МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

МАТЕРИАЛЫ

VI Международной молодежной научной конференции

«МАТЕМАТИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ИНФОРМАЦИОННЫХ, ТЕХНИЧЕСКИХ И ЭКОНОМИЧЕСКИХ СИСТЕМ»

Томск, 24–26 мая 2018 г.

Под общей редакцией
кандидата технических наук И.С. Шмырина

Томск
Издательский Дом Томского государственного университета
2018

3. *Ширяев А.Н.* Вероятность-2. Суммы и последовательности случайных величин – стационарные, мартингалы, марковские цепи : [учеб. для вузов по физ.-мат. направления и специальностям]. Изд. 5-е, стер. – М.: Изд-во МЦНМО, 2011. – С. 555–967.

СКОРЕЙШЕЕ ОБНАРУЖЕНИЕ РАЗЛАДКИ АВТОРЕГРЕССИОННЫХ ПРОЦЕССОВ ПРИ НЕИЗВЕСТНОЙ КОНЕЧНОЙ МОДЕЛИ

А.В. Пупков
Томский государственный университет
andrewpupkov@gmail.com

Введение

В теории стохастических процессов особую роль занимает задача обнаружения разладки. Разладкой называют любое событие, при котором происходит резкое изменение поведения случайного процесса. Например, в качестве такого изменения могут выступать смена математического ожидания случайного процесса, его дисперсии, корреляционной зависимости элементов, смена распределения шума и т.д. В реальной жизни таким изменениям может соответствовать обвал на рынке ценных бумаг, авария на производстве, несанкционированное вторжение в компьютерную сеть, обнаружение космического объекта и многое другое.

Целью задачи обнаружения разладки является наискорейшая реакция на изменение поведения случайного процесса. Для решения этой проблемы создаются статистические процедуры, одной из таких процедур является алгоритм кумулятивных сумм (CUSUM), разработанный Е.С. Пэйджем в 1954 г. в работе [1]. Оптимальность данного алгоритма была доказана Г. Лорденом в 1971 г. [2] для последовательности независимых наблюдений. Для случая зависимых случайных величин доказательство оптимальности привел Т.Л. Лай в 1998 г. [3]. Алгоритм рассматривался в параметрической постановке и требовал знания распределения случайного процесса до и после момента разладки для построения статистик логарифмического правдоподобия. Поскольку алгоритм кумулятивных сумм, в классической постановке, требует большого априорного знания о изучаемом процессе, возникает желание уменьшить это знание без потери эффективности алгоритма. Решение данной проблемы предложили В.В. Конев и С.Э. Воробейчиков в статье [4] в 2017 г. Основная идея предложенных авторами модификаций алгоритма заключается в замене статистик логарифмического правдоподобия на другую систему статистик, которая не зависит от распределения случайного процесса до и после момента разладки и обладает аналогичными статистическими свойствами, что и статистики логарифмического правдоподобия. В работе предполагалось, что в некоторый момент времени происходит резкое изменение значений параметров модели стохастической регрессии, значения которых известны до и после момента разладки. Данное допущение крайне редко выполняется на практике, поскольку до разладки идентификация параметров модели может быть произведена, а после разладки это не представляется возможным. В данной статье рассматривается модификация алгоритма, предложенного в [4], для обнаружения разладки процесса авторегрессии первого порядка AR(1) при неизвестном значении параметра процесса после момента отклонения.

1. Случай известного параметра модели после разладки

Рассмотрим процесс авторегрессии первого порядка AR(1), в котором происходит смена значения корреляционного коэффициента в момент времени ν

$$\begin{aligned} x_n &= \theta_0 x_{n-1} + \varepsilon_n, \ n = 1, 2, \ldots, \nu - 1, \\ x_n &= \theta_1 x_{n-1} + \varepsilon_n, \ n \geq \nu, \end{aligned} \tag{1}$$

где θ_0, θ_1 ($\theta_0 \neq \theta_1$) – значения параметров процесса до и после момента разладки соответственно, $\{\varepsilon_n\}_{n\geq 1}$ – белый шум, распределение которого предполагается неизвестным. Начальное значение x_0 и процесс $\{\varepsilon_n\}_{n\geq 1}$ независимы. Предполагается, что процесс $\{x_n\}_{n\geq 0}$ измерим относительно фильтрации $\{F_n\}_{n\geq 0}$ такой, что $F_0 = \sigma(x_0)$, $F_n = \sigma(x_0, \varepsilon_1, ..., \varepsilon_n)$, $n \geq 1$. Заметим также, что $E(\varepsilon_n \mid F_{n-1}) = 0$, $E(\varepsilon_n^2 \mid F_{n-1}) = 1$.

Необходимо по наблюдениям процесса наискорейшим образом среагировать на изменение параметра. Рассмотрим упрощенный вариант процедуры, предложенной в [4]. Для решения поставленной задачи рассматривается функционал вида

$$J(m) = 2\sum_{n=1}^{m-1} x_{n-1}(\theta_1 - \theta_0)\left(x_n - \frac{x_{n-1}(\theta_1 + \theta_0)}{2}\right) + C_N,$$

где

$$C_N = \sum_{n=1}^{N}(x_n - \theta_1 x_{n-1})^2.$$

Приращения функционала имеют вид

$$\Delta J(m) = J(m+1) - J(m) = 2x_{m-1}(\theta_1 - \theta_0)\left(x_m - \frac{x_{m-1}(\theta_1 + \theta_0)}{2}\right). \quad (2)$$

Статистики вида (2) обладает свойствами, аналогичными свойствам статистик логарифмического правдоподобия. В частности, математическое ожидание данных статистик до момента разладки имеет значение меньше нуля, а после – больше. Именно этот факт позволяет использовать данные статистики в процедуре CUSUM, вместо статистик логарифмического правдоподобия. Данные статистики является более предпочтительными, поскольку для их использования не нужно знать распределение случайного процесса до и после момента отклонения.

Процедура CUSUM имеет следующий вид

$$N = \inf\left(n \geq 1 : \max_{1\leq k\leq n} \sum_{i=k}^{n} \Delta J(i) \geq C\right), \quad (3)$$

где C – некоторый порог.

Процедура хорошо реагирует на изменение коэффициента корреляции, но требует знание значения этого параметра после разладки, что на практике не всегда выполнимо. Следовательно, возникает естественное желание модифицировать процедуру таким образом, чтобы она позволяла отслеживать разладку, не зная значение параметра θ_1. Предложенная модификация рассматривается в следующем разделе.

2. Случай неизвестного параметра после разладки

Рассмотрим модель типа (1) при условии, что параметр авторегрессии θ_1 после момента разладки неизвестен и переобозначим его как θ. Значение неизвестного параметра принадлежит некоторому множеству $\theta \in \Theta \subset (-\infty, +\infty)$. Рассмотрим функцию от параметра θ следующего вида

$$\Delta J_m(\theta) = J_{m+1}(\theta) - J_m(\theta) = 2x_{m-1}(\theta - \theta_0)\left(x_m - \frac{x_{m-1}(\theta + \theta_0)}{2}\right). \quad (4)$$

Для модификации процедуры используем метод, аналогичный тому, который применяется в обобщенном методе кумулятивных сумм (Generalized CUSUM) [5]. Основная идея этого метода заключается в максимизации статистик логарифмического правдоподобия по неизвестному параметру. Используем аналогичный подход для статистик (4). Модифицированные статистки примут вид

$$\Lambda_n^{k+1} = \sup_{\theta\in\Theta} \sum_{i=k+1}^{n} \Delta J_i(\theta).$$

Модифицированная процедура кумулятивных сумм имеет вид

$$N^* = \inf\left(n \geq 1 : \max_{1\leq k\leq n} \Lambda_n^k \geq C\right), \tag{5}$$

где C – порог.

Данная модификация является привлекательной с практической точки зрения, поскольку при её использовании требуется только знание функционального вида модели, используемой для аппроксимации случайного процесса, и знание значения параметра модели до разладки.

В следующем разделе рассмотрим реализацию предложенной процедуры.

3. Численное моделирование

Рассмотрим процесс AR(1) вида

$$x_n = 0.4x_{n-1} + \varepsilon_n, \ n = 1,2,...,\nu-1,$$

$$x_n = -0.5x_{n-1} + \varepsilon_n, \ n \geq \nu,$$

где $\nu = 100$, $\varepsilon_n \sim N(0,1)$, т.е. в качестве функции распределения шума выступает нормальное распределение с нулевым средним и единичной дисперсией.

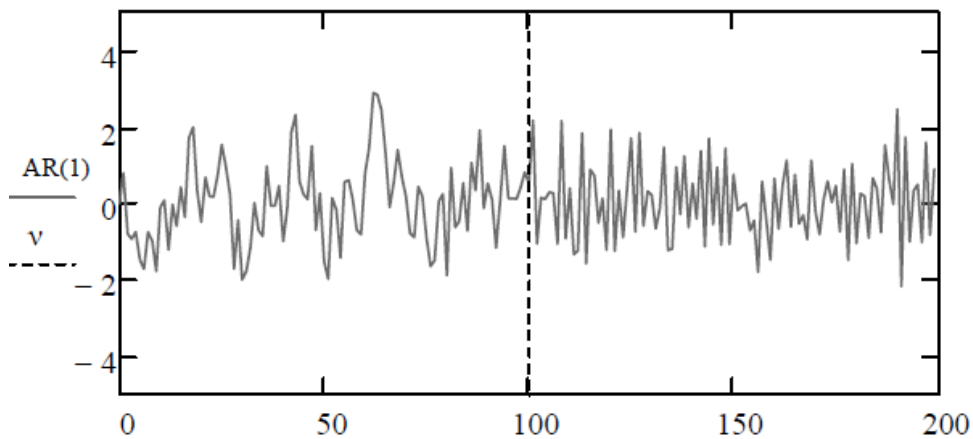

Рис. 1. Процесс AR(1) с разладкой ($\theta_0 = 0.4$, $\theta_1 = -0.5$, $\nu = 100$)

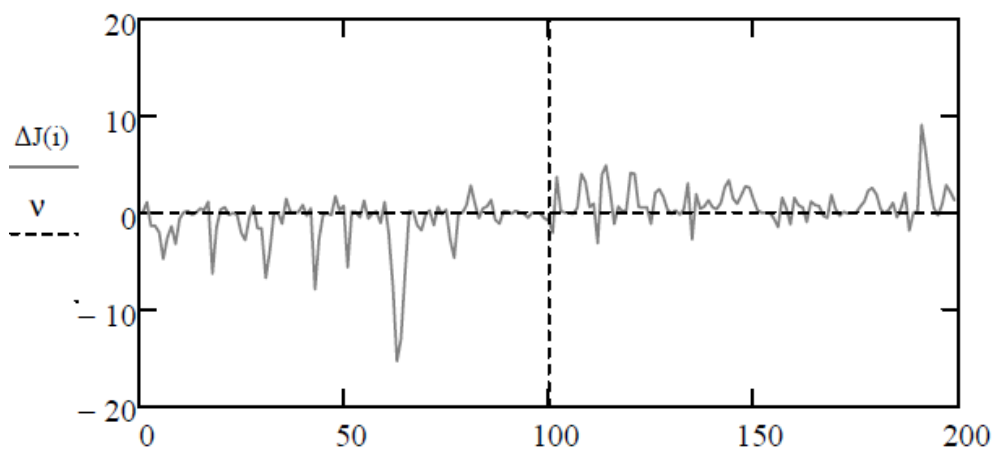

Рис. 2. Последовательность статистик $\Delta J(m)$

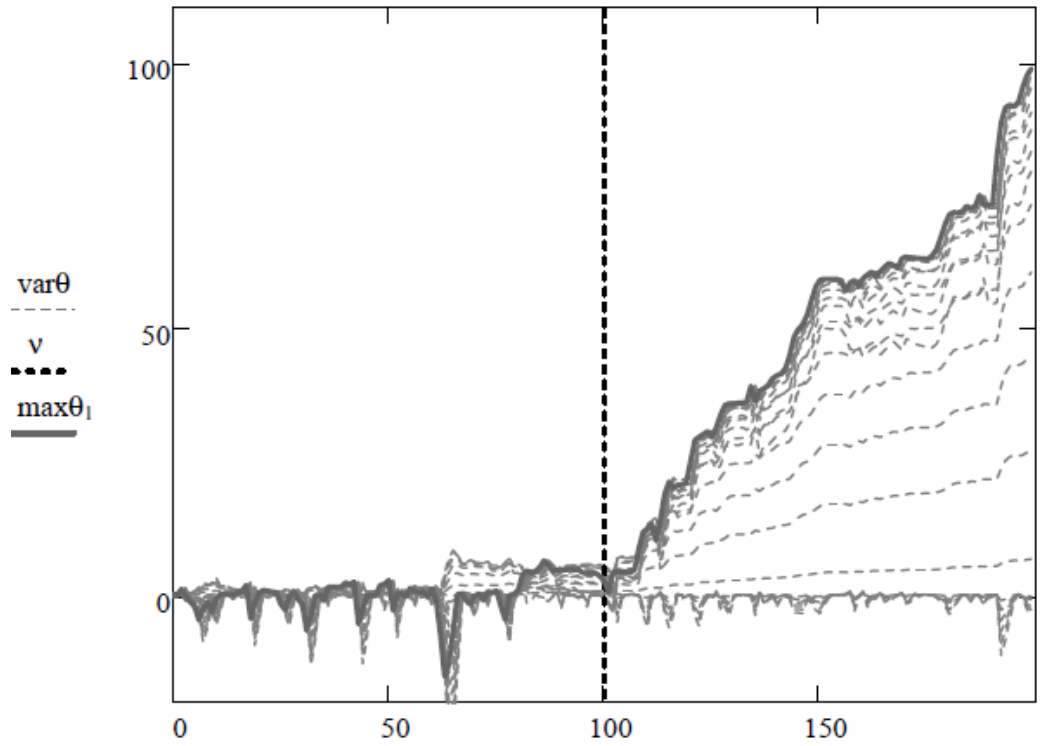

Рис. 3. Реализация процедуры при вариации неизвестного параметра θ

На рис. 1 представлена реализация процесса AR(1) с разладкой в момент времени $\nu = 100$, в который происходит смена значения параметра процесса. На рис. 2 изображена реализация последовательности статистик (2). Видно, что среднее статистик до момента разладки меньше нуля, а после разладки – больше нуля. На рис. 3 изображена реализация процедуры (5). В частности, сплошной темной линией ($\max\theta_1$) изображена последовательность максимумов сумм статистик до момента *n*, при условии, что известно значение параметра процесса после момента отклонения. Последовательность представлена статистиками вида

$$\left\{\max_{1\le k\le n}\sum_{i=k}^{n}\Delta J(i)\right\}_{n\ge 1}.$$

Тонкими пунктирными линиями ($\mathrm{var}\,\theta$) изображены последовательности максимумов сумм статистик до момента *n* при вариации неизвестного параметра, т.е.

$$\left\{\max_{1\le k\le n}\sum_{i=k}^{n}\Delta J_i(\theta)\right\}_{n\ge 1},\ \theta\in\{-1,-1+\Delta x,...,1\},$$

где Δx – шаг дискретизации. В данной реализации $\Delta x = 0.1$. Иными словами, производится параллельный запуск непараметрических процедур вида (3) для всех предполагаемых значений неизвестного параметра θ и на каждом шаге *n* максимизируется значение по всем полученным реализациям; т.е. результирующая последовательность представляет из себя верхнюю границу всех реализаций при вариации неизвестного параметра. Видно, что последовательность $\max\theta_1$ (реализация процедуры при известном значении параметра после разладки) приближается к верхней границе реализаций процедуры при вариации неизвестного параметра. Можно сделать предварительный вывод, что предложенная процедура по эффективности близка к процедуре, рассмотренной в [4].

Таблица 1

Частотные характеристики процедуры

C	δFA	ADD	C	δFA	ADD
10	0,13	12,264	20	0	22,28
15	0,01	16,202	25	0	26.84

где C – пороговое значение процедуры, δFA – доля ложных тревог (алгоритм сигнализирует о разладке, когда её нет) в последовательности реализаций процедуры (количество реализаций $m = 100$), ADD – среднее время запаздывания процедуры

$$ADD = \frac{\sum_{t=1}^{m} \left(N_t^* - \nu + 1 \right) \chi_{\left\{ N_t^* \geq \nu \right\}}}{\sum_{t=1}^{m} \chi_{\left\{ N_t^* \geq \nu \right\}}},$$

где N_t^* – момент остановки процедуры при реализации на данных из набора t, χ – индикаторная функция.

Заключение

В данной статье предложена модификация непараметрической процедуры CUSUM, позволяющая отслеживать разладку при неизвестном параметре процесса авторегрессии первого порядка AR(1) после момента отклонения. Представлена реализация процедуры при гауссовости шума и приведены частотные характеристики, демонстрирующие работоспособность алгоритма.

ЛИТЕРАТУРА

1. *Page E.S.* Continuous inspection scheme // Biometrika. – 1954. – Vol. 41. – P. 100–115.
2. *Lorden G.* Procedures for reacting to a change in distribution // The annals of mathematical statistics. – 1971. – Vol. 42. – № 6. – P. 1897–1908.
3. *Lai T.L.* Information bounds and quick detection of parameter changes in stochastic system // IEEE transaction on information theory. – 1998. – Vol. 44. – № 7. – P. 2917–2929.
4. *Konev V., Vorobeychikov S.* Quickest detection of parameter changes in stochastic regression: nonparametric CUSUM // IEEE transaction on information theory. – 2017. – Vol. 63. – № 9 – P. 5588–5602.
5. *Tartakovsky A., Nikiforov I., Basseville M.* Sequentional analysis: hypothesis testing and changpoint detection. – Boca Raton: A CHAPMAN & HALL BOOK, 2015. – 584 p.

ОБРАБОТКА МЕДИЦИНСКИХ ДАННЫХ, СОДЕРЖАЩИХ ИНФОРМАЦИЮ О ФАКТОРАХ СЕРДЕЧНО-СОСУДИСТЫХ ЗАБОЛЕВАНИЙ

Т.Е. Малахова[1], Т.В. Кабанова[1], А.А. Горлова[2]

[1]*Томский государственный университет*

[2]*Научно-исследовательский институт кардиологии*

malahova.tanny@yandex.ru, tvk@bk.ru

Введение

В последние годы очень востребованными являются статистические методы обработки медицинских данных, поскольку математическая статистика разрабатывает методы статистической обработки и анализа данных, занимается обоснованием и проверкой их достоверности, эффективности, условий применения, устойчивости к нарушению условий применения и т.п. [1,2].

В данной работе изучаются факторы эмболического инсульта, характеризующегося закупоркой сосудов головного мозга или других внутренних органов тромбом либо эмболом, который формируется в сердце, а при определённых условиях, с током крови попадает в сосуды.