УДК

Потешкин Е. П.

Выбор параметров в методе Монте-Карло SSA

Рекомендовано к публикации доцентом

1. Введение. Обнаружения сигнала в красном шуме (красным шумом называется процесс авторегрессии первого порядка с положительным коэффициентом) с целью его дальнейшего выделения является важной задачей, например для анализа поведения глобального потепления [1]. Одним из методов, используемым для этой цели, является метод Monte-Carlo SSA (MC-SSA) [2], который проверяет гипотезу о том, что отсутствует сигнал, а в случае, если она отвергается, позволяет приближенно выделить этот сигнал. При использовании метода MC-SSA возникают следующие проблемы. Используемый на практике вариант критерия в методе MC-SSA является радикальным, поэтому для его использования необходимы поправки. В [3] предлагается поправка на множественное тестирование, частично исправляющая радикальность. Однако применяемая на практике модификация критерия все еще остается радикальной. Эту радикальность сложно исправить теоретически, поэтому используется метод эмпирической поправки критерия, основанный на оцененных ошибках первого рода в зависимости от уровня значимости, см., например, [4], где этот подход используется в контексте метода MC-SSA.

Используемый в той же работе [4] подход через построение ROC-кривых позволяет сравнивать критерии по мощности. Соответственно возникает вопрос о выборе параметра метода MC-SSA, называемого длиной окна, дающего оптимальную мощность. Ответу на этот вопрос мешает то, что если исходно критерий сильно радикальный, то поправку, исправляющую радикальность, очень трудоемко или даже невозможно сделать.

Данная статья посвящена поиску длины окна, при которой критерий наиболее мощный среди не слишком радикальных критериев.

Потешкин Егор Павлович – студент, Санкт-Петербургский государственный университет; email: egor.poteshkin@yandex.ru, тел.: +7(921)560-32-02

Работа выполнена при финансовой поддержке РНФ, проект № 23-21-00222

В разделе 2 приведены известные результаты. В разделе 3 приведены результаты численного исследования зависимости ошибки первого рода и мощности от длины окна. На основе этих результатов в разделе 4 предложен алгоритм поиска длины окна и продемонстрирована его работа.

2. Известные результаты. В этом разделе введем нужные обозначения и приведем известные результаты, чтобы в дальнейшем их использовать.

2.1. SSA [5]. Пусть $\mathsf{X} = (x_1, \ldots, x_N)$, $x_i \in \mathbb{R}$, – временной ряд длины N. Зафиксируем параметр L, $1 < L < N$, называемый длиной окна. Построим матрицу $\mathbf{X} = [X_1 : \ldots : X_K]$, состоящую из $K = N - L + 1$ векторов вложения $X_i = (x_i, \ldots, x_{i+L-1})^{\mathrm{T}} \in \mathbb{R}^L$.

Следующий шаг – разложение в сумму матриц единичного ранга $\mathbf{X} = \sum_{i=1}^{d} \mathbf{X}_i$. В базовом SSA используется сингулярное разложение матрицы $\mathbf{X}$.

Далее компоненты полученного матричного разложения разумно группируются, и каждая сгруппированная матрица преобразуется во временной ряд. Таким образом, результатом SSA является разложение временного ряда.

2.2. Toeplitz SSA [5]. Модификация базового SSA, Toeplitz SSA, использует вместо сингулярного разложения разложение матрицы $\mathbf{X}$, основанное на собственных векторах $\{P_i\}_{i=1}^{L}$ тёплицевой матрицы $\mathbf{C}$ с элементами

$$c_{ij} = \frac{1}{N - |i-j|} \sum_{m=1}^{N-|i-j|} x_m x_{m+|i-j|}, 1 \leqslant i, j \leqslant L. \tag{1}$$

Тогда

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

где $Q_i = \mathbf{X}P_i/\sigma_i$, $\sigma_i = \|\mathbf{X}P_i\|$. Такое разложение представляется более естественным для стационарных временных рядов, когда $\mathbf{C}$ является оценкой автоковариационной матрицы.

2.3. Monte Carlo SSA [2]. Рассмотрим задачу поиска сигнала во временном ряде. Модель:

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

где S – сигнал, $\boldsymbol{\xi}$ – красный шум с параметрами φ и δ. Тогда нулевая гипотеза $H_0 : \mathsf{S} = 0$ и альтернатива $H_1 : \mathsf{S} \neq 0$.

Зафиксируем длину окна L и обозначим траекторную матрицу ряда $\boldsymbol{\xi}$ как $\boldsymbol{\Xi}$. Рассмотрим вектор $W \in \mathbb{R}^L$ единичной длины, называемый проекционным вектором. Введем величину

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

Статистикой критерия является величина

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

Распределение статистики критерия оценивается с помощью моделирования согласно нулевой гипотезе; отсюда и название метода.

Если вектор W – синусоида с частотой ω, то $\widehat{p}$ отражает вклад частоты ω в исходный ряд. Так как частота ожидаемого сигнала неизвестна, но необходимо рассматривать несколько векторов W_k, $k = 1, \ldots, H$. Решение возникающей при этой проблемы множественного тестирования рассматривается в [3]. Гипотеза об отсутствии сигнала отвергается, если хотя бы для одного вектора $W = W_k$ значение $\widehat{p}$ оказывается значимым.

Еще одним параметром MC-SSA является способ выбора векторов W_k. В данной работе в качестве векторов для проекции берутся собственные векторы матрицы $\mathbf{C}$, определенная в (1). Такой способ выбора самый распространенный, поскольку, если есть значимые векторы, можно восстановить сигнал с помощью SSA на их основе. Но этот вариант, вообще говоря, дает радикальный критерий, поскольку W_k зависят от ряда X, в котором мы ищем сигнал. С помощью эмпирической поправки неточных критериев можно бороться с этой проблемой. Опишем ее в следующем разделе.

2.4. Поправка неточных критериев. Данный алгоритм позволяет преобразовывать радикальные и консервативные статистические критерии в точные.

Зафиксируем нулевую гипотезу H_0, уровень значимости α^*, количество выборок M для оценки $\alpha_I(\alpha)$ и их объем N (в случае временных рядов N – длина ряда).

Сначала моделируется M выборок объема N при верной H_0. Затем по моделированным данным строится зависимость ошибки первого рода от уровня значимости $\alpha_I(\alpha)$. Результатом работы алгоритма является формальный уровень значимости $\widetilde{\alpha}^* = \alpha_I^{-1}(\alpha^*)$. Крите-

рий с таким уровнем значимости является асимптотически точным при $M \to \infty$.

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

2.5. ROC-кривая. 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 потенциальной мощности неточные критерии. Отметим, что для точного (в частности, поправленного) критерия ROC-кривая совпадает с графиком мощности, так как $\alpha_I(\alpha) = \alpha$.

3. Зависимость радикальности и мощности MC-SSA 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.

Рассмотрим пример: пусть дана модель

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

где $\mathsf{S} = \{A\cos(2\pi\omega n)\}_{n=1}^{N}$ – сигнал с частотой $\omega = 0.075$, а $\boldsymbol{\xi}$ – красный шум с параметрами $\varphi = 0.7$ и $\delta = 1$. Длина ряда равна $N = 100$. Рассмотрим следующие нулевую гипотезу и альтернативу: $H_0 : A = 0$, $H_1 : A = 1$. В работе [4] было показано, что оценка параметров модели почти не искажает критерий после применения алгоритма поправки, поэтому будем предполагать, что параметры красного шума нам известны.

По графику ошибок первого рода на рис. 1 видно, что чем больше L, тем более радикальным становится критерий. На рис. 2 изображены ROC-кривые критериев, наибольшую мощность дает критерий с $L = 90$. На этом примере видно, что, в целом, чем сильнее радикальность критерия, тем более мощным он является после поправки.

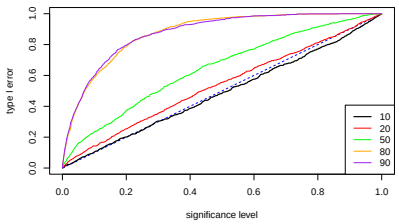

Рис. 1. Ошибка I рода ($N = 100$, $\varphi = 0.7$)

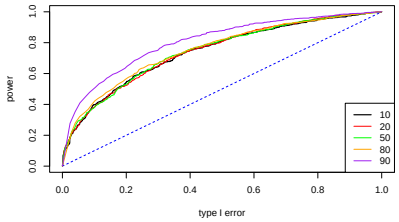

Рис. 2. ROC-кривая ($N = 100$, $\varphi = 0.7$)

Посмотрим на другой пример. Рассмотрим теперь красный шум с параметрами $\varphi = 0.35$ и $\delta = 1$ и увеличим длину ряда до $N = 200$. Для удобства сравнения ROC-кривых, уменьшим амплитуду сигнала до $A = 0.6$ при верной H_1.

ROC-кривую для этого примера на рис. 4 не удалось построить полностью для $L = 160$ и $L = 190$ из-за сильной радикальности, это видно на рис. 3.

Рядом других численных примеров было подтверждено, что с увеличением L радикальность критерия сильно растет, а мощность, в целом, растет, но несильно. Это верно для разных длин ряда, для разных параметров красного шума и сигнала.

4. Алгоритм поиска оптимального L. В примере с $\varphi = 0.35$ из-за слишком сильной радикальности было невозможно применить поправку для $L = 160, 190$. Поэтому идея выбора длины окна L состоит в том, чтобы найти такое L, при котором ошибка первого

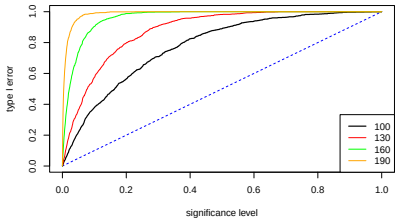

Рис. 3. Ошибка I рода ($N = 200$, $\varphi = 0.35$)

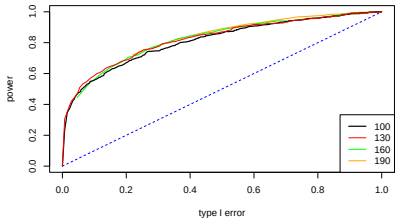

Рис. 4. ROC-кривая ($N = 200$, $\varphi = 0.35$)

рода не больше, чем заданная.

Учитывая, что с увеличением L растет радикальность исходного критерия и мощность поправленного, предлагаем алгоритм поиска оптимальной длины окна, для которой ошибка I рода критерия не превосходит порогового значения. Фактически, сам алгоритм представляет собой бинарный поиск, однако для определенности мы его опишем формально.

Вход алгоритма. Временной ряд X, уровень значимости α^*, порог ошибки первого рода $\widehat{\alpha}_I$ при уровне значимости α^*, интервал $[L_{\text{left}}, L_{\text{right}}]$ поиска оптимального L, желаемая точность ε по α_I и точность δ_L по L.

Выход алгоритма. Оптимальная длина окна L_{optim}:

$$\left| L_{\text{optim}} - \operatorname{argmax} \left\{ L \in [L_{\text{left}}, L_{\text{right}}] : |\alpha_I^{(L)} - \widehat{\alpha}_I| < \varepsilon \right\} \right| \leqslant \delta_L.$$

1. Вычислить $\alpha_I^{(L_{\text{left}})}$ и $\alpha_I^{(L_{\text{right}})}$. Если $\alpha_I^{(L_{\text{right}})} \leqslant \widehat{\alpha}_I$, завершить алгоритм с $L_{\text{optim}} = L_{\text{right}}$. Иначе, если $|\alpha_I^{(L_{\text{left}})} - \widehat{\alpha}_I| < \varepsilon$, завершить алгоритм с $L_{\text{optim}} = L_{\text{left}}$. Иначе, если $\alpha_I^{(L_{\text{left}})} > \widehat{\alpha}_I$, оптимальной длины окна на этом интервале нет, завершить алгоритм.

2. Вычислить $\alpha_I^{(L_{\text{mid}})}$, где $L_{\text{mid}} = \lfloor (L_{\text{left}} + L_{\text{right}})/2 \rfloor$. Если $|\alpha_I^{(L_{\text{mid}})} - \widehat{\alpha}_I| < \varepsilon$, завершить алгоритм с $L_{\text{optim}} = L_{\text{mid}}$. Иначе, если $\alpha_I^{(L_{\text{mid}})} < \widehat{\alpha}_I$, $L_{\text{left}} = L_{\text{mid}}$. Если $\alpha_I^{(L_{\text{mid}})} > \widehat{\alpha}_I$, $L_{\text{right}} = L_{\text{mid}}$.

3. Если $|\alpha_I^{(L_{\text{mid}})} - \widehat{\alpha}_I| \geqslant \varepsilon$, повторить пункт 2 до тех пор, пока $L_{\text{left}} < L_{\text{right}} - \delta_L$.

4. Если $L_{\text{left}} \geqslant L_{\text{right}} - \delta_L$, завершить алгоритм с $L_{\text{optim}} = L_{\text{left}}$.

4.1. Проверка работы алгоритма. Проверим работу алго-

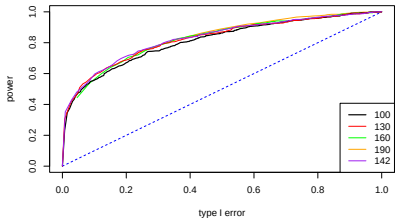

Рис. 5. ROC-кривая ($N = 200$, $\varphi = 0.35$), $L_{\text{optim}} = 142$

ритма на примере с $\varphi = 0.35$ и $N = 200$ из раздела 3. Поиск L_{optim} велся на интервале $[100, 190]$ со следующими параметрами: $\alpha^* = 0.05$, $\widehat{\alpha}_I = 0.5$, $\varepsilon = 0.01$ и $\delta_L = 5$. Результат алгоритма: $L_{\text{optim}} = 142$ с $\alpha_I^{(L_{\text{optim}})} = 0.49$. По рис. 5 изображены ROC-кривые критериев с добавлением L_{optim}. Как видно, ROC-кривую для L_{optim} уже удалось построить и такая длина окна действительно оптимальна.

5. Заключение. В работе предложен способ выбора параметра «длина окна» в методе MC-SSA, позволяющем обнаруживать сигнал в красном шуме. На основе численно выявленных закономерностей

был построен алгоритм, позволяющий выбрать длину окна, приводящую к построению точного критерия, возможного на практике, и при этом наиболее мощного.

При демонстрации алгоритма мы предполагали, что значения параметров красного шума известны, однако их оценивание не искажает выявленные закономерности, лежащие в основе алгоритма.

Литература

1. Селиверстова К. А. Применение статистических методов для оценки глобального потепления // Процессы управления и устойчивость. 2023. Т. 10. № 1. С. 195–199.

2. Allen M., Smith L. Monte Carlo SSA: detecting irregular oscillations in the presence of coloured noise // Journal of Climate. 1996. Vol. 9. P. 3373—3404.

3. Golyandina N. Detection of signals by Monte Carlo singular spectrum analysis: multiple testing // Statistics and Its Interface. 2023. Vol 16. № 1. P. 147–157.

4. Ларин Е. С. Применимость исправления статистических критериев на примере задачи обнаружения сигнала во временных рядах // Процессы управления и устойчивость. 2022. Т. 9. № 1. С. 267–271.

5. Golyandina N., Nekrutkin V., Zhigljavsky A. Analysis of Time Series Structure: SSA and Related Techniques. Chapman and Hall/CRC, 2001.