

Санкт-Петербургский государственный университет
Прикладная математика и информатика

Отчет по учебной практике 3 (научно-исследовательской работе) (семестр 6)

МЕТОД МОНТЕ-КАРЛО SSA ДЛЯ МНОГОМЕРНЫХ ВРЕМЕННЫХ РЯДОВ

Выполнил:

Потешкин Егор Павлович

группа 20.Б04-мм

Научный руководитель:

к. ф.-м. н., доцент

Голяндина Нина Эдуардовна

Кафедра Статистического Моделирования

Санкт-Петербург

2023

Оглавление

Введение	3
Глава 1. Метод MSSA	4
1.1. Описание метода	4
1.2. Модификации метода	6
1.3. Выбор длины окна	7
Глава 2. Метод Monte-Carlo MSSA	9
2.1. Постановка задачи	9
2.2. Одиночный тест	9
2.3. Множественный тест	10
2.4. Выбор векторов для проекции	11
2.5. Численное сравнение методов	12
Заключение	17
Список литературы	18

Введение

TODO

Глава 1

Метод MSSA

1.1. Описание метода

Метод Multivariate Singular Spectrum Analysis (сокращенно MSSA) состоит из четырех этапов: *вложения*, *разложения*, *группировки* и *диагонального усреднения*. Пусть $N_d > 2$, $d = 1, \dots, D$. Рассмотрим вещественнозначные ненулевые одномерные временные ряды $F^{(d)} = (f_1^{(d)}, f_2^{(d)}, \dots, f_{N_d}^{(d)})$. Составим из этих рядов $F = \{F^{(d)}\}_{d=1}^D$ — D -канальный временной ряд с длинами N_d , $d = 1, \dots, D$.

1.1.1. Вложение

Выберем параметр L , называемый *длиной окна*, $1 < L < \min(N_1, \dots, N_D)$. Для каждого ряда $F^{(d)}$ рассмотрим $K_d = N - L + 1$ векторов вложения $X_i^{(d)} = (f_i^{(d)}, \dots, f_{i+L-1}^{(d)})^T$, $1 \leq j \leq K_d$ и составим *траекторную матрицу* $\mathbf{X}^{(d)} = [X_1^{(d)} : \dots : X_{K_d}^{(d)}]$. Обозначим $K = \sum_{d=1}^D K_d$. Результатом этапа вложения является матрица размера $L \times K$

$$\mathbf{X} = [\mathbf{X}^{(1)} : \dots : \mathbf{X}^{(D)}]. \quad (1.1)$$

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 $\mathbf{S} = \mathbf{X}\mathbf{X}^T$. 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

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

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

1.1.3. Группировка

На этом шаге множество индексов $I = \{1, \dots, p\}$ 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 $\mathbf{X}_{I_k} = \sum_{i \in I_k} \mathbf{X}_i$.

1.1.4. Диагональное усреднение

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

Пусть $\mathbf{Y} = (y_{ij})$ — матрица размера $L \times K$. Положим $L^* = \min(L, K)$, $K^* = \max(L, K)$ и $N = L + K - 1$. Пусть $y_{ij}^* = y_{ij}$, если $L < K$, и $y_{ij}^* = y_{ji}$ иначе. *Диагональное усреднение* переводит матрицу \mathbf{Y} в ряд g_1, \dots, g_N по формуле

$$g_k = \begin{cases} \frac{1}{k} \sum_{m=1}^k y_{m,k-m+1}^*, & \text{при } 1 \leq k < L^* \\ \frac{1}{L^*} \sum_{m=1}^{L^*} y_{m,k-m+1}^*, & \text{при } L^* \leq k \leq K^* \\ \frac{1}{N-k+1} \sum_{m=k-K^*+1}^{N-K^*+1} y_{m,k-m+1}^*, & \text{при } K^* < k \leq N \end{cases}$$

Из (1.1) следует, что \mathbf{X}_{I_k} можно представить в следующем виде:

$$\mathbf{X}_{I_k} = [\mathbf{X}_{I_k}^{(1)} : \dots : \mathbf{X}_{I_k}^{(D)}].$$

Тогда, чтобы получить D -канальный временной ряд, применим диагональное усреднение к каждой матрице $\mathbf{X}_{I_k}^{(d)}$, $d = 1, \dots, D$.

1.1.5. Частный случай

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

1.2. Модификации метода

1.2.1. Тёплицев MSSA

В случае анализа стационарных рядов можно улучшить базовый метод, используя другое разложение. Для начала введем следующее понятие.

Определение 1. Пусть $F = (f_1, \dots, f_N)$ — одномерный временной ряд и L — фиксированное. **Тёплицевой L -ковариационной матрицей** называют матрицу $\tilde{\mathbf{C}}$ с элементами

$$\tilde{c}_{ij} = \frac{1}{N - |i - j|} \sum_{n=1}^{N-|i-j|} f_n f_{n+|i-j|}, \quad 1 \leq i, j \leq L.$$

Пусть теперь $F = \{F^{(d)}\}_{d=1}^D$ — D -канальный временной ряд, каждый канал которого имеет одинаковую длину N , $K = N - L + 1$. Тогда можно получить разложение \mathbf{X} двумя способами:

1. Пусть $\tilde{\mathbf{C}}_1, \dots, \tilde{\mathbf{C}}_D$ — тёплицевы матрицы для каждого канала. Рассмотрим $\tilde{\mathbf{C}} = \sum_{d=1}^D \tilde{\mathbf{C}}_d$. Найдем ортонормированные собственные векторы H_1, \dots, H_L матрицы $\tilde{\mathbf{C}}$ и разложим траекторную матрицу \mathbf{X} следующим образом:

$$\mathbf{X} = \sum_{i=1}^L \sigma_i H_i Q_i^T, \quad (1.2)$$

где $Z_i = \mathbf{X}^T U_i$, $Q_i = Z_i / \|Z_i\|$ и $\sigma_i = \|Z_i\|$.

2. Можно рассмотреть блочную матрицу размера $DK \times DK$:

$$\mathbf{T} = \begin{pmatrix} \mathbf{T}_{1,1} & \mathbf{T}_{1,2} & \cdots & \mathbf{T}_{1,D} \\ \mathbf{T}_{2,1} & \mathbf{T}_{2,2} & \cdots & \mathbf{T}_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{T}_{D,1} & \mathbf{T}_{D,D} & \cdots & \mathbf{T}_{D,D} \end{pmatrix}.$$

Элементы каждого блока \mathbf{T}_{lk} имеют вид

$$t_{ij}^{(lk)} = \frac{1}{\tilde{N}} \sum_{n=\max(1, 1+i-j)}^{\min(N, N+i-j)} f_n^{(l)} f_{n+j-i}^{(k)}, \quad 1 \leq i, j \leq K,$$

где $\tilde{N} = \min(N, N + i - j) - \max(1, 1 + i - j) + 1$. Найдя ортонормированные собственные векторы Q_1, \dots, Q_{DK} матрицы \mathbf{T} , получаем разложение

$$\mathbf{X} = \sum_{i=1}^{DK} \sigma_i H_i Q_i^T, \quad (1.3)$$

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

Шаги группировки и диагонального усреднения можно оставить в том виде, в котором они представлены в разделе 1.1.3 и в разделе 1.1.4.

Для конкретности, будем называть первый метод Sum, а второй — Block. Стоит отметить, что в Sum собственные векторы матрицы $\tilde{\mathbf{C}}$ — аналоги левых сингулярных векторов матрицы \mathbf{X} , в то время как в Block собственные векторы матрицы \mathbf{T} — аналоги правых сингулярных векторов.

1.3. Выбор длины окна

Посмотрим на точность базового и модифицированных методов, для разных значений параметра L , на подобию работы [2]. Рассмотрим следующий двухканальный временной ряд: $(F^{(1)}, F^{(2)}) = (H^{(1)}, H^{(2)}) + (N^{(1)}, N^{(2)})$, где $H^{(1)}$, $H^{(2)}$ — гармоники, а $N^{(1)}$, $N^{(2)}$ — независимые реализации гауссовского белого шума. *Гауссовский белый шум* — стационарный случайный процесс, имеющий нормальное распределение. Как и в [2], пусть $N = 71$, дисперсия шумовых компонент $\sigma^2 = 25$, число повторений равно 10000. Рассмотрим 2 случая:

Случай 1	$L = 12$	$L = 24$	$L = 36$	$L = 48$	$L = 60$
MSSA	3.18	1.83	1.59	1.47	2.00
SSA	3.25	2.01	2.00	2.01	3.25
Sum	3.17	1.75	1.44	1.32	1.33
Block	1.39	1.26	1.25	1.33	1.97
Случай 2	$L = 12$	$L = 24$	$L = 36$	$L = 48$	$L = 60$
MSSA	6.91	3.77	3.07	2.88	3.84
SSA	3.23	2.01	2.00	2.01	3.23
Sum	6.88	3.65	2.64	2.37	2.27
Block	4.47	3.67	3.22	3.23	3.8

Таблица 1.1. MSE восстановления сигнала.

1. Одинаковые периоды:

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

2. Разные периоды:

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

В таблице 1.1 представлены результаты восстановления сигнала для разных L . Данные для методов SSA и MSSA были взяты из работы [2]. Наиболее точные результаты для каждого метода были выделены жирным шрифтом. Как видим из таблицы 1.1, в обоих случаях метод Sum показывал наилучший результат для $L > (N + 1)/2$, в то время как метод Block наиболее точен при длине окна, близкой к половине длины ряда, причем оба метода в случае одинаковых периодов показывают лучше результат, чем MSSA.

Глава 2

Метод Monte-Carlo MSSA

2.1. Постановка задачи

Рассмотрим задачу поиска сигнала (не случайной составляющей) в многоканальном временном ряде. Нулевая гипотеза H_0 — отсутствие сигнала (ряд состоит из чистого шума). Тогда альтернатива H_1 — ряд содержит сигнал, например, периодическая составляющая.

Определение 2. Случайный вектор $\boldsymbol{\xi} = (\xi_1, \dots, \xi_N)$ 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 данной главе под шумом будем подразумевать именно красный. Также будем рассматривать только односторонние критерии.

2.2. Одиночный тест

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

$$p = \|\boldsymbol{\Theta}^T W_k\|^2.$$

Статистикой критерия является величина

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

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

Рассмотрим алгоритм статистического критерия проверки наличия сигнала в ряде с проекцией на один вектор W , описанный в работе [3].

Алгоритм 1. Одиночный тест [3]

1. Построить статистику критерия \hat{p} .

2. Построить доверительную область случайной величины p : интервал от нуля до γ -квантиля.
3. Если \hat{p} не попадает в построенный интервал — H_0 отвергается.

Построенная доверительная область называется *прогнозируемым интервалом* с уровнем доверия γ .

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

2.3. Множественный тест

Пусть теперь частоты периодических компонент неизвестны (что не редкость на практике), но известен диапазон частот и нужно проверить, что в ряде присутствует сигнал с хотя бы одной частотой из заданного диапазона. Тогда нулевая гипотеза H_0 о том, что ряд не содержит сигнала ни на одной из частот из рассматриваемого диапазона, а альтернатива H_1 — ряд содержит сигнал с хотя бы одной частотой, принадлежащей рассматриваемому диапазону.

Пусть W_1, \dots, W_H — вектора для проекции. В таком случае нужно построить H предсказательных интервалов по выборкам $P_k = \{p_{ki}\}_{i=1}^G$ с элементами

$$p_{ki} = \|\Xi_i^T W_k\|^2, \quad i = 1, \dots, G; \quad k = 1, \dots, H, \quad (2.1)$$

где G — количество суррогатных реализаций ξ , Ξ_i — траекторная матрица i -й реализации ξ .

В работе [3] подробно описана проблема многократного 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 [3]

1. Для $k = 1, \dots, H$ вычисляется статистика \hat{p}_k , выборка $P_k = \{p_{ki}\}_{i=1}^G$, ее среднее μ_k и стандартное отклонение σ_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_k как выборочный $(1 - \alpha)$ -квантиль η .
4. Нулевая гипотеза не отвергается, если

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

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

2.4. Выбор векторов для проекции

Для начала отметим, что в одномерном случае можно рассматривать как проекции на собственные вектора, так и на факторные — не имеет значения, поскольку это ни на что кроме размерности не влияет. А в многомерном случае это не так по построению матрицы (1.1), поэтому их нужно рассматривать по-отдельности.

Перечислим основные способы выбора векторов для проекции. Первый вариант — рассматривать собственные вектора теоретической матрицы красного шума. При рассмотрении собственных векторов матрица, разложение которой дает эти собственные векторы имеет вид $\sum_{d=1}^D \{\varphi^{|i-j|}\}_d$, а при рассмотрении факторных векторов матрица имеет вид $\text{diag}_{d=1, \dots, D} \{\varphi^{|i-j|}\}_d$. Такой вариант в обоих случаях дает точный критерий при любой длине окна.

Второй вариант — рассматривать собственные или факторные векторы матрицы **X**. Этот вариант вообще радикальный, но, используя поправку, описанную в работе [4], можно сделать такой критерий точным. Если рассматривать собственные векторы, то их длина равна L , а сами векторы отображают общую структуру для всех рядов. Если же рассматривать факторные, то их длина равна $D(N - L + 1)$, а сами векторы имеют составную структуру, где каждому ряду соответствует вектор размера $N - L + 1$.

Определение 3. 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 неточных (в частности радикальных) критериев. Отметим, что для точного критерия ROC-кривая совпадает с графиком мощности.

2.5. Численное сравнение методов

В одномерном случае было установлено, что если вместо SVD разложения матрицы \mathbf{X} использовать тёплицево, то радикальность критерия уменьшается. Установим, что будет в многомерном случае, если использовать модификации, описанные в разделе 1.2.1. Пусть количество каналов равно двум, количество суррогатных реализаций красного шума $G = 1000$. Для оценки ошибки первого рода, будем рассматривать красный шум с параметрами $\varphi = 0.7$ и $\delta = 1$, а для оценки мощности будем рассматривать

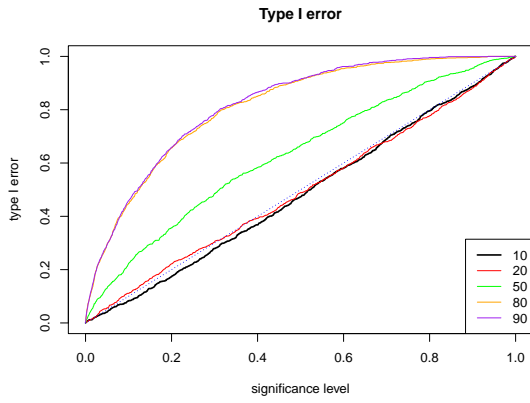
$$f_n^{(1)} = f_n^{(2)} = \cos(2\pi\omega n), \quad n = 1, \dots, 100,$$

где $\omega = 0.075$.

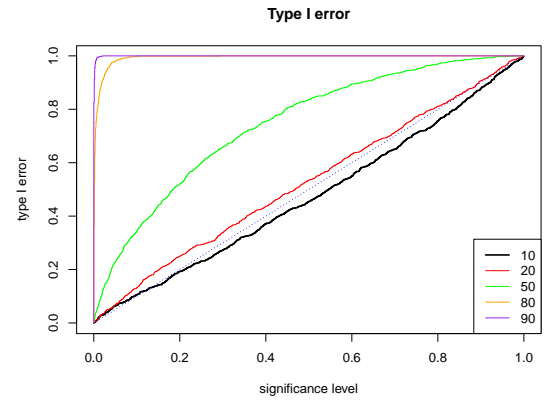
Построим графики ошибки первого рода и ROC-кривые для каждой длины окна $L = 10, 20, 50, 80, 90$. Будем воспринимать ROC-кривую как график мощности критерия, к которому была применена поправка, описанная в работе [4].

На рис. 2.1 и 2.2 векторы для проекции были взяты из разложения (1.2). На рис. 2.1 видно, что при $L > 20$ метод радикальный, а наибольшая мощность достигается при $L = 90$. На рис. 2.2 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.

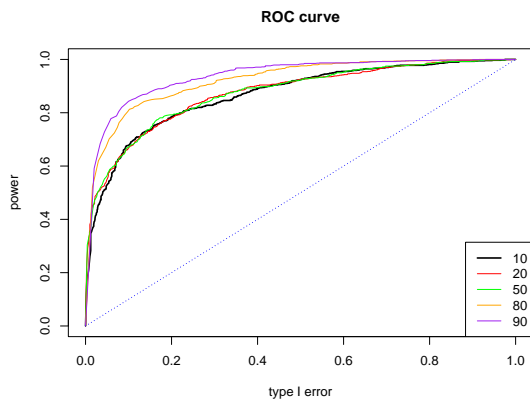
На рис. 2.3 и 2.4 векторы для проекции были взяты из разложения (1.3). Если рассматривать проекцию на собственные векторы, то на рис. 2.3 видно, что метод радикальный, а наиболее оптимальным значением длины окна будет $L = 20$. Проекция



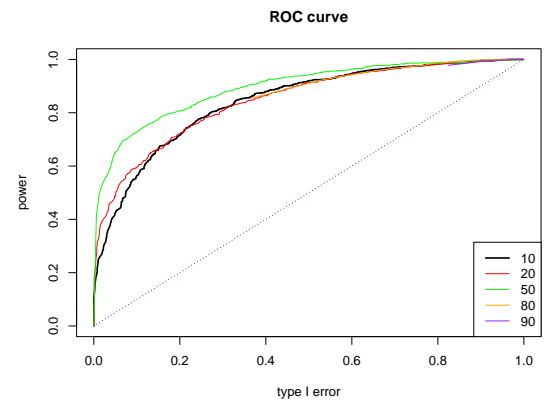
(a) Ошибка первого рода (Sum).



(б) Ошибка первого рода (базовый MSSA).



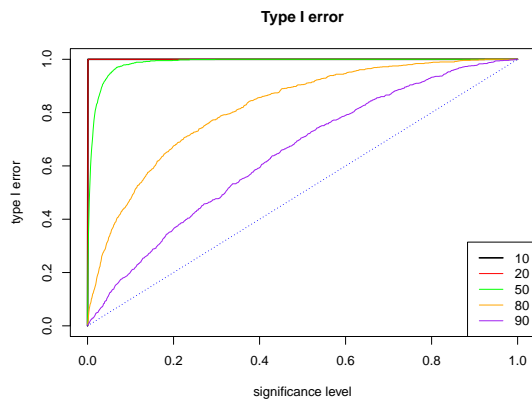
(в) ROC-кривая (Sum).



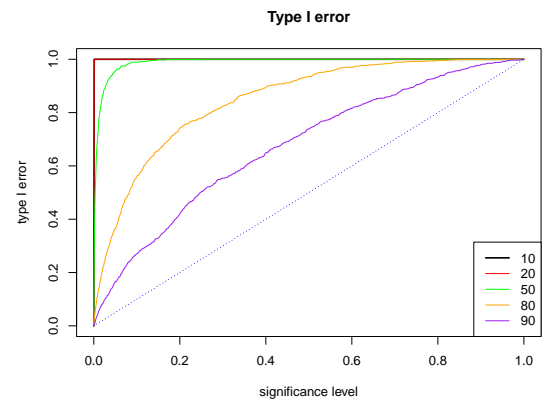
(г) ROC-кривая (базовый MSSA).

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

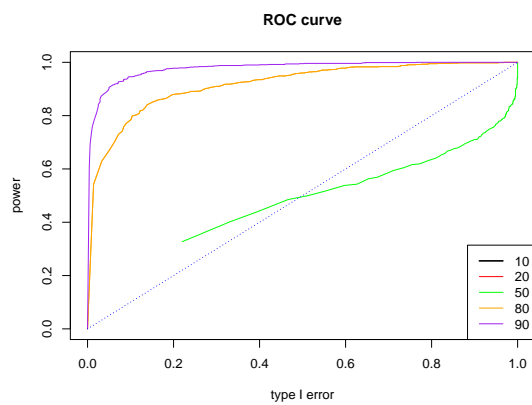
на факторные векторы также дает радикальный критерий, как видно на рис. 2.4. Наибольшая мощность наблюдается при $L = 80$, но из-за слишком большой ошибки первого рода ROC-кривую для $L = 10$ и $L = 20$, для которых метод, предположительно, имеет большую мощность, удалось построить не на всем промежутке.



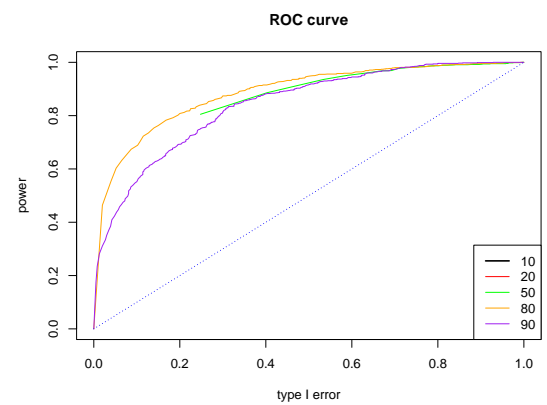
(a) Ошибка первого рода (Sum).



(б) Ошибка первого рода (базовый MSSA).

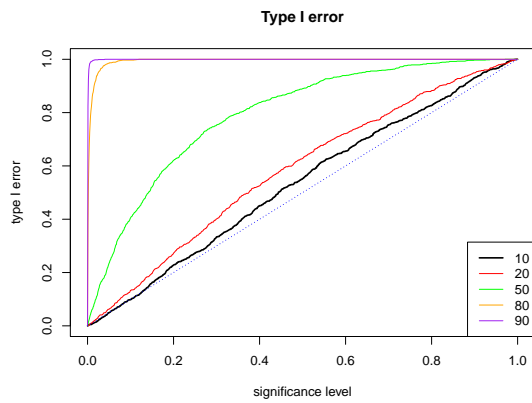


(в) ROC-кривая (Sum).

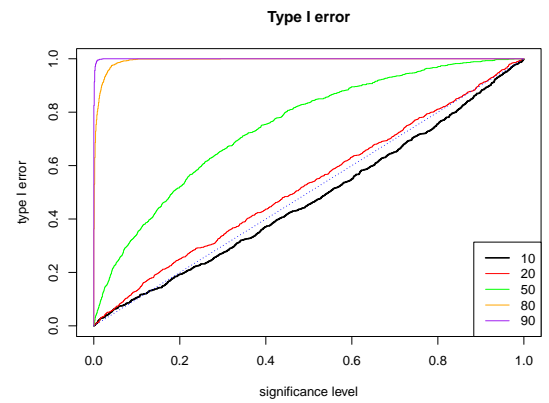


(г) ROC-кривая (базовый MSSA).

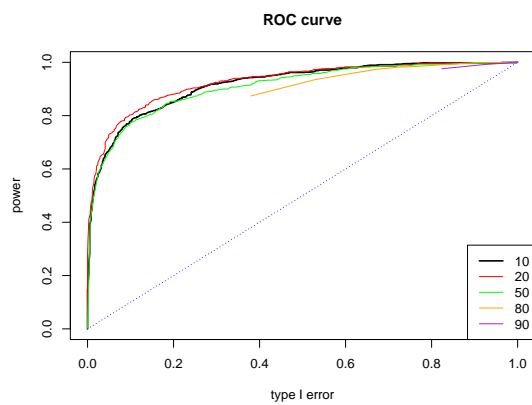
Рис. 2.2. Сравнение методов Sum и базового MSSA (проекция на факторные векторы).



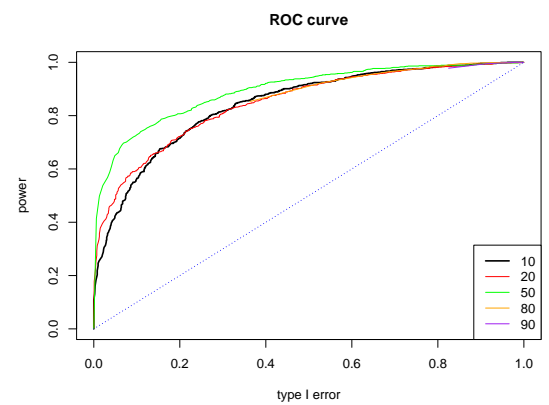
(a) Ошибка первого рода (Block).



(б) Ошибка первого рода (базовый MSSA).

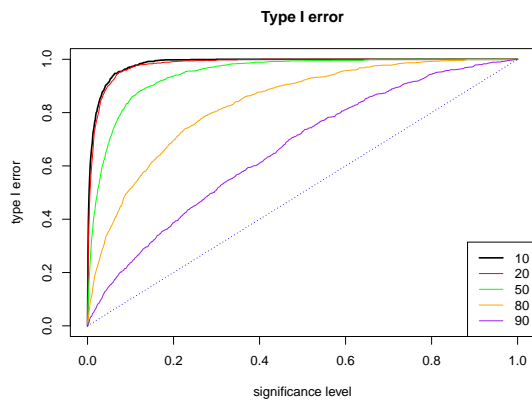


(в) ROC-кривая (Block).

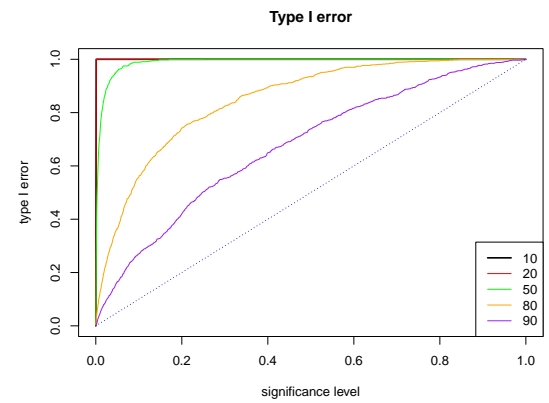


(г) ROC-кривая (базовый MSSA).

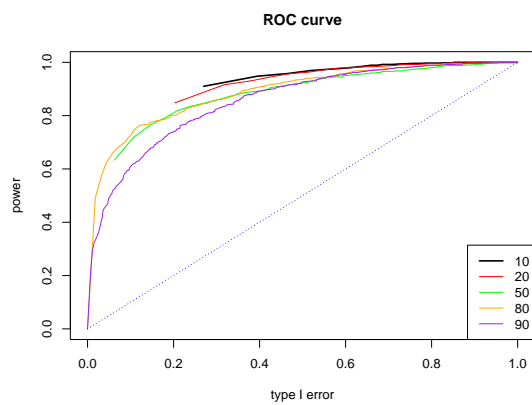
Рис. 2.3. Сравнение методов Block и базового MSSA (проекция на собственные векторы).



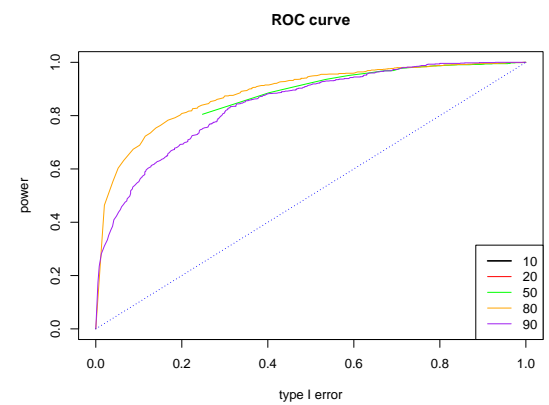
(a) Ошибка первого рода (Block).



(б) Ошибка первого рода (базовый MSSA).



(в) ROC-кривая (Block).



(г) ROC-кривая (Block).

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

Заключение

TODO

Список литературы

1. Голяндина Н. Э. Метод «Гусеница»-SSA: анализ временных рядов. — СПбГУ, 2004. — Учебное пособие.
2. Multivariate and 2D Extensions of Singular Spectrum Analysis with the Rssa Package / Golyandina N., Korobeynikov, A., Shlemov A., and Usevich K. // Journal of Statistical Software. — 2015.
3. Golyandina N. Detection of signals by Monte Carlo singular spectrum analysis: Multiple testing // arXiv preprint. — 2021. — arXiv : stat.ME/arXiv:1903.01485v2.
4. Ларин Е. С. Метод SSA для проверки гипотезы о существовании сигнала во временном ряде : квалификационная работа магистра ; СПбГУ. — 2022.