

УДК 512.647.2

Костеров М. А.

Вычисление коэффициентов параболических аппроксимационно-оценочных критериев при помощи QR-разложения

Рекомендовано к публикации профессором Котиной Е. Д.

1. Введение. Пусть $\mathcal{T} = \{1, \dots, m\}$, семейство $\xi = \{\xi_t, t \in \mathcal{T}\}$ случайных величин $\xi_t = \xi_t(\omega)$, заданных для любого $t \in \mathcal{T}$ на вероятностном пространстве $(\Omega, \mathcal{F}, \mathbf{P})$, называют дискретным случайным процессом [1], который можно рассматривать как функцию двух переменных $\xi_t(\omega) = \xi(t, \omega)$, где $t \in \mathcal{T}$, $\omega \in \Omega$. Если зафиксировать $t \in \mathcal{T}$, получим случайную величину ξ_t . Если зафиксировать $\omega_0 \in \Omega$, получим последовательность $y_t = \xi_t(\omega_0)$, которую называют траекторией случайного процесса ξ . Далее будем рассматривать только процессы с монотонно возрастающими траекториями.

Задача, которую решают аппроксимационно-оценочные критерии (АОК), заключается в определении момента t_0 , в котором траектория y_t переходит от линейного роста к нелинейному [2]. Причем для определения этого момента используются только k значений траектории $y_{t_0-k+1}, y_{t_0-k+2}, \dots, y_{t_0}$, из левой полукрестности точки t_0 . Эти критерии можно рассматривать как аналитическую формулировку эвристики «метода локтя» [3]. Определим величины

$$\delta_f^2(k_0) = \min_{f \in \mathcal{Y}} \sum_{i=1}^k (f(i) - y_{k_0-k+i})^2, \quad \delta_l^2(k_0) = \min_{a,b} \sum_{i=1}^k (ai + b - y_{k_0-k+i})^2,$$

где $\mathcal{Y}$ – класс функций вида $a\varphi(x) + b$, φ – непрерывная монотонно возрастающая нелинейная функция и $y_{k_0-k+1}, y_{k_0-k+2}, \dots, y_{k_0}$ – скользящее окно размера k . Минимумы существуют, поскольку рассматриваются положительно квазиопределенные квадратичные формы [2].

Квадратичные формы $\delta_f^2(k_0)$ и $\delta_l^2(k_0)$ характеризуют «близость» траектории y_t в рассматриваемом окне к функциям из класса $\mathcal{Y}$ и к

Костеров Максим Андреевич – студент, Санкт-Петербургский государственный университет; e-mail: maxim.kosterv@gmail.com, тел.: +7(951)668-77-69

линейным функциям вида $ax + b$ соответственно. Определим теперь величину $\delta^2(k_0) = \delta_l^2(k_0) - \delta_f^2(k_0)$ и значение $t_0 = \min\{k_0 \mid \delta^2(k_0) > 0\}$. Согласно АОК, процесс переходит от линейного роста к нелинейному между точками $t_0 - 1$ и t_0 , поскольку окно $y_{t_0-k}, \dots, y_{t_0-1}$ «ближе» к линейной функции, а окно $y_{t_0-k+1}, \dots, y_{t_0}$ — уже к нелинейной из класса $\mathcal{Y}$.

В данной работе рассмотрим параболические АОК, т. е. выберем $\varphi(x) = (x - 1)^2$. Эти критерии могут быть использованы в механике твердого тела [4, 5], в многомерных статистических методах [6], для нахождения особых точек кинетической кривой ПЦР [7] и автоматического обнаружения аномалий сетевого трафика при DDoS-атаках методами последовательного статистического анализа [8].

2. Квадратичные формы АОК и QR-разложение. Приведем величину $\delta^2(k_0)$ к более удобному виду. Зафиксируем размер окна k и функцию φ . Введем $e = (1, \dots, 1)^T \in \mathbb{R}^k$ и определим функцию $\rho^2 : \mathbb{R}^k \times \mathbb{R}^k \rightarrow \mathbb{R}$,

$$\rho^2(g, x) = \min_{\alpha, \beta} (\alpha g + \beta e - x)^2 = \min_{\alpha, \beta} \sum_{i=1}^k (\alpha g_i + \beta - x_i)^2.$$

Зададим векторы $l, f \in \mathbb{R}^k$ через их компоненты: $l_i = i$, $f_i = \varphi(i)$. Обозначим $z = (y_{k_0-k+1}, \dots, y_{k_0})^T$. Тогда

$$\begin{aligned} \delta_l^2(k_0) &= \min_{a, b} \sum_{i=1}^k (ai + b - y_{k_0-k+i})^2 = \min_{\alpha, \beta} (\alpha l + \beta e - z)^2 = \rho^2(l, z), \\ \delta_f^2(k_0) &= \min_{g \in \mathcal{Y}} \sum_{i=1}^k (g(i) - y_{k_0-k+i})^2 = \min_{a, b} \sum_{i=1}^k (a\varphi(i) + b - y_{k_0-k+i})^2 = \\ &= \min_{\alpha, \beta} (\alpha f + \beta e - z)^2 = \rho^2(f, z). \end{aligned}$$

Положим, наконец, что $\rho_2 : \mathbb{R}^k \rightarrow \mathbb{R}$, $\rho_2(x) = \rho^2(l, x) - \rho^2(f, x)$ и получим

$$\delta^2(k_0) = \delta_l^2(k_0) - \delta_f^2(k_0) = \rho^2(l, z) - \rho^2(f, z) = \rho_2(z). \quad (1)$$

Таким образом, определение значения величины $\delta^2(k_0)$ сводится к вычислению $\rho_2(x)$ по $x \in \mathbb{R}^k$.

Изучим $\rho^2(g, x)$ как функцию от x при фиксированном g . Пусть $\{g, e\}$ линейно независимы. Введем матрицы $M = (g, e) \in \mathbb{R}^{k \times 2}$ и $S = E - M(M^T M)^{-1} M^T$, где E – единичная матрица. Согласно [9], S – матрица ортогонального проектирования на линейное подпространство, ортогональное $\{g, e\}$, $S^2 = S = S^T$ и

$$\begin{aligned} \rho^2(g, x) &= \min_{\alpha, \beta} (\alpha g + \beta e - x)^2 = (Sx)^2 = x^T S^T S x = x^T S x = \\ &= x^2 - x^T M (M^T M)^{-1} M^T x. \end{aligned} \quad (2)$$

Для решения задач линейного метода наименьших квадратов используют QR-разложение [10], поэтому рассмотрим

$$\tilde{g} = g - \frac{e^T g}{e^2} e, \quad Q = (\tilde{g}, e), \quad L = \begin{pmatrix} 1 & 0 \\ e^T g / e^2 & 1 \end{pmatrix}.$$

Заметим, что $\tilde{g}^T e = 0$ и L невырождена. Нетрудно убедиться, что $M = QL$. Тогда

$$\begin{aligned} M(M^T M)^{-1} M^T &= QL(L^T(Q^T Q)L)^{-1} L^T Q^T = \\ &= QLL^{-1}(Q^T Q)^{-1} L^{-T} L^T Q^T = Q(Q^T Q)^{-1} Q^T, \end{aligned}$$

$$(Q^T Q)^{-1} = \begin{pmatrix} \tilde{g}^2 & \tilde{g}^T e \\ \tilde{g}^T e & e^2 \end{pmatrix}^{-1} = \begin{pmatrix} \tilde{g}^2 & 0 \\ 0 & e^2 \end{pmatrix}^{-1} = \begin{pmatrix} 1/\tilde{g}^2 & 0 \\ 0 & 1/e^2 \end{pmatrix}.$$

Поскольку $\{g, e\}$ линейно независимы, то $\tilde{g} \neq 0$ и $(Q^T Q)^{-1}$ определена. Таким образом, по формуле (2) получим

$$\begin{aligned} \rho^2(g, x) &= x^2 - x^T Q(Q^T Q)^{-1} Q^T x = \\ &= x^2 - \begin{pmatrix} \tilde{g}^T x & e^T x \end{pmatrix} \begin{pmatrix} 1/\tilde{g}^2 & 0 \\ 0 & 1/e^2 \end{pmatrix} \begin{pmatrix} \tilde{g}^T x \\ e^T x \end{pmatrix} = \\ &= x^2 - \frac{(\tilde{g}^T x)^2}{\tilde{g}^2} - \frac{(e^T x)^2}{e^2}, \\ \rho_2(x) &= \rho^2(l, x) - \rho^2(f, x) = \left(x^2 - \frac{(e^T x)^2}{e^2} - \frac{(\tilde{l}^T x)^2}{\tilde{l}^2} \right) - \\ &- \left(x^2 - \frac{(e^T x)^2}{e^2} - \frac{(\tilde{f}^T x)^2}{\tilde{f}^2} \right) = \frac{(\tilde{f}^T x)^2}{\tilde{f}^2} - \frac{(\tilde{l}^T x)^2}{\tilde{l}^2}, \end{aligned} \quad (3)$$

где $\tilde{f}$ и $\tilde{l}$ определены по аналогии с $\tilde{g}$. Нетрудно убедиться, что предположение о линейной независимости $\{f, e\}$ и $\{l, e\}$ выполняется.

3. Вычислительная схема. Воспользуемся формулой (3) для составления вычислительной схемы для параболического АОК. Рассмотрим $g \in \mathbb{R}^k$,

$$\tilde{g}^2 = \left(g - \frac{e^T g}{e^2} e \right)^2 = g^2 + \frac{(e^T g)^2}{e^2} - 2 \frac{(e^T g)^2}{e^2} = g^2 - \frac{(e^T g)^2}{e^2}.$$

Введем

$$g^\Sigma = \frac{e^T g}{e^2} = \frac{1}{k} \sum_{i=1}^k g_i, \quad g^V = \tilde{g}^2 = \sum_{i=1}^k g_i^2 - k(g^\Sigma)^2.$$

Заметим, что тогда $\tilde{g}_i = g_i - g^\Sigma$. По формулам из [11] определим многочлены

$$F_1(n) = \frac{1}{2}n^2 + \frac{1}{2}n = \sum_{i=1}^n i, \quad F_2(n) = \frac{1}{3}n^3 + \frac{1}{2}n^2 + \frac{1}{6}n = \sum_{i=1}^n i^2,$$

$$F_4(n) = \frac{1}{5}n^5 + \frac{1}{2}n^4 + \frac{1}{3}n^3 - \frac{1}{30}n = \sum_{i=1}^n i^4.$$

Определим по аналогии с g^Σ и g^V значения

$$l^\Sigma = \frac{1}{k} \sum_{i=1}^k i = \frac{1}{k} F_1(k), \tag{4}$$

$$l^V = \sum_{i=1}^k i^2 - k(l^\Sigma)^2 = F_2(k) - \frac{F_1^2(k)}{k}, \tag{5}$$

$$f^\Sigma = \frac{1}{k} \sum_{i=1}^k (i-1)^2 = \frac{1}{k} F_2(k-1), \tag{6}$$

$$f^V = \sum_{i=1}^k (i-1)^4 - k(f^\Sigma)^2 = F_4(k-1) - \frac{F_2^2(k-1)}{k}. \tag{7}$$

По формуле (3) получим

$$\rho_2(x) = \frac{1}{f^V} \left(\sum_{i=1}^k ((i-1)^2 - f^\Sigma) x_i \right)^2 - \frac{1}{l^V} \left(\sum_{i=1}^k (i - l^\Sigma) x_i \right)^2.$$

Тогда, согласно (1),

$$\begin{aligned} \delta^2(k_0) = & \frac{1}{f^V} \left(\sum_{i=1}^k ((i-1)^2 - f^\Sigma) y_{k_0-k+i} \right)^2 - \\ & - \frac{1}{l^V} \left(\sum_{i=1}^k (i - l^\Sigma) y_{k_0-k+i} \right)^2. \end{aligned} \quad (8)$$

Вычислительная схема состоит из двух этапов: вычисление вспомогательных значений l^Σ , l^V , f^Σ , f^V по формулам (4)–(7) и вычисление $\delta^2(k_0)$ по формуле (8).

Например, для $k = 5$

$$l^\Sigma = 3, \quad l^V = 10, \quad f^\Sigma = 6, \quad f^V = 174,$$

$$\begin{aligned} \delta^2(k_0) = & \frac{1}{174} \left(\sum_{i=1}^5 ((i-1)^2 - 6) y_{k_0-k+i} \right)^2 - \\ & - \frac{1}{10} \left(\sum_{i=1}^5 (i - 3) y_{k_0-k+i} \right)^2 = \\ & = \frac{1}{174} (-6y_{k_0-4} - 5y_{k_0-3} - 2y_{k_0-2} + 3y_{k_0-1} + 10y_{k_0})^2 - \\ & - \frac{1}{10} (-2y_{k_0-4} - y_{k_0-3} + y_{k_0-1} + 2y_{k_0})^2. \end{aligned} \quad (9)$$

Зафиксировав $y_{k_0-4} = 0$, раскрыв скобки, получим

$$\begin{aligned} \delta^2(k_0) = & \frac{1}{435} (19y_{k_0-3}^2 + 10y_{k_0-2}^2 - 21y_{k_0-1}^2 + 76y_{k_0}^2 - 24y_{k_0-1}y_{k_0} + \\ & + y_{k_0-3}(50y_{k_0-2} + 12y_{k_0-1} - 76y_{k_0}) + y_{k_0-2}(-30y_{k_0-1} - 100y_{k_0})), \end{aligned} \quad (10)$$

что после переобозначения y_{k_0-5+i} как y_i совпадает с формулой, полученной в работе [2]. На практике рекомендуется не раскрывать

скобки, а пользоваться более простой формулой (9) или, в общем случае, формулой (8).

4. Заключение. В работе [2] предложен алгоритм вычисления $\delta^2(k_0)$, в котором используется метод скользящего окна и преобразование $\tilde{y}_i = y_{k_0-k+1+i} - y_{k_0-k+1}$. Заметим, что δ^2 совпадает для окон $\tilde{y}_0, \dots, \tilde{y}_{k-1}$ и $y_{k_0-k+1}, \dots, y_{k_0}$. Значения $\tilde{y}_1, \dots, \tilde{y}_{k-1}$ (известно, что $\tilde{y}_0 = 0$) подставляют в заранее выведенные формулы, аналогичные (10). Эти формулы получают отдельно для каждого k .

Сравним этот алгоритм с построенной вычислительной схемой. Предложенный ранее алгоритм требует хранения вычисленных коэффициентов, количество которых растет с ростом k , а новая вычислительная схема при любом k использует только четыре значения: $l^\Sigma, l^V, f^\Sigma, f^V$. Сложность вычисления коэффициентов разложения растет с ростом k . Получение четырех чисел схемы сводится к нахождению значений многочленов по степеням k , причем число операций не зависит от k . Алгоритм в [2] подразумевает реализацию отдельной формулы для каждого используемого k , а новая вычислительная схема позволяет реализовать процесс вычисления величины $\delta^2(k_0)$ для произвольного k и открывает возможность применения АОК для больших k .

Литература

1. Булинский А. В., Ширяев А. Н. Теория случайных процессов. М.: Физматлит, 2004. 401 с.
2. Orekhov A. V. Quasi-deterministic processes with monotonic trajectories and unsupervised machine learning // Mathematics. 2021. Vol. 9. No 18. Art. no 2301.
3. Thorndike R. L. Who belongs in the family? // Psychometrika. 1953. Vol. 18. P. 267–276.
4. Orekhov A. V. Statistical criteria for the limits of application of Hooke's law // Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes. 2020. Vol. 16. No 4. P. 391–401.

5. Орехов А. В. Машинное обучение и определение момента перехода к следующей стадии ползучести // X Поляховские чтения: Материалы международной научной конференции по механике. СПб.: ООО «Издательство ВВМ», 2024. С. 713–717.
6. Орехов А. В., Васильев Е. В. Бинарные метрические деревья и иерархия вложенных кластеров // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2024. Т. 20. № 4. С. 487–499.
7. Orekhov A. V., Potekhina M. A. Unsupervised machine learning methods for determining special points of the polymerase chain reaction fluorescence accumulation curve // Moscow University Physics Bulletin. 2023. Vol. 78. No 1. P. S169–S179.
8. Орехов А. В., Орехов А. А. Автоматическое обнаружение аномалий сетевого трафика при DDoS-атаках // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2023. Т. 19. № 2. С. 251–263.
9. Буре В. М., Парилина Е. М. Теория вероятностей и математическая статистика. СПб.: Лань, 2013. 416 с.
10. Nocedal J., Wright S. J. Numerical Optimization. New York: Springer, 2006. 664 p.
11. Двайт Г. Б. Таблицы интегралов и другие математические формулы. Изд. 5-е. М.: Наука, 1978. 224 с.