УДК

Потешкин Е. П.

1. Введение. Существует задача обнаружения сигнала в красном шуме с целью его дальнейшего выделения, например для анализа поведения глобального потепления [TODO]. Метод Monte-Carlo SSA (MC-SSA) [TODO] проверяет гипотезу о том, что во временном ряде присутствует сигнал, но самый распространенный вариант является радикальным. Поправка неточных критериев [TODO] делает этот критерий точным. В данной работе рассмотрены две проблемы: выбор параметра L для получения максимально возможной мощности и невозможность применения поправки на практике, если исходно критерий был слишком радикальным.

2. Известные результаты. Опишем уже известные результаты, чтобы в дальнейшем их использовать.

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

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. Модификация базового SSA, Toeplitz SSA, использует вместо SVD тёплицево разложение матрицы $\mathbf{X}$:

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

Потешкин Егор Павлович – студент, Санкт-Петербургский государственный университет; email: ..., тел.: ...

Работа выполнена при финансовой поддержке РФФИ, проект №

где $\{P_i\}_{i=1}^{L}$ — собственные векторы матрицы $\widetilde{\mathbf{C}}$ с элементами

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

Такое разложение имеет преимущество для стационарных временных рядов.

2.3. Monte Carlo SSA. 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

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

Если вектор W – синусоида с частотой ω, то $\widehat{p}$ отражает вклад частоты ω в исходный ряд.

Далее строится доверительный интервал случайной величины p: в большинстве случаев распределение p неизвестно, поэтому оно оценивается методом Монте-Карло.

Построив доверительный интервал, проверяется, лежит ли $\widehat{p}$ в нем. Если нет, гипотеза отвергается.

Стоит отметить, что данный критерий является несостоятельным против H_1, если частота ω сигнала S неизвестна, что не редкость на практике.

Поэтому вместо одного вектора W рассматривают набор W_k и для каждого такого вектора строят доверительные интервалы. Без поправки на множественные сравнения данный вариант дает радикальный критерий. В работе [TODO] был представлен модифицированный алгоритм построения критерия в случае множественного тестирования.

Параметром MC-SSA является способ выбора векторов W_k. В данной работе в качестве векторов для проекции берутся собственные векторы матрицы $\widetilde{\mathbf{C}}$ (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 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.

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 $\widetilde{\alpha}^* = \alpha_I^{-1}(\alpha^*)$. Критерий с таким уровнем значимости является асимптотически точным при $M \to \infty$.

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 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 критерий MC-SSA в самом распространенном варианте радикальный, существует проблема выбора такой длины окна L, чтобы получить наилучший по мощности критерий, 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 $\omega = 0.075$, а $\boldsymbol{\xi}$ – красный шум с параметрами $\varphi = 0.7$ и $\delta = 1$. Длина ряда равна $N = 100$. Рассмотрим следующую нулевую гипотезу с альтернативой: $H_0 : A = 0$, $H_1 : A = 1$.

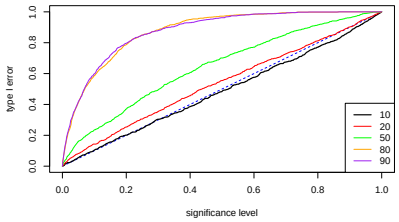

Рис. 1. Ошибка I рода ($N = 100$, $\varphi = 0.7$)

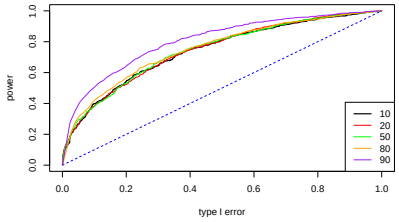

Рис. 2. ROC-кривая ($N = 100$, $\varphi = 0.7$)

По графику ошибкок 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 чем сильнее радикальность критерия, тем более мощным он является после поправки.

Посмотрим на другой пример. Рассмотрим теперь красный шум с параметрами $\varphi = 0.3$ и $\delta = 1$ и увеличим длину ряда до $N = 200$. Для удобства сравнения ROC-кривых, уменьшим амплитуду сигнала до $A = 0.7$ при верной H_1.

ROC-кривую для этого примера на рис. 4 не удалось построить

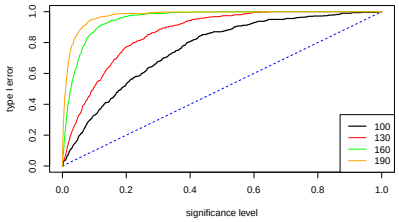

Рис. 3. Ошибка I рода ($N = 200$, $\varphi = 0.3$)

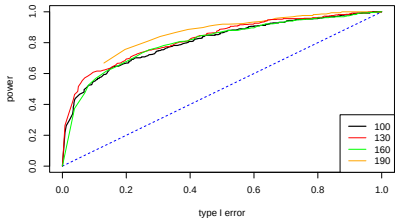

Рис. 4. ROC-кривая ($N = 200$, $\varphi = 0.3$)

полностью для $L = 190$ из-за сильной радикальности, это видно на рис. 3.

Для многочисленных примеров было получено, 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 предыдущем примере из-за слишком сильной радикальности невозможно применить поправку для $L = 190$. Поэтому нужно найти такое L, чтобы ошибка первого рода была не слишком большой.

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.

4.1. Вход алгоритма. Временной ряд X, уровень значимости

α^*, порог ошибки первого рода $\hat{\alpha}_I$ при уровне значимости α^*, интервал $[L_{\text{left}}, L_{\text{right}}]$ поиска оптимального L, желаемая точность ε.

4.2. Выход алгоритма. Оптимальная длина окна L_{optim}:

$$L_{\text{optim}} = \max\left\{L \in [L_{\text{left}}, L_{\text{right}}] : |\alpha_I^{(L)} - \hat{\alpha_I}| < \varepsilon\right\}.$$

4.3. Алгоритм.

1. Вычислить $\alpha_I^{(L_{\text{left}})}$ и $\alpha_I^{(L_{\text{right}})}$. Если $\alpha_I^{(L_{\text{right}})} \leqslant \hat{\alpha}_I$, завершить алгоритм с $L_{\text{optim}} = L_{\text{right}}$. Иначе, если $|\alpha_I^{(L_{\text{left}})} - \hat{\alpha_I}| < \varepsilon$, завершить алгоритм с $L_{\text{optim}} = L_{\text{left}}$. Иначе, если $\alpha_I^{(L_{\text{left}})} > \hat{\alpha_I}$, 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 $|\alpha_I^{(L_{\text{mid}})} - \hat{\alpha}_I| < \varepsilon$, завершить алгшоритм с $L_{\text{optim}} = L_{\text{mid}}$. Иначе, если $\alpha_I^{(L_{\text{mid}})} < \hat{\alpha_I}$, $L_{\text{left}} = L_{\text{mid}}$. Если $\alpha_I^{(L_{\text{mid}})} > \hat{\alpha_I}$, $L_{\text{right}} = L_{\text{mid}}$.

3. Если $|\alpha_I^{(L_{\text{mid}})} - \hat{\alpha}_I| \geqslant \varepsilon$, повторить пункт 2 до тех пор, пока $L_{\text{left}} < L_{\text{right}} - 5$.

4. Если $L_{\text{left}} \geqslant L_{\text{right}} - 5$, завершить алгоритм с $L_{\text{optim}} = L_{\text{left}}$.

5. Проверка работы алгоритма. Проверим работу алгоритма

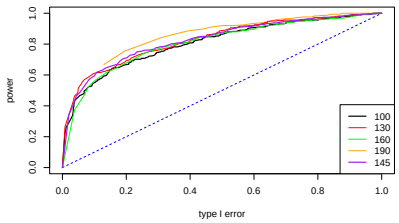

Рис. 5. ROC-кривая ($N = 200$, $\varphi = 0.3$), $L_{\text{optim}} = 145$

на втором примере из раздела 3. Поиск L_{optim} велся на интервале $[100, 190]$ со следующими параметрами: $\alpha^* = 0.05$, $\hat{\alpha}_I = 0.5$, $\varepsilon = 0.05$.

Результат алгоритма: $L_{\text{optim}} = 145$ с $\alpha_I^{(L_{\text{optim}})} = 0.494$. 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.

Литература