

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

МАТЕРИАЛЫ
VII Международной молодежной
научной конференции
«МАТЕМАТИЧЕСКОЕ
И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ
ИНФОРМАЦИОННЫХ,
ТЕХНИЧЕСКИХ
И ЭКОНОМИЧЕСКИХ СИСТЕМ»

Томск, 23–25 мая 2019 г.

Под общей редакцией
кандидата технических наук И.С. Шмырина

Томск
Издательский Дом Томского государственного университета
2019

использования предложенной в работе программы для вычисления нормированной тестовой статистики, что позволяет использовать асимптотические результаты работы [1] и позволяет успешно и корректно применять БКР-критерий на практике.

ЛИТЕРАТУРА

1. Blum J.R., Kiefer J., Rosenblatt M. Distribution free tests of independence based on the sample distribution function // Ann. Math. Statist. – 1961. – N 21. P. 485–498.
2. Оуэн Д.Б. Сборник статистических таблиц. Изд. 2-е испр. – М.: ВЦ АН СССР, 1973. – 586 с.
3. Холлендер М., Вулф Д. Непараметрические методы статистики. – М.: Финансы и статистика, 1983. – 518 с.
4. Шуленин В.П. Дополнительные главы математической статистики (курс лекций). – Томск: Изд-во НТЛ, 2018. – 516 с.

АСИМПТОТИЧЕСКИ ЭФФЕКТИВНОЕ ОЦЕНИВАНИЕ В НЕПАРАМЕТРИЧЕСКОЙ ГЕТЕРОСКЕДАСТИЧНОЙ РЕГРЕССИИ*

С.С. Перелевский, Е.А. Пчелинцев

Томский государственный университет

Введение

Рассмотрим задачу асимптотически эффективного оценивания неизвестной функции $S(\cdot)$ гетероскедастичной регрессии, описываемую уравнением

$$y_j = S(x_j) + \sigma_j \xi_j, \quad 1 \leq j \leq n, \quad (1)$$

где $x_j = j/n$, $(\xi_j)_{1 \leq j \leq n}$ – последовательность независимых одинаково распределённых случайных величин с нулевым средним и дисперсией равной единице, $(\sigma_j)_{1 \leq j \leq n}$ – коэффициенты волатильности, зависящие от x_j и функции S , n – число наблюдений.

Понятие асимптотической эффективности связано с оптимальной скоростью сходимости минимаксного риска, т.е. важным вопросом в результатах оптимальности является изучение точной асимптотики минимаксного риска, определенного ниже в (5). Проблема асимптотического непараметрического оценивания в модели гетероскедастичной регрессии изучались Эфроймовичем [1] и Пинскером [2]. Для доказательства асимптотической эффективности процедуры, необходимо показать, что ее асимптотический квадратичный риск совпадает с нижней границей, определяемой константой Пинскера [3]. В данной работе поставленная задача решается с использованием подхода, основанного на методах выбора модели и точных оракульных неравенств. Напомним, что метод выбора моделей появился в пионерских работах Акайке [4] и Мэллоуза [5], которые предложили ввести пенализационное слагаемое в критерий максимального правдоподобия. Далее, Барон, Бирже и Массар [6], Массар [7] и Кнайп [8] развили этот метод для получения неасимптотических оракульных неравенств в непараметрических регрессионных моделях с гауссовскими шумами в дискретном времени. К сожалению, этот метод не может быть применен в нашем случае для доказательства эффективности, поскольку коэффициент в основном слагаемом, в полученных оракульных неравенствах больше единицы. По этой причине в данной работе для эффективного оценивания функции в модели (1) применяется метод построения оракульных неравенств, развитый в [9]. В данной работе исследуется свойство асимптотической эффективности улучшенной процедуры выбора модели, предложенной в [10], для оценивания функции S в (1).

* Работа выполнена в рамках государственного задания Министерства науки и высшего образования РФ (проект № 2.3208.2017/4.6).

1. Оракульное неравенство для квадратического риска адаптивной процедуры выбора модели

Для оценивания неизвестной функции S в модели (1) предлагается использовать улучшенные взвешенные оценки МНК, определенные в [10],

$$S_{\lambda}^*(x) = \sum_{j=1}^n \lambda(j) \theta_{j,n}^* \phi_j(x), \quad (2)$$

где $\theta_{j,n}^* = \left(1 - \frac{c_n}{\|\tilde{\theta}_n\|}\right) \hat{\theta}_{j,n}$, $\|\tilde{\theta}_n\|^2 = \sum_{j=1}^d \hat{\theta}_{j,n}^2$, $\hat{\theta}_{j,n} = \theta_j + \frac{1}{n} \xi_{j,n}$, $\xi_{j,n} = \frac{1}{\sqrt{n}} \sum_{l=1}^n \sigma_l \xi_l \phi_j(x_l)$. Здесь

коэффициент $d \approx n^\delta / \ln n$, $0 < \delta < 1$, $c_n \approx d/n$. Ясно, что $c_n \rightarrow 0$ при $n \rightarrow \infty$.

Метод выбора модели заключается в описании правила выбора процедуры из семейства оценок (2), т.е. нужно выбрать вектор весовых коэффициентов $\lambda \in \Lambda$ в (2). Для этого естественно минимизировать эмпирическую квадратическую ошибку вида

$$Err_n(\lambda) = \|S_{\lambda}^* - S\|_n^2 = \frac{1}{n} \sum_{l=1}^n (S_{\lambda}^*(x_l) - S(x_l))^2 = \sum_{j=1}^n \lambda^2(j) \theta_{j,n}^{*2} - 2 \sum_{j=1}^n \lambda(j) \theta_{j,n}^* \theta_j + \|S\|^2.$$

Однако, поскольку во втором слагаемом коэффициенты θ_j неизвестны, то минимизация этого выражения, не приводит к нахождению коэффициентов λ . Необходимо величины $\theta_{j,n}^* \theta_j$ заменить некоторыми оценками. В качестве таких оценок предлагаются

$\bar{\theta}_{j,n} = \theta_{j,n}^* \hat{\theta}_{j,n} - \frac{1}{n} \hat{\varsigma}_n$, где $\hat{\varsigma}_n$ – некоторая оценка интегрированной дисперсии шума

$\varsigma_n = \frac{1}{n} \sum_{j=1}^n \sigma_j^2$. При осуществлении этой замены в эмпирической ошибке нужно заплатить "штраф". Определим платежную функцию как

$$J_n(\lambda) = \sum_{j=1}^n \lambda^2(j) \theta_{j,n}^{*2} - 2 \sum_{j=1}^n \lambda(j) \bar{\theta}_{j,n} + \rho \hat{P}_n(\lambda). \quad (3)$$

Здесь $0 < \rho < 1$ – некоторый положительный коэффициент, зависящий от n и такой, что

$\rho(n) \rightarrow 0$ при $n \rightarrow \infty$ и пенализационное слагаемое $\hat{P}_n(\lambda) = \frac{|\lambda|^2 \hat{\varsigma}_n}{n}$, $|\lambda|^2 = \sum_{j=1}^n \lambda^2(j)$. От-

метим, что в случае, когда последовательность $(\sigma_j)_{1 \leq j \leq n}$, известна, то $P_n(\lambda) = \frac{|\lambda|^2 \varsigma_n}{n}$.

Полагая $\lambda^* = \arg \min_{\lambda \in \Lambda} J_n(\lambda)$, определим процедуру выбора модели равенством

$$S^* = S_{\lambda^*}^*. \quad (4)$$

Заметим, что λ^* существует, поскольку множество Λ конечно. В случае, когда λ^* не единственное, берем любое из них.

Обозначим через P_n семейство распределений p в $\mathbb{R}^n$ векторов $(\xi_1, \dots, \xi_n)$ в модели (1) таких, что компоненты ξ_j совместно независимы, центрированы с единичной дисперсией и $\max_{1 \leq k \leq n} E(\xi_k^4) \leq l_n^*$, где l_n^* – медленно возрастающая последовательность, т.е.

$\lim_{n \rightarrow \infty} l_n^* = \infty$, $\lim_{n \rightarrow \infty} l_n^*/n^\delta = 0$ для любого $\delta > 0$. Легко видеть, что для любого $n \geq 1$ центри-

рованное гауссовское распределение в $\mathbb{R}^n$ с единичной ковариационной матрицей принадлежит семейству P_n . Обозначим это распределение через q . Для оценки S^* определим следующий робастный среднеквадратический риск

$$\mathfrak{R}(S^*, S) = \sup_{p \in P_n} \mathbf{E}_{S,p} \|S^* - S\|_n^2, \quad (5)$$

где $\mathbf{E}_{S,p}$ – математическое ожидание относительно распределения $\mathbf{P}_{S,p}$ по наблюдениям $(y_1, \dots, y_n)$ с фиксированной функцией S и фиксированным распределением $p \in P_n$ случайных величин $(\xi_j)_{1 \leq j \leq n}$ в модели (1).

В [10] получено следующее оракульное неравенство, определяющее неасимптотическую верхнюю границу для среднеквадратического риска (5) процедуры оценивания (4).

Теорема 1. Пусть наблюдения описываются уравнением (1). Тогда для любого $0 < \rho < 1/3$ среднеквадратический риск (5), предложенной процедуры выбора модели (4) для функции S удовлетворяет оракульному неравенству: $\mathfrak{R}(S^*, S) \leq \frac{1-\rho}{1-2\rho} \min_{\lambda \in \Lambda} \mathfrak{R}(S_\lambda^*, S) + \frac{1}{n\rho} \Psi_n$, где $\lim_{n \rightarrow \infty} \Psi_n / n^\delta = 0$ для любого $\delta > 0$.

2. Асимптотическая эффективность

Пусть $C(\mathbb{R})$ – множество 1-периодических k раз дифференцируемых $\mathbb{R} \rightarrow \mathbb{R}$ функций. Предположим, что S принадлежит следующему множеству:

$$W_r^k = \left\{ f \in C_{per,1}^k(\mathbb{R}) : \sum_{j=0}^k \|f^{(j)}\|^2 \leq r \right\},$$

где $\|\cdot\|$ обозначает норму в $L_2[0,1]$. Кроме того, предположим, что $r > 0$ и $k \geq 1$ – неизвестные параметры. Заметим, что множество W_r^k может быть представлено как эллипс в $L_2[0,1]$, т.е.

$$W_r^k = \left\{ f \in L_2[0,1] : \sum_{j=1}^{\infty} a_j \theta_j^2 \leq r \right\}, \quad \theta_j = (f, \phi_j) = \int_0^1 f(t) \phi_j(t) dt, \quad a_j = \sum_{l=0}^k \|\phi_j^{(l)}\|^2 = \sum_{i=0}^k (2\pi[j/2])^{2i}.$$

Обозначим Σ_n – множество всех оценок $\hat{S}_n$ измеримых относительно наблюдений $(y_1, \dots, y_n)$. Предположим, что коэффициенты волатильности $(\sigma_j(S))_{j \geq 1}$ удовлетворяют следующим условиям.

C₁) $\sigma_j(S) = g(x_j, S)$ для некоторой неизвестной функции, $g : [0, 1] \times L_1[0, 1] \rightarrow \mathbb{R}_+$ которая квадратично интегрируема по x и такая, что

$$\limsup_{n \rightarrow \infty} \left| \frac{1}{n} \sum_{j=1}^n g^2(x_j, S) - \zeta(S) \right| = 0,$$

где $\zeta(S) = \int_0^1 g^2(x, S) dx$. Кроме того, $g_* = \inf_{0 \leq x \leq 1} \inf_{W_r^k} g^2(x, S) > 0$, $\zeta(S) < \infty$.

C₂) Для некоторого $x \in [0, 1]$, оператор $g^2(x, \cdot) : C[0, 1] \rightarrow \mathbb{R}$ дифференцируема в смысле Фреше при любой фиксированной функции f_0 из $C[0, 1]$, т.е. для некоторого f из некоторой окрестности f_0 в $C[0, 1]$,

$$g^2(x, f) = g^2(x, f_0) + L_{x, f_0}(f - f_0) + Y(x, f_0, f),$$

где производная по Фреше $L_{x, f_0}(f - f_0) : C[0, 1] \rightarrow \mathbb{R}$ – ограниченный линейный оператор, а остаточный член $Y(x, f_0, f)$ для каждого $x \in [0, 1]$, удовлетворяет следующему свойству:

$$\lim_{\|f-f_0\|_\infty \rightarrow 0} \frac{\|Y(x, f_0, f)\|}{\|f-f_0\|_\infty} = 0, \quad \|f\|_\infty = \sup_{0 \leq t \leq 1} |f(t)|.$$

C₃) Существует некоторая положительная постоянная C^* такая, что для некоторой функции S из $C[0,1]$, оператор $L_{x,S}$, определенный в C₂), удовлетворяет следующему неравенству для любой функции f из $C[0,1]$:

$$|L_{x,S}(f)| \leq C^* (|S(x)f(x)| + |f|_1 + \|S\| \|f\|), \quad |f|_1 = \int_0^1 |f(t)| dt.$$

C₄) Функция $g_0(\cdot) = g(\cdot, S_0)$, соответствующая $S_0 \equiv 0$, является непрерывной на интервале $[0,1]$. Кроме того,

$$\limsup_{\delta \rightarrow \infty} \sup_{0 \leq x \leq 1, \|S\|_\infty \leq \delta} |g(x, S) - g(x, S_0)| = 0.$$

Справедливы следующие асимптотические результаты.

Теорема 2. Пусть для модели (1) выполнены условия C1) – C4). Тогда робастный риск (5) для всех $k \geq 1$ и $r > 0$ удовлетворяет следующему асимптотическому неравенству

$$\liminf_{n \rightarrow \infty} n^{\frac{2k}{2k+1}} \inf_{S_0 \in \Sigma_1} \sup_{S \in W_r^k} \frac{\mathfrak{R}(\hat{S}_n, S)}{\gamma_k(S)} \geq 1,$$

где $\gamma_k(S) = \Gamma_k^*(\zeta(S))^{2k/(2k+1)}$, $\Gamma_k^* = ((2k+1)r)^{1/(2k+1)} (k/(\pi(k+1)))^{2k/(2k+1)}$ – константа Пинскера.

Теорема 3. Пусть для модели (1) выполнены условия C1) – C4). Тогда робастный риск (5) процедуры выбора модели S^* для всех $k \geq 1$ и $r > 0$ удовлетворяет следующей асимптотической верхней границе

$$\limsup_{n \rightarrow \infty} n^{\frac{2k}{2k+1}} \sup_{S \in W_r^k} \frac{\mathfrak{R}(S_{\lambda^*}^*, S)}{\gamma_k(S)} \leq 1.$$

Следствие 1. Из теорем 2 и 3 следует, что процедура выбора модели $S_{\lambda^*}^*$, определенная в (4), является асимптотически эффективной, т.е.

$$\lim_{n \rightarrow \infty} n^{\frac{2k}{2k+1}} \sup_{S \in W_r^k} \frac{\mathfrak{R}(S_{\lambda^*}^*, S)}{\gamma_k(S)} = 1.$$

3. Численное моделирование

В этом разделе проиллюстрируем теоретически установленные результаты с помощью численного моделирования в среде Matlab. В качестве функции S в модели (1) выберем

$$S_1(x) = x \sin(2\pi x) + x^2(1-x) \cos(4\pi x), \quad S_2(x) = \begin{cases} 0.5 - |x - 0.5|, & 0 \leq x \leq 1, \\ 0, & \text{иначе,} \end{cases}$$

$(\xi_j)_{1 \leq j \leq n}$ – последовательность независимых $N(0,1)$ случайных величин, коэффициенты волатильности $\sigma_j^2 = 2 + S(x_j)$. Для вычисления весовых коэффициентов $(\lambda_\alpha)_{\alpha \in A}$, определенных в [10], положим

$$k^* = 100 + \sqrt{\ln n}, \quad \varepsilon = \frac{1}{\ln n}, \quad m = \lfloor \ln^2 n \rfloor, \quad \rho = \frac{1}{3 + \ln^2 n}, \quad \omega_\alpha = 100 + (A_\beta m)^{\frac{1}{2\beta+1}}.$$

В табл. 1 приведены результаты поведения эмпирических среднеквадратических рисков по $N = 1000$ реализациям процедур выбора модели (4).

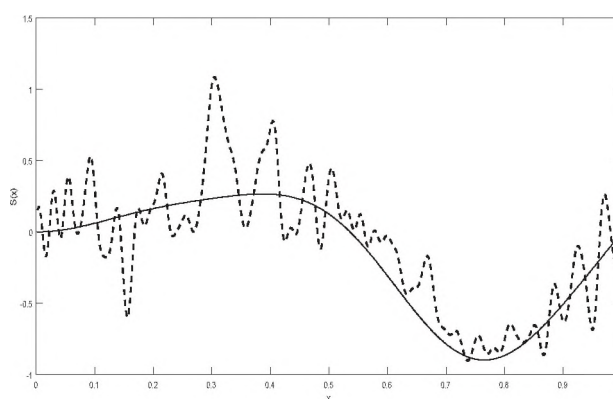
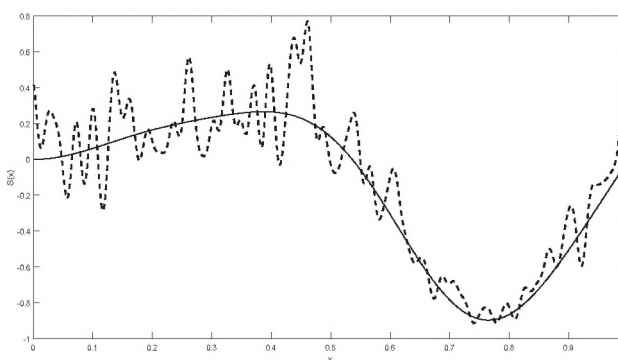
Таблица 1

Эмпирические квадратичные асимптотические риски

n	101	501	1001	2001	10001
$\frac{\sum_{i=1}^n F(S_i^*, S_i)}{n \gamma_e(S_i)}$	8.1458	5.7257	3.0856	1.0009	0.9125
$\frac{\sum_{i=1}^n F(S_i^*, S_i)}{n \gamma_e(S_i)}$	11.1257	6.1254	3.1587	1.0015	0.9321

Из табл. 1 видно, что с ростом числа наблюдений n нормированные эмпирические среднеквадратические риски стремятся к единице, что численно подтверждает следствие 1.

Далее на рисунках представлены графики истинных функций S_1 и S_2 (сплошная линия) и их оценок (4) при $n = 501$; 1001; 10001.

Рис. 1. Графики функции регрессии и ее оценок при $n = 501$ Рис. 2. Графики функции регрессии и ее оценок при $n = 1001$

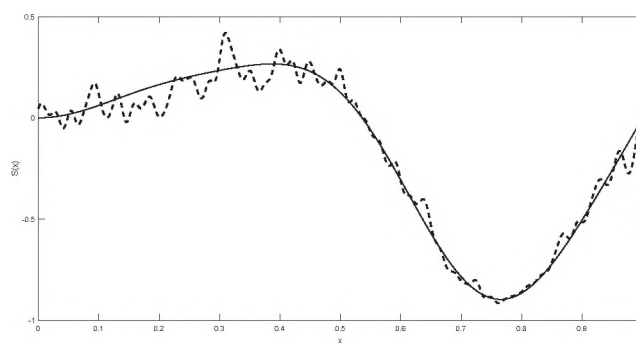


Рис. 3. Графики функции регрессии и ее оценок при $n = 10001$

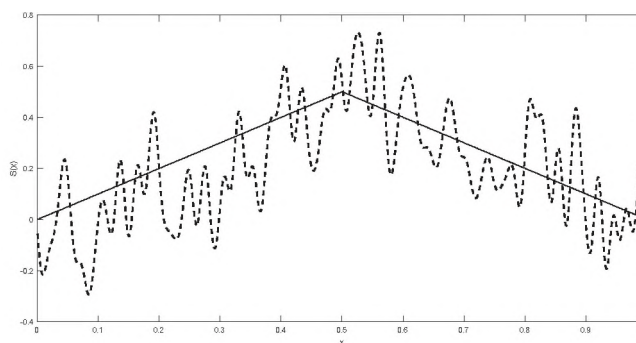


Рис. 4. Графики функции регрессии и ее оценок при $n = 501$

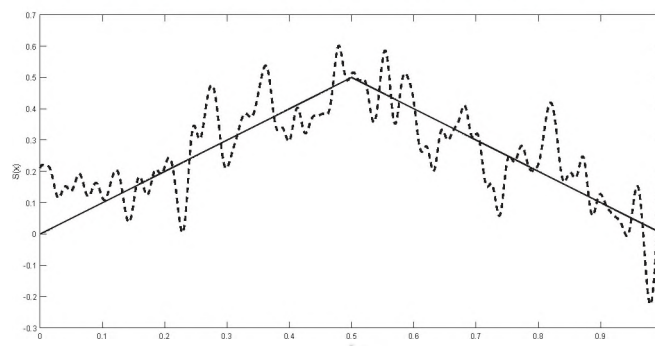


Рис. 5. Графики функции регрессии и ее оценок при $n = 1001$

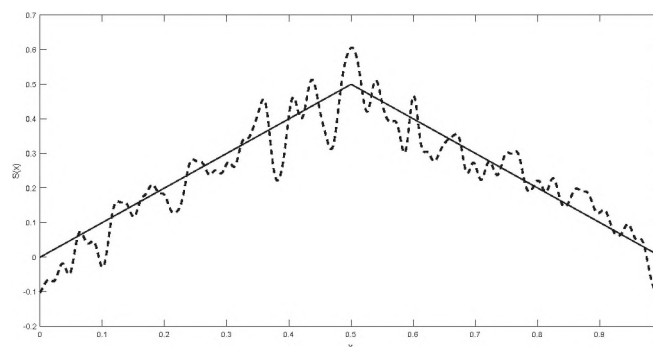


Рис. 6. Графики функции регрессии и ее оценок при $n = 10001$

ЛИТЕРАТУРА

1. Efroimovich S. Sequential design and estimation in heteroscedastic nonparametric regression // Sequential Analysis. – 2007. – N 26. – P. 3-25.
2. Pinsker M.S. Optimal filtration of square integrable signals in gaussian white noise // Problems Transimis. information. – 1981. – N 17. – P. 120-133.
3. Galtchouk L., Pergamenshchikov S. Asymptotic efficient sequential kernel estimates of the drift coefficient in ergodic diffusion processes // Statistical Inference for Stochastic Process. – 2006. – V. 9. – N 1. – P. 1-16.
4. Akaike H. A new look at the statistical model identification // IEEE Trans. on Automatic Control. – 1974. – V. 19. – N 7. – P. 716-723.
5. Mallows C. Some comments on Cp // Technometrics. – 1973. – V. 15. – P. 661-675.
6. Barron A., Birge L., Massart P. Risk bounds for model selection via penalization // Probab. Theory Relat. Fields. – 1999. – V. 113. – P. 301-415.
7. Massart P. A non-asymptotic theory for model selection // European Congress of Mathematics. Zurich: Eur. Math. Soc., 2005.
8. Kneip A. Ordered linear smoothers // Annals of Statistics. – 1994. – V. 22. – P. 835-866.
9. Galtchouk L., Pergamenshchikov S. Adaptive asymptotically efficient estimation in heteroscedastic nonparametric regression // Journal of the Korean Statistical Society. – 2009. – V. 38. – N 4. – P. 305-322.
10. Пчелинцев Е.А., Перелевский С.С. Адаптивное оценивание в гетероскедастичной непараметрической регрессии // Вестник ТГУ. Математика и механика. – 2019. – № 57. – С. 38-52.

ОБНАРУЖЕНИЕ РАЗЛАДКИ В ГАУССОВСКОМ ПРОЦЕССЕ AR(1) С НЕИЗВЕСТНЫМ ПАРАМЕТРОМ

А.В. Пупков

Томский государственный университет

Введение

В данной работе рассматривается подход к задаче обнаружения разладки в процессе AR(1) с неизвестным параметром после момента отклонения. Основная идея заключается в последовательной оценке параметра гауссовской авторегрессии по особой схеме. Используя эту оценку, можно получить статистику с неасимптотическим гауссовским распределением, дисперсию которой мы можем контролировать. В момент, когда происходит разладка процесса авторегрессии, меняется распределение полученной статистики. На основе этого изменения отслеживается разладка.

1. Постановка задачи

Рассмотрим процесс AR(1), с разладкой в момент времени v , вида

$$\begin{aligned}x_k &= \theta_0 x_{k-1} + \varepsilon_k, \quad k = \overline{1, v-1}, \\x_k &= \theta x_{k-1} + \varepsilon_k, \quad k \geq v,\end{aligned}$$

где $x_0 = 0$, $\varepsilon_k \sim N(0, 1)$, параметры принадлежат промежутку устойчивости процесса θ , $\theta_0 \in (-1, 1)$.

Введем последовательность моментов остановки

$$\tau_i(h) = \inf \left\{ n > \tau_{i-1} : \sum_{k=\tau_{i-1}+1}^n x_{k-1}^2 \geq h \right\}, \quad \inf \{\emptyset\} = \infty, \quad (1)$$

где $\tau_0 = \tau_0(h) \equiv 0$. Далее, на каждом полученном интервале $\Delta \tau_i = (\tau_{i-1}, \tau_i]$, оценим значение параметра по следующей схеме

$$\hat{\theta}_i = \frac{1}{\tilde{h}_i} \sum_{k=\tau_{i-1}+1}^{\tau_i} \sqrt{\alpha_i(k)} x_{k-1} x_k, \quad (2)$$

$$\text{где } \tilde{h}_i = \sum_{k=\tau_{i-1}+1}^{\tau_i} \sqrt{\alpha_i(k)} x_{k-1}^2,$$