

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

<b>Введение</b>	4
<b>Глава 1. Метод MSSA и его модификации</b>	5
1.1. Вспомогательные определения	5
1.2. Метод MSSA	6
1.2.1. Вложение	6
1.2.2. Разложение	6
1.2.3. Группировка	7
1.2.4. Диагональное усреднение	7
1.3. Этап разложения	7
1.3.1. Basic MSSA	7
1.3.2. Toeplitz Block MSSA	8
1.3.3. Toeplitz Sum MSSA	9
1.4. Сравнение методов MSSA	9
1.4.1. Теоретическое сравнение методов	9
1.4.2. Численное сравнение методов	9
<b>Глава 2. Метод Monte-Carlo (M)SSA</b>	12
2.1. Проверка статистических гипотез	12
2.1.1. Проблема критериев с неизвестным распределением статистики	13
2.1.2. Поправка неточных критериев	14
2.2. Monte Carlo SSA	15
2.2.1. Постановка задачи	15
2.2.2. Одиночный тест	15
2.2.3. Множественный тест	16
2.2.4. Выбор векторов для проекции	17
2.2.5. Численное сравнение MC-SSA с другими критериями	18
2.3. Monte-Carlo MSSA	20
2.3.1. Отличие от одномерного случая	20
2.3.2. Численное сравнение модификаций MC-MSSA	20

<b>Глава 3. Метод Monte-Carlo SSA для реальных задач . . . . .</b>	<b>26</b>
3.1. Зависимость радикальности и мощности от параметра $L$ . . . . .	26
3.2. Оценка параметров красного шума . . . . .	30
3.2.1. Реакция оценок на наличие сигнала . . . . .	31
3.2.2. Искажение критерия при использовании оценок . . . . .	31
3.3. Наличие мешающего сигнала . . . . .	32
3.3.1. Периодическая компонента . . . . .	34
3.3.2. Тренд . . . . .	34
3.3.3. Другой вариант алгоритма . . . . .	37
3.4. Применение к реальным временным рядам . . . . .	38
3.4.1. Niño 3.4 . . . . .	38
<b>Заключение . . . . .</b>	<b>44</b>
<b>Список литературы . . . . .</b>	<b>45</b>
<b>Приложение А. . . . .</b>	<b>47</b>
A.1. $\varphi = 0.7, \omega = 0.075$ . . . . .	47
A.2. $\varphi = 0.3, \omega = 0.075$ . . . . .	48
A.3. $\varphi = 0.7, \omega = 0.225$ . . . . .	49
<b>Приложение Б. . . . .</b>	<b>52</b>
Б.1. $\varphi = 0.3, \omega = 0.075$ . . . . .	52
Б.2. $\varphi = 0.7, \omega = 0.225$ . . . . .	52

## 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 с 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  $\mathbf{X}$  — одномерный временной ряд длины  $N$ . Выберем параметр  $L$ , называемый *длиной окна*,  $1 < L < N$ . Рассмотрим  $K = N - L + 1$  векторов вложения  $X_i = (x_i, \dots, x_{i+L-1})^T$ ,  $1 \leq i \leq K$ . Определим оператор вложения  $\mathcal{T}$  следующим образом:

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

**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  $\mathcal{H}$ :

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

где  $s = i+j-1$ ,  $A_s = \{(l, k) : l+k = s+1, 1 \leq l \leq L, 1 \leq k \leq K\}$  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  $\xi = (\xi_1, \dots, \xi_n, \dots)$  называется стационарным, если  $\forall k \geq 1 \ E\xi_k = \text{const}$  и  $\forall k, l \geq 1$

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

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

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

$$R_x^{(N)}(k, l) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{m=1}^N x_{k+m} x_{l+m} \xrightarrow[N \rightarrow \infty]{} R_x(k - l).$$

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

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

## 1.2. Метод MSSA

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

### 1.2.1. Вложение

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

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

### 1.2.2. Разложение

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

**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  $I = \{1, \dots, p\}$  разбивается на  $m$  непересекающихся множеств  $I_1, \dots, I_m$  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  $\mathbf{Y} = [\mathbf{Y}^{(1)} : \dots : \mathbf{Y}^{(M)}]$  — некоторая составная матрица, тогда оператор ганкелизации для составной матрицы

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

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

$$\tilde{\mathbf{X}}_{I_k} = \mathcal{T}_{\text{MSSA}}^{-1} \circ \mathcal{H}_{\text{stacked}}(\mathbf{X}_{I_k}), \quad (1.5)$$

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

**Замечание 1.** При  $D = 1$   $\mathbf{X}$  — 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  $\mathbf{S} = \mathbf{X}\mathbf{X}^T$ . Пусть  $\lambda_i$  — собственные числа, а  $U_i$  — ортонормированная система векторов матрицы  $\mathbf{S}$ . Упорядочим  $\lambda_i$  по убыванию и найдем  $p$  такое, что  $\lambda_p > 0$ , а  $\lambda_{p+1} = 0$ . Тогда

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

где  $V_i = \mathbf{X}^T U_i / \sqrt{\lambda_i}$ . 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  $\mathbf{X}\mathbf{X}^T$ , а правые  $V_i$  в свою очередь — матрицы  $\mathbf{X}^T\mathbf{X}$ . 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  $\mathbf{X}\mathbf{X}^T = \mathbf{X}^{(1)}(\mathbf{X}^{(1)})^T + \mathbf{X}^{(2)}(\mathbf{X}^{(2)})^T$ , а

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

### 1.3.2. Toeplitz Block MSSA

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

**Определение 10.** Пусть  $\mathbf{X} = \{\mathbf{X}^{(d)}\}_{d=1}^D$  —  $D$ -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

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

**Замечание 3.** Если ряд  $\mathbf{X}$  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  $\mathbf{X}^T\mathbf{X}$  матрицу

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

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

$$\mathbf{X} = \sum_{i=1}^{DK} \sigma_i P_i Q_i^T = \mathbf{X}_1 + \dots + \mathbf{X}_{DK}, \quad (1.7)$$

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  $\mathbf{X}\mathbf{X}^T$  матрицу  $\mathbf{T}_{\text{Sum}} = \sum_{d=1}^D \mathbf{T}_{d,d}^{(L)} \in \mathbb{R}^{L \times L}$ . Найдем ортонормированные собственные векторы  $P_1, \dots, P_L$  матрицы  $\mathbf{T}_{\text{Sum}}$  и разложим траекторную матрицу  $\mathbf{X}$  следующим образом:

$$\mathbf{X} = \sum_{i=1}^L \sigma_i P_i Q_i^T = \mathbf{X}_1 + \dots + \mathbf{X}_L, \quad (1.8)$$

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

## 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  $\mathcal{O}(D^3 K^3)$  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

$N = 71$ :

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

где  $S^{(1)}, S^{(2)}$  — некоторые сигналы, а  $N^{(1)}, N^{(2)}$  — 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:

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

2. Косинусы с разными частотами:

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

3. Полиномы первой степени (нестационарные ряды):

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

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	<b>2.01</b>	<b>2.00</b>	<b>2.01</b>	3.25
Toeplitz SSA	3.2	1.87	1.63	<b>1.59</b>	1.67
MSSA	3.18	1.83	1.59	<b>1.47</b>	2.00
Toeplitz Sum MSSA	3.17	1.75	1.44	<b>1.32</b>	<b>1.33</b>
Toeplitz Block MSSA	1.39	<b>1.26</b>	<b>1.25</b>	1.33	1.97
Случай 2 ( $\omega_1 \neq \omega_2$ )	$L = 12$	$L = 24$	$L = 36$	$L = 48$	$L = 60$
SSA	3.25	<b>2.01</b>	<b>2.00</b>	<b>2.01</b>	3.25
Toeplitz SSA	3.2	1.87	1.63	<b>1.59</b>	1.67
MSSA	6.91	3.77	3.07	<b>2.88</b>	3.84
Toeplitz Sum MSSA	6.88	3.65	2.64	2.37	<b>2.27</b>
Toeplitz Block MSSA	4.47	3.67	<b>3.22</b>	<b>3.23</b>	3.8
Случай 3 (тренд)	$L = 12$	$L = 24$	$L = 36$	$L = 48$	$L = 60$
SSA	3.65	2.08	<b>1.96</b>	2.08	3.65
Toeplitz SSA	3.33	<b>2.43</b>	3.74	7.84	16.29
MSSA	3.42	1.94	1.63	<b>1.57</b>	2.27
Toeplitz Sum MSSA	3.32	<b>2.24</b>	3.04	5.91	11.95
Toeplitz Block MSSA	12.55	6.18	2.97	<b>1.78</b>	1.97

## Глава 2

## Метод Monte-Carlo (M)SSA

Эта глава делится на две части. В разделе 2.2 рассматривается метод Monte-Carlo SSA (сокращенно MC-SSA), а в разделе 2.3 его многомерное обобщение (MC-MSSA). В названии методов присутствует (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 2.2.2 и 2.2.3 приведены алгоритмы метода MC-SSA и его модификации с поправкой на множественные тестирования [1]. В разделе 2.2.4 рассмотрен выбор параметров метода, который используется в дальнейших численных исследованиях. Известное решение проблемы радикальности критерия при таком выборе параметров приведено в разделе 2.1.2. В разделе 2.2.5 проведено численное сравнение MC-SSA с другими статистическими критериями. Завершает эту главу раздел ??, где происходит численное сравнение модификаций метода MC-MSSA на нескольких примерах.

## 2.1. Проверка статистических гипотез

Рассмотрим правосторонний критерий со статистикой  $T$ . Введем некоторые обозначения.

**Определение 11.**  $p$ -value — такое значение  $p$ , что при значениях уровня значимости  $\alpha$ , больших  $p$ ,  $H_0$  отвергается (статистика попадает в критическую область  $A_{\text{крит}}(\alpha)$ ), а при меньших — не отвергается.

**Определение 12.** Ошибка первого рода — вероятность отвергнуть нулевую гипотезу, если она верна:  $\alpha_I(\alpha) = P_{H_0}(T \in A_{\text{крит}}(\alpha))$ .

**Определение 13.** Мощность критерия — вероятность отвергнуть нулевую гипотезу, если верна альтернативная:  $\beta(\alpha) = P_{H_1}(T \in A_{\text{крит}}(\alpha))$ .

**Замечание 4.** По определению  $p$ -value  $\alpha_I(\alpha) = P_{H_0}(p < \alpha)$  и  $\beta(\alpha) = P_{H_1}(p < \alpha)$ .

**Определение 14.** ROC-кривая — это кривая, задаваемая параметрически

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

**Замечание 5.** 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  $\alpha_I(\alpha) = \alpha$ .

### 2.1.1. Проблема критериев с неизвестным распределением статистики

Если распределение  $T$  при верной  $H_0$  неизвестно, оно оценивается с помощью  $G$  суррогатных выборок.

**Алгоритм 1.** Критерий, использующий суррогатные выборки

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

**Замечание 6.** Пусть  $T$  имеет функцию распределения  $F$  при верной  $H_0$ . Тогда p-value вычисляется как  $p = 1 - F(t)$ . Если распределение  $T$  неизвестно, вместо  $F$  используют ее оценку

$$\hat{F}_G(x) = \frac{1}{G} \sum_{i=1}^G \mathbb{1}_{t_i \leq x},$$

и  $\hat{p}^{(G)} = 1 - \hat{F}_G(t)$  — оценка  $p$ .

**Предположение 1.** Обозначим за  $\delta(\alpha, G) = \mathbb{E}\hat{t}_\alpha - t_\alpha$  смещение оценки  $\hat{t}_\alpha$ . Тогда  $\delta(\alpha, G) < 0$  и  $|\delta(\alpha, G)|$  монотонно убывает по  $\alpha$  и  $G$ .

*Обоснование.* Проверим это предположение с помощью моделирования: оценим 0.999-квантиль стандартного нормального распределения с помощью выборки размера 100. Тогда истинное значение квантиля равно 3.09, а в среднем оценка равна 2.482. Увеличив размер выборки до 1000 получаем уже более точную оценку: в среднем 2.944.

Общие соображения: если  $\alpha = 0$ , то  $t_\alpha = +\infty$ , а ее оценка  $\hat{t}_\alpha$  конечна и  $\delta(0, G) = -\infty$ .  $\square$

**Замечание 7.** Из предположения 1 следует, что критерий, описанный в алгоритме 1 радикальнее исходного критерия: поскольку  $\hat{t}_\alpha < t_\alpha$ , значения  $t$  будут чаще попадать в критическую область.

**Предположение 2.** При таких  $x$ , что  $F(x) > 0.5$ ,  $\text{med } \hat{F}_G(x) > F(x)$ .

*Обоснование.*  $G\hat{F}_G(x)$  имеет распределение  $\text{Bin}(F(x), G)$ . Если  $F(x) > 0.5$ , то коэффициент асимметрии меньше нуля. Поэтому  $\text{med } \hat{F}_G(x) > \mathbb{E}\hat{F}_G(x) = F(x)$ .  $\square$

### 2.1.2. Поправка неточных критериев

Зафиксируем некоторый неточный (консервативный или радикальный) правосторонний критерий и уровень значимости  $\alpha^*$ . Пусть дана зависимость ошибки первого рода от уровня значимости  $\alpha_I(\alpha) = \mathbf{P}_{H_0}(\mathbf{p} < \alpha) = \mathbf{P}_{H_0}(1 - F(t) < \alpha)$ . Тогда критерий с формальным уровнем значимости  $\tilde{\alpha}^* = \alpha_I^{-1}(\alpha^*)$  является точным: ошибка первого рода  $\alpha_I(\tilde{\alpha}^*) = \alpha^*$ .

**Замечание 8.** Если критерий строится по суррогатным данным (алгоритм 1), то ошибка первого рода  $\hat{\alpha}_I^{(G)}(\alpha) = \mathbf{P}_{H_0}(\hat{\mathbf{p}}^{(G)} < \alpha) = \mathbf{P}_{H_0}(1 - \hat{F}_G(t) < \alpha)$  и формальный уровень значимости  $\tilde{\alpha}_G^* = \left(\hat{\alpha}_I^{(G)}\right)^{-1}(\alpha^*)$ , который меньше  $\tilde{\alpha}^*$  по замечанию 9.

**Предположение 3.** Функция распределения  $\hat{\mathbf{p}}_G$  — скачкообразная и для очень радикального критерия первый скачок очень большой. Найти обратную функцию для скачкообразной функции невозможно, поэтому нужно использовать интерполяцию, но результат будет зависеть от выбора этой интерполяции.

**Замечание 9.** Рассуждения выше показывают, что критерий из алгоритма 1 не должен быть слишком радикальным для применения поправки.

Если зависимость  $\alpha_I(\alpha)$  неизвестна, она оценивается с помощью моделирования. Приведем алгоритм поправки в этом случае. Помимо критерия и уровня значимости, зафиксируем количество выборок  $M$  для оценки  $\alpha_I(\alpha)$  и их объем  $N$ .

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

1. Моделируется  $M$  выборок объема  $N$  при верной  $H_0$ .

2. По моделированным данным строится оценка зависимости ошибки первого рода от уровня значимости  $\alpha_I(\alpha)$ .
3. Рассчитывается формальный уровень значимости:  $\tilde{\alpha}^* = \alpha_I^{-1}(\alpha^*)$ . Критерий с таким уровнем значимости является асимптотически точным при  $M \rightarrow \infty$ .

## 2.2. Monte Carlo SSA

### 2.2.1. Постановка задачи

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

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

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

**Определение 15.** Случайный процесс  $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n, \dots)$  называют красным шумом с параметрами  $\varphi$  и  $\delta$ , если  $\xi_n = \varphi \xi_{n-1} + \delta \varepsilon_n$ , где  $0 < \varphi < 1$ ,  $\varepsilon_n$  — белый гауссовский шум с дисперсией 1 и  $\xi_1$  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.

### 2.2.2. Одиночный тест

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

$$p = \|\Xi^T W\|^2.$$

Статистикой критерия является величина

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

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

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

**Алгоритм 3.** Одиночный тест [1]

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

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

**Замечание 10.** В большинстве случаев, распределение  $p$  неизвестно. Поэтому оно оценивается методом Монте-Карло: берется  $G$  реализаций случайной величины  $\xi$ , для каждой вычисляется  $p$  и строится эмпирическое распределение. В связи с этим описанный выше алгоритм называют методом Monte-Carlo SSA.

**Замечание 11.** Если частота  $\omega$  сигнала  $S$  известна, то в качестве  $W$  можно взять синусоиду с частотой  $\omega$ . Но на практике  $\omega$  редко бывает известна, что делает данный критерий несостоятельным против  $H_1$ .

### 2.2.3. Множественный тест

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

$$\hat{p}_k = \|\mathbf{X}^T W_k\|^2, \quad k = 1, \dots, H. \quad (2.1)$$

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

$$p_{ki} = \|\Xi_i^T W_k\|^2, \quad i = 1, \dots, G; \quad k = 1, \dots, H, \quad (2.2)$$



где  $G$  — количество суррогатных реализаций  $\xi$ ,  $\Xi_i$  — траекторная матрица  $i$ -й реализации  $\xi$ .

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 4. Multiple MC-SSA [1]

1. Для  $k = 1, \dots, H$  вычисляется статистика  $\hat{p}_k$ , выборка  $P_k = \{p_{ki}\}_{i=1}^G$ , ее среднее  $\mu_k$  и стандартное отклонение  $\sigma_k$ .
2. Вычисляется  $\eta = (\eta_1, \dots, \eta_G)$ , где

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

3. Находится  $q$  как выборочный  $(1 - \alpha)$ -квантиль  $\eta$ , где  $\alpha$  — уровень значимости.
4. Нулевая гипотеза не отвергается, если

$$t = \max_{1 \leq k \leq H} (\hat{p}_k - \mu_k) / \sigma_k < q.$$

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

**Замечание 12.** p-value для этого критерия вычисляется как  $1 - \hat{F}_G(t)$ , где  $\hat{F}_G$  — эмпирическая функция распределения  $\eta$ .

#### 2.2.4. 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  $W_1, \dots, W_H$  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. Борьба с этой проблемой позволяет метод эмпирической поправки критерия, основанный на оцененных ошибках первого рода в зависимости от уровня значимости.

### 2.2.5. Численное сравнение MC-SSA с другими критериями

Рассмотрим другой подход проверки ряда на наличие сигнала. Пусть дан временной ряд  $X = S + \xi$ , где  $S$  — сигнал,  $\xi$  — красный шум с параметрами  $\varphi$  и  $\delta$ . Подход заключается в следующем:

1. Провести отбеливание: вычислить  $Y = \Sigma^{-1/2}X$ , где  $\Sigma$  — 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  $S = \{A \cos(2\pi n\omega)\}_{n=1}^N$  и сравним мощность методов с отбеливанием и MC-SSA при помощи ROC-кривых для разных  $\omega$ .

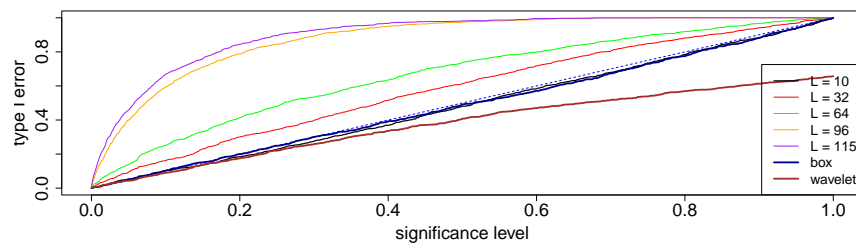


Рис. 2.1. Сравнение ошибки I рода с другими методами

На рис. 2.2, а, 2.2, б, 2.2, в изображены ROC-кривые методов при разных  $\omega$ . 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  $\omega$  MC-SSA при всех длинах окна мощнее, чем *box* и *wavelet*, кроме случая с высокой частотой

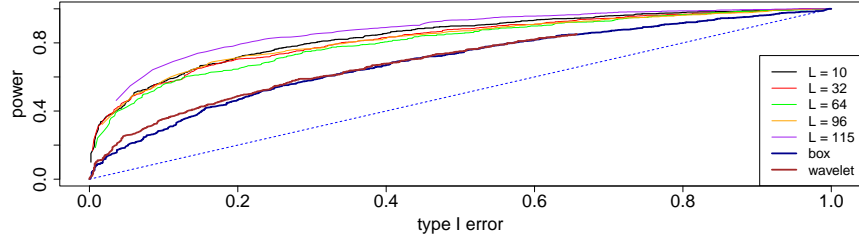
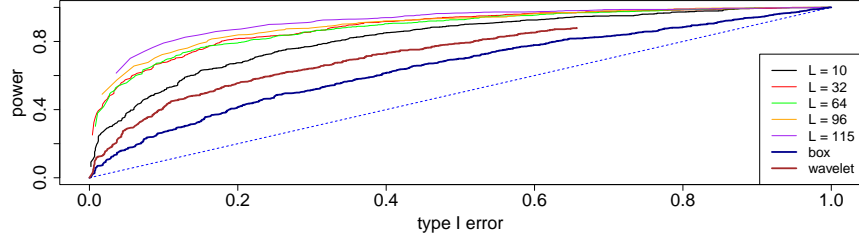
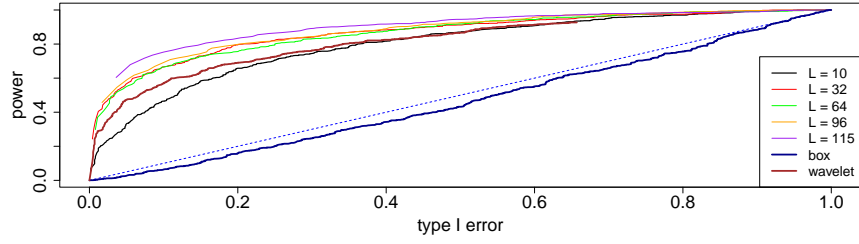
(a) ROC-кривая ( $\omega = 0.025$ )(б) ROC-кривая ( $\omega = 0.125$ )(в) ROC-кривая ( $\omega = 0.225$ )

Рис. 2.2. Сравнение с другими методами

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

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

( $\omega = 0.225$ ), где wavelet оказался немного мощнее MC-SSA с  $L = 10$ . Отметим также, что на рис. 2.2, в ROC-кривая метода box лежит ниже прямой  $y = x$ , поэтому этот тест не имеет смысла 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  $\alpha^* = 0.1$ .

## 2.3. Monte-Carlo MSSA

### 2.3.1. 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  $W_1, \dots, W_H$  — левые векторы матрицы  $\mathbf{X}$ , метод совпадает с алгоритмом 4. Если рассматривать в качестве векторов для проекции правые векторы, то в формулах (2.1) и (2.2) нужно заменить  $\mathbf{X}$  на  $\mathbf{X}^T$  и  $\Xi_i$  на  $\Xi_i^T$  соответственно.

### 2.3.2. Численное сравнение модификаций MC-MSSA

Алгоритм 2 поправки радикальных критериев плохо работает для сильно радикальных критериев. Поскольку рассматриваемый вариант критерия MC-(M)SSA является радикальным, нужно найти длину окна  $L$ , дающую максимально мощный критерий, но при этом не слишком радикальный, чтобы можно было применить поправку.

Введем понятие условно равномошных критериев. Поскольку оценка мощности критерия  $\hat{\beta}$  является долей отвергнутых нулевых гипотез при верной альтернативе, соответствующий доверительный интервал для настоящего значения  $\beta$  выглядит следующим образом:

$$\hat{\beta} \pm z_\alpha \sqrt{\frac{\hat{\beta}(1 - \hat{\beta})}{M}},$$

где  $z_\alpha$  —  $(1 - \alpha/2)$ -квантиль  $N(0, 1)$  и  $M$  — количество моделирований. В худшем случае ( $\beta = 0.5$ ) 95%-ный доверительный интервал при  $M = 1000$  примерно равен  $\hat{\beta} \pm 0.03$ .

Поэтому будем условно считать критерии равномошными, если их мощности будут отличаться не более чем на 0.03.

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 тѐпицево, 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  $D = 2$ , длина каждого канала равна  $N = 100$ , количество суррогатных реализаций красного шума с параметрами  $\varphi$  и  $\delta = 1$  равно  $G = 1000$ . В качестве альтернативы рассмотрим сигнал  $\mathbf{S}$  с элементами

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

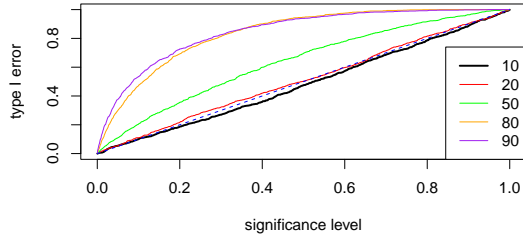
Будем смотреть на графики ошибок первого рода и ROC-кривые критериев для длин окна  $L = 10, 20, 50, 80, 90$ . Будем воспринимать ROC-кривую как график мощности критерия, к которому был применен алгоритм 2.

Рассмотрим несколько примеров. В данном разделе графики ошибок I рода и ROC-кривые представлены только для первого примера, для остальных графики можно найти в приложении [TODO].

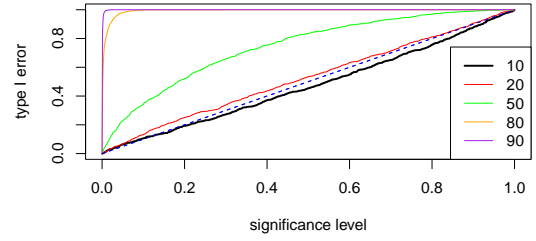
**Пример 1.** Пусть  $\varphi = 0.7$  и параметры сигнала  $A = 1$ ,  $\omega = 0.075$ .

На рис. 2.3 и 2.4 векторы для проекции были взяты из разложения (1.8). На рис. 2.3, *a* видно, что при  $L > 20$  метод радикальный, а наибольшая мощность достигается при  $L = 90$ . На рис. 2.4, *a* отчетливо заметно, что метод радикальный для всех  $L$ . Наибольшая мощность наблюдается при  $L = 80$ , но отметим, что из-за слишком большой ошибки первого рода построить ROC-кривую на промежутке  $[0, 0.22)$  для  $L = 50$  и на всем промежутке для  $L = 10$  и  $L = 20$  не получилось.

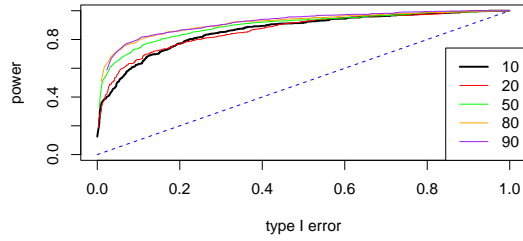
На рис. 2.5 и 2.6 векторы для проекции были взяты из разложения (1.7). Если рассматривать проекцию на левые векторы, то на рис. 2.5, *a* видно, что метод радикальный, а наибольшая мощность достигается при  $L = 20$ . Проекция на правые векторы также дает радикальный критерий, как видно на рис. 2.6, *a*. Наибольшая мощность наблюдается при  $L = 80$ , но из-за слишком большой ошибки первого рода ROC-кривую для



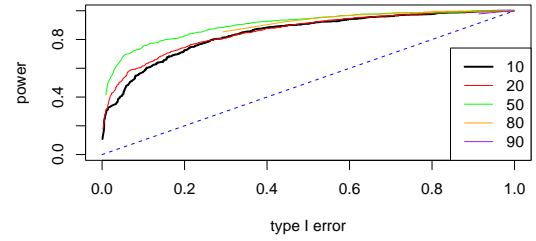
(a) Ошибка первого рода (Sum)



(б) Ошибка первого рода (SVD)



(в) ROC-кривая (Sum)



(г) ROC-кривая (SVD)

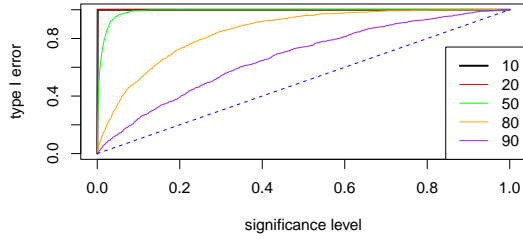
Рис. 2.3. Сравнение методов с проекцией на левые векторы ( $\varphi = 0.7$ ,  $\omega = 0.075$ )

$L = 10$  и  $L = 20$ , для которых метод, предположительно, имеет бóльшую мощность, удалось построить не на всем промежутке.

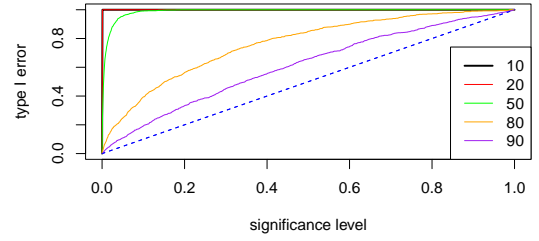
**Пример 2.** Пусть  $\varphi = 0.3$  и  $\omega = 0.075$ . Поскольку при уменьшении  $\varphi$  уменьшается дисперсия шума и это приводит к увеличению мощности критериев, уменьшим амплитуду сигнала до  $A = 0.5$ .

**Пример 3.** В условиях примера 1 увеличим частоту сигнала  $\omega$  до 0.225 и посмотрим на ROC-кривые при такой альтернативе. Поскольку спектральная плотность красного шума монотонно убывает, чем выше частота сигнала, тем он меньше смешивается с красным шумом, следовательно, тем больше мощность критерия. Поэтому уменьшим амплитуду  $A$  до 0.4.

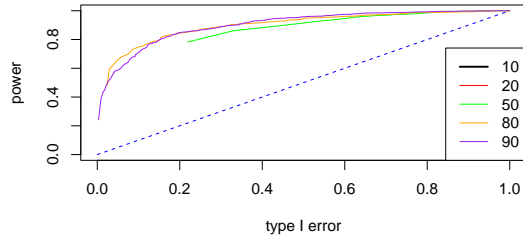
В таблице 2.2 представлены мощность поправленных критериев при уровне значимости  $\alpha^* = 0.1$  для оптимальной длины окна для всех рассмотренных примеров, черным выделены наибольшая мощность и примерно равные ей. В таблице 2.3 представлены размеры матриц методов, у которых находятся вектора для проекции для тех длин окна, мощность которых примерно равна мощности при оптимальном  $L$ , но



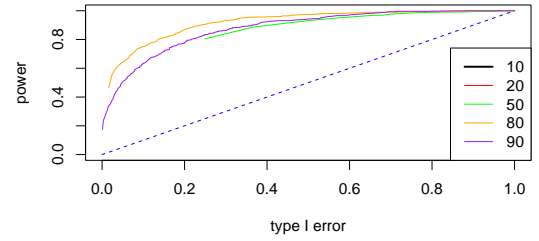
(a) Ошибка первого рода (Sum)



(б) Ошибка первого рода (SVD)



(в) ROC-кривая (Sum)



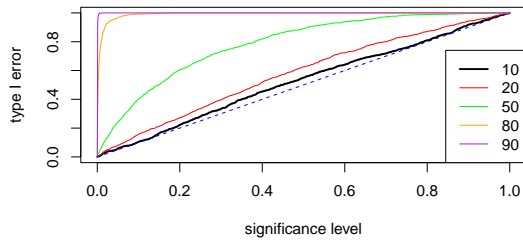
(г) ROC-кривая (SVD)

Рис. 2.4. Сравнение методов с проекцией на правые векторы ( $\varphi = 0.7$ ,  $\omega = 0.075$ )

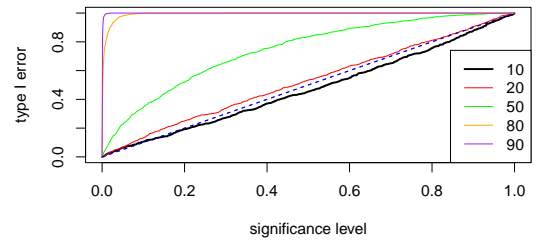
критерии с такой длиной окна более эффективные (матрица меньшего размера).

## Выводы

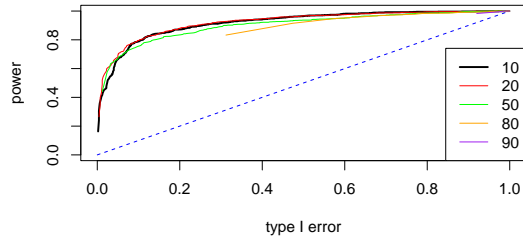
Подведем итоги. Для рассмотренных примеров методы Block и Sum с проекцией на левые векторы показывают наибольшую мощность среди всех рассмотренных модификаций MC-MSSA, однако метод Sum во всех трех случаях численно эффективнее метода Block в нахождении векторов для проекции. Также хочется отметить, что метод Sum с проекцией на левые векторы дает наименее радикальный критерий, что важно при построении поправки (см. замечание 9). Поэтому на данный момент, беря в учет рассмотренные примеры, рекомендуется использовать метод Sum с проекцией на левые векторы с  $L = 90$  для рядов длины  $N = 100$ .



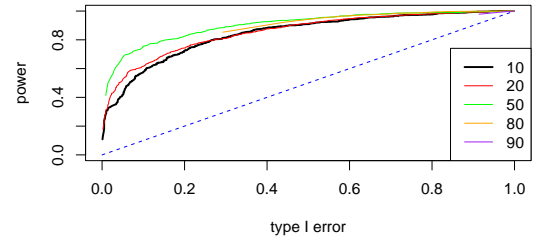
(a) Ошибка первого рода (Block)



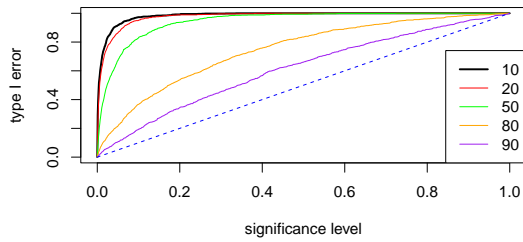
(б) Ошибка первого рода (SVD)



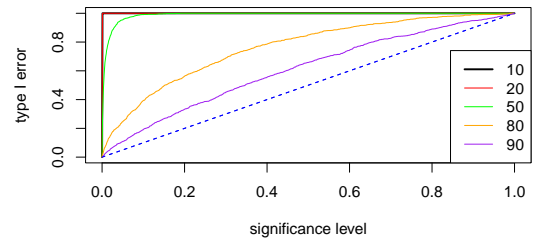
(в) ROC-кривая (Block)



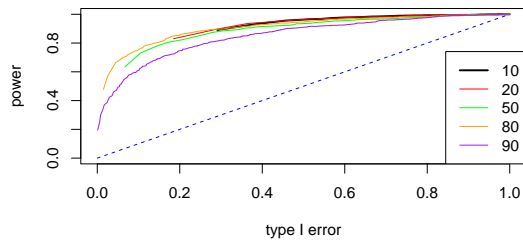
(г) ROC-кривая (SVD)

Рис. 2.5. Сравнение методов с проекцией на левые векторы ( $\varphi = 0.7$ ,  $\omega = 0.075$ )

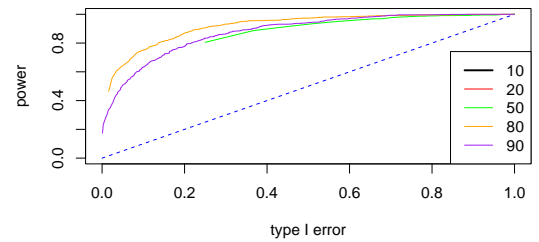
(a) Ошибка первого рода (Block)



(б) Ошибка первого рода (SVD)



(в) ROC-кривая (Block)



(г) ROC-кривая (SVD)

Рис. 2.6. Сравнение методов с проекцией на правые векторы ( $\varphi = 0.7$ ,  $\omega = 0.075$ )



Таблица 2.2. Мощность методов для оптимальных  $L$ 

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

Таблица 2.3. Трудоемкость 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

### 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 проверке 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.

В этой главе будем использовать MC-SSA с проекцией на векторы теплицева разложения (1.7) или (1.8) (для  $D = 1$  это не важно) матрицы  $\mathbf{X}$ .

#### 3.1. Зависимость радикальности и мощности от параметра $L$

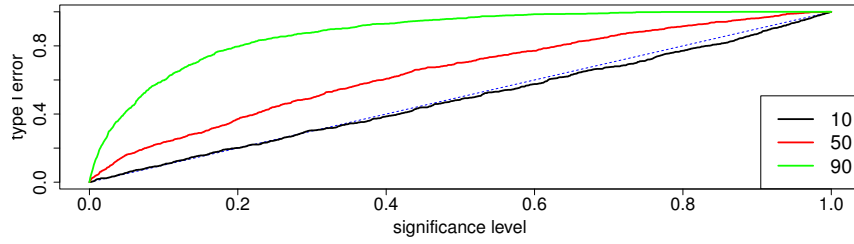
Поскольку рассматриваемый вариант критерия MC-SSA является радикальным, существует проблема выбора такой длины окна  $L$ , которая дает максимально мощный критерий, но при этом не слишком радикальный, чтобы можно было применить поправку. Однако, в зависимости от длины ряда  $N$  и параметров красного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

Рассмотрим несколько примеров. Пусть дана модель

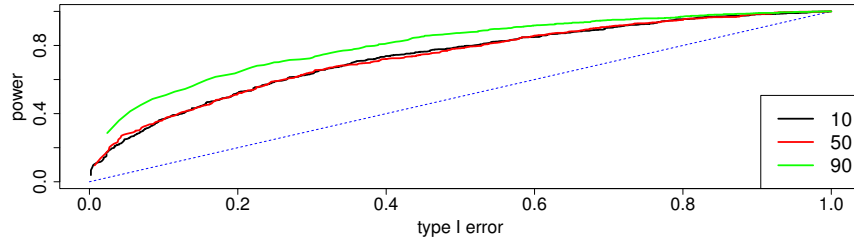
$$\mathbf{X} = \mathbf{S} + \xi,$$

где  $\mathbf{S} = \{A \cos(2\pi\omega n)\}_{n=1}^N$ , а  $\xi$  — красный шум с параметрами  $\varphi$  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 4.** Пусть  $\varphi = 0.7$ ,  $N = 100$ . По графику ошибок первого рода на рис. 3.1,  $a$



(a) Ошибка I рода

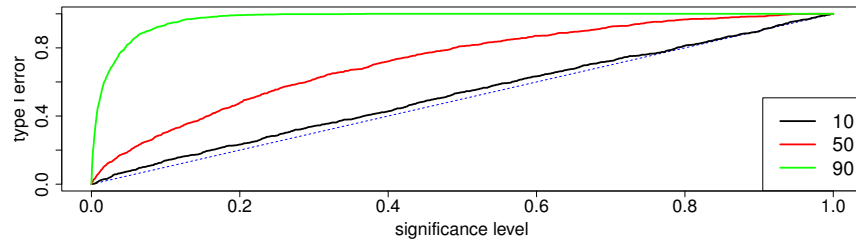
(б) ROC-кривая ( $\omega = 0.075$ )Рис. 3.1. Пример 1 ( $\varphi = 0.7$ ,  $N = 100$ )

видно, что чем больше  $L$ , тем более радикальным становится критерий. На рис. 3.1, б изображены ROC-кривые критериев, наибольшую мощность дает критерий с  $L = 90$ . На этом примере видно, что самым мощным является самый радикальный критерий.

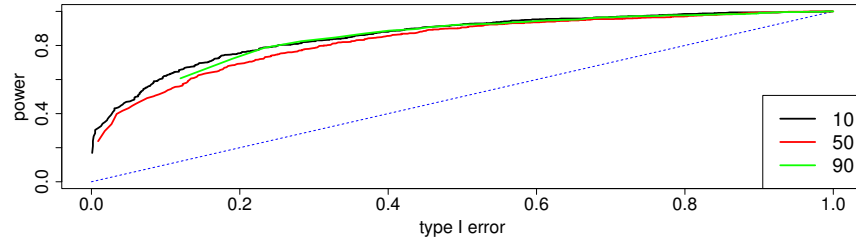
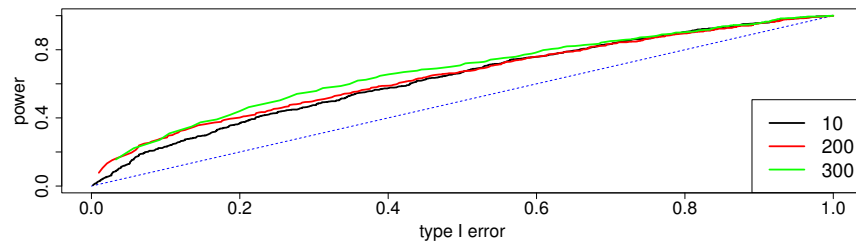
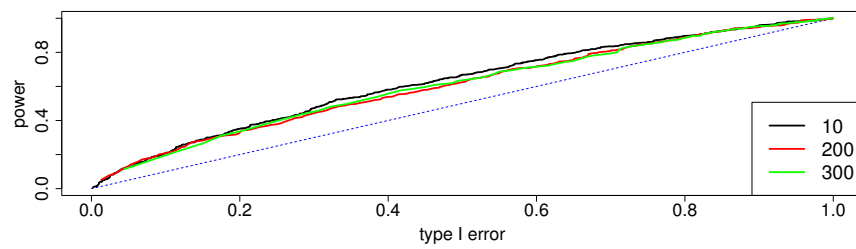
**Пример 5.** Пусть  $\varphi = 0.3$ ,  $N = 100$ . На рис. 3.2, а изображен график ошибок первого рода. По нему видно, что, как и в примере 1, чем больше  $L$ , тем больше радикальность критерия. Если взглянуть на ROC-кривые на рис. 3.2, б, то видно, что с уменьшением параметра  $\varphi$  уменьшается 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 6.** Теперь увеличим длину ряда до  $N = 400$  и посмотрим на ROC-кривые для примеров 4, 5. На рис. 3.3, а, 3.3, б видим, что для обоих примеров с увеличением



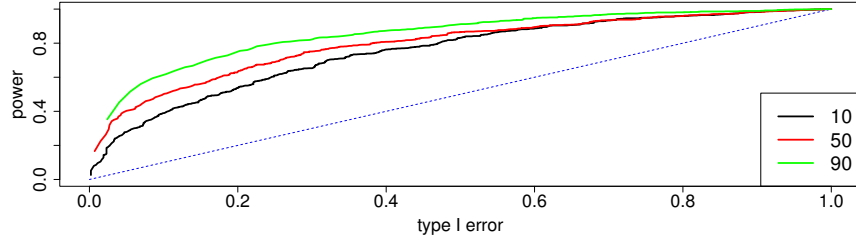
(a) Ошибка I рода

(б) ROC-кривая ( $\omega = 0.075$ )Рис. 3.2. Пример 2 ( $\varphi = 0.3$ ,  $N = 100$ )(a) ROC-кривая ( $\varphi = 0.7$ )(б) ROC-кривая ( $\varphi = 0.3$ )Рис. 3.3. Пример 3 ( $N = 400$ ,  $\omega = 0.075$ )

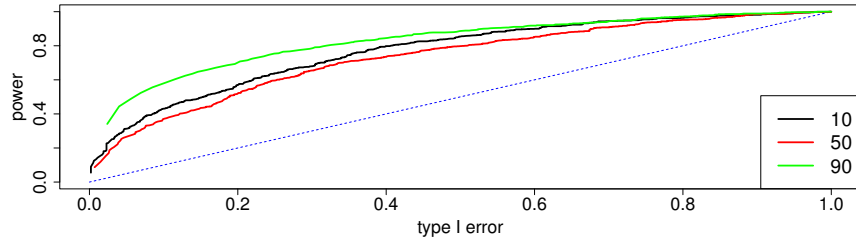
длины ряда уменьшается различие в мощностях после поправки в зависимости от длины окна.

**Пример 7.** В условиях примера 4 рассмотрим разные частоты  $\omega$  сигнала  $S$  и зависи-

мость упорядоченности критериев по мощности от  $L$ . На рис. 3.4, б, 3.4, а изображены



(а) 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  $\omega$ , то видна динамика по соотношению ROC-кривых для  $L = 10$  и  $L = 50$ .

**Пример 8.** 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 2.1.2, 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

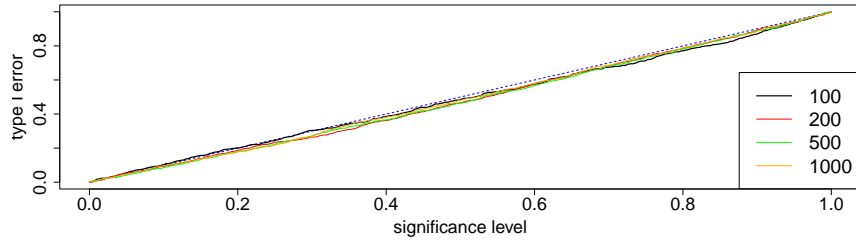
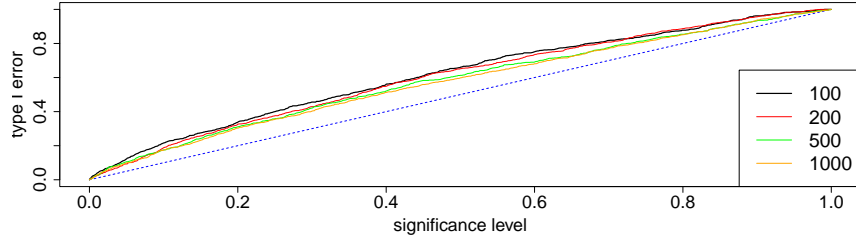
(a) Ошибка I рода ( $L = 10$ )(б) Ошибка I рода ( $L = 40$ )

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

MC-SSA только с  $L = 10$ . Это нетрудозатратно, но возможна некоторая потеря 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 предполагалось, что параметры красного шума  $\varphi$  и  $\delta$  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]. Тогда параметры оцениваются на основе исходного

ряда методом максимального правдоподобия, где для нахождения начальных значений используется метод CSS [15].

### 3.2.1. Реакция оценок на наличие сигнала

Посмотрим, как испортится оценка параметров, если помимо красного шума будет присутствовать сигнал. В качестве сигнала возьмем  $S = \{A \cos(2\pi\omega n)\}_{n=1}^N$ , где  $A = 1$ . Пусть  $N = 100$ ,  $\varphi = 0.7$ ,  $\delta = 1$ . Обозначим оценку параметров за  $\hat{\varphi}$  и  $\hat{\delta}$ . В качестве оценки точности было взято среднеквадратичное отклонение от истинного значения.

	$\omega = 0.075$	$\omega = 0.175$	$\omega = 0.275$	$\omega = 0.375$	$\omega = 0.475$
$\text{MSE}\hat{\varphi}$	0.00416	0.0084	0.0411	0.0986	0.1565
$\text{D}\hat{\varphi}$	0.0031	0.0042	0.0078	0.0122	0.0176
$\text{Bias}(\hat{\varphi}, \varphi)$	0.033	-0.0657	-0.1828	-0.2941	-0.373
$\text{MSE}\hat{\delta}$	0.00706	0.03816	0.1073	0.1714	0.2141
$\text{D}\hat{\delta}$	0.0046	0.005	0.0082	0.0061	0.009
$\text{Bias}(\hat{\delta}, \delta)$	0.05	0.1822	0.315	0.4066	0.453

Таблица 3.1. Оценка параметров красного шума при наличии сигнала ( $\varphi = 0.7$ ,  $\delta = 1$ )

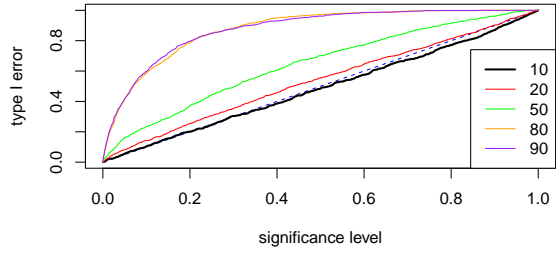
В таблице 3.1 представлены результаты на основе 100 реализаций шума. Поскольку  $\text{MSE}\hat{\theta} = \text{D}\hat{\theta} + \text{Bias}(\hat{\theta}, \theta)^2$ , в таблице также представлены значения дисперсии и смещения оценок. По таблице видно, что основной вклад в ошибку оценок вносит смещение, причем при увеличении частоты  $\omega$  сигнала это смещение увеличивается по модулю.

**Замечание 13.** В таблице 3.1 были взяты такие  $\omega$  с целью показать общий случай, поскольку если  $\tilde{L}\omega$  — целое, то SSA точнее выделяет сигнал [11] и, следовательно, сделает оценку параметров лучше.

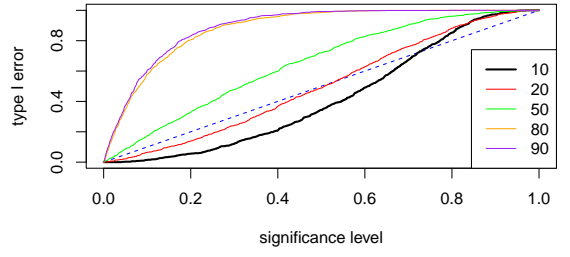
### 3.2.2. Искажение критерия при использовании оценок

В условиях примеров 1, 2 и 3 из раздела 2.3.2 и посмотрим, насколько сильно искажается критерий MC-SSA, если оценивать параметры красного шума.

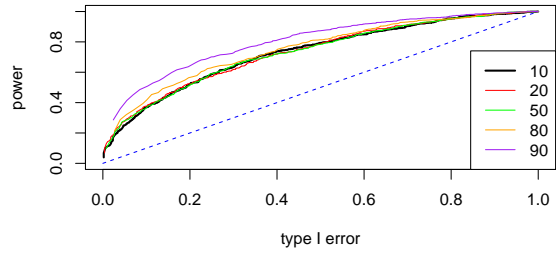
На рис. 3.6 представлено сравнение критерия с известными и критерия с оцененными параметрами красного шума. На рис. 3.6, а и 3.6, б изображены графики ошибок первого рода для разных длин окна, а на рис. 3.6, в и 3.6, г — ROC-кривые. По ним



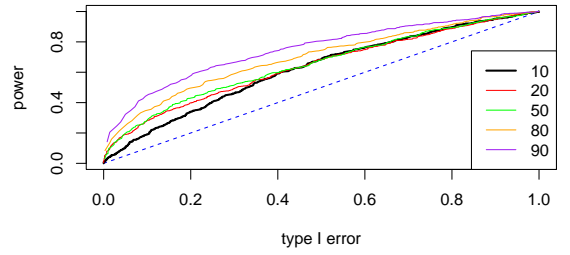
(a) Ошибка I рода (известные параметры)



(б) Ошибка I рода (оцененные параметры)



(в) ROC-кривая (известные параметры)



(з) ROC-кривая (оцененные параметры)

Рис. 3.6. Сравнение критериев ( $\varphi = 0.7$ ,  $\omega = 0.075$ )

видно, что оценка параметров снижает радикальность критерия, делая его консервативным при  $L = 10$  и  $L = 20$  для небольших  $\alpha$ . Мощность поправленного критерия при этом тоже снижается. Графики для остальных примеров можно найти в приложении [TODO].

В таблице 3.2 представлены результаты численного сравнения методов: длина окна, дающая наибольшую мощность, ошибка первого рода и мощность поправленного критерия при уровне значимости  $\alpha^* = 0.1$ . Как видно из таблицы, при маленьких  $\varphi$  оценка параметров довольно сильно искажает критерий: мощность уменьшилась довольно сильно. Однако при больших  $\varphi$  уменьшение мощности незначительное.

### 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:

$$X = F + S + \xi,$$



Таблица 3.2. Мощность методов для оптимальных  $L$  ( $\alpha^* = 0.1$ )

Пример 1 ( $\varphi = 0.7, \omega = 0.075$ )	$L$	$\alpha_I(\alpha^*)$	$\beta(\tilde{\alpha}^*)$
Известные параметры	90	0.598	0.509
Оцененные параметры	90	0.605	0.454
Пример 2 ( $\varphi = 0.3, \omega = 0.075$ )	$L$	$\alpha_I(\alpha^*)$	$\beta(\tilde{\alpha}^*)$
Известные параметры	80	0.814	0.487
Оцененные параметры	80	0.839	0.229
Пример 3 ( $\varphi = 0.7, \omega = 0.225$ )	$L$	$\alpha_I(\alpha^*)$	$\beta(\tilde{\alpha}^*)$
Известные параметры	90	0.598	0.373
Оцененные параметры	90	0.605	0.352

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

$$H_0 : \mathbf{S} = 0,$$

$$H_1 : \mathbf{S} \neq 0.$$

#### Алгоритм 5. MC-SSA с мешающим сигналом

1. Находится приближенное значение мешающего сигнала  $\hat{\mathbf{F}}$  и оцениваются параметры  $\boldsymbol{\xi}$  на основе остатка  $\tilde{\mathbf{X}} = \mathbf{X} - \hat{\mathbf{F}}$ .
2. Находятся левые векторы  $P_1, \dots, P_L$  траекторной матрицы временного ряда  $\tilde{\mathbf{X}}$ , полученные из разложения (1.8).
3. Применяется MC-SSA к исходному ряду  $\mathbf{X}$  с проекцией на векторы  $P_1, \dots, P_L$ , при этом суррогатными рядами являются реализации случайной величины  $\boldsymbol{\eta}$ :

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

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:

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

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

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

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  $\omega$  частотами.

На рис. 3.7 представлены графики ошибок первого рода и ROC-кривые следующих критериев: когда мешающий сигнал и параметры шума известны точно, когда  $F$  известен 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 2.1.2 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  $L = 90$ .

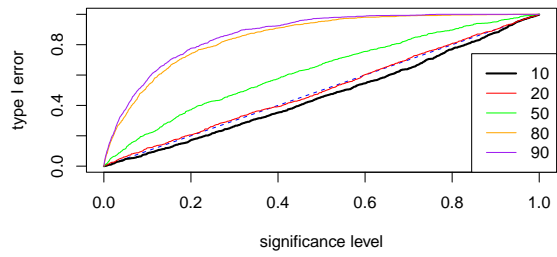
На рис. 3.8 представлена ROC-кривая критериев для оптимального  $L$ . Как видно из графика, при оценке параметров шума и мешающего сигнала мощность падает, но незначительно (примерно на 10%), что важно при применении алгоритма на практике.

### 3.3.2. Тренд

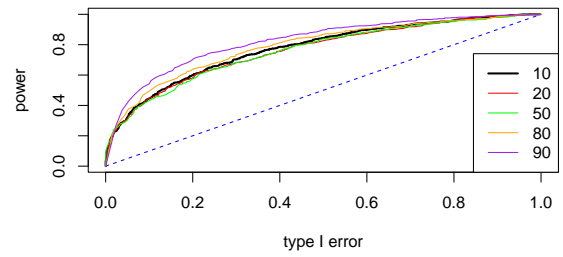
Отдельно рассмотрим вариант, когда мешающий сигнал — тренд, т.е. медленно меняющаяся компонента. Рассмотрим следующий экспоненциальный ряд:

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

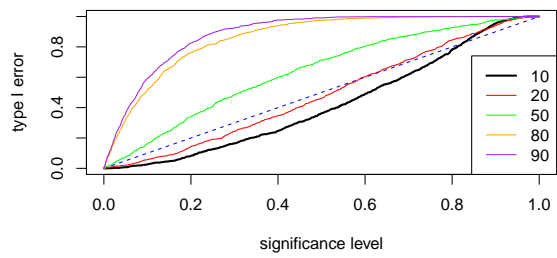
где  $A = 0.2$ ,  $\alpha = 0.05$ .



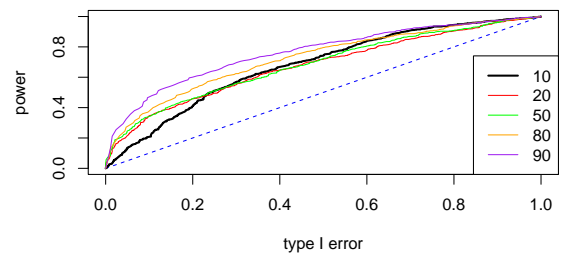
(a) Ошибка I рода



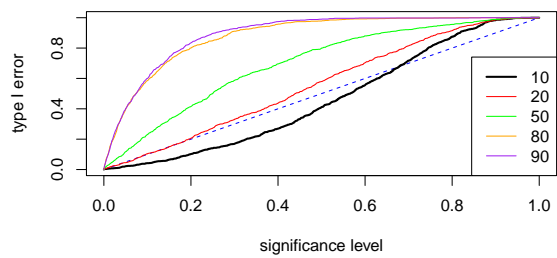
(б) ROC-кривая



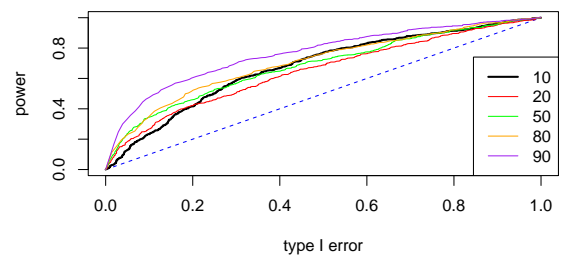
(в) Ошибка I рода (оцененные параметры шума)



(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

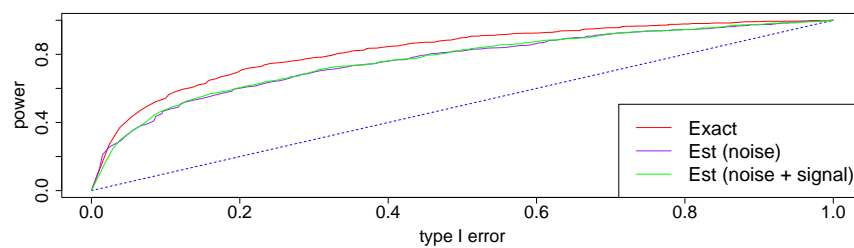
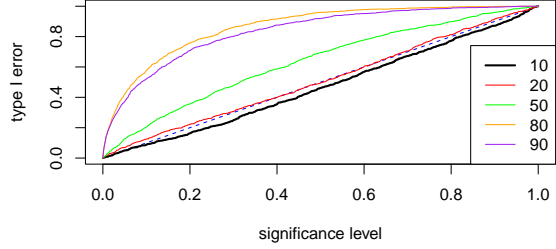
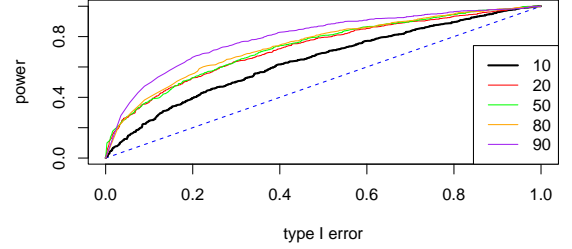


Рис. 3.8. Сравнение ROC-кривых критериев (мешающий сигнал — периодическая компонента)

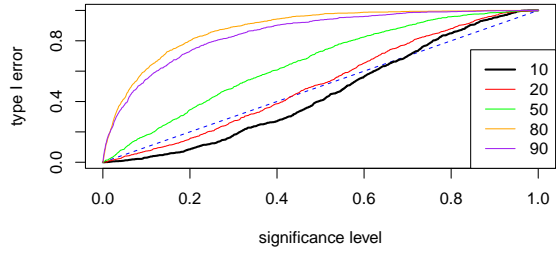
Выделять тренд будем с помощью SSA: поскольку в SVD разложении (1.6) сингулярные числа, соответствующие тренду, будут самыми большими среди всех сингулярных чисел, на шаге группировки (раздел 1.2.3) будем брать первые  $r$  элементарных компонент, где  $r$  — ранг тренда. В данном случае  $r = 1$ .



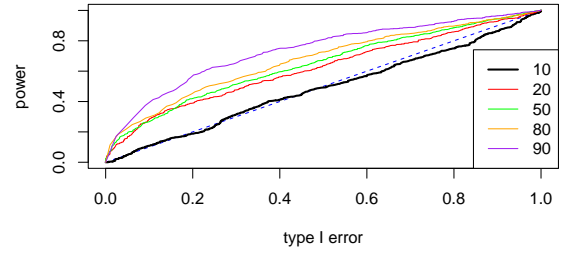
(a) Ошибка I рода



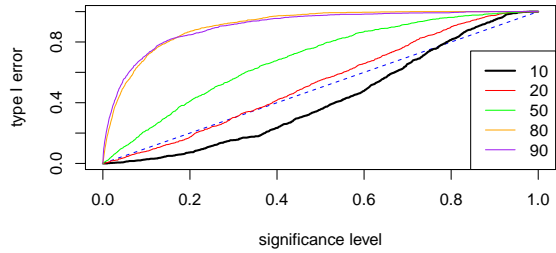
(б) ROC-кривая



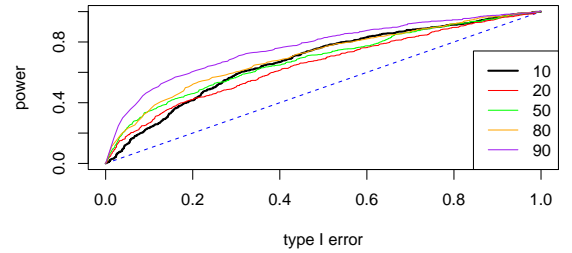
(в) Ошибка I рода (оцененные параметры шума)



(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

На рис. 3.9 представлены графики ошибок первого рода и ROC-кривые следующих критериев: когда тренд и параметры шума  $\varphi$  и  $\delta$  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

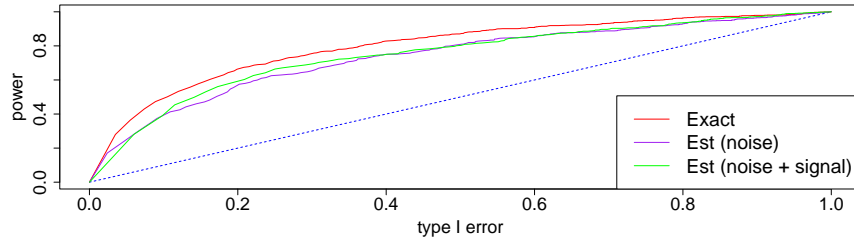


Рис. 3.10. Сравнение ROC-кривых критериев (мешающий сигнал — тренд)

и 3.9,  $\delta$  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.

### 3.3.3. Другой вариант алгоритма

Рассмотрим другой вариант MC-SSA с мешающим сигналом. Отличается он от алгоритма 5 тем, что рассматривается проекция на левые векторы траекторной матрицы исходного временного ряда, а не его остатка.

**Алгоритм 6.** MC-SSA с мешающим сигналом (проекция на векторы исходного ряда)

1. Находится приближенное значение мешающего сигнала  $\hat{F}$  и оцениваются параметры  $\xi$  на основе остатка  $\tilde{X} = X - \hat{F}$ .
2. Находятся левые векторы  $P_1, \dots, P_L$  траекторной матрицы временного ряда  $X$ , полученные из разложения (1.8).
3. Применяется MC-SSA к исходному ряду  $X$  с проекцией на векторы  $P_1, \dots, P_L$ , при этом суррогатными рядами являются реализации случайной величины  $\eta$ :

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

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 5 для  $L = 90$ . Выделять мешающий сигнал будем также, как и в разделе 3.3.1 с помощью SSA.

Таблица 3.3. Сравнение алгоритма 5 и алгоритма 6 при  $\alpha^* = 0.1$

Алгоритм 5	$\alpha_I(\alpha^*)$	$\beta(\tilde{\alpha}^*)$
Точная модель	0.57	0.542
Оцененные параметры шума	0.593	0.48
Оценные параметры шума и мешающий сигнал	0.6	0.475
Алгоритм 6	$\alpha_I(\alpha^*)$	$\beta(\tilde{\alpha}^*)$
Точная модель	0.594	0.532
Оцененные параметры шума	0.588	0.468
Оценные параметры шума и мешающий сигнал	0.624	0.521

В таблице 3.3 представлены результаты сравнения двух алгоритмов, а именно ошибка I рода и мощность поправленного критерия при уровне значимости  $\alpha^* = 0.1$ . Были рассмотрены варианты как с точной моделью, так и варианты, когда параметры шума и мешающий сигнал оцениваются. По таблице 3.3 видно, что алгоритм 6 немного радикальнее алгоритма 5, но при оценивании параметров шума и мешающего сигнала алгоритм 6 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 за-

метное влияние на погодные условия во всем мире, поэтому важно изучить их поведение.

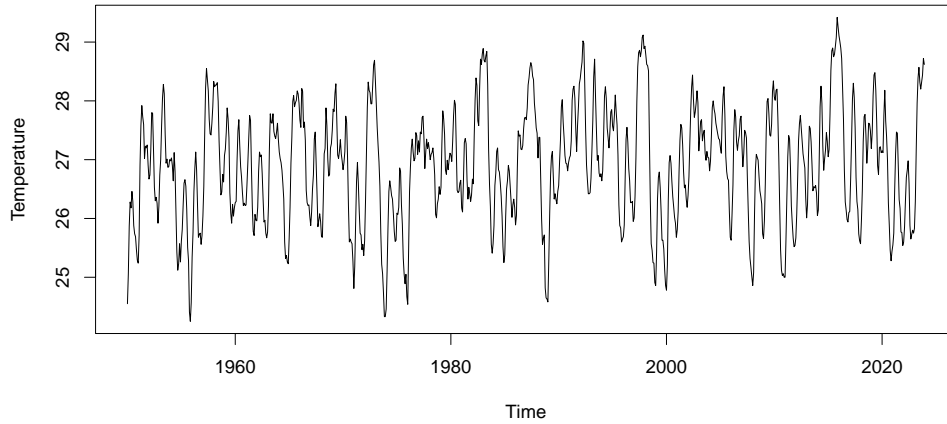


Рис. 3.11. Температура поверхности моря в центральной тропической части Тихого океана

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 соответствует тренду. Посмотрев на двумерные графики собственных векторов на рис. 3.13, видно, что вторая и третья компоненты образуют двенадцатиугольник. Это означает, 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 компоненты 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  $\hat{p}_k$ ,  $k = 1, \dots, L$  (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

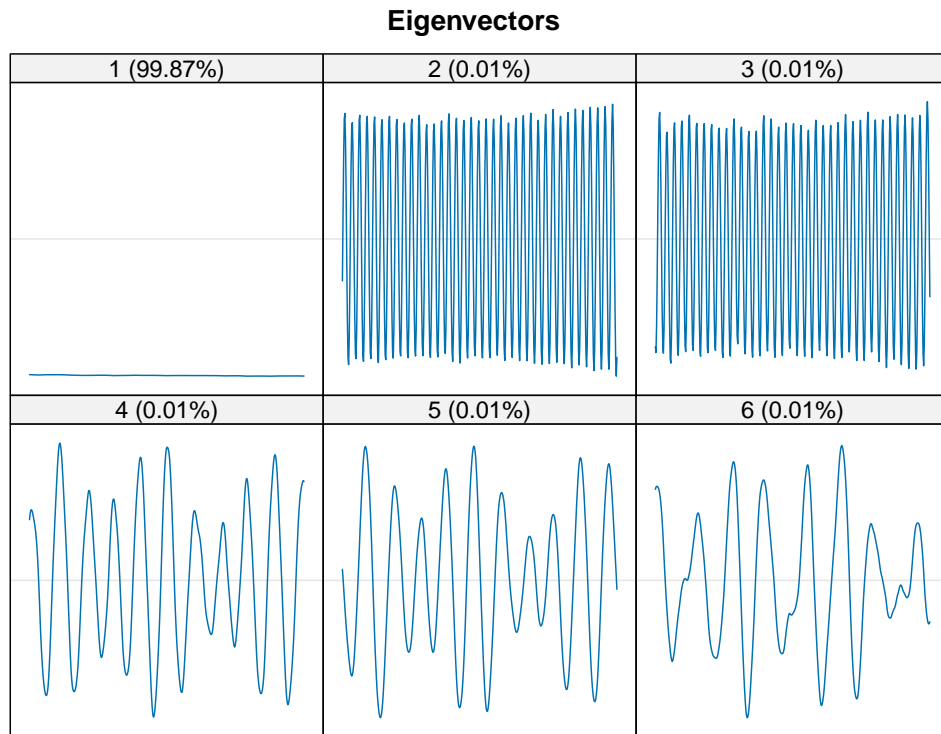


Рис. 3.12. Графики собственных векторов

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 статистик  $\hat{p}_k$  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.



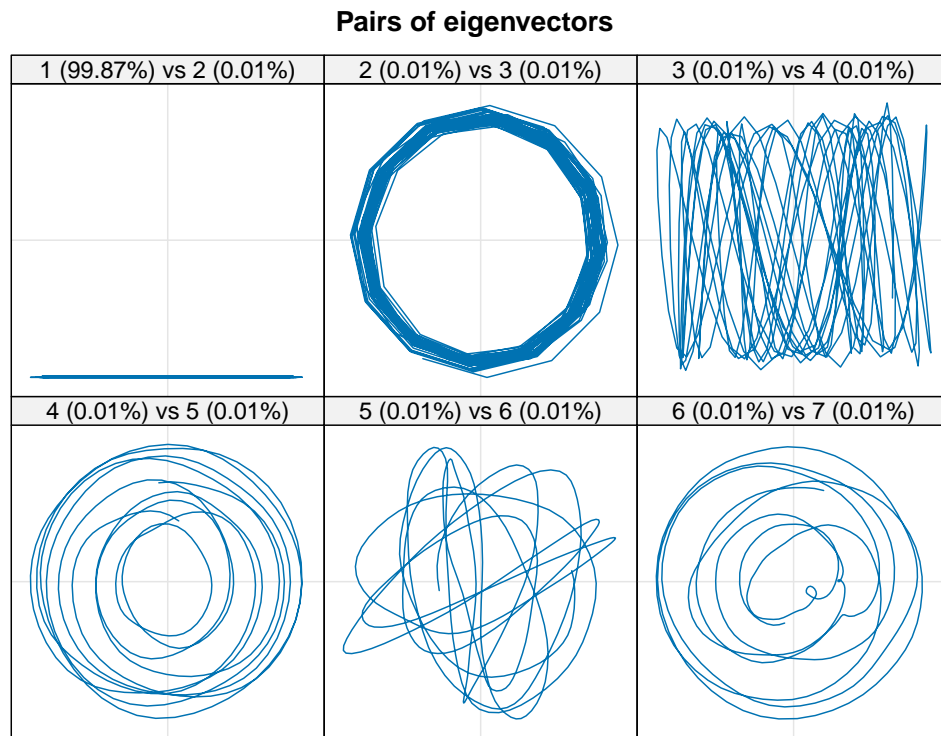


Рис. 3.13. Двумерные графики собственных векторов

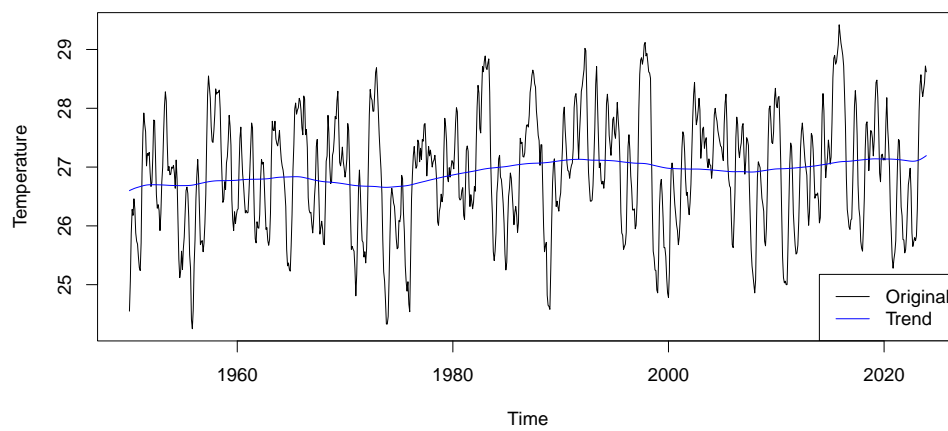


Рис. 3.14. Выделенный тренд

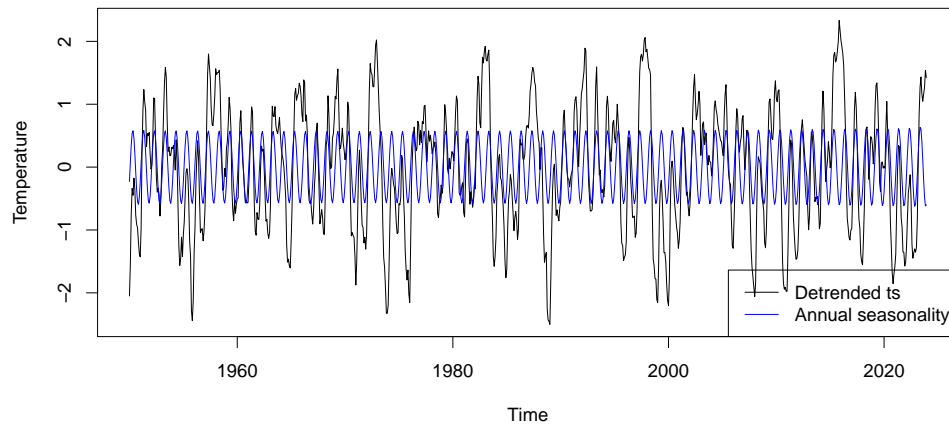


Рис. 3.15. Выделенная годовая сезонность

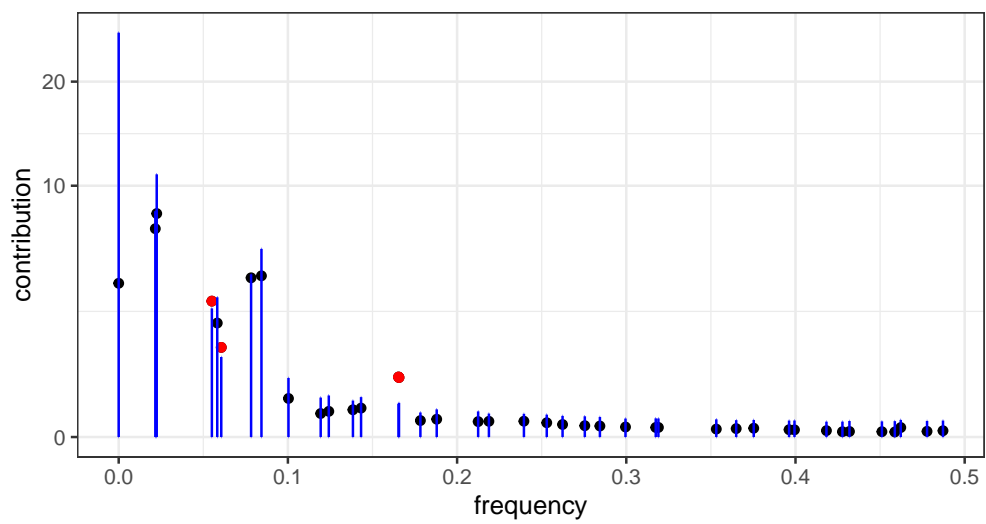


Рис. 3.16. Результат работы MC-SSA

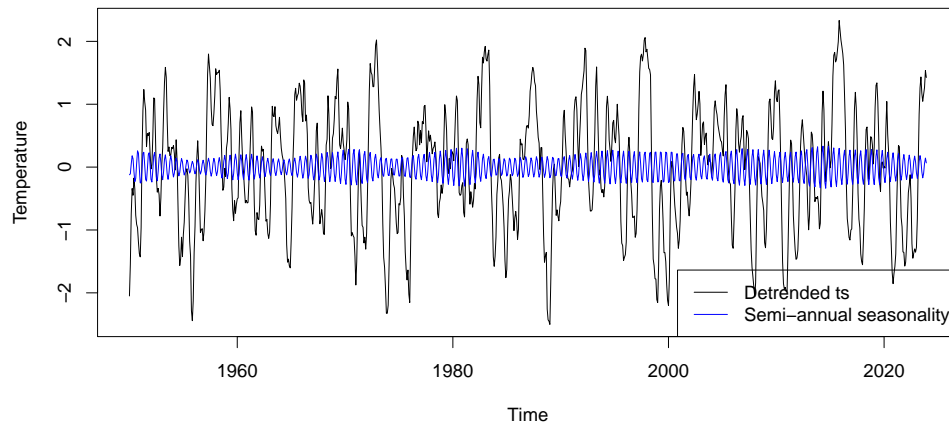


Рис. 3.17. Выделенная полугодовая сезонность

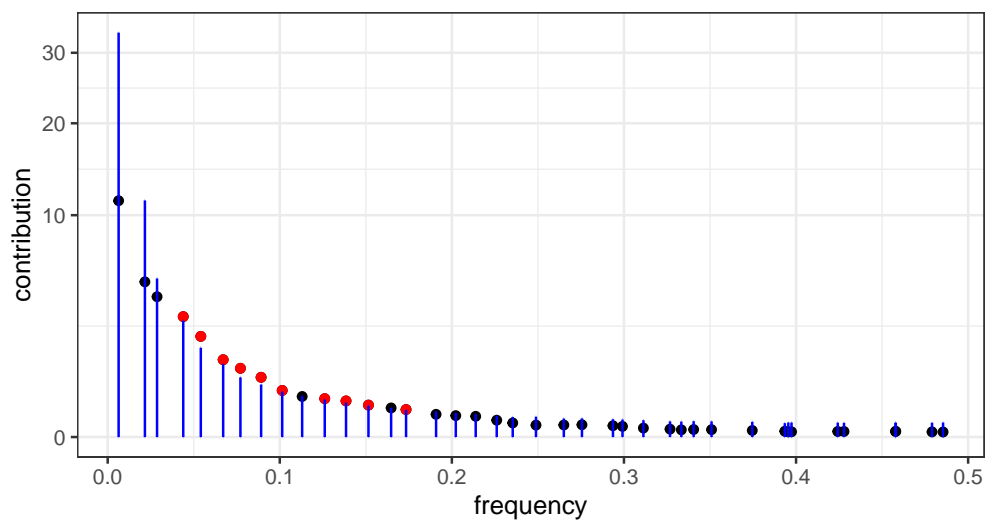


Рис. 3.18. Результат работы MC-SSA для модели ARMA(1, 2)

## 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 M., Smith L. Monte Carlo SSA: Detecting irregular oscillations in the Presence of Colored Noise // *Journal of Climate*. — 1996. — Vol. 9. — P. 3373–3404.
5. GáMiz-Fortis S. R., Pozo-Vázquez D., Esteban-Parra M. J. et al. Spectral characteristics and predictability of the NAO assessed through Singular Spectral Analysis // *Journal of Geophysical Research (Atmospheres)*. — 2002. — Vol. 107, no. D23.
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 G., Aldrich C. 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. Golyandina N., Korobeynikov A., Shlemov A et al. Multivariate and 2D Extensions of Singular Spectrum Analysis with the Rssa Package // *Journal of Statistical Software*. — 2015. — Vol. 67, no. 2.
10. Plaut G., Vautard R. 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 G., Savchev D. White noise testing using wavelets // Stat. — 2014. — Vol. 3. — P. 351–362.
  14. Groth A., Ghil M. 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 N., Korobeynikov A., Zhigljavsky A. 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.

## Приложение А

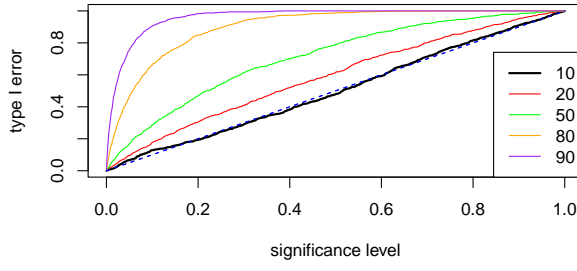
Здесь приведены результаты к разделу 2.3.2 основного текста.

### A.1. $\varphi = 0.7$ , $\omega = 0.075$

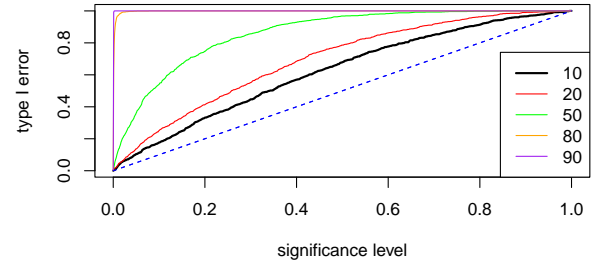
Таблица A.1. Численное сравнение методов для оптимальных  $L$  ( $\varphi = 0.7$ ,  $\omega = 0.075$ )

Метод	левые/правые векторы	$L$	$KD$	длина векторов	количество векторов	$\alpha_I(\alpha^*)$	$\beta(\tilde{\alpha}^*)$
SVD*	левые	50	102	50	50	0.316	0.754
SVD*	правые	80	42	42	42	0.394	0.754
Block*	левые	20, 50	162, 102	162	20	0.157	<b>0.796</b>
Block*	правые	80	42	80	42	0.367	0.717
Sum	левые	80, 90	42, 22	90	22	0.535	<b>0.806</b>
Sum*	правые	80	42	42	42	0.51	0.748

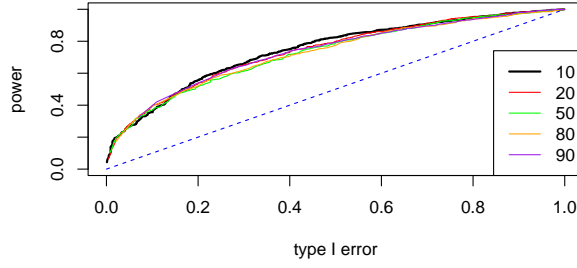
В таблице A.1 для каждого метода указана оптимальная по мощности длина окна (для которой удалось построить ROC-кривую, звездочкой помечены те методы, для которых эта  $L$  может не являться оптимальной), значение  $KD$ , длина проекционных векторов, их количество, а также ошибка первого рода и мощность поправленного критерия при уровне значимости  $\alpha^* = 0.1$ . Черным выделены максимальная мощность и примерно равные ей. Также для каждого критерия дополнительно добавлена длина окна, дающая примерно ту же мощность, что и оптимальная, но метод с таким  $L$  более эффективный в нахождении левых или правых векторов автоковариационной матрицы (такая длина окна выделена синим). Видно, что MC-MSSA с проекцией на левые или правые векторы обеих модификаций мощнее, чем с проекцией на векторы базового MSSA. Самыми мощными являются методы Block и Sum с проекцией на левые векторы, причем метод Sum более эффективнее (матрица размера 80 против матрицы размера 162).



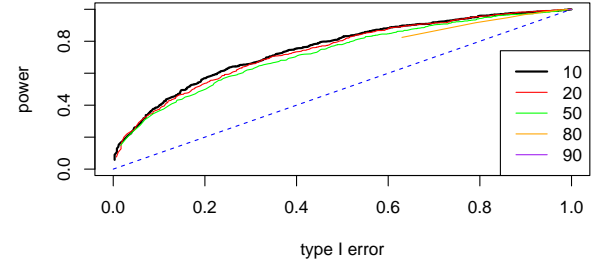
(a) Ошибка I рода (Sum)



(б) Ошибка I рода (Block)



(в) ROC-кривая (Sum)



(г) ROC-кривая (Block)

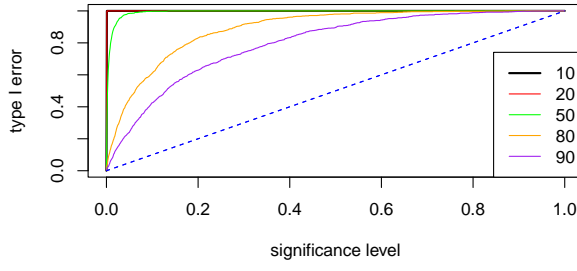
Рис. А.1. Сравнение методов с проекцией на левые векторы ( $\varphi = 0.3$ ,  $\omega = 0.075$ )

## А.2. $\varphi = 0.3$ , $\omega = 0.075$

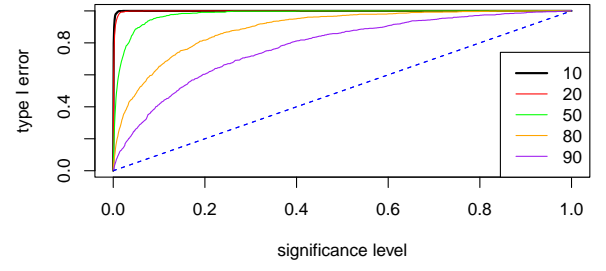
На рис. А.1 изображены графики ошибок первого рода и ROC-кривые методов Sum и Block с проекцией на левые векторы, а на рис. А.2 — на правые векторы. Как видно по рисункам, при уменьшении параметра  $\varphi$ , во-первых, увеличивается радикальность критериев для любой длины окна, а во-вторых уменьшается разброс мощности после поправки.

В таблице А.2 представлены результаты для этого примера. По таблице видно, что различие в мощности у методов совсем небольшое. Если сравнивать методы с наибольшей мощностью по трудоемкости, самым эффективным является метод Sum с проекцией на левые векторы с  $L = 20$ .

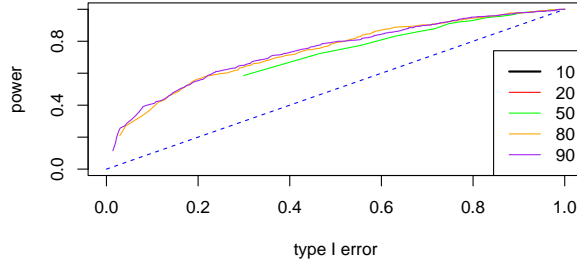




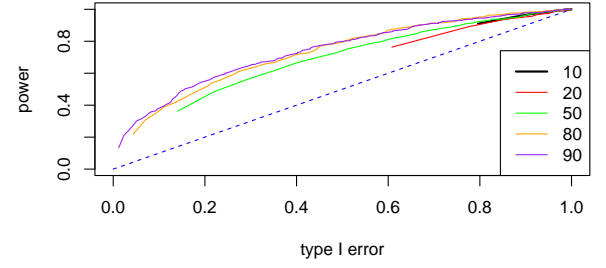
(a) Ошибка I рода (Sum)



(б) Ошибка I рода (Block)



(в) ROC-кривая (Sum)



(г) ROC-кривая (Block)

Рис. A.2. Сравнение методов с проекцией на правые векторы ( $\varphi = 0.3$ ,  $\omega = 0.075$ )

### A.3. $\varphi = 0.7$ , $\omega = 0.225$

На рис. A.4, A.3 изображены ROC-кривые критериев. Если сравнить получившиеся графики с графиками на рис 2.4, в, 2.4, в, 2.5, в и 2.6, в, то видно, что с ростом частоты сигнала в альтернативе увеличивается различие в мощностях критериев в зависимости от параметра  $L$ .

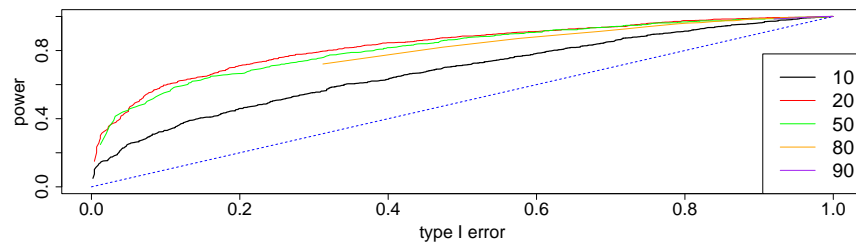
Также для наглядного сравнения методов в таблице A.3 представлены результаты. Метод Block с проекцией на левые и метод Sum с проекцией на левые и правые векторы в этом случае дают наибольшую мощность, причем метод Sum с проекцией на левые или правые векторы при  $L = 80$  наиболее эффективный среди наиболее мощных критериев.

Таблица А.2. Численное сравнение методов для оптимальных  $L$  ( $\varphi = 0.3$ ,  $\omega = 0.075$ )

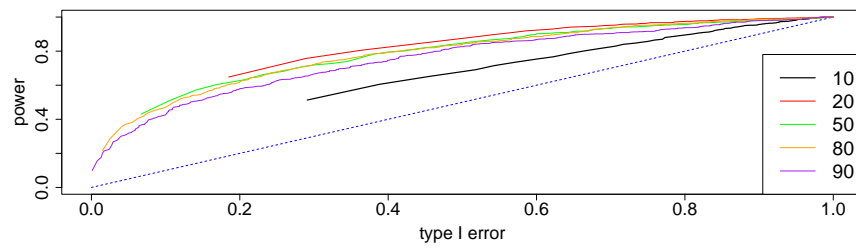
Метод	левые/правые векторы	$L$	$KD$	длина векторов	количество векторов	$\alpha_I(\alpha^*)$	$\beta(\tilde{\alpha}^*)$
SVD*	левые	10, 20	182, 162	20	20	0.199	<b>0.399</b>
SVD*	правые	80, 90	42, 22	22	22	0.449	0.382
Block*	левые	10, 20	182, 162	182	10	0.177	<b>0.398</b>
Block*	правые	80, 90	42, 22	90	22	0.414	0.389
Sum	левые	20, 90	162, 22	20	20	0.905	<b>0.421</b>
Sum*	правые	80, 90	42, 22	22	22	0.425	<b>0.412</b>

Таблица А.3. Численное сравнение методов для оптимальных  $L$  ( $\varphi = 0.7$ ,  $\omega = 0.225$ )

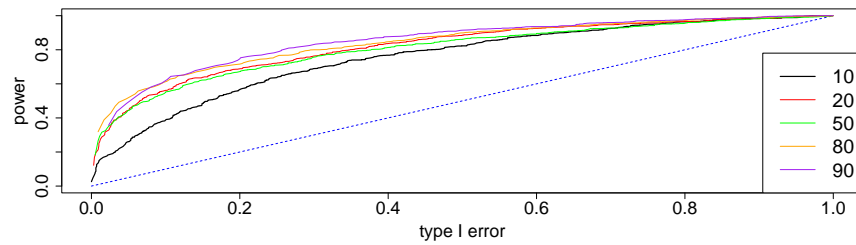
Метод	левые/правые векторы	$L$	$KD$	длина векторов	количество векторов	$\alpha_I(\alpha^*)$	$\beta(\tilde{\alpha}^*)$
SVD*	левые	20	162	20	20	0.122	0.573
SVD*	правые	80	42	80	42	0.394	0.442
Block*	левые	20	162	162	20	0.157	<b>0.597</b>
Block*	правые	50, 80	102, 42	80	42	0.83	0.509
Sum	левые	80, 90	42, 22	10	90	0.535	<b>0.625</b>
Sum*	правые	80	42	42	42	0.51	<b>0.613</b>



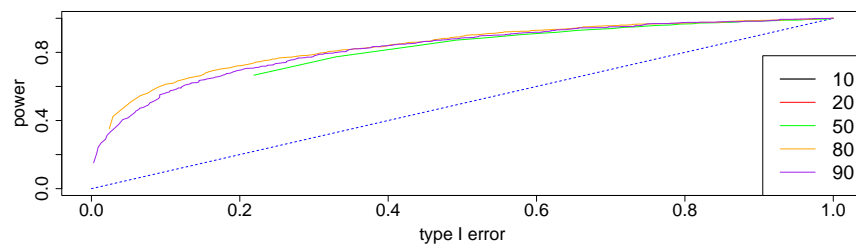
(a) Проекция на левые векторы



(б) Проекция на правые векторы

Рис. А.3. ROC-кривые метода Block ( $\varphi = 0.7$ ,  $\omega = 0.225$ )

(a) Проекция на левые векторы



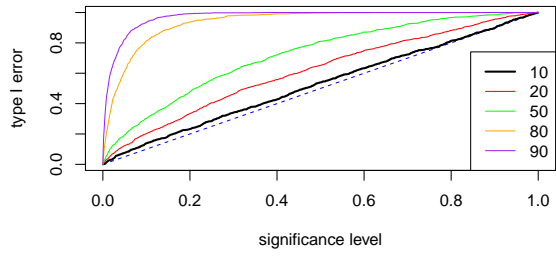
(б) Проекция на правые векторы

Рис. А.4. ROC-кривые метода Sum ( $\varphi = 0.7$ ,  $\omega = 0.225$ )

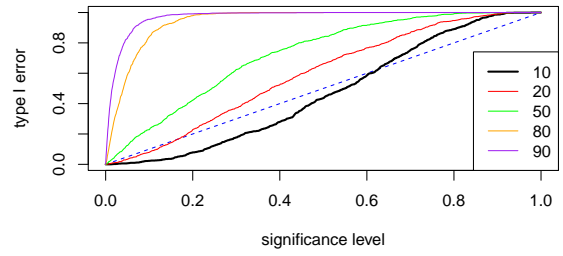
## Приложение Б

Здесь приведены результаты к разделу 3.2.2 основного текста.

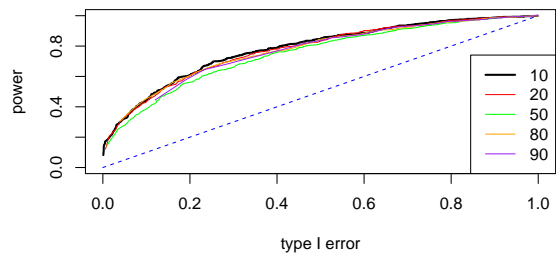
### Б.1. $\varphi = 0.3, \omega = 0.075$



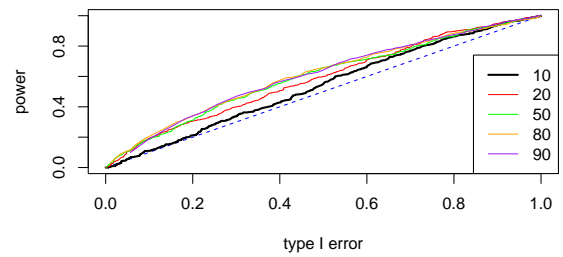
(a) Ошибка I рода (известные параметры)



(б) Ошибка I рода (оцененные параметры)



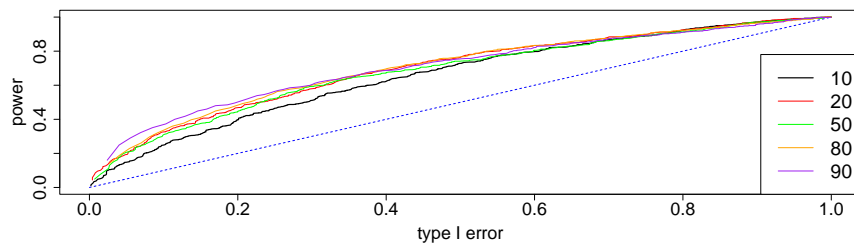
(в) ROC-кривая (известные параметры)



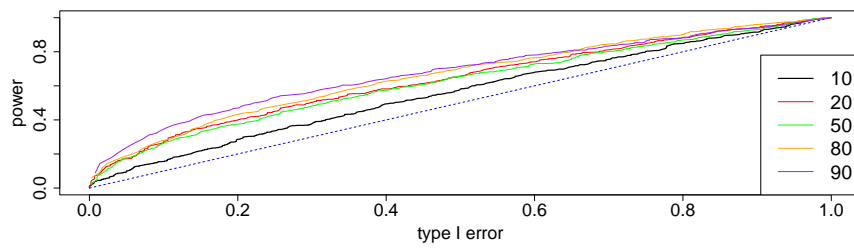
(г) ROC-кривая (оцененные параметры)

Рис. Б.1. Сравнение критериев ( $\varphi = 0.3, \omega = 0.075$ )

### Б.2. $\varphi = 0.7, \omega = 0.225$



(a) ROC-кривая (известные параметры)



(б) ROC-кривая (оцененные параметры)

Рис. Б.2. Сравнение критериев ( $\varphi = 0.7$ ,  $\omega = 0.225$ )