УДК 519.254, 519.688

Потешкин Е. П.

Выбор параметров в методе Монте-Карло SSA

Рекомендовано к публикации доцентом Голяндиной Н.Э.

1. Введение. Обнаружение 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 течения Эль-Ниньо [1] или глобального потепления [2]. Одним из методов, используемым для этой цели, является метод Monte Carlo SSA (MC-SSA) [3], который проверяет гипотезу об отсутствии сигнала, а в случае если она отвергается, позволяет приближенно выделить этот сигнал. При применении метода MC-SSA возникают следующие проблемы. Самый распространенный на практике вариант критерия MC-SSA является радикальным, поэтому для его использования необходимы поправки. В [4] 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. Например, в [5] этот подход используется в контексте метода MC-SSA.

Используемый в той же работе [5] 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. Сложность состоит в том, что если исходно критерий сильно радикальный, то поправку, исправляющую радикальность, очень трудоемко или даже невозможно сделать.

Данная статья посвящена построению подхода к поиску длины окна, при которой критерий наиболее мощный среди ограниченно радикальных критериев.

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

2. Вспомогательные результаты. Введем обозначения и приведем известные результаты.

2.1. Метод SSA [6]. 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 и построим так называемую траекторную 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 X.

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

2.2. Метод Toeplitz SSA [6]. Этот метод является модификацией базового SSA и использует вместо сингулярного разложения разложение матрицы $\mathbf{X}$, основанное на собственных векторах $\{P_i\}_{i=1}^{L}$ тёплицевой матрицы $\mathbf{C}$ с элементами

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

Тогда

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

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

2.3. Метод Monte Carlo SSA [3]. Рассмотрим задачу поиска сигнала во временном ряде. Модель временного ряда имеет вид

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

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, то необходимо рассматривать несколько векторов W_k, $k = 1, \ldots, H$. Решение возникающей при этом проблемы множественного тестирования рассматривается в [4]. 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 $\mathbf{C}$, определенной 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 ищется сигнал. С помощью эмпирической поправки неточных критериев можно бороться с этой проблемой.

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 алгорит-

ма является формальный уровень значимости $\widetilde{\alpha}^* = \alpha_I^{-1}(\alpha^*)$. Критерий 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. Это кривая, задаваемая параметрически:

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

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. Однако, в зависимости от длины ряда N и параметров красного шума $\boldsymbol{\xi}$ наблюдаются разные зависимости мощности от L.

Рассмотрим несколько примеров. Пусть дана модель

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

где $\mathsf{S} = \{A\cos(2\pi\omega n)\}_{n=1}^N$, а $\boldsymbol{\xi}$ – красный шум с параметрами φ и $\delta = 1$. Рассмотрим следующие нулевую гипотезу и альтернативу: $H_0 : A = 0$, $H_1 : A \neq 0$. В работе [5] показано, что оценка параметров модели почти не искажает критерий после применения алгоритма поправки, поэтому будем предполагать, что параметры красного шума известны. В первых трех примерах рассмотрим частоту сигнала $\omega = 0{,}075$.

Пример 1. Пусть $\varphi = 0{,}7$, $N = 100$. 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,

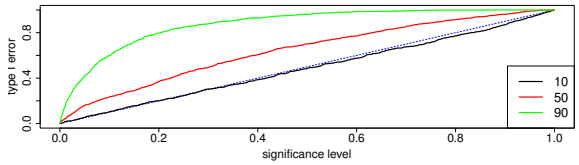

Рис. 1. Ошибка I рода ($\varphi = 0{,}7$, $N = 100$)

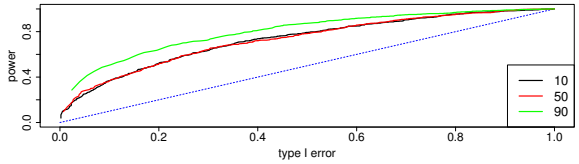

Рис. 2. ROC-кривая ($\varphi = 0{,}7$, $N = 100$, $\omega = 0{,}075$)

наибольшую мощность дает критерий с $L = 90$. На этом примере видно, что самым мощным является самый радикальный критерий.

Пример 2. Пусть $\varphi = 0{,}3$, $N = 100$. На рис. 3 изображен гра-

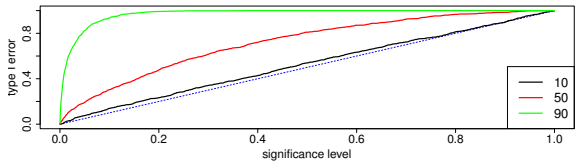

Рис. 3. Ошибка I рода ($\varphi = 0{,}3$, $N = 100$)

фик ошибок первого рода. По нему видно, что, как и в примере 1, чем больше L, тем больше радикальность критерия. Если взглянуть на ROC-кривые на рис. 4, то видно, что с уменьшением параметра φ уменьшается разброс мощности критериев после поправки в зависимости от длины окна. Лучшей из рассмотренных в этом случае является $L = 10$, хотя разница с $L = 50$ совсем небольшая, а для $L = 90$ для небольших ошибок I рода поправку сделать не удалось

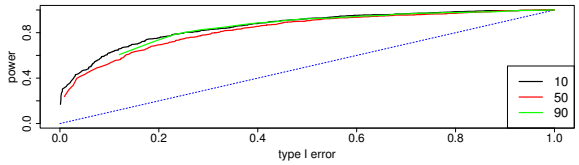

Рис. 4. ROC-кривая ($\varphi = 0{,}3$, $N = 100$, $\omega = 0{,}075$)

из-за радикальности.

Пример 3. Теперь увеличим длину ряда до $N = 400$ и посмотрим на ROC-кривые для примеров 1, 2. На рис. 5, 6 видим, что для

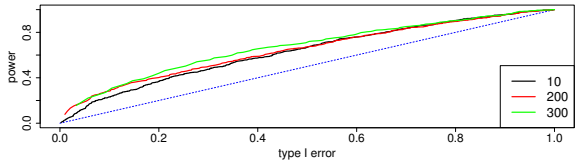

Рис. 5. ROC-кривая ($\varphi = 0{,}7$, $N = 400$, $\omega = 0{,}075$)

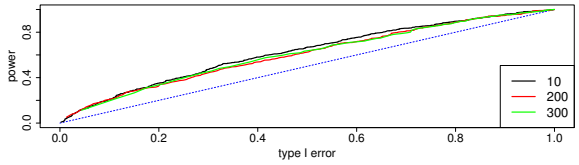

Рис. 6. ROC-кривая ($\varphi = 0{,}3$, $N = 400$, $\omega = 0{,}075$)

обоих примеров с увеличением длины ряда уменьшается различие в мощности после поправки в зависимости от длины окна.

Пример 4. В условиях примера 1 рассмотрим разные частоты ω сигнала S и зависимость упорядоченности критериев по мощности от L. На рис. 7, 8 изображены ROC-кривые критериев при разных альтернативах. Видно, что упорядоченность L нарушается при ма-

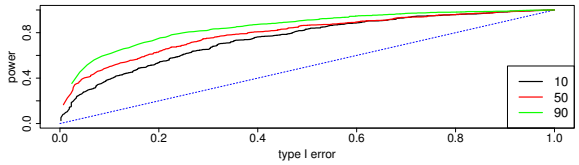

Рис. 7. ROC-кривая ($\varphi = 0{,}7$, $N = 100$, $\omega = 0{,}175$)

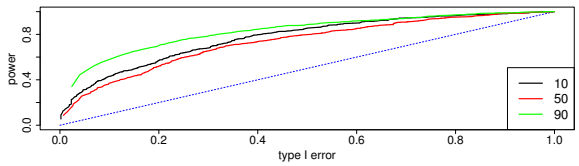

Рис. 8. ROC-кривая ($\varphi = 0{,}7$, $N = 100$, $\omega = 0{,}025$)

леньких частотах сигнала. Если упорядочить рис. 7, рис. 2 и рис. 8 по частоте ω, то видна динамика по соотношению ROC-кривых для $L = 10$ и $L = 50$.

4. Выбор длины окна. Численные эксперименты показали, что длина окна L, дающая максимальную мощность критерия после поправки, зависит от параметров шума, длины ряда и, главное, от частоты сигнала в альтернативной гипотезе. Поэтому при выборе длины окна возможны следующие варианты:

1. Использовать поправленный критерий MC-SSA с маленькой длиной окна, например, $L = 10$ при длине ряда $N = 100$. Это нетрудозатратно, а также критерий не является сильно радикальным (в рассмотренных примерах он близок к точному). Недостаток состоит в том, что такой выбор L может являться неоптимальным, т. е. возможна некоторая потеря в мощности.

2. В поведении оптимальной по мощности длины окна L в зависимости от параметров ряда наблюдается некоторая регулярность. Поэтому можно было бы построить зависимость оптимальной длины окна от параметров ряда с помощью численного моделирования, оценив параметры красного шума. Однако

было показано, что упорядоченность критериев по мощности зависит от частоты сигнала в альтернативе, эта рекомендация имеет практический смысл, только если есть дополнительная информация о диапазоне возможных частот в альтернативе.

5. Заключение. В работе выработан подход к выбору параметра «длина окна» в методе MC-SSA, позволяющем обнаруживать сигнал в красном шуме. Оказалось, что из-за того, что возможная частота сигнала, вообще говоря, неизвестна, использование численного моделирования не является решением задачи выбора оптимальной длины окна. В целом, численное моделирование позволяет дать рекомендации по выбору L при уточнении диапазона возможный частоты сигнала. В любом случае, численное моделирование является весьма трудоемким, поэтому, вполне возможно, что наиболее разумной альтернативой является выбор маленькой длины окна (например, $L = 10$ при длине ряда $N = 100$). Как показано на примерах, MC-SSA с такой длиной окна не всегда дает максимальную мощность, однако обладает небольшой трудоемкостью и малой радикальностью.

Литература

1. Jevrejeva S., Moore J.C., Grinsted A. Oceanic and atmospheric transport of multiyear El Niño–Southern Oscillation (ENSO) signatures to the polar regions // Geophysical Research Letters. 2004. Vol. 31. Art. no L24210.

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

3. 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.

4. 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.

5. 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.

6. Golyandina N., Nekrutkin V., Zhigljavsky A. Analysis of Time Series Structure: SSA and Related Techniques. Boca Raton: Chapman and Hall/CRC, 2001. 320 p.