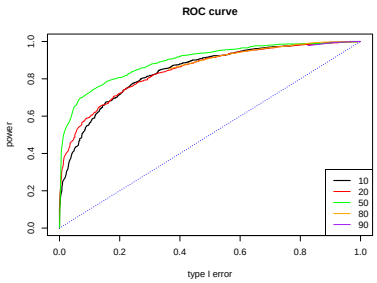
(*а*) Ошибка первого рода (Block).



(*б*) Ошибка первого рода (базовый MSSA).

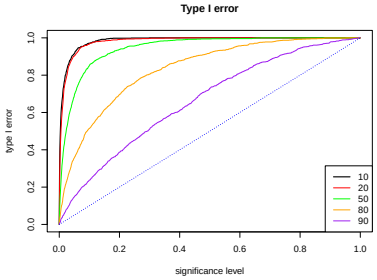


(*в*) ROC-кривая (Block).

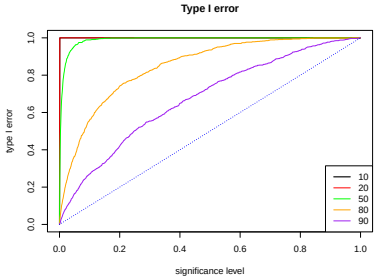


(*г*) ROC-кривая (базовый MSSA).

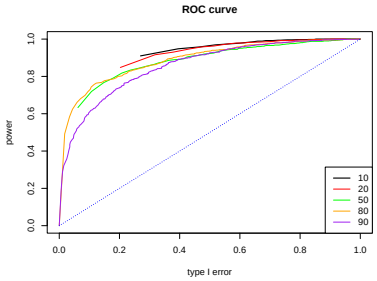
Рис. 2.5. Сравнение методов Block и базового MSSA (проекция на левые векторы).



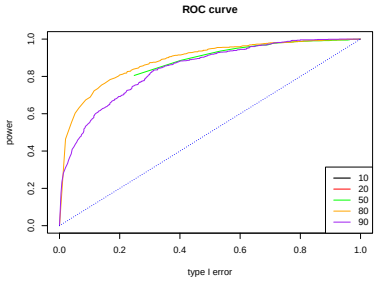
($а$) Ошибка первого рода (Block).



($б$) Ошибка первого рода (базовый MSSA).



($в$) ROC-кривая (Block).



($г$) ROC-кривая (базовый MSSA).

Рис. 2.6. Сравнение методов Block и базового MSSA (проекция на правые векторы).

дает радикальный критерий, как видно на рис. 2.6, $а$. Наибольшая мощность наблюдается при $L = 80$, но из-за слишком большой ошибки первого рода ROC-кривую для $L = 10$ и $L = 20$, для которых метод, предположительно, имеет бóльшую мощность, удалось построить не на всем промежутке.

Таблица 2.2. Результаты численного сравнения методов для оптимальных длин окна

Метод	левые/правые векторы	L	длина векторов	количество векторов	$\alpha_I(\alpha)$ AUC	ROC AUC
MSSA	левые	50*	50	50	0.7455	0.405
MSSA	правые	80*	42	42	0.849	0.3954
Block	левые	20*	162	20	0.5823	0.4326
Block	правые	80*	80	42	0.8328	0.3982
Sum	левые	90	90	22	0.8185	0.4402
Sum	правые	90*	22	22	0.6441	0.4415

В таблице 2.2 для каждого метода указана оптимальная длина окна (для которой удалось простроить ROC-кривую, звездочкой помечены L, которые могут не являться

оптимальными), длина векторов для проекции, их количество, площадь под кривой ошибки первого рода, а также площадь под ROC-кривой при $\alpha_I \in [0, 0.5]$. Видно, что MC-MSSA с проекцией на левые или правые векторы обеих модификаций мощнее, чем с проекцией на векторы базового MSSA.

Выводы

Подведем итоги. Для данного примера для метода Sum оптимальной длиной окна является $L = 90$, если рассматривать проекцию как на левые, так и на правые векторы. Для метода Block оптимальной длиной окна является $L = 20$, если рассматривать проекцию на левые векторы, и $L = 80$, если рассматривать проекцию на правые векторы.

Также все методы, кроме Sum, с проекцией на левые вектора сильно радикальные. Поэтому рекомендуется использовать вариант Sum с проекцией на левые векторы с $L = 90$.

Глава 3

Метод Monte-Carlo SSA для реальных задач

В главе 2 предполагалось, что параметры шума известны и нет мешающего сигнала (например, сезонности или тренда). Мешающий сигнал — это сигнал, который уже не нужно обнаруживать при проверки гипотезы о наличии сигнала. В этой главе рассмотрим случаи, которые более близки к реальным задачам.

В разделе 3.1 исследована зависимость радикальности и мощности MC-SSA от длины окна на различных примерах. В разделе 3.2 рассмотрены два способа оценки параметров красного шума и их численное сравнение в точности оценивания и мощности MC-SSA. В разделе 3.3 приведены два алгоритма модификации MC-SSA с мешающим сигналом и проведено численное исследование методов для разных примеров мешающего сигнала. В разделе 3.4 рассмотрены примеры реальных временных рядов и их анализ с помощью SSA и MC-SSA.

3.1. Зависимость радикальности и мощности от параметра L

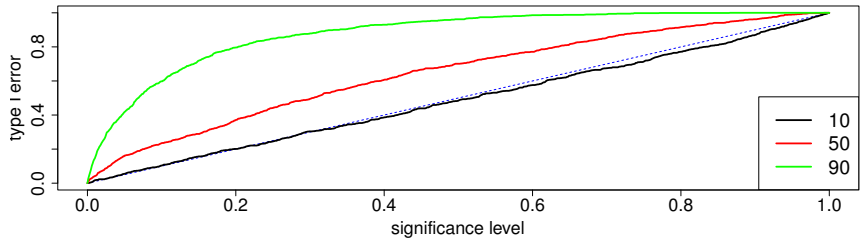
Поскольку рассматриваемый вариант критерия MC-SSA является радикальным, существует проблема выбора такой длины окна L, которая дает максимально мощный критерий, но при этом не слишком радикальный, чтобы можно было применить поправку. Однако, в зависимости от длины ряда N и параметров красного шума $\boldsymbol{\xi}$ наблюдаются разные зависимости мощности от L.

Рассмотрим несколько примеров. Пусть дана модель

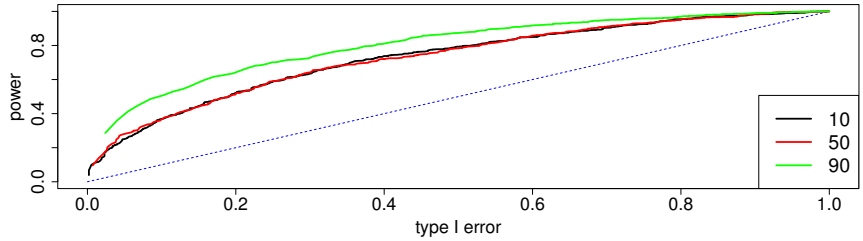
$$\mathsf{X} = \mathsf{S} + \boldsymbol{\xi},$$

где $\mathsf{S} = \{A\cos(2\pi\omega n)\}_{n=1}^{N}$, а $\boldsymbol{\xi}$ — красный шум с параметрами φ и $\delta = 1$. Рассмотрим следующие нулевую гипотезу и альтернативу: $H_0 : A = 0$, $H_1 : A \neq 0$. В этом разделе будем предполагать, что параметры красного шума известны. В первых трех примерах рассмотрим частоту сигнала $\omega = 0.075$.

Пример 1. Пусть $\varphi = 0.7$, $N = 100$. По графику ошибок первого рода на рис. 3.1, *а* видно, что чем больше L, тем более радикальным становится критерий. На рис. 3.1, *б*



(a) Ошибка I рода

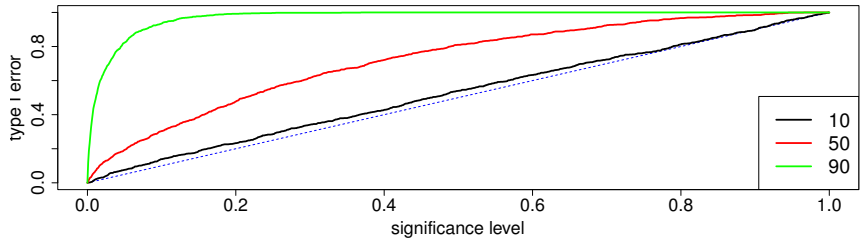


($б$) ROC-кривая ($\omega = 0.075$)

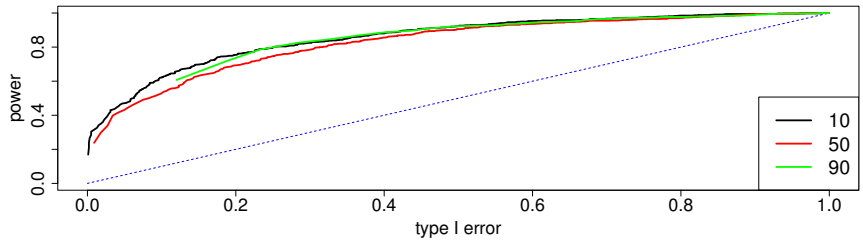
Рис. 3.1. Пример 1 ($\varphi = 0.7$, $N = 100$)

изображены ROC-кривые критериев, наибольшую мощность дает критерий с $L = 90$. На этом примере видно, что самым мощным является самый радикальный критерий.

Пример 2. Пусть $\varphi = 0.3$, $N = 100$. На рис. 3.2, $а$ изображен график ошибок первого



($а$) Ошибка I рода



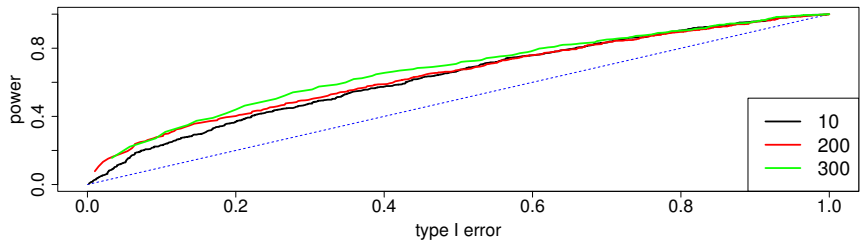
($б$) ROC-кривая ($\omega = 0.075$)

Рис. 3.2. Пример 2 ($\varphi = 0.3$, $N = 100$)

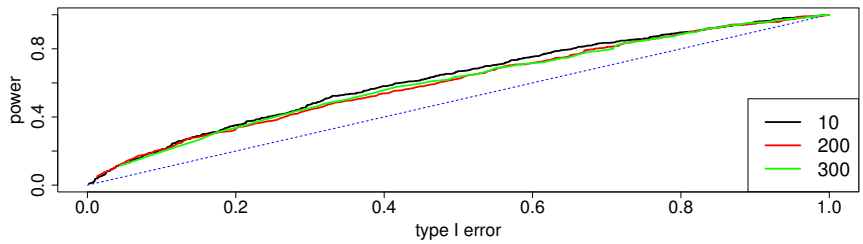
рода. По нему видно, что, как и в примере 1, чем больше L, тем больше радикальность критерия. Если взглянуть на ROC-кривые на рис. 3.2, *б*, то видно, что с уменьшением параметра φ уменьшается разброс мощности критериев после поправки в зависимости от длины окна. Лучшей из рассмотренных в этом случае является $L = 10$, хотя разница с $L = 50$ совсем небольшая, а для $L = 90$ для небольших ошибок I рода поправку сделать не удалось из-за радикальности.

Объяснить такое поведение радикальности в зависимости от длины окна L можно теоретически: с увеличением L увеличивается размер автоковариационной матрицы и тем самым увеличивается количество оцениваемых параметров, что влечет за собой большую подгонку собственных векторов к конкретной реализации шума.

Пример 3. Теперь увеличим длину ряда до $N = 400$ и посмотрим на ROC-кривые для примеров 1, 2. На рис. 3.3, *а*, 3.3, *б* видим, что для обоих примеров с увеличением



(*а*) ROC-кривая

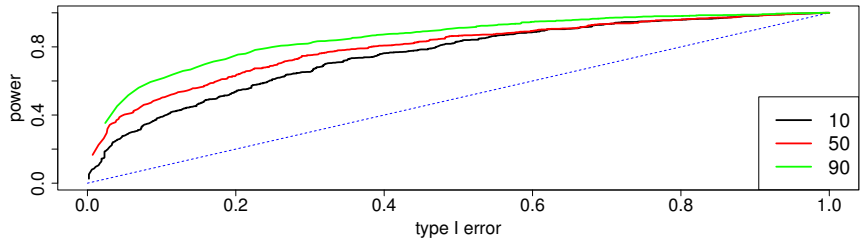


(*б*) ROC-кривая

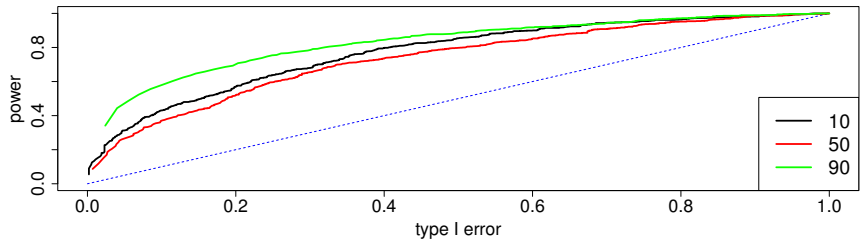
Рис. 3.3. Пример 3 ($\varphi = 0.3$, $N = 400$, $\omega = 0.075$)

длины ряда уменьшается различие в мощности после поправки в зависимости от длины окна.

Пример 4. В условиях примера 1 рассмотрим разные частоты ω сигнала S и зависимость упорядоченности критериев по мощности от L. На рис. 3.4, *б*, 3.4, *а* изображены



(a) ROC-кривая ($\omega = 0.175$)



($б$) ROC-кривая ($\omega = 0.025$)

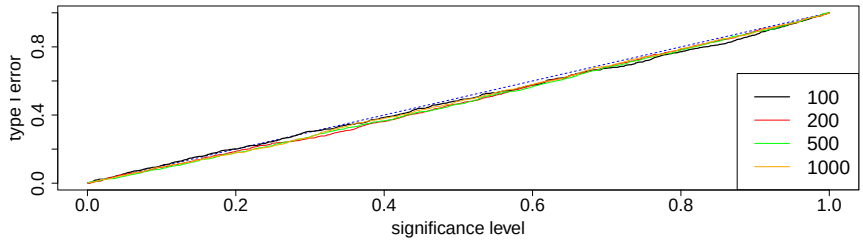
Рис. 3.4. Пример 4 ($\varphi = 0.7$, $N = 100$)

ROC-кривые критериев при разных альтернативах. Видно, что упорядоченность L нарушается при маленьких частотах сигнала. Если упорядочить рис. 3.4, $б$, рис. 3.1, $б$ и рис. 3.4, $а$ по частоте ω, то видна динамика по соотношению ROC-кривых для $L = 10$ и $L = 50$.

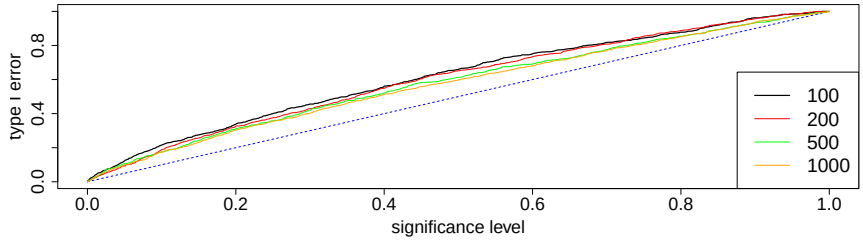
Пример 5. Рассмотрим $N = 100, 200, 500, 1000$ и посмотрим на графики ошибок I рода при разных L. Из рис 3.5, $а$ видно, что критерий с $L = 10$ примерно точный для всех N. На рис. 3.5, $б$ наблюдается очень медленное уменьшение радикальности критерия с ростом N.

Численные эксперименты показали, что длина окна L, дающая максимальную мощность критерия после поправки, зависит от параметров шума, длины ряда и, главное, от частоты сигнала в альтернативной гипотезе. Поэтому при выборе длины окна возможны следующие варианты:

1. При больших N применение поправки, описанной в разделе 2.1.5, является трудоемкой задачей даже для небольших L. Также было замечено, что с ростом N радикальность критерия при фиксированном L едва заметно уменьшается. Поэтому из рассмотренных примеров получено, что без поправки можно использовать MC-SSA только с $L = 10$. Это нетрудозатратно, но возможна некоторая потеря



($а$) Ошибка I рода ($L = 10$)



($б$) Ошибка I рода ($L = 40$)

Рис. 3.5. Пример 5 (поведение ошибки первого рода в зависимости от длины ряда)

в мощности. Но, поскольку с увеличением длины ряда уменьшается различие в мощности в зависимости от длины окна, для больших N эта потеря будет не такая заметная.

2. В поведении оптимальной по мощности длины окна L в зависимости от параметров ряда наблюдается некоторая регулярность. Поэтому можно построить зависимость оптимальной длины окна от параметров ряда с помощью численного моделирования, оценив параметры красного шума. Однако было показано, что упорядоченность критериев по мощности зависит от частоты сигнала в альтернативе, поэтому эта рекомендация имеет практический смысл, только если есть дополнительная информация о диапазоне возможных частот в альтернативе.

3.2. Оценка параметров красного шума

До сих пор мы предполагали, что параметры красного шума φ и δ известны, но в реальных задачах редко возникает такая ситуация. В этой ситуации можно воспользоваться методом bootstrapping, который позволяет использовать оцененные параметры шума для построения критерия [1]. Рассмотрим два способа оценивания неизвестных параметров φ и δ:

1. Оценка параметров на основе исходного ряда методом максимального правдоподобия, где для нахождения начальных значений используется метод CSS [15].

2. Двухступенчатая оценка параметров, предлагаемая в этой работе:

   a) Находится ОМП на основе исходного ряда;

   b) Применяется MC-SSA с поправкой;

   c) Если сигнал обнаруживается, он выделяется;

   d) Находится ОМП на основе «остатка».

Проверим на практике, что двухступенчатая оценка параметров даст результаты лучше, чем обычная без выделения сигнала. За альтернативу возьмем

$$s_n = A\cos(2\pi\omega n), \quad n = 1, \ldots, N, \tag{3.1}$$

с амплитудой $A = 1.5$ и частотой $\omega \in (0, 0.5)$. Пусть $N = 100$, $\varphi = 0.7$, $\delta = 1$. Для двухступенчатой оценки длина окна $\widetilde{L} = 50$, $G = 1000$, $\alpha = 0.1$

Будем оценивать параметр φ, обозначим оценку за $\hat{\varphi}$. В качестве оценки точности было взято среднеквадратичное отклонение от истинного значения. В таблице 3.1 представлены результаты на основе 100 реализаций шума. Поскольку $\mathsf{MSE}\hat{\varphi} = \mathsf{D}\hat{\varphi} + \mathsf{bias}^2\hat{\varphi}$, в таблице также представлены значения дисперсии и смещения оценки.

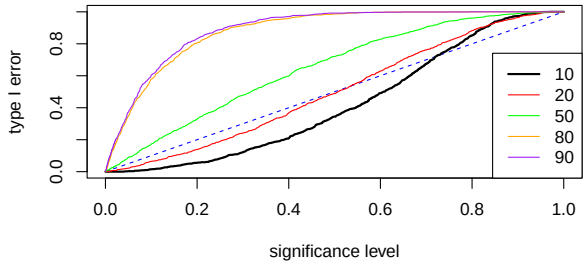
Обычная оценка	$\omega = 0.075$	$\omega = 0.175$	$\omega = 0.275$	$\omega = 0.375$	$\omega = 0.475$
$\mathsf{MSE}\hat{\varphi}$	0.0053	0.0124	0.1185	0.3109	0.4018
$\mathsf{D}\hat{\varphi}$	0.0023	0.0036	0.007	0.0189	0.0204
$\mathsf{bias}\hat{\varphi}$	0.055	−0.0938	−0.3341	−0.5406	−0.6178
Двухступенчатая оценка	$\omega = 0.075$	$\omega = 0.175$	$\omega = 0.275$	$\omega = 0.375$	$\omega = 0.475$
$\mathsf{MSE}\hat{\varphi}$	0.0091	0.0057	0.0038	0.0098	0.0129
$\mathsf{D}\hat{\varphi}$	0.0084	0.0056	0.0037	0.0085	0.0101
$\mathsf{bias}\hat{\varphi}$	−0.0284	−0.0119	−0.0107	−0.0375	−0.0538

Таблица 3.1. Оценка параметров красного шума ($\varphi = 0.7$)

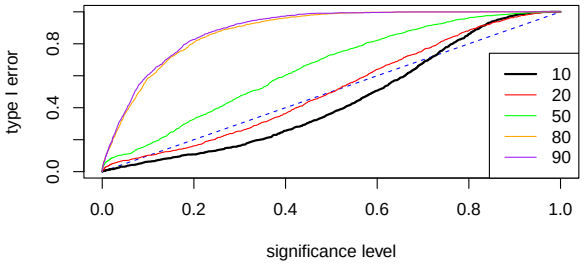
По таблице 3.1 видно, что основной вклад в ошибку оценки вносит смещение, описанная выше процедура это смещение сильно уменьшает, делая слабо-отрицательным.

Замечание 6. В таблице 3.1 были взяты такие ω с целью показать общий случай, поскольку если $\widetilde{L}\omega$ — целое, то SSA точнее выделяет сигнал [11] и, следовательно, сделает оценку параметров лучше.

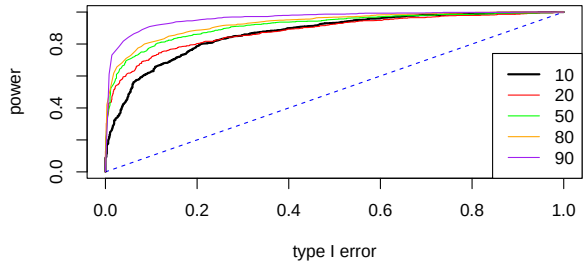
Теперь сравним графики ошибок первого рода и ROC-кривые критерия MC-SSA против альтернативы (3.1) с $\omega = 0.075$. Длины окна L будем брать те же, что и в разделе 2.2.2.



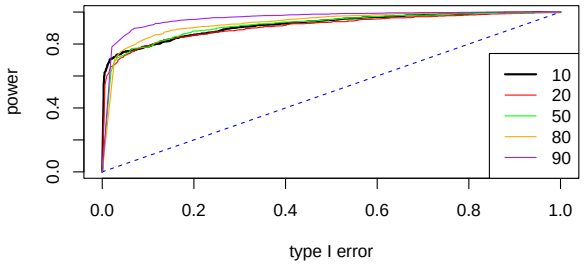
(*а*) Ошибка первого рода (обычная оценка)



(*б*) Ошибка первого рода (двухступенчатая оценка)



(*в*) ROC-кривая (обычная оценка)



(*г*) ROC-кривая (двухступенчатая оценка)

Рис. 3.6. Сравнение обычной и двухступенчатой оценок

По рис. 3.6, *а* и 3.6, *б* видно, графики ошибок первого рода примерно одинаковые для всех длин окна, что естественно, поскольку сигнала нет, и, следовательно, оценки параметров приблизительно одинаковые. А если посмотреть на ROC-кривые на рис. 3.6, *в* и 3.6, *г*, заметно повышение мощности при $L = 10, 20, 50$. Поскольку для оптимальной длины окна $L = 90$ разницы в мощности нет, и для рядов большой длины применение поправки численно трудоемкая задача, далее будем оценивать параметры шума обычным, не двухступенчатым, способом.

3.3. Наличие мешающего сигнала

Пусть известно, что во временном ряде присутствует некоторый сигнал, но, возможно, еще есть какой-то другой. Тогда модель выглядит следующим образом:

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

где F — мешающий сигнал, S — неизвестный сигнал и $\boldsymbol{\xi}$ — красный шум. Тогда проверяется следующая нулевая гипотеза с альтернативой:

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

Алгоритм 4. MC-SSA с мешающим сигналом

1. Находится приближенное значение мешающего сигнала $\hat{\mathsf{F}}$ и оцениваются параметры $\boldsymbol{\xi}$ на основе остатка $\tilde{\mathsf{X}} = \mathsf{X} - \hat{\mathsf{F}}$.

2. Находятся левые векторы $P_1, \ldots, P_L$ траекторной матрицы временного ряда $\tilde{\mathsf{X}}$, полученные из разложения (1.8).

3. Применяется MC-SSA к исходному ряду X с проекцией на векторы $P_1, \ldots, P_L$, при этом суррогатными рядами являются реализации случайной величины $\boldsymbol{\eta}$:

$$\boldsymbol{\eta} = \boldsymbol{\xi} + \hat{\mathsf{F}}.$$

Цель данной модификации алгоритма — устранить влияние мешающего сигнала на вектора, на которые делается проекция.

Рассмотрим два примера мешающего сигнала, для которых будем рассматривать следующую альтернативу:

$$H_1 : \mathsf{S} = \{\cos(2\pi\omega n)\}_{n=1}^{N},$$

где $\omega = 0.075$. Параметры красного шума и длину ряда N возьмем такими же, как в разделе 3.2: $\varphi = 0.7$, $\delta = 1$, $N = 100$.

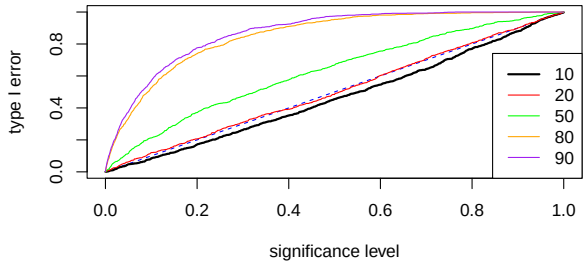
3.3.1. Периодическая компонента

Рассмотрим в качестве мешающего сигнала синусоиду

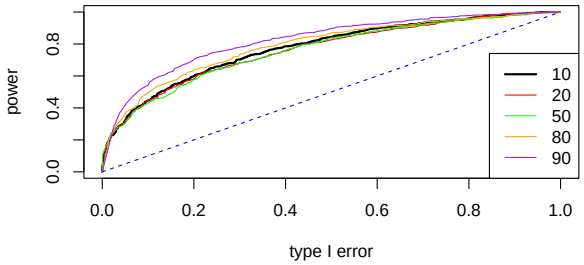
$$f_n = A\cos(2\pi\omega n), \quad n = 1, \ldots, N,$$

с амплитудой $A = 3$ и частотой $\omega = 0.25$.

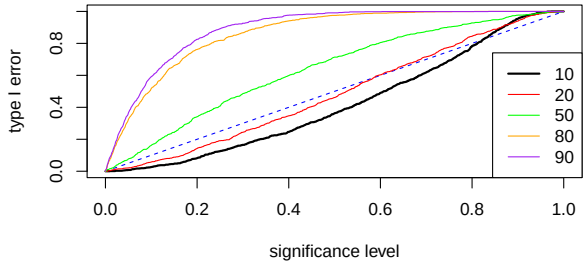
Будем выделять периодическую компоненту при помощи SSA: будем оценивать доминирующую частоту левых векторов с помощью метода ESPRIT [16, Раздел 3.1] и на шаге группировки (раздел 1.2.3) будем брать две компоненты с наиболее близкими к ω частотами.



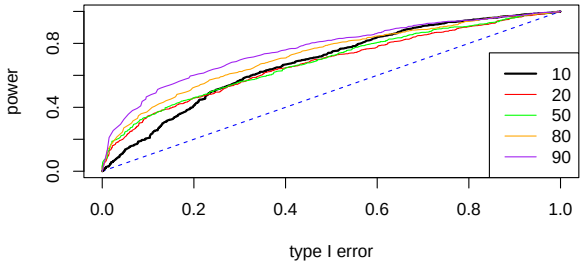
($а$) Ошибка I рода



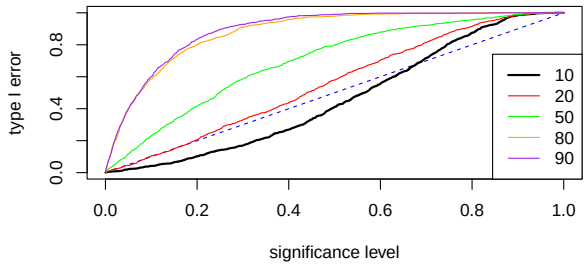
($б$) ROC-кривая



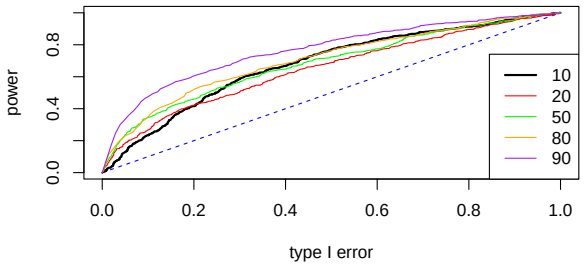
($в$) Ошибка I рода (оцененные параметры шума)



($г$) ROC-кривая (оцененные параметры шума)



($д$) Ошибка I рода (оцененный мешающий сигнал и параметры шума)



($е$) ROC-кривая (оцененный мешающий сигнал и параметры шума)

Рис. 3.7. Анализ метода, когда мешающий сигнал — периодическая компонента

На рис. 3.7 представлены графики ошибок первого рода и ROC-кривые следующих критериев: когда мешающий сигнал и параметры шума известны точно, когда F изве-

стен точно, но параметры шума оцениваются, и когда и мешающий сигнал, и параметры шума оцениваются. Графики ошибок первого рода на рис. 3.7, *а*, 3.7, *в* и 3.7, *д* похожи друг на друга, а отклонение от случая, когда все известно, можно объяснить погрешностью при оценке неизвестных параметров. После применения поправки из раздела 2.1.5 критерии становятся точными для любой длины окна и ROC-кривые на рис. 3.7, *б*, 3.7, *г* и 3.7, *е* представляют собой графики мощности этих критериев. Таким образом, наибольшая мощность во всех трех случаях достигается при $L = 90$.

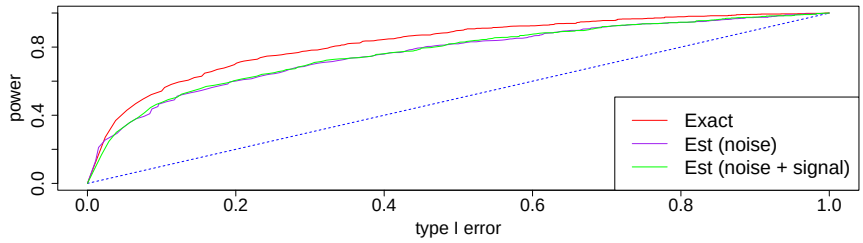


Рис. 3.8. Сравнение ROC-кривых критериев (мешающий сигнал — периодическая компонента)

На рис. 3.8 представлена ROC-кривая критериев для оптимального L. Как видно из графика, при оценке параметров шума и мешающего сигнала мощность падает, но незначительно (примерно на 10%), что важно при применении алгоритма на практике.

3.3.2. Тренд

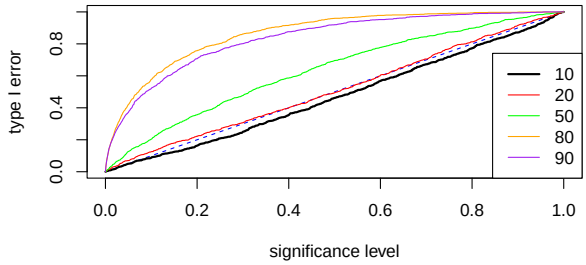
Отдельно рассмотрим вариант, когда мешающий сигнал — тренд, т.е. медленно меняющаяся компонента. Рассмотрим следующий экспоненциальный ряд:

$$f_n = Ae^{\alpha n}, \quad n = 1, \ldots, N,$$

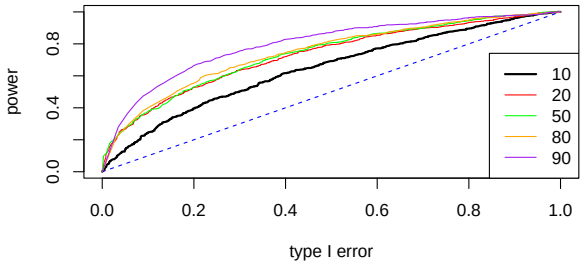
где $A = 0.2$, $\alpha = 0.05$.

Выделять тренд будем с помощью SSA: поскольку в SVD разложении (1.6) сингулярные числа, соответствующие тренду, будут самыми большими среди всех сингулярных чисел, на шаге группировки (раздел 1.2.3) будем брать первые r элементарных компонент, где r — ранг тренда. В данном случае $r = 1$.

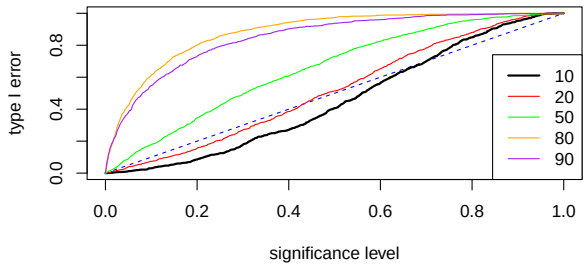
На рис. 3.9 представлены графики ошибок первого рода и ROC-кривые следующих критериев: когда тренд и параметры шума φ и δ известны точно, когда тренд известен точно, но параметры шума оцениваются, и когда и тренд, и параметры шума оцениваются. Как и в разделе 3.3.1, графики ошибок первого рода на рис. 3.9, *а*, 3.9, *в*



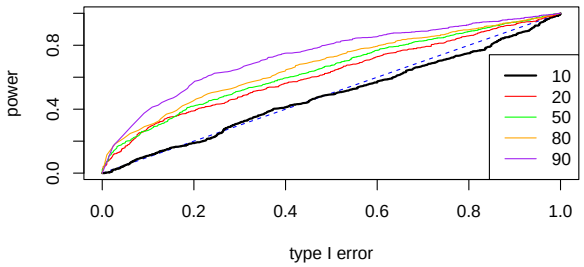
(*а*) Ошибка I рода



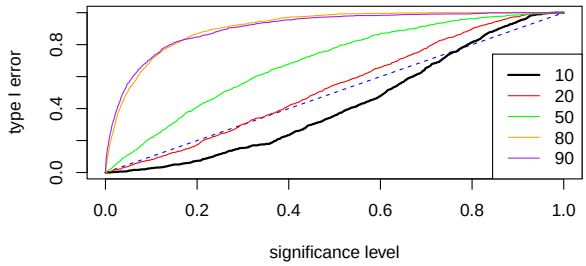
(*б*) ROC-кривая



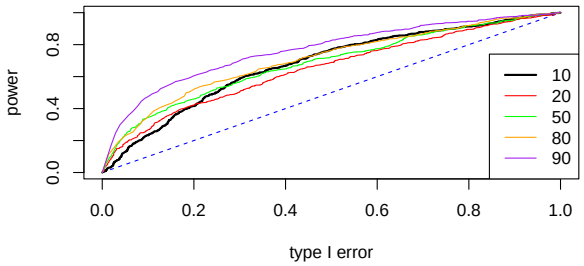
(*в*) Ошибка I рода (оцененные параметры шума)



(*г*) ROC-кривая (оцененные параметры шума)



(*д*) Ошибка I рода (оцененный мешающий сигнал и параметры шума)



(*е*) ROC-кривая (оцененный мешающий сигнал и параметры шума)

Рис. 3.9. Анализ метода, когда мешающий сигнал — тренд

и 3.9, *д* сохраняют общую тенденцию при оценке параметров/тренда. По ROC-кривым на рис. 3.9, *б*, 3.9, *г* и 3.7, *е* видно, что оптимальной длиной окна является $L = 90$. Стоит также отметить странное поведение мощности при $L = 10$.

На рис. 3.10 представлена ROC-кривая критериев для оптимального L. Аналогично случаю периодической компоненты, при оценке параметров шума и мешающего сигнала мощность падает незначительно.

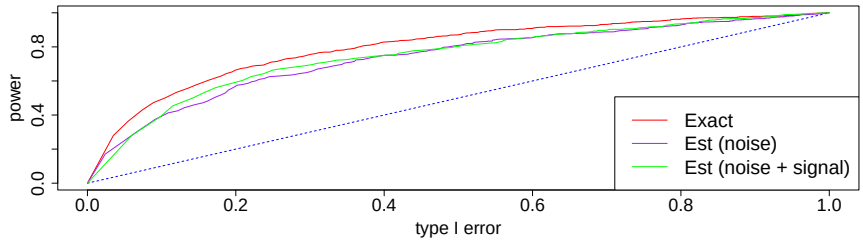


Рис. 3.10. Сравнение ROC-кривых критериев (мешающий сигнал — тренд)

3.3.3. Другой вариант алгоритма

Рассмотрим дургой вариант MC-SSA с мешающим сигналом. Отличается он от алгоритма 4 тем, что рассматривается проекция на левые векторы траекторной матрицы исходного временного ряда, а не его остатка.

Алгоритм 5. MC-SSA с мешающим сигналом (проекция на векторы исходного ряда)

1. Находится приближенное значение мешающего сигнала $\hat{\mathsf{F}}$ и оцениваются параметры $\boldsymbol{\xi}$ на основе остатка $\tilde{\mathsf{X}} = \mathsf{X} - \hat{\mathsf{F}}$.

2. Находятся левые векторы $P_1, \ldots, P_L$ траекторной матрицы временного ряда X, полученные из разложения (1.8).

3. Применяется MC-SSA к исходному ряду X с проекцией на векторы $P_1, \ldots, P_L$, при этом суррогатными рядами являются реализации случайной величины $\boldsymbol{\eta}$:

$$\boldsymbol{\eta} = \boldsymbol{\xi} + \hat{\mathsf{F}}.$$

Отметим, что так как к исходному временному ряду применяется Toeplitz SSA, использовать в качестве мешающего сигнала тренд нельзя. Поэтому в качестве мешающего сигнала рассмотрим синусоиду из раздела 3.3.1 и сравним описанный алгоритм с алгоритмом 4 для $L = 90$. Выделять мешающий сигнал будем также, как и в разделе 3.3.1 с помощью SSA.

В таблице 3.2 представлены результаты сравнения двух алгоритмов, а именно ошибка I рода и мощность поправленного критерия при уровне значимости $\alpha^* = 0.1$. Были рассмотрены варинты как с точной моделью, так и варианты, когда параметры шума и мешающий сигнал оцениваются. По таблице 3.2 видно, что алгоритм 5 немного

Таблица 3.2. Сравнение алгоритма 4 и алгоритма 5 при $\alpha^* = 0.1$

Алгоритм 4	$\alpha_I(\alpha^*)$	$\beta(\widetilde{\alpha}^*)$
Точная модель	0.57	0.542
Оцененные параметры шума	0.593	0.48
Оценные параметры шума и мешающий сигнал	0.6	0.475
Алгоритм 5	$\alpha_I(\alpha^*)$	$\beta(\widetilde{\alpha}^*)$
Точная модель	0.594	0.532
Оцененные параметры шума	0.588	0.468
Оценные параметры шума и мешающий сигнал	0.624	0.521

радикальнее алгоритма 4, но при оценивании параметров шума и мешающего сигнала алгоритм 5 дает значимо бóльшую мощность. Поэтому рекомендуется использовать именно этот алгоритм, когда мешающий сигнал — стационарные колебания с какой-то частотой.

3.4. Применение к реальным временным рядам

Рассмотрим несколько реальных временных рядов и применим к ним MC-SSA.

3.4.1. Niño 3.4

На рис. 3.11 представлена ежемесячная температура поверхности моря в центральной тропической части Тихого океана в период с 1950 по 2024 год (888 месяцев). В данном регионе происходит явление под названием Эль-Ниньо, характеризующееся аномальным потеплением поверхностных вод. Эти колебания температуры оказывают заметное влияние на погодные условия во всем мире, поэтому важно изучить их поведение.

Сразу заметим, что в этом временном ряде присутствует небольшой тренд, поэтому перед применением MC-SSA удалим его. Для начала воспользуемся базовым SSA с длиной окна $L = N/2 = 444$, как рекомендуется в [11]. На рис. 3.12 изображены первые 6 собственных векторов сингулярного разложения. Видно, что первый вектор соответ-

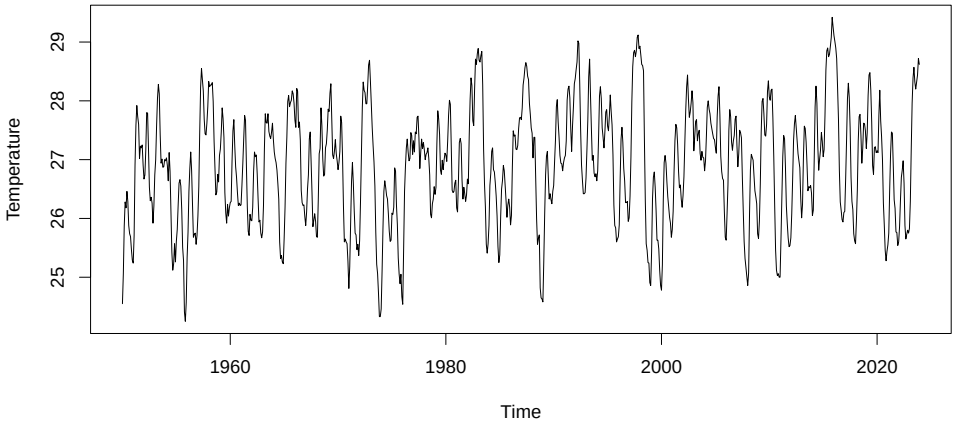


Рис. 3.11. Температура поверхности моря в центральной тропической части Тихого океана

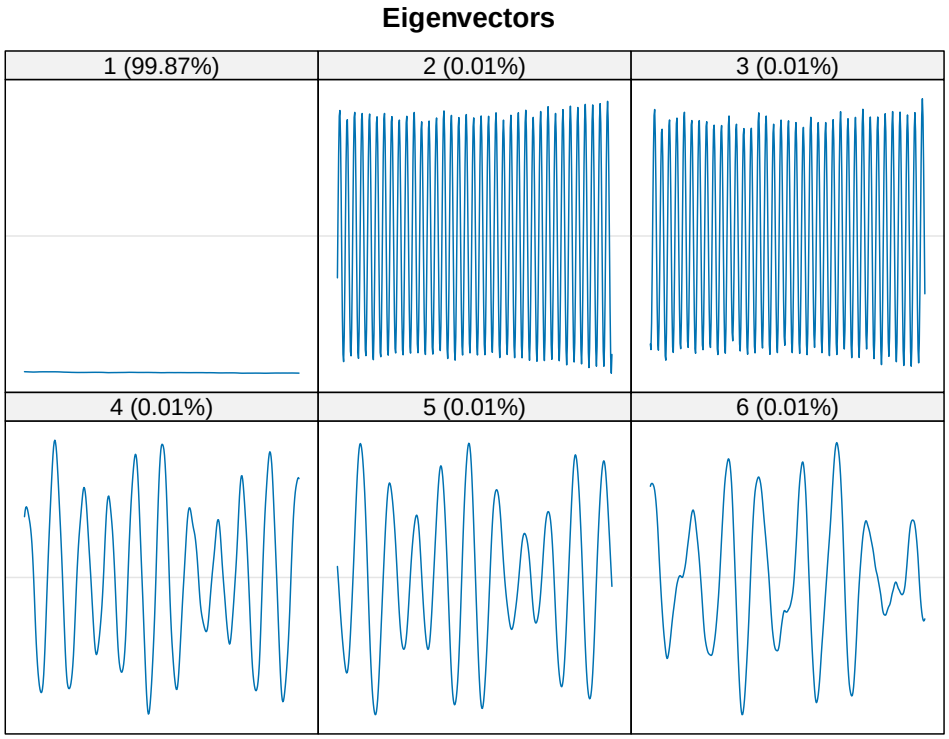


Рис. 3.12. Графики собственных векторов

ствует тренду. Посмотрев на двумерные графики собственных векторов на рис. 3.13, видно, что вторая и третья компоненты образуют двенадцатиугольник. Это означает,

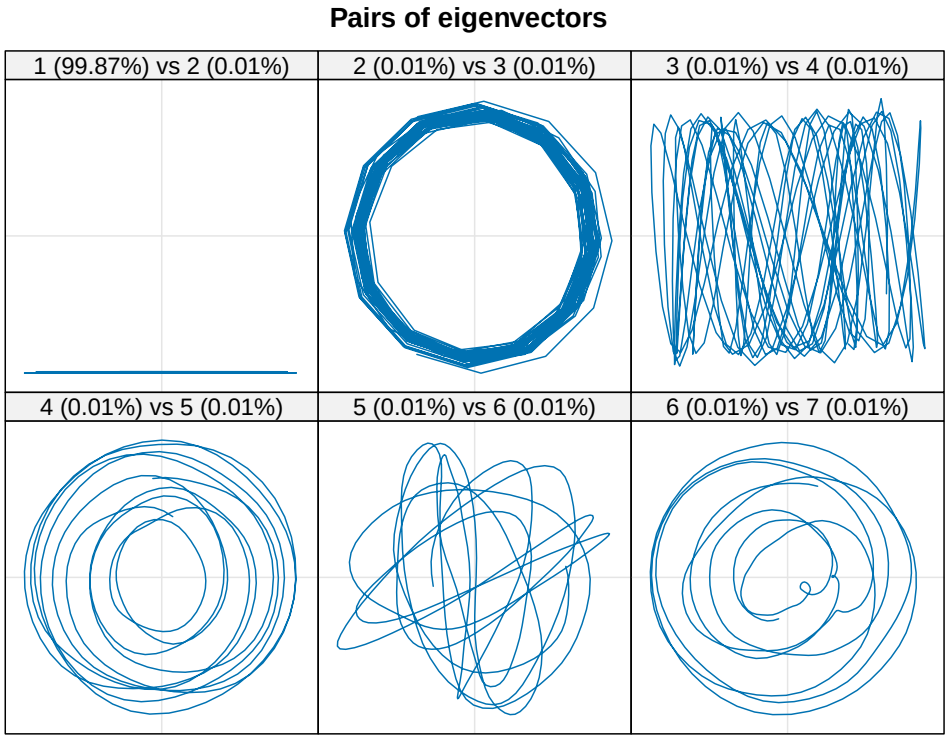


Рис. 3.13. Двумерные графики собственных векторов

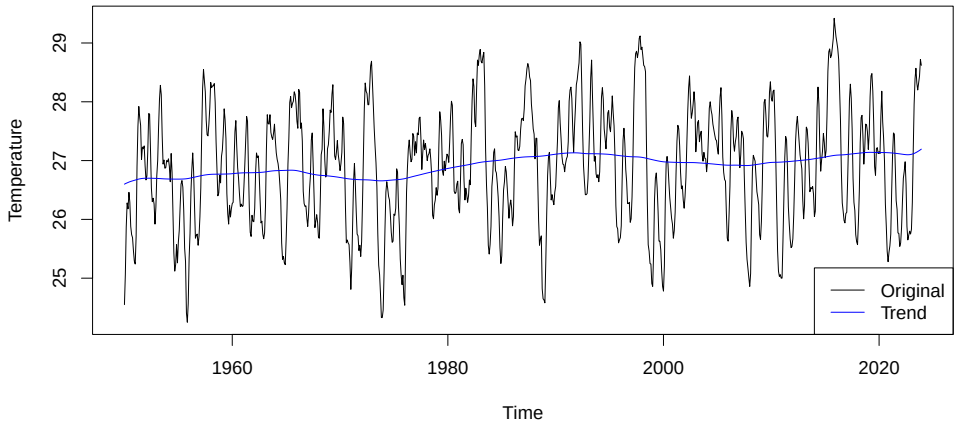


Рис. 3.14. Выделенный тренд

что они соответствуют периодике с периодом 12 [11]. С учетом всей полученной информации, возьмем длину окна для выделения тренда небольшой, но делящейся на период

периодической компоненты для обеспечения разделимости [11]. На рис. 3.14 изображен выделенный тренд при $L = 120$. Для наиболее точного выделения периодической ком-

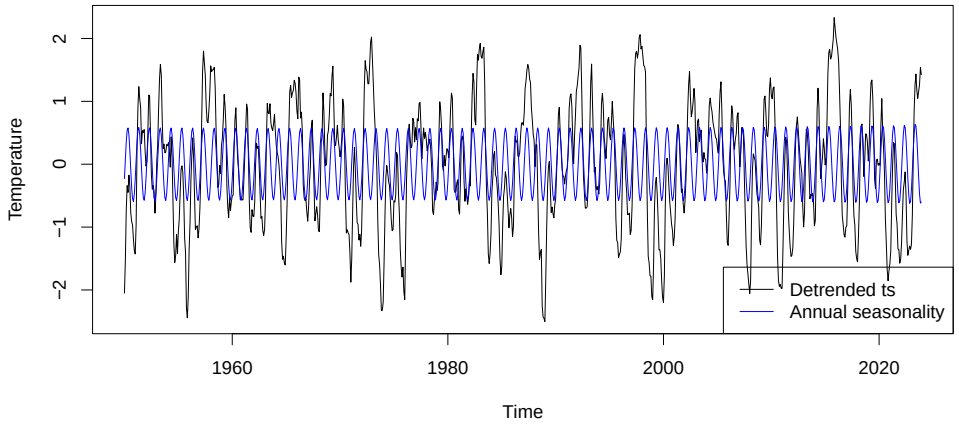


Рис. 3.15. Выделенная годовая сезонность

поненты длина окна должна быть близкой к половине длины ряда и должна делиться на ее период [11], поэтому применим Toeplitz SSA с длиной окна $L = 444$. На рис. 3.15 изображена выделенная годовая сезонность.

Применим поправленный MC-SSA с $L = 40$ к ряду без тренда с годовой периодичностью в качестве мешающего сигнала. Оцененные параметры красного шума следующие: $\varphi = 0.94$ и $\delta = 0.305$. На рис. 3.16 изображены 95%-ные доверительные интервалы статистики $\hat{p}_k$, $k = 1, \ldots, L$ (2.1). Значимыми являются четыре компоненты, две компоненты, имеющие период приблизительно 6, легко интерпретируются — это замеченная полугодовая сезонность. С помощью Toeplitz SSA с той же длиной окна эта сезонность была выделенна, ее вид изображен на рис 3.17. Оставшиеся значимые компоненты имеют периоды 18.16 и 16.5, которые довольно сложно интерпретировать.

Значимые векторы, интерпретация которых не представляется возможной, нельзя относить к сигналу, поскольку MC-SSA проверяет гипотезу о том, что временной ряд представляет собой реализацию красного шума, то есть возможно модель является неверной. Предположим, что рассматриваемый ряд без тренда и годовой периодичности является моделью ARMA(p, q), где p — порядок авторегрессии, q — порядок скользящего среднего. Тогда наиболее подходящей моделью ARMA [17] является ARMA$(1, 2)$. Отметим, что красному шуму соответствует модель ARMA$(1, 0)$. Промоделировав ряд в соответствии с полученной моделью и посмотрев на доверительные интервалы стати-

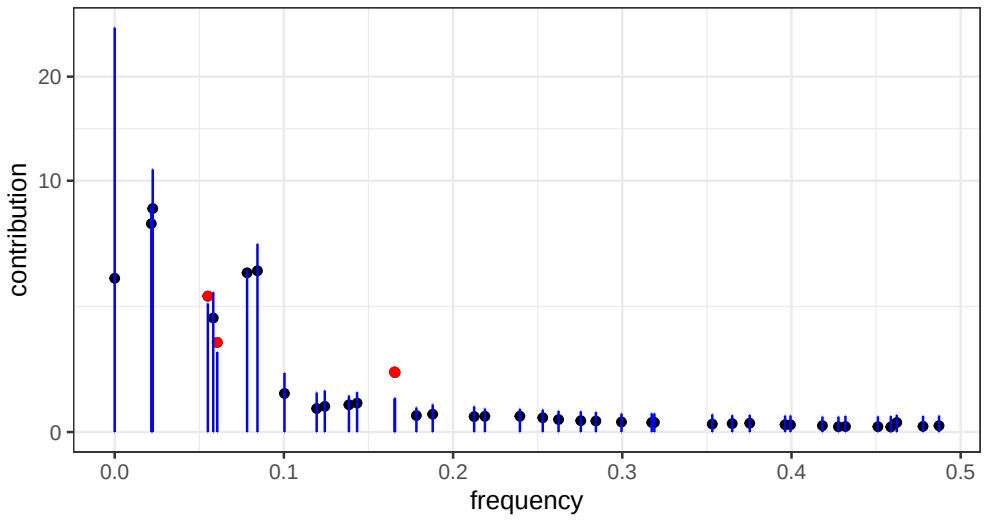


Рис. 3.16. Результат работы MC-SSA

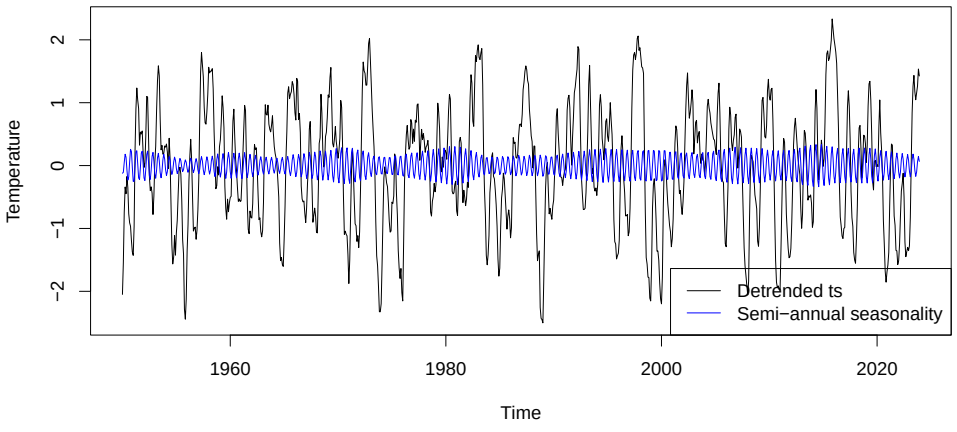


Рис. 3.17. Выделенная полугодовая сезонность

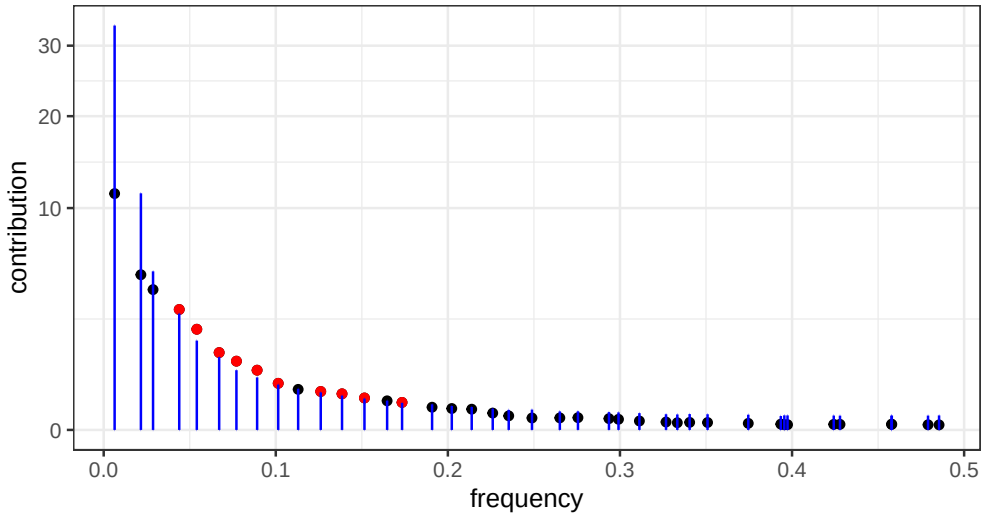


Рис. 3.18. Результат работы MC-SSA для модели ARMA$(1, 2)$

стик $\hat{p}_k$ на рис. 3.18, получаем много значимых компонент. Таким образом, неправильно выбранная модель может исказить выводы, полученные в результате применения MC-SSA, поэтому важно внимательно относиться к выбору модели при проверке гипотезы.

Заключение

В ходе данной работы для реализации двух методов Тёплицева MSSA был использован язык программирования R. Было получено, что в точности восстановления сигнала оба метода в большинстве случаев показывают лучший результат, чем обычный MSSA. Но в Monte-Carlo SSA метод Sum более предпочтителен, чем метод Block, что важно ввиду его простоты в реализации и структуры, подходящей под пакет Rssa [3].

Также было рассмотрено два способа оценки неизвестных параметров красного шума: обычным bootstrap'ом и двухступенчатой оценкой, описанной в данной работе. Было получено, что двухступенчатая оценка точнее оценивает истинные параметры красного шума, если, помимо красного шума, во временном ряде присутствует сигнал.

Еще были разобраны два примера Monte-Carlo SSA с мешающим сигналом. Были рассмотрены случаи, когда мешающий сигнал и параметры красного шума известны, когда оценивались только параметры красного шума, и когда оценивались и мешающий сигнал, и параметры красного шума.

В дальнейшем предполагается расширить набор примеров мешающих сигналов и выработать более общие рекомендации, а также рассмотрение алгоритмов в случае многомерных временных рядов.

Список литературы

1. Golyandina N. Detection of signals by Monte Carlo singular spectrum analysis: multiple testing // Statistics and Its Interface. — 2023. — Vol. 16, no. 1. — P. 147–157.
2. Ларин Е. С. Метод SSA для проверки гипотезы о существовании сигнала во временном ряде : квалификационная работа магистра ; СПбГУ. — 2022.
3. Rssa: A Collection of Methods for Singular Spectrum Analysis. — R package version 1.0.5. Access mode: `https://CRAN.R-project.org/package=Rssa`.
4. Allen Myles, Smith Leonard. Monte Carlo SSA: Detecting irregular oscillations in the Presence of Colored Noise // Journal of Climate. — 1996. — Vol. 9. — P. 3373–3404.
5. Spectral characteristics and predictability of the NAO assessed through Singular Spectral Analysis / Gámiz-Fortis S., Pozo-Vazquez D., Esteban-Parra Maria-Jesus, and Castro-Díez Yolanda // Journal of Geophysical Research. — 2002. — Vol. 107.
6. Paluš M., Novotná D. Enhanced Monte Carlo Singular System Analysis and detection of period 7.8 years oscillatory modes in the monthly NAO index and temperature records // Nonlinear Processes in Geophysics. — 2004. — Vol. 11, no. 5/6. — P. 721–729.
7. Jemwa Gorden, Aldrich Chris. Classification of process dynamics with Monte Carlo singular spectrum analysis // Computers & Chemical Engineering. — 2006. — Vol. 30. — P. 816–831.
8. Xu Chang, Yue Dongjie. Monte Carlo SSA to detect time-variable seasonal oscillations from GPS-derived site position time series // Tectonophysics. — 2015. — Vol. 665. — P. 118–126.
9. Multivariate and 2D Extensions of Singular Spectrum Analysis with the Rssa Package / Golyandina Nina, Korobeynikov Anton, Shlemov Alex, and Usevich Konstantin // Journal of Statistical Software. — 2015. — Vol. 67, no. 2.
10. Plaut Guy, Vautard Robert. Spells of Low-Frequency Oscillations and Weather Regimes in the Northern Hemisphere. // Journal of the Atmospheric Sciences. — 1994. — Vol. 51. — P. 210–236.
11. Голяндина Н. Э. Метод «Гусеница»-SSA: анализ временных рядов. — СПбГУ, 2004. — Учебное пособие.
12. Box G. E. P., Pierce David A. Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models // Journal of the Amer-

ican Statistical Association. — 1970. — Vol. 65, no. 332. — P. 1509–1526.

13. Nason Guy, Savchev Delyan. White noise testing using wavelets // Stat. — 2014. — Vol. 3.
14. Groth Andreas, Ghil Michael. Monte Carlo Singular Spectrum Analysis (SSA) Revisited: Detecting Oscillator Clusters in Multivariate Datasets // Journal of Climate. — 2015. — Vol. 28. — P. 7873–7893.
15. Gardner G., Harvey A., Phillips G. An Algorithm for Exact Maximum Likelihood Estimation of Autoregressive-Moving Average Models by Means of Kalman Filtering // Applied Statistics. — 1980. — Vol. 29. — P. 311–322.
16. Golyandina Nina, Korobeynikov Anton, Zhigljavsky Anatoly. Singular Spectrum Analysis with R. — 2018. — ISBN: 978-3-662-57378-5.
17. Hyndman Rob J., Khandakar Yeasmin. Automatic Time Series Forecasting: The forecast Package for R // Journal of Statistical Software. — 2008. — Vol. 27, no. 3. — P. 1–22.