«Метод SSA для проверки гипотезы о существовании сигнала во временном ряде»

Презентация ВКР

Ларин Евгений Сергеевич, группа 20.М03-мм

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель: к.ф.-м.н., доцент Голяндина Н.Э.
Рецензент: Программист, Майкрософт Израиль, Шлемов А.Ю.

Санкт-Петербург
2022г.

Введение

Дано: Ряд $\mathsf{X} = (x_1, \ldots, x_N)$, удовлетворяющий модели $\mathsf{X} = S + R$, где S — сигнал, а R — шум.

Общая проблема:
Для того, чтобы проанализировать какую-либо зависимость, необходимо убедиться, что в ряде есть сигнал. В противном случае можно ошибочно анализировать что-то похожее на сигнал, что на самом деле — часть шума.

Задачи работы:

1. Разработать систему корректировки критериев и сравнения результатов по ROC-кривым. При этом предлагается и вариант с построением γ-гарантированного точного критерия.
2. Сравнить разные модификации критериев MC-SSA, выработать рекомендации по выбору параметров, рассмотреть вариант применения системы из п.1.

Часть I. Техника сравнения критериев по мощности.

Дано: Набор данных, *нулевая гипотеза* H_0, *альтернативная гипотеза* H_1 и уровень значимости α^*.

Задача: Проверить истинность H_0 против альтернативы H_1.

Способ решения: Построить *статистику критерия* (обозначение: t) и разбить диапазон её значений на *доверительную* и *критическую* (обозначение: T_{α^*}) области так, чтобы $P_{H_0}(t \in T_{\alpha^*}) = \alpha^*$.

Получение ответа: Если $t(X) \in T_{\alpha^*}$, то H_0 отвергается; иначе нет оснований отвергать гипотезу H_0.

Ошибка первого рода α_I. $\alpha_I = P_{H_0}(t \in T_{\alpha^*})$. При множественной проверке гипотез применяется $FWER = P_{H_0}(\bigvee_i t_i \in T_{\alpha^* i})$.

Мощность β. $\beta = P_{H_1}(t \in T_{\alpha^*})$.

Классификация критериев на основе α_I:

1. *Радикальный:* $\alpha_I > \alpha^*$. Не может применяться на практике.
2. *Точный:* $\alpha_I = \alpha^*$. Может применяться без поправки.
3. *Консервативный:* $\alpha_I < \alpha^*$. Может применяться на практике, но мощность меньше, чем у точного критерия.

Дано:

1. Нулевая гипотеза H_0: $\xi \in \mathcal{P}(0, \Theta)$;
2. Уровень значимости α^*;
3. Радикальный или консервативный критерий $t(X)$ (если $t(X) \in T_{\alpha^*}$, то H_0 отвергается);
4. Количество моделируемых данных M_1;
5. Параметры модели Θ;
6. Объём выборки N;

Выход алгоритма: Формальный уровень значимости $\widetilde{\alpha}^*$. Критерий с этим уровнем значимости (если $t(X) \in T_{\widetilde{\alpha}^*}$, то H_0 отвергается) является асимптотически точным при $M_1 \to \infty$, т.е. ошибка первого рода асимптотически по M_1 стремится к уровню значимости α^*.

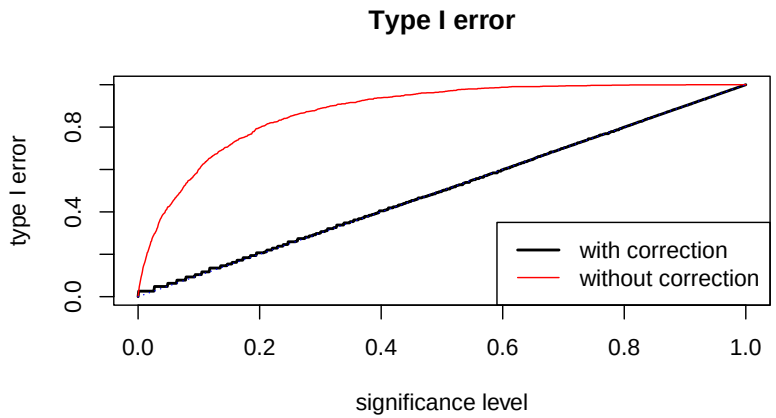
1. Моделируется M_1 выборок объёма N согласно распределению $\mathcal{P}(0, \Theta)$;
2. По смоделированным данным строится зависимость ошибки первого рода от уровня значимости $\alpha_I(\alpha)$. Для каждой выборки X_i строится $t(X_i)$ и находится вероятностный уровень p_i. Зависимость $\alpha_I(\alpha)$ оценивается как эмпирическая функция распределения полученной выборки $p_1, \ldots, p_{M_1}$;
3. Рассчитывается формальный уровень значимости с помощью функции, обратной к $\alpha_I(\alpha)$, полученной на предыдущем шаге: $\tilde{\alpha}^* = \alpha_I^{-1}(\alpha^*)$;

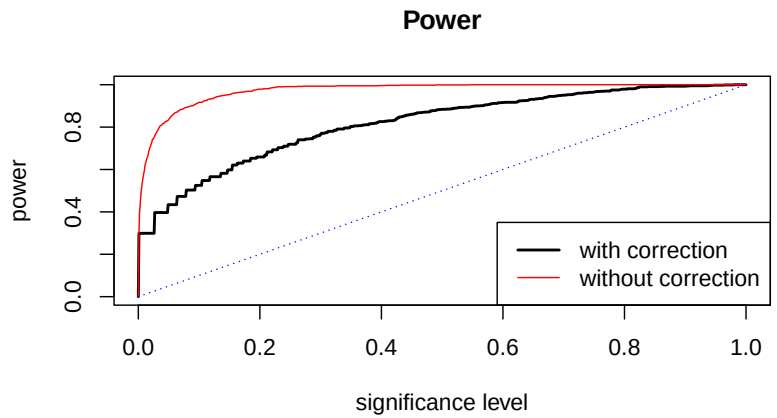
Примечания:

1. Полученный критерий (если $t(X) \in T_{\tilde{\alpha}^*}$, то H_0 отвергается) является асимптотически точным при $M_1 \to \infty$.
2. Если параметры Θ не заданы, то перед применением алгоритма строится оценка $\widehat{\Theta}$ по исходным данным X, которая и передаётся в алгоритм вместо параметров Θ. Асимптотическая точность при этом теряется.

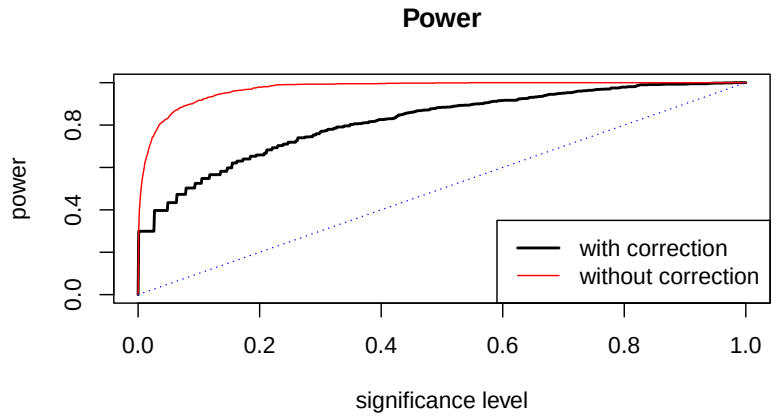
Часть I. Корректировка критерия и потенциальная мощность по ROC-кривой

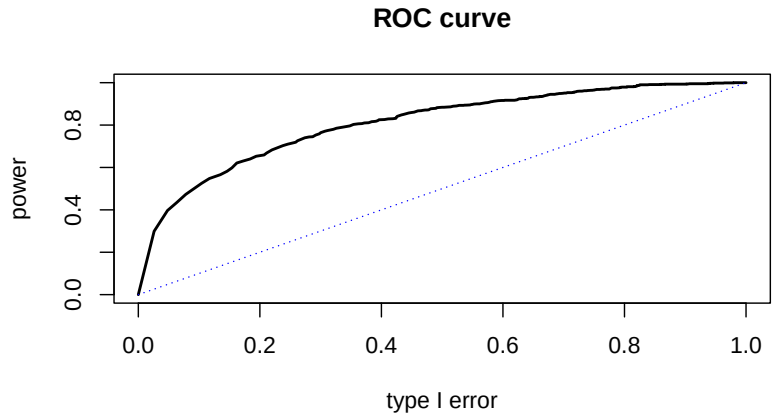
Корректировка уровня значимости: $\tilde{\alpha}^* = \alpha_I^{-1}(\alpha^*)$.

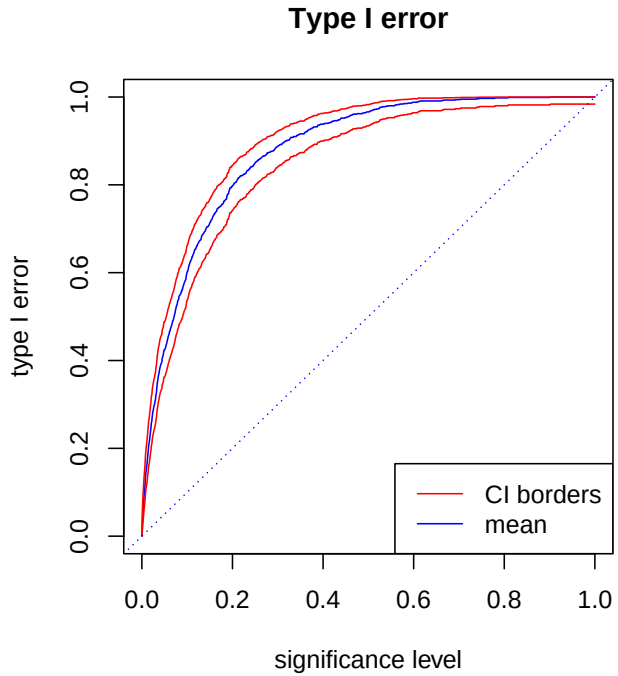


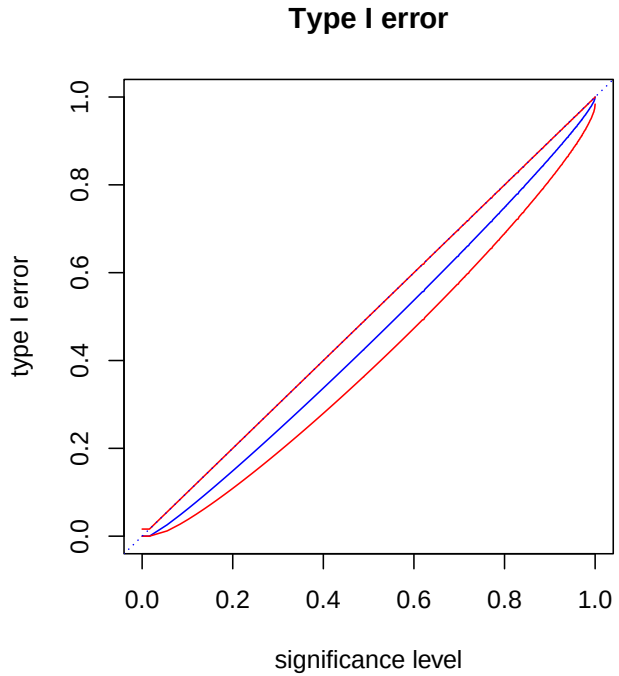


- *Ошибка первого рода*: $\alpha_I(\alpha) = P_{H_0}(t \in T_\alpha)$;
- *Мощность*: $\beta(\alpha) = P_{H_1}(t \in T_\alpha)$;
- ROC-кривая: $y = \beta(\alpha)$, $x = \alpha_I(\alpha)$.









Замечание

Пусть в алгоритме корректировки получен γ-доверительный интервал $(f_{upp}(\alpha), f_{low}(\alpha))$ для $\alpha_I(\alpha)$. Тогда критерий, у которого ошибка первого рода равна α^* с вероятностью γ, можно получить при формальном уровне значимости $\widetilde{\alpha}^* = f_{upp}^{-1}(\alpha^*)$.

Соответственно, это приведет к изменению ROC-кривой и уменьшению мощности скорректированного критерия.

Критерий с корректировкой $\widetilde{\alpha}^* = f_{upp}^{-1}(\alpha^*)$ будем называть γ-гарантированным, имея в виду, что он будет нерадикальным с вероятностью γ.

Часть II. Применение к задаче обнаружения сигнала методом Monte Carlo SSA

Перед тем как сформулировать задачу, кратко опишем метод SSA.

Дано: ряд $\mathsf{X} = (x_1, \ldots, x_N)$, предполагая, что $\mathsf{X} = S + E$.

Параметры: длина окна L, способ группировки элементарных матриц.

1. *Вложение*: $\mathbf{X} = \begin{pmatrix} x_1 & x_2 & \ldots & x_K \\ x_2 & x_3 & \ldots & x_{K+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & \ldots & x_N \end{pmatrix}$ — траекторная матрица.
2. *Разложение*: $\mathbf{X} = \sum_{i=1}^{d} \mathbf{X}_i = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$
   - **SVD**: $\sqrt{\lambda_i}$ — с.ч. $\mathbf{X}\mathbf{X}^{\mathrm{T}}$, U_i — с.в. $\mathbf{X}\mathbf{X}^{\mathrm{T}}$, $V_i = \mathbf{X}U_i/\sqrt{\lambda_i}$
   - **Toepliz**: U_i — с.в. $\widetilde{\mathbf{C}} = \{\frac{1}{N-|i-j|-1} \sum_{m=0}^{N-|i-j|} x_{m+1} x_{m+1+|i-j|}\}_{1 \le i,j \le L}$, $\lambda_i = ||\mathbf{X}^{\mathrm{T}} U_i||^2$, $V_i = \frac{\mathbf{X}^{\mathrm{T}} U_i}{\sqrt{\lambda_i}}$
3. *Группировка*: $\widetilde{\mathbf{X}} = \sum_{i \in I} \mathbf{X}_i$, где I — набор индексов элементарных матриц, соответствующих сигналу.
4. *Восстановление*: $\widetilde{S} = \mathcal{H}\widetilde{\mathbf{X}}$

Результат: Восстанровленный сигнал $\tilde{S}$.

Примечание: В обоих вариантах шага Разложение $d = \min(L, N - L + 1)$.

Дано:

- X $= S + E$;
- Сигнал:
  - Модулированная синусоида $s_n = Ae^{\alpha n}\cos(2\pi\omega n + \phi)$.
- Шум:
  - Авторегрессия первого порядка $\xi_n = \varphi\xi_{n-1} + \varepsilon_n$, $0 < \varphi < 1$.
  - ε_n — белый шум с нулевым средним и дисперсией δ^2.
  - ε_1 имеет нулевое среднее и дисперсию $\frac{\delta^2}{1-\varphi^2}$.

Задача: Проверка гипотезы H_0, что ряд состоит из чистого шума $(S = 0)$.

Способ решения: Monte-Carlo SSA [Allen et al. (1996)]:

- Выбрать векторы для проекции W_i, $i = 1, \ldots, H$, соответствующие разным частотам;
- Построить статистику критерия $\widehat{p} = ||\mathbf{X}^{\mathrm{T}} W_i||^2$, где $\mathbf{X}$ — траекторная матрица;
- Построить предсказательные интервалы с помощью метода Монте-Карло (суррогатных данных).

Связь с методом SSA: стандартный способ выбора W — с.в. U_i SSA разложения. По свойствам SSA, если в ряде есть сигнал, то пара векторов соответствует частоте сигнала.

Проблемы применения критерия:

- Из-за проблемы множественного тестирования нужна поправка: если поправка Бонферрони — критерий консервативный, если Multiple MC-SSA [Бояров (2012), Golyandina (2021)] — точный, при определённом способе построения W_i.
- Оценка параметров шума меняет критерий в сторону консервативности, делает критерий менее мощным.
- Критерии с проекцией на собственные вектора SSA радикальные, так как набор собственных векторов исходного ряда строится по ряду и значения статистики критерия завышены.

Решение: Применение методики коррекции из части I. Теперь можно использовать любые критерии после коррекции.

Есть несколько модификаций MC-SSA. Так как есть алгоритм коррекции, то не важно точные они или нет, можно сравнить их точные модификации.

Обозначения для критериев:

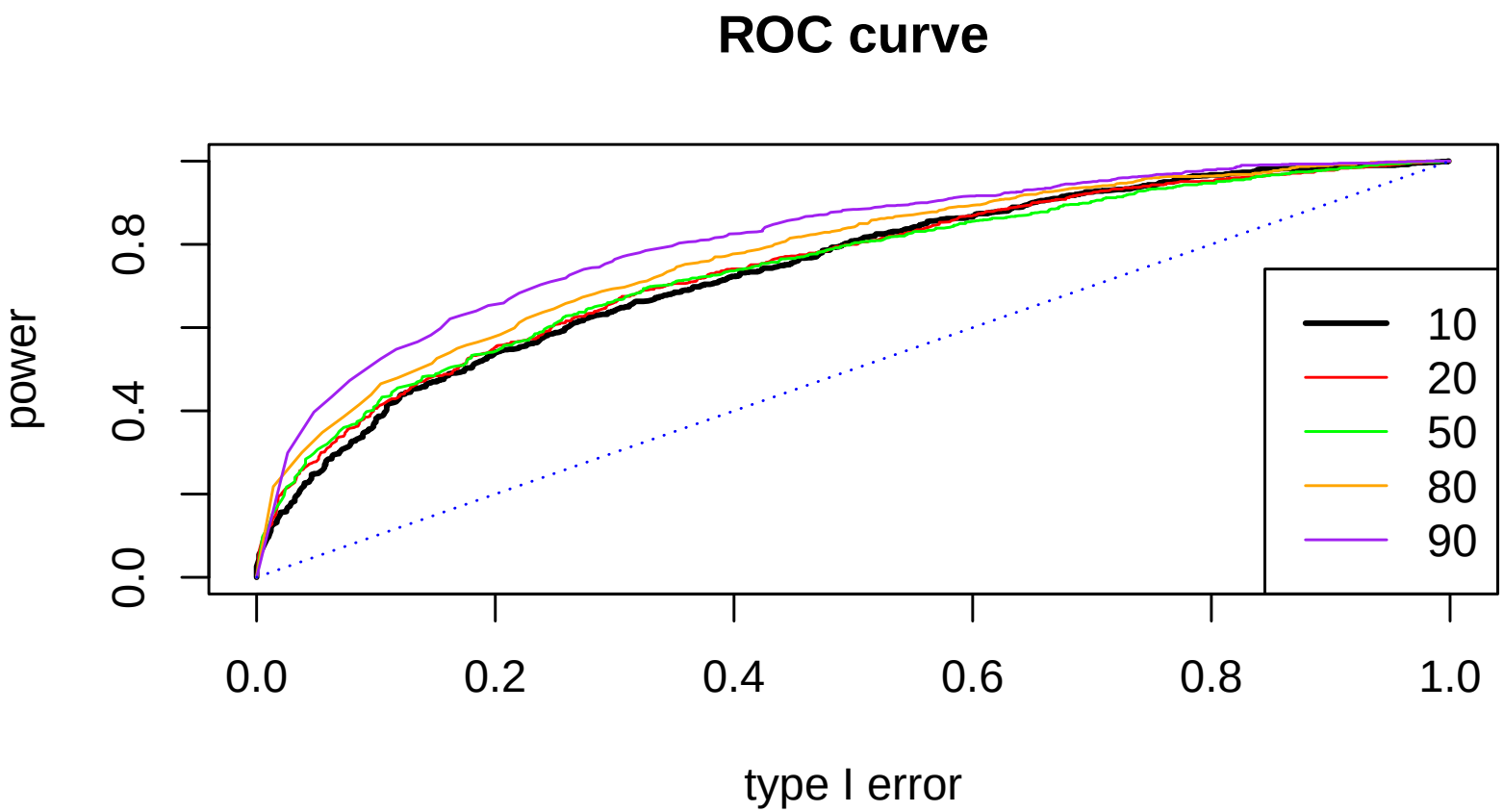
- ***me1*** Односторонний Multiple MC-SSA с проекцией на собственные векторы U_i SSA разложения.
  Особенности: Радикальный. Позволяет извлечь сигнал после обнаружения с помощью SSA.
- ***mt1*** Односторонний Multiple MC-SSA с проекцией на собственные векторы матрицы $\{\phi^{|i-j|}\}_{i,j=1}^{L}$.
  Особенности: Точный.
- ***be1*** Односторонний Single MC-SSA с проекцией на собственные векторы U_i SSA разложения с поправкой Бонферрони.
  Особенности: Радикальный. Позволяет извлечь сигнал после обнаружения с помощью SSA. Часто используется на практике.
- ***bt1*** Односторонний Single MC-SSA с проекцией на собственные векторы матрицы $\{\phi^{|i-j|}\}_{i,j=1}^{L}$.
  Особенности: Консервативный.

Часть II. План численного сравнения критериев

- Сравнить на основном примере ряда:
  сигнал имеет вид $s_n = A\cos(2\pi\omega n)$
  длина ряда $N = 100$, амплитуда сигнала $A = 1$, частота сигнала $\omega = 0.075$, параметр шума $\varphi = 0.7$
  метод: $M = 1000$, в me1 и be1 с Toepliz SSA,
  в mt1 и me1 $G = 1000$, в bt1 и be1 $G = 10000$.
  - Для каждого метода найти оптимальную длину окна L (длину векторов для проекции).
  - Сравнить разные версии критериев с оптимальной длиной окна L.
- Изменить длину ряда, частоту сигнала, параметры шума, чтобы проверить устойчивость выводов.
- Посмотреть, важно ли при сравнении учитывать γ-гарантированные версии поправки.

Выбор длины окна на примере критерия me1

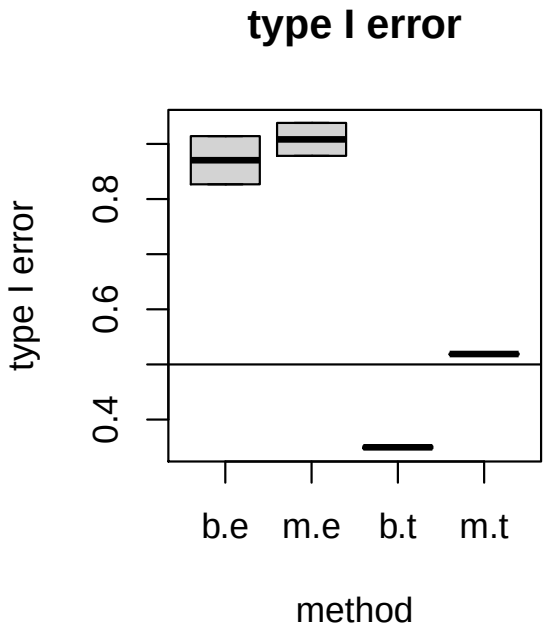
me1, $G = 1000$, $M = 1000$, $N = 100$, $\varphi = 0.7$, $A = 1$, $\omega = 0.075$

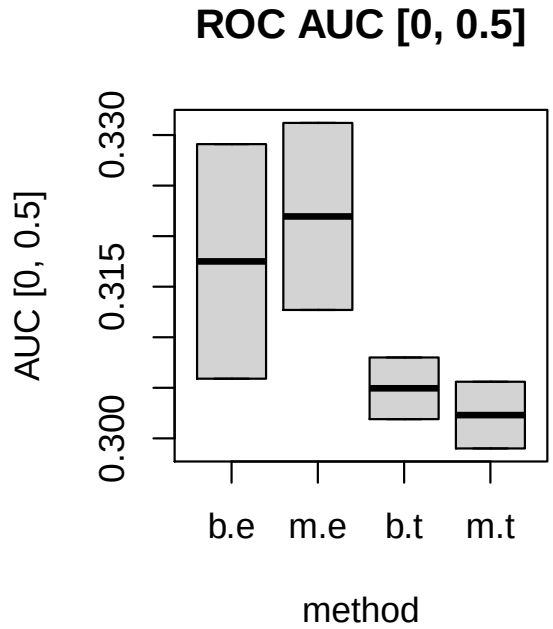


Получаем: длина окна $L = 90$ наилучшая (AUC значимо больше).

$N = 100$, $\varphi \in [0.5, 0.7]$, $\mathrm{SNR} = 0.255$, $A = [0.34, 1]$, $\omega = 0.075$.

Сравнение модификаций MC-SSA (Toeplitz) по AUC.





Basic SSA или Toeplitz SSA при построении W_i в me1.

W_i	L	количество векторов	$\alpha_I(\alpha)$ AUC	ROC AUC [0,0.5]	Левая граница ROC AUC [0,0.5]	Правая граница ROC AUC [0,0.5]
SVD	80	21	0.9962	—	—	—
SVD	50	50	0.7645	0.2872	0.2741	0.3004
SVD	20	20	0.5415	0.2851	0.2715	0.2986
Toepliz	80	21	0.8761	0.3024	0.2896	0.3162
Toepliz	50	50	0.6435	0.2846	0.271	0.2973
Toepliz	20	20	0.5352	0.2828	0.2692	0.2965

Часть II. Доверительные интервалы $\alpha_I(\alpha)$ при гарантированно нерадикальной поправке при оценке параметров шума

Пусть при получении зависимости $\alpha_I(\alpha)$ используются оценки параметров шума. Тогда получаемая зависимость тоже имеет случайную погрешность. Построим эмпирические доверительные интервалы для $\alpha_I(\alpha)$.

Параметры: $G = 1000$, $M = 1000$, $N = 100$, $\varphi = 0.7$, $A = 1$, $\omega = 0.075$

Вывод: радикальность критерия почти не влияет на размер доверительного интервала.

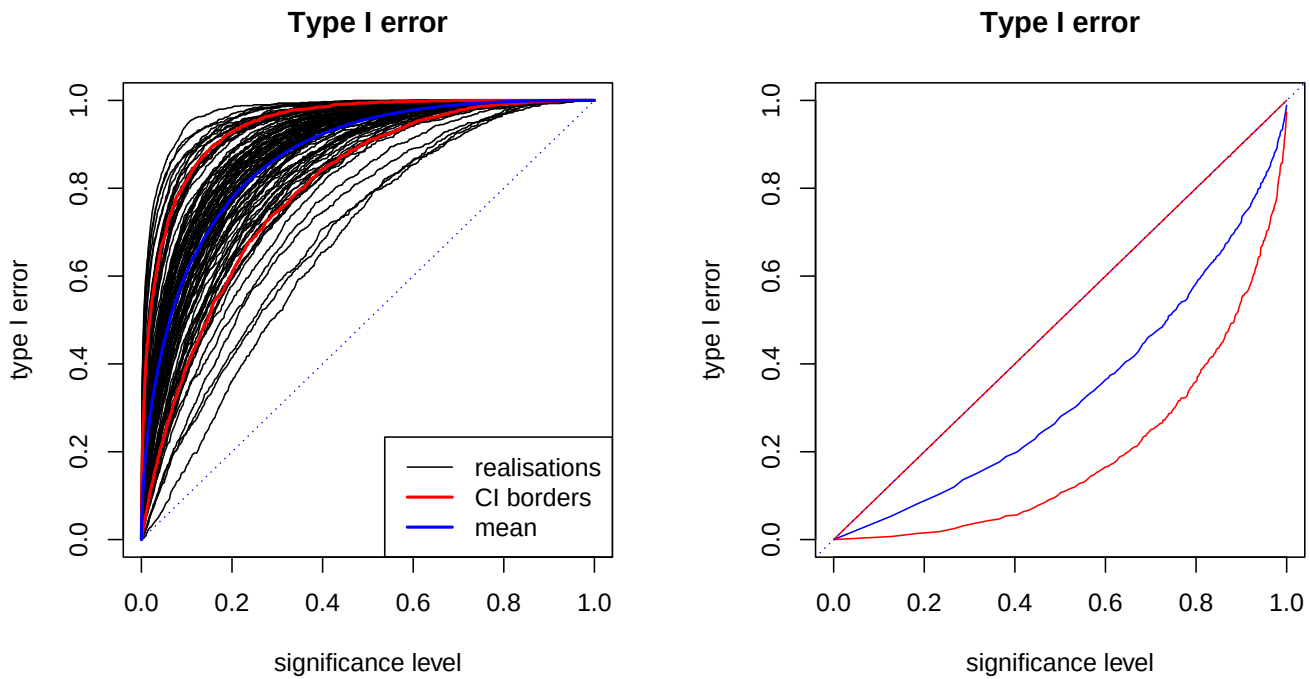


Рис.: Радикальный me1, $L = 90$

Часть II. Доверительные интервалы $\alpha_I(\alpha)$ при гарантированно нерадикальной поправке при оценке параметров шума

Пусть при получении зависимости $\alpha_I(\alpha)$ используются оценки параметров шума. Тогда получаемая зависимость тоже имеет случайную погрешность. Построим эмпирические доверительные интервалы для $\alpha_I(\alpha)$.

Параметры: $G = 1000$, $M = 1000$, $N = 100$, $\varphi = 0.7$, $A = 1$, $\omega = 0.075$

Вывод: радикальность критерия почти не влияет на размер доверительного интервала.

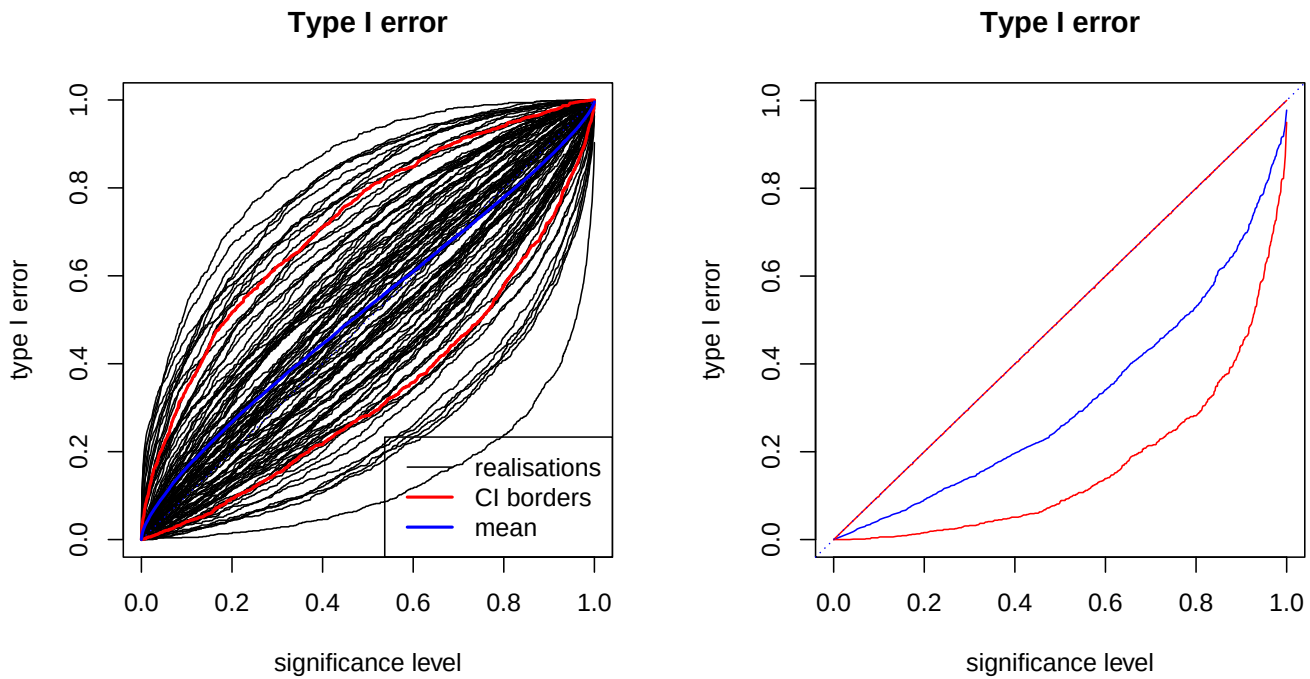


Рис.: Точный mt1, $L = 80$

Дано: Ряд $\mathbf{F} = [\mathsf{X}_1, \ldots, \mathsf{X}_D]$, D — количество каналов.

Метод MSSA: Аналогичен SSA, только траекторная матрица $\mathbf{X} = [\mathbf{X}_1 : \ldots : \mathbf{X}_D]$.

Метод Monte Carlo MSSA: Аналогичен MC-SSA.

Отличие от Monte Carlo SSA: В одномерном случае факторные векторы не рассматривались, т. к., собственные векторы при длине окна L совпадают с факторными при длине окна $N - L + 1$. Однако в многомерном случае это не так: траекторная матрица составная, а значит структура строк и столбцов разная.

Замечание

Т.к. в Rssa нет реализации Toepliz SSA для многомерного случая, то из-за проблемы слишком радикального критерия здесь рассмотрена лишь часть вариантов L.

Заключение

Проделанная работа:

1. Разработана система корректировки критериев и сравнения результатов по ROC-кривым.
2. Система из п.1 применена к неточным критериям MC-SSA.
3. Проведено всесторонне сравнение разных модификаций критериев MC-SSA.
4. Рассмотрено обобщение критерия на многомерный случай.
5. Представлена реализация на R.

Рекомендации по применению MC-SSA:

1. Можно рекомендовать использовать большую длину окна L для увеличения мощности.
2. Рекомендуется использовать скорректированные методы, у которых W_i — с.в. из Toeplitz SSA (me и be), хотя они исходно радикальные.
3. Использование с.в. из Basic SSA приводит к чрезмерной радикальности критерия при больших L, поэтому рекомендуется использовать Toeplitz SSA.
4. Для выработки рекомендаций для MC-MSSA необходимы дополнительные исследования, в частности, реализация Toeplitz MSSA.