Санкт-Петербургский государственный университет

Прикладная математика и информатика

Отчет по учебной практике 4 (научно-исследовательской работе) (семестр 7)

Метод Монте-Карло SSA для 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.

Глава 1

Метод MSSA и его модификации

Метод Multivariate Singular Spectrum Analysis (сокращенно MSSA) состоит из четырех этапов: *вложения*, *разложения*, *группировки* и *диагонального усреднения*. Давайте начнем с общих частей всех версий алгоритмов SSA. Этими общими частями являются процедура вложения и диагонального усреднения (ганкелизации).

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.

1.1. Метод MSSA

Рассмотрим вещественнозначные одномерные временные ряды $\mathsf{X}^{(d)} = (x_1^{(d)}, x_2^{(d)}, \ldots, x_{N_d}^{(d)})$ длины $N_d > 2$, $d = 1, \ldots, D$. Составим из этих рядов $\mathsf{X} = \{\mathsf{X}^{(d)}\}_{d=1}^{D}$ — D-канальный временной ряд с длинами N_d.

1.1.1. 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

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

1.1.2. Разложение

Задача этапа разложения — разбить траекторную матрицу $\mathbf{X}$ в сумму матриц ранга 1. В базовой версии MSSA используется сингулярное разложение (SVD).

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

$$\mathbf{X} = \sum_{i=1}^{p} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}} = \sum_{i=1}^{p} \mathbf{X}_i, \tag{1.4}$$

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. Отметим, что левые сингулярные векторы имеют размерность L, а правые сингулярные вектора — размерность K.

1.1.3. Группировка

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

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

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

1.1.4. Диагональное усреднение

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

$$\mathcal{H}_{\text{stacked}}(\mathbf{Y}) = [\mathcal{H}(\mathbf{Y}^{(1)}) : \ldots : \mathcal{H}(\mathbf{Y}^{(M)})]. \tag{1.5}$$

Финальным шагом MSSA является преобразование каждой матрицы $\mathbf{X}_{I_k}$, составленной в разделе 1.1.3, в D-канальный временной ряд:

$$\widetilde{\mathsf{X}}_{I_k} = \mathcal{T}^{-1}_{\text{MSSA}} \circ \mathcal{H}_{\text{stacked}} \left(\mathbf{X}_{I_k} \right), \tag{1.6}$$

где $\mathcal{T}_{\text{MSSA}}$ — оператор вложения (1.3), $\mathcal{H}_{\text{stacked}}$ — оператор ганкелизации (1.5).

Замечание 1. При $D = 1$ X — одномерный временной ряд, и приведенный выше алгоритм совпадает с алгоритмом Basic SSA, описанный в [4].

1.2. Toeplitz MSSA

В случае анализа стационарных рядов можно улучшить базовый метод, используя тёплицево разложение матрицы $\mathbf{X}$.

Определение 4. Случайный процесс $\xi = (\xi_1, \ldots, \xi_n, \ldots)$ называется стационарным, если $\forall k \geqslant 1$ $\mathsf{E}\xi_k = 0$ и $\forall k, l \geqslant 1$

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

Определение 5. 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 6. Пусть $\mathsf{X} = \{\mathsf{X}^{(d)}\}_{d=1}^{D}$ — D-канальный временной ряд. Ряд X называют стационарным, если каждый канал $\mathsf{X}^{(d)}$ — стационарный.

Определение 7. Пусть $\mathsf{X} = \{\mathsf{X}^{(d)}\}_{d=1}^{D}$ — D-канальный временной ряд с $N_d = N$. Зафиксируем $1 < M < N$. Матрица $\mathbf{T}_{l,k}^{(M)} \in \mathbb{R}^{M \times M}$ с элементами

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

является оценкой ковариационной матрицы l и k-го каналов.

Toeplitz MSSA отличается от MSSA только другим разложением $\mathbf{X}$ в сумму матриц ранга 1 (1.4). В работе [5] предложен один способ разложения $\mathbf{X}$, который мы назовем Block. Вместе с ним рассмотрим другой вариант разложения, который назовем Sum.

1.2.1. Toeplitz Block MSSA

Рассмотрим блочную матрицу

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

где $K = N - L + 1$. Найдя ортонормированные собственные векторы $Q_1, \ldots, Q_{DK}$ матрицы $\mathbf{T}$, 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 $Z_i = \mathbf{X}Q_i$, $P_i = Z_i/\|Z_i\|$, $\sigma_i = \|Z_i\|$.

1.2.2. Toeplitz Sum MSSA

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

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

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

Замечание 2. Toeplitz Sum MSSA можно использовать для временных рядов с разными длинами каналов, в отличие от Toeplitz Block MSSA.

Замечание 3. Пусть $\mathbf{X} = \sum_i \sigma_i P_i Q_i^T$ — любое из разложений (1.4), (1.7) или (1.8). Будем называть P_i *левыми*, а Q_i — *правыми векторами* матрицы $\mathbf{X}$.

1.3. 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 $\sigma = 5$. Рассмотрим 3 случая, первые два из которых рассматривались ранее в [6]:

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 $L \approx 2N/3$ строится матрица размера $2N/3 \times 2N/3$, в случае Block $L \approx N/2$ и матрица размера $DN/2 \times DN/2$. 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 MSSA

2.1. Постановка задачи

Рассмотрим задачу поиска сигнала (не случайной составляющей) в многоканальном временном ряде. Модель выглядит следующим образом:

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

где S — сигнал, $\boldsymbol{\xi}$ — какой-то шум. 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 8. Случайный вектор $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)$ называют красным шумом с параметрами φ и δ, если $\xi_n = \varphi\xi_{n-1} + \delta\varepsilon_n$, где $0 < \varphi < 1$, ε_n — белый гауссовский шум со средним значением 0 и 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.

2.2. 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.

2.3. 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_k 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.

2.4. 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.

2.5. MC-MSSA: 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.

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.

2.6. Поправка неточных критериев

Приведем алгоритм поправки, преобразовывающий радикальные и консервативные критерии в точные. Зафиксируем уровень значимости α^*, количество выборок M_1 для оценки $\alpha_I(\alpha)$ и их объем N.

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

1. Моделируется M_1 выборок объема N при верной H_0.
2. По моделированным данным строится зависимость ошибки первого рода от уровня значимости $\alpha_I(\alpha)$.
3. Рассчитывается формальный уровень значимости: $\widetilde{\alpha}^* = \alpha_I^{-1}(\alpha^*)$. Критерий с таким уровнем значимости является асимптотически точным при $M_1 \to \infty$.

Определение 9. 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 $\alpha_I(\alpha) = \alpha$.

2.7. Численное сравнение методов

Алгоритм 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 (или правые) векторы SVD разложения матрицы $\mathbf{X}$ (1.4) дает очень радикальный критерий для больших 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 главе 1.2.

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

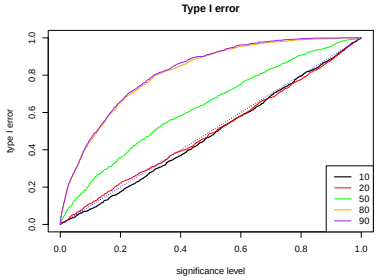
$$s_n^{(1)} = s_n^{(2)} = \cos(2\pi\omega n), \quad n = 1, \ldots, N,$$

где $\omega = 0.075$, $N = 100$.

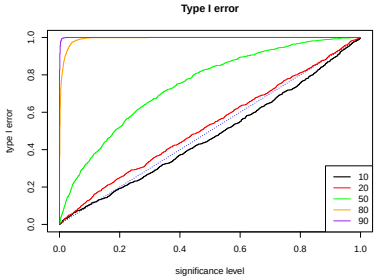
Построим графики ошибки первого рода и ROC-кривые для каждой длины окна $L = 10, 20, 50, 80, 90$. Будем воспринимать ROC-кривую как график мощности критерия, к которому был применен алгоритм 3.

На рис. 2.1 и 2.2 векторы для проекции были взяты из разложения (1.8). На рис. 2.1, *а* видно, что при $L > 20$ метод радикальный, а наибольшая мощность достигается при $L = 90$. На рис. 2.2, *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.

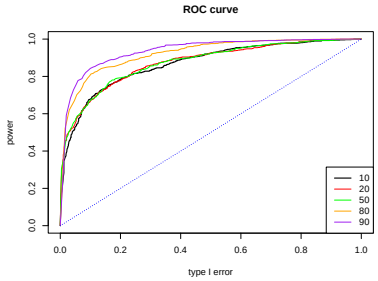
На рис. 2.3 и 2.4 векторы для проекции были взяты из разложения (1.7). Если рассматривать проекцию на левые векторы, то на рис. 2.3, *а* видно, что метод радикальный, а наибольшая мощность достигается при $L = 20$. Проекция на правые векторы также дает радикальный критерий, как видно на рис. 2.4, *а*. 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,



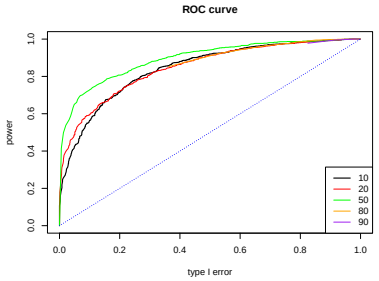
(*а*) Ошибка первого рода (Sum).



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

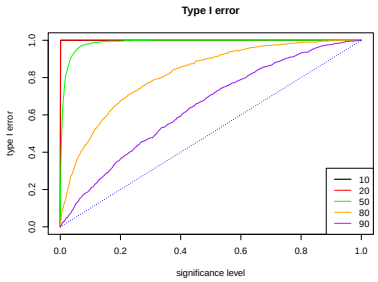


(*в*) ROC-кривая (Sum).

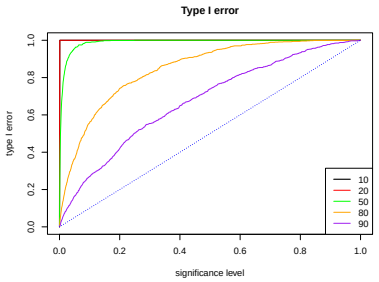


(*г*) ROC-кривая (базовый MSSA).

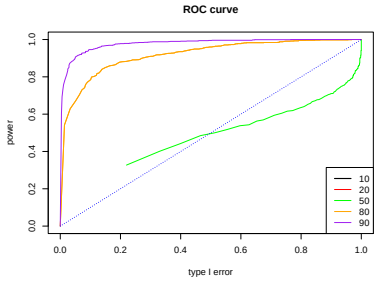
Рис. 2.1. Сравнение методов Sum и базового MSSA (проекция на левые векторы).



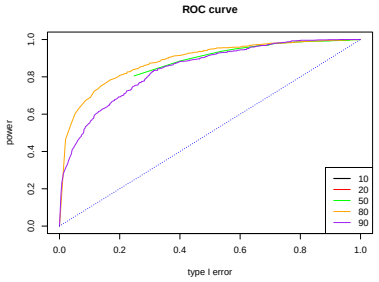
(*а*) Ошибка первого рода (Sum).



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

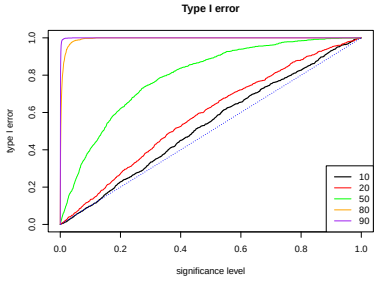


(*в*) ROC-кривая (Sum).

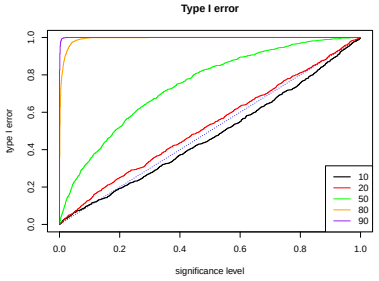


(*г*) ROC-кривая (базовый MSSA).

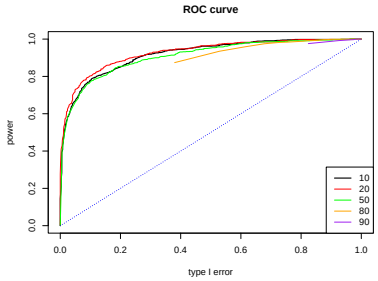
Рис. 2.2. Сравнение методов Sum и базового MSSA (проекция на правые векторы).



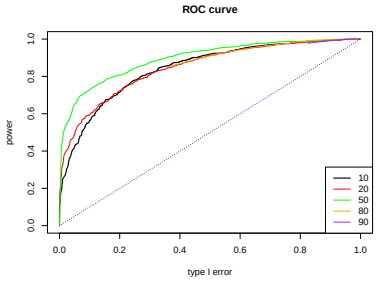
(*а*) Ошибка первого рода (Block).



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

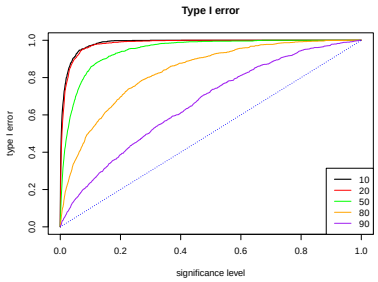


(*в*) ROC-кривая (Block).

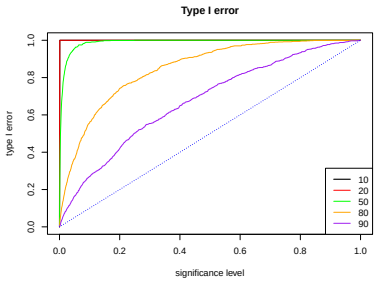


(*г*) ROC-кривая (базовый MSSA).

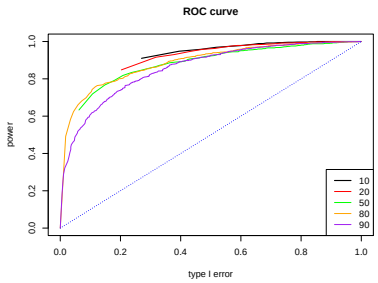
Рис. 2.3. Сравнение методов Block и базового MSSA (проекция на левые векторы).



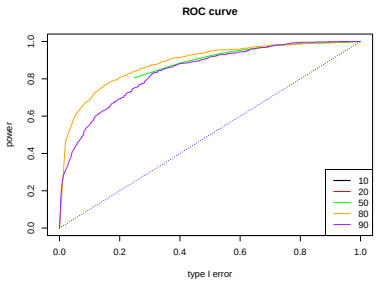
(*а*) Ошибка первого рода (Block).



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



(*в*) ROC-кривая (Block).



(*г*) ROC-кривая (базовый MSSA).

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

удалось построить не на всем промежутке.

Таблица 2.1. 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 2.1 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.

2.8. Выводы

Подведем итоги. На данный момент 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 мы предполагали, что параметры шума известны и нету мешающего сигнала (например, сезонности или тренда). В этой главе рассмотрим случаи, которые более близки к реальным задачам.

3.1. 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]. Параметры красного шума оценивались функцией arima с параметром method="CSS-ML" из пакета stats на языке программирования R.

В этом разделе рассмотрим следующую двухступенчатую оценку параметров: сначала оцениваются параметры на основе исходного ряда и применяется MC-SSA с поправкой, затем выделяется сигнал, если он обнаружится, и оцениваются параметры на основе «остатка». Такая оценка точнее оценивает неизвестные параметры и увеличивает мощность MC-SSA, если верна H_1.

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

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

с амплитудой $A = 1.5$ и частотой $\omega \in (0, 0.5)$. Параметры красного шума и длину ряда N возьмем такими же, как в разделе 2.7. 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. .

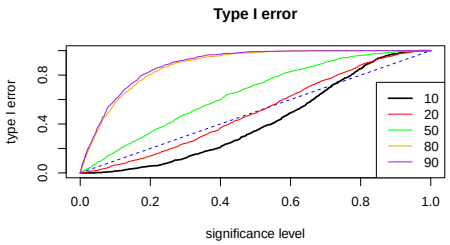
По таблице 3.1 видно, что основной вклад в ошибку оценки вносит смещение, описанная выше процедура это смещение сильно уменьшает, делая слабо-отрицательным.

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

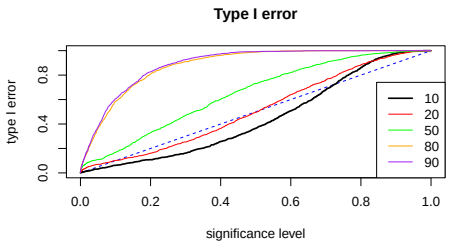
Таблица 3.1. Оценка параметров красного шума

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 [4] и, следовательно, сделает оценку параметров лучше.

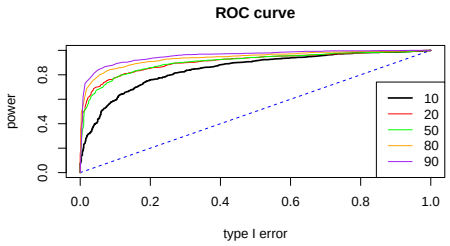
Теперь сравним графики ошибок первого рода и ROC-кривые критерия MC-SSA против альтернативы (3.1) с $\omega = 0.075$. Длины окна L будем брать те же, что и в разделе 2.7.



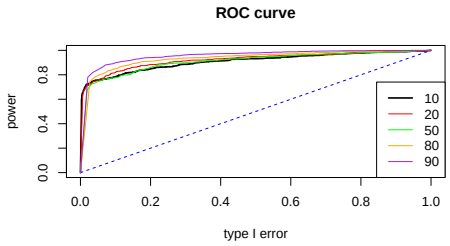
(*а*) Ошибка первого рода (обычная оценка)



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



(*в*) ROC-кривая (обычная оценка)



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

Рис. 3.1. Сравнение обычной и двухступенчатой оценок

По рис. 3.1, *а* и 3.1, *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. 3.1, *в* и 3.1, *г*, заметно повышение мощности при $L = 10, 20, 50$. Поскольку для оптимальной длины окна $L = 90$ разницы в мощности нет, далее будем оценивать параметры шума обычным, не двухступенчатым, способом.

3.2. 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. Будем проверять следующую нулевую гипотезу с альтернативой:

$$H_0 : \mathsf{S} = 0,$$
$$H_1 : \mathsf{S} = \{\cos(2\pi\omega n)\}_{n=1}^{N},$$

где $\omega = 0.075$.

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 $\boldsymbol{\eta}$:

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

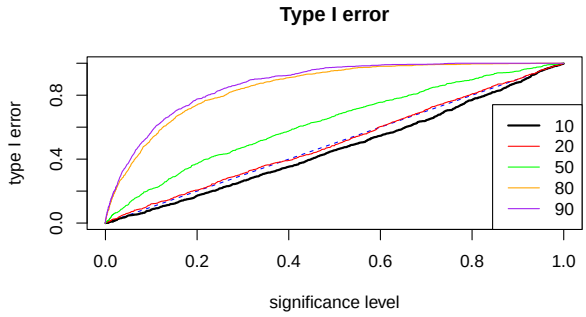
3.2.1. Периодическая компонента

Рассмотрим в качестве мешающего сигнала синусоиду

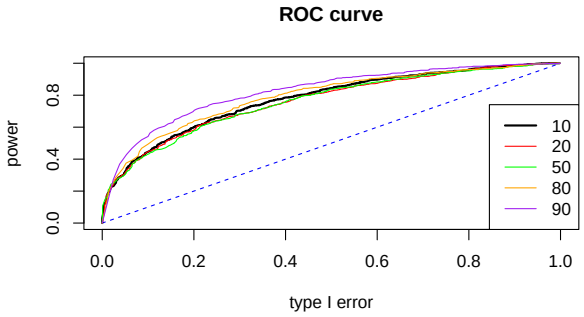
$$f_n = A\cos(2\pi\omega n), \quad n = 1, \ldots, N,$$

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

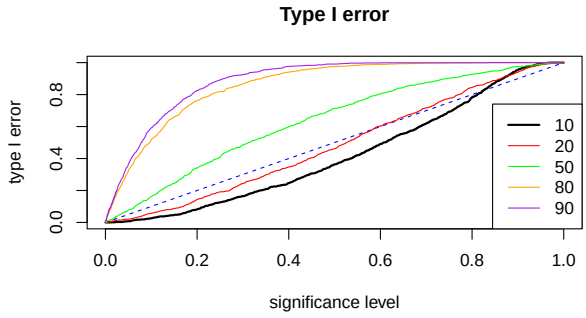
Будем выделять периодическую компоненту при помощи SSA: будем оценивать доминирующую частоту левых векторов с помощью метода ESPRIT [7, Раздел 3.1] и на шаге группировки (раздел 1.1.3) будем брать две компоненты с наиболее близкими к ω частотами.



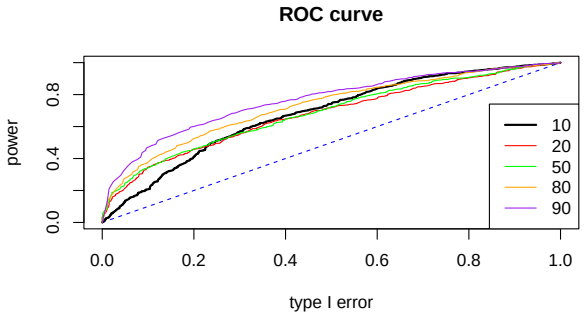
(*а*) Ошибка первого рода



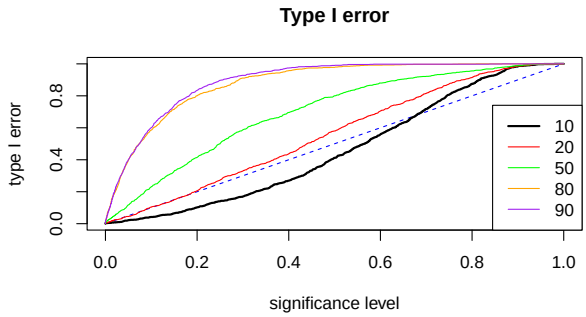
(*б*) ROC-кривая



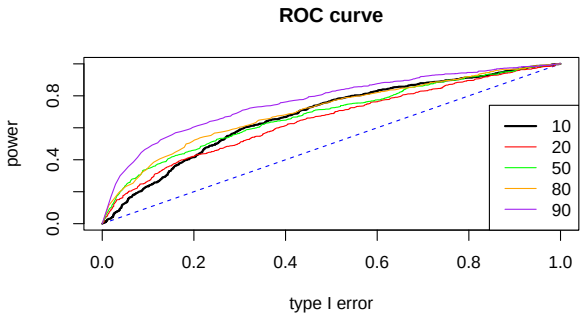
(*в*) Ошибка первого рода (оцененные параметры шума)



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



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



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

Рис. 3.2. Анализ метода, когда мешающий сигнал — периодическая компонента

На рис. 3.2 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. 3.2, *а*, 3.2, *в* и 3.2, *д* похожи друг на друга, а отклонение от случая, когда все известно, можно объяснить погрешностью при оценке неизвестных параметров. После применения поправки из раздела 2.6 критерии становятся точными для любой длины окна и ROC-кривые на рис. 3.2, *б*, 3.2, *г* и 3.2, *е* представляют собой графики мощности этих критериев. Таким образом, наибольшая мощность во всех трех случаях достигается при $L = 90$.

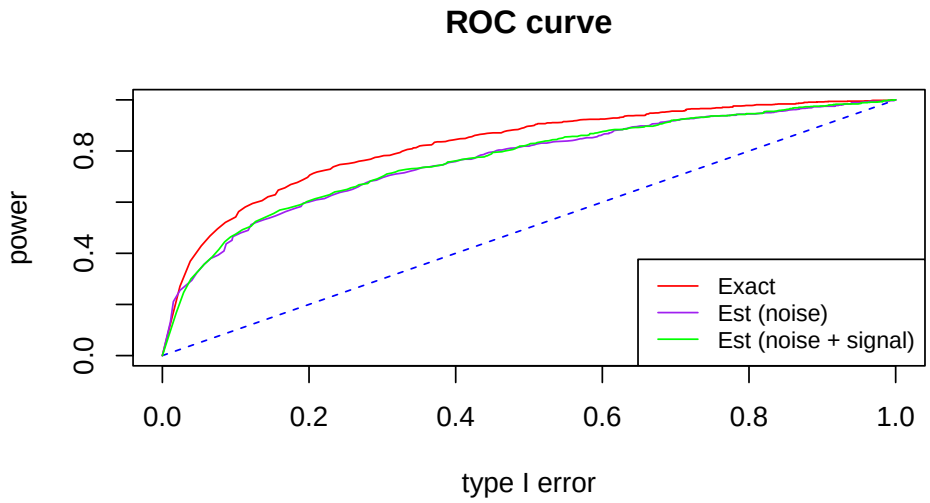


Рис. 3.3. Сравнение мощности критериев (мешающий сигнал — периодическая компонента)

На рис. 3.3 представлена ROC-кривая критериев для оптимального L. Как видно из графика, при оценке параметров шума/мешающего сигнала мощность падает, но незначительно (примерно на 10%). Также отметим, что оценка мешающего сигнала никак не повлияла на мощность.

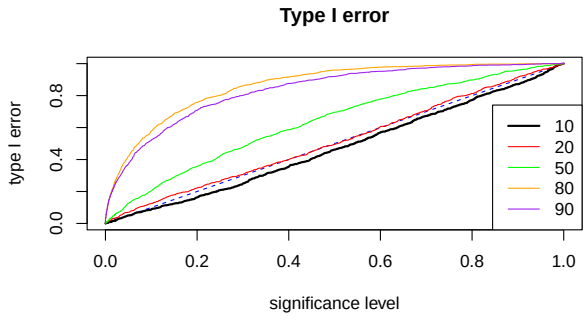
3.2.2. Тренд

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

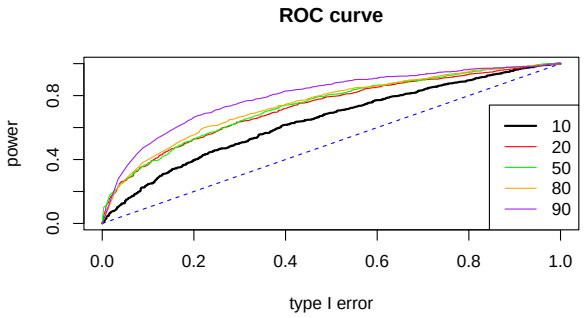
$$k_n = Ae^{\alpha n}, \quad n = 1, \ldots, N,$$

где $A = 0.2$, $\alpha = 0.05$.

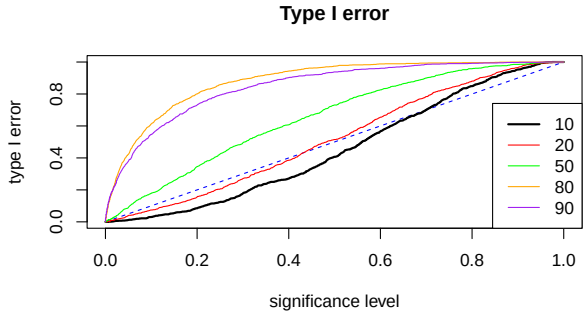
Выделять тренд будем с помощью SSA: поскольку в SVD разложении (1.4) сингулярные числа, соответствующие тренду, будут самыми большими среди всех сингулярных чисел, на шаге группировки (раздел 1.1.3) будем брать первые r элементарных компонент, где r — ранг тренда. В данном случае $r = 1$.



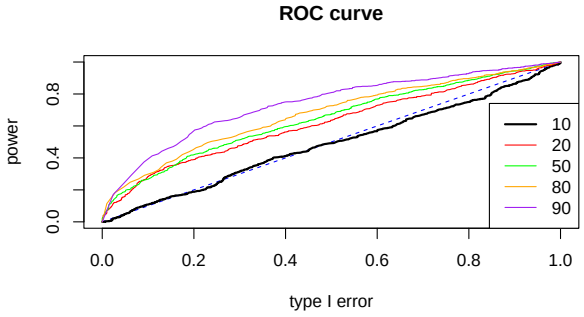
($а$) Ошибка первого рода



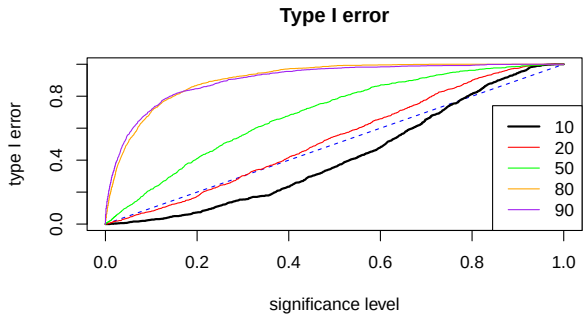
($б$) ROC-кривая



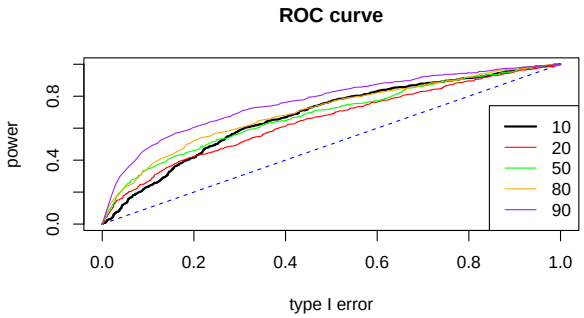
($в$) Ошибка первого рода (оцененные параметры шума)



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



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



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

Рис. 3.4. Анализ метода, когда мешающий сигнал — тренд

На рис. 3.4 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 3.2.1, графики ошибок первого рода на рис. 3.4, *а*, 3.4, *в* и 3.4, *д* сохраняют общую тенденцию при оценке параметров/тренда. По ROC-кривым на рис. 3.4, *б*, 3.4, *г* и 3.2, *е* видно, что оптимальной длиной окна является $L = 90$. Стоит также отметить странное поведение мощности при $L = 10$.

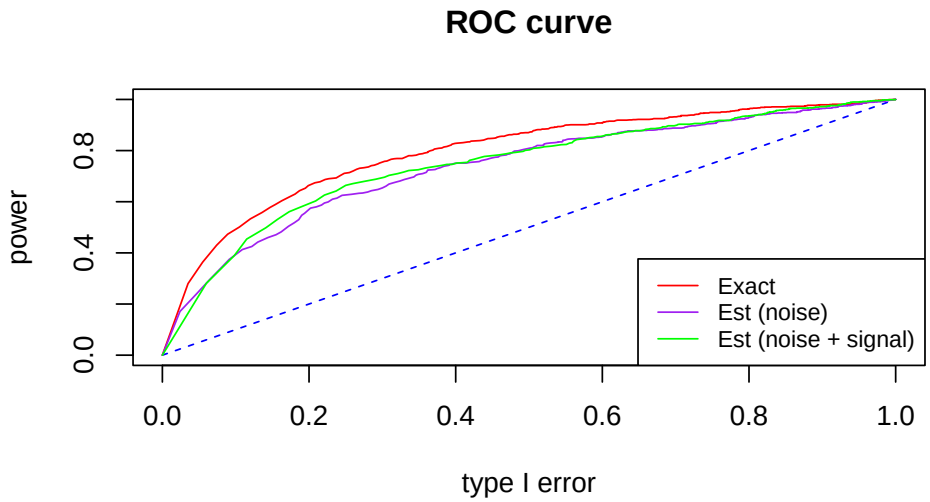


Рис. 3.5. Сравнение мощности критериев (мешающий сигнал — тренд)

На рис. 3.5 представлена ROC-кривая критериев для оптимального L. Аналогично случаю периодической компоненты, при оценке параметров шума/мешающего сигнала мощность падает незначительно. Также отметим, что критерий с оценкой тренда и параметров шума оказался немого мощнее, чем критерий с оценкой только параметров шума.

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. Голяндина Н. Э. Метод «Гусеница»-SSA: анализ временных рядов. — СПбГУ, 2004. — Учебное пособие.
5. 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.
6. 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.
7. Golyandina Nina, Korobeynikov Anton, Zhigljavsky Anatoly. Singular Spectrum Analysis with R. — 2018. — 01. — ISBN: 978-3-662-57378-5.