ЛОКАЛЬНО-МЕДИАННЫЙ МЕТОД ПРОГНОЗИРОВАНИЯ И ЕГО ОБОБЩЕНИЕ

А. В. Авдей

1. Введение

Прогнозирование с использованием регрессионной модели данных актуально в экономике, биологии, инженерной и других сферах деятельности. Если имеет место линейная модель регрессии, то для прогнозирования используется стандартный метод наименьших квадратов. Этот метод достаточно хорошо работает с гипотетическими моделями данных, однако в реальности среди экспериментальных данных встречаются выбросы, что значительно снижает эффективность метода наименьших квадратов [3]. Поэтому весьма важно разработать робастные алгоритмы, устойчивые к искажениям данных [3]. Локально-медианный метод прогнозирования, исследованный в [1], имеет несколько преимуществ перед другими робастными методами прогнозирования. В этой работе рассматривается обобщение локально-медианного метода прогнозирования, и исследуется, какой метод из заданного семейства обеспечивает наибольшую устойчивость прогнозов.

2. Локально-медианный метод прогнозирования регрессионной модели при наличии выбросов и его модификации

Пусть наблюдения $\{x_t\}$ исследуемой стохастической динамической системы описываются уравнением регрессии:

$$x_t = \theta^{o'}\varphi(z_t) + u_t + \xi_t v_t, \qquad (1)$$

где $t \in \mathbb{Z}$ – дискретный момент времени, $z_t \in \mathbb{R}^M$ – неслучайный вектор факторов в момент времени t, $\varphi(z) = \big(\varphi_i(z)\big) : \mathbb{R}^M \to \mathbb{R}^m$ – вектор из m линейно независимых функций, $\theta^o = (\theta_i^0) \in \mathbb{R}^m$ – m-мерный вектор неизвестных параметров модели, $u_t \in \mathbb{R}$ – случайная ошибка в момент времени t, $v_t \in \mathbb{R}$ – выброс в момент времени t, $\{\xi_t\}$ – независимые одинаково распределенные Бернулевские случайные величины, которые определяют наличие выбросов, $P\{\xi_t = 1\} = \varepsilon$, $P\{\xi_t = 0\} = 1 - \varepsilon$, $\varepsilon \in [0; 0.5)$ – вероятность появления выброса. Случайные ошибки $\{u_t\}$ – независимые одинаково распределенные случайные величины, $E\{u_t\} = 0$,

$D\{u_t\}=\sigma^2<+\infty$; $\{v_t\}$ – независимые одинаково распределенные случайные величины, $E\{v_t\}=a_t$, $D\{v_t\}=K\sigma^2<+\infty$, $K\geq 0$. Случайные величины $\{u_t\}$, $\{v_t\}$, $\{\xi_t\}$ независимы в совокупности.

Введем обозначения: T – время наблюдения; $\{t_1^l,...,t_n^l\}\subset\{1,2,...,T\}$ – подмножество из n $(m\leq n\leq T)$ моментов времени наблюдения $(l=1,...,L)$, где L – количество различных подмножеств моментов времени $\left(m\leq L\leq L_+=C_T^n\right)$; $\Psi=(\varphi_j(z_t))$, $j=1,...,m$, $t=1,...,T$; $\Psi_n^{(l)}=(\varphi_j(z_{t_i^{(l)}}))$, $i=1,...,n$, $j=1,...,m$, – $(n\times m)$-подматрица $(T\times m)$-матрицы Ψ, $\left|\Psi_n^{(l)\prime}\Psi_n^{(l)}\right|\neq 0$; $X=(x_1,x_2,...,x_T)'\in\mathbb{R}^T$, $X_n^{(l)}=(x_{t_1^{(l)}},x_{t_2^{(l)}},...,x_{t_n^{(l)}})'\in\mathbb{R}^n$ – подвыборка выборки наблюдений X.

Определим l-ую локальную МНК-оценку для θ^o, построенную по l-ой подвыборке $X_n^{(l)}$:

$$\hat{\theta}^{(l)}=\left(\Psi_n^{(l)\prime}\Psi_n^{(l)}\right)^{-1}\Psi_n^{(l)\prime}X_n^{(l)},\ l=1,...,L, \tag{2}$$

и семейство из L локальных прогнозов будущего состояния $x_{T+\tau}$ для $\tau\geq 1$, построенных по локальным оценкам $\hat{\theta}^{(l)}$ из (2):

$$\hat{x}_{T+\tau}^{(l)}=\hat{\theta}^{(l)\prime}\varphi(z_{T+\tau}),\ l=1,...,L. \tag{3}$$

Локально-медианным прогнозом называется [1] медиана выборки из L локальных прогнозов (3):

$$\hat{x}_{T+\tau}=S(X)=med\left\{\hat{x}_{T+\tau}^{(1)},...,\hat{x}_{T+\tau}^{(L)}\right\}. \tag{4}$$

Рассмотрим модификации локально-медианного метода прогнозирования. В качестве прогноза будущего состояния $x_{T+\tau}$ возьмем δ – усеченное среднее упорядоченных локальных прогнозов:

$$\hat{x}_{T+\tau}=S(X,\delta)=\frac{1}{L-2s}\sum_{j=s+1}^{L-s}\hat{x}_{T+\tau,(j)}^{(l)}, \tag{5}$$

где $s=[\delta L]$, $\hat{x}_{T+\tau,(j)}^{(l)}$ – j-ая порядковая статистика вариационного ряда L локальных прогнозов (3), $0\leq\delta<\frac{1}{2}$ – параметр усечения. В частности при $\delta=0$ получаем среднее арифметическое локальных прогнозов, а при нечетном L и $\delta=\frac{1}{2}$ медиану (4) выборки локальных прогнозов.

Рассмотрим еще один подход к прогнозированию будущего состояния $x_{T+\tau}$, аналогичный (5) с той лишь разницей, что вместо МНК-оценки $\hat{\theta}^{(l)}$, строится робастная медианная оценка $\tilde{\theta}^{(l)}$, которая мини-мизирует сумму $\sum_{i=1}^{n}\left|\theta'\varphi(z_{t_i^{(l)}}) - x_{t_i^{(l)}}\right|$. Семейство из L локальных прогнозов по θ определяется как и прежде: $\tilde{x}_{T+\tau}^{(l)} = \tilde{\theta}^{(l)'}\varphi(z_{T+\tau})$, $l = 1,...,L$.

Выбираем уровень усечения $0 \le \delta < \frac{1}{2}$ и строим прогноз будущего состояния $x_{T+\tau}$:

$$\tilde{x}_{T+\tau} = \frac{1}{L-2s}\sum_{j=s+1}^{L-s}\tilde{x}_{T+\tau,(j)}^{(l)}, \qquad (6)$$

где $s = [\delta L]$, $\tilde{x}_{T+\tau,(j)}^{(l)}$ – j-ая порядковая статистика вариационного ряда L локальных прогнозов $\tilde{x}_{T+\tau}^{(l)}$.

Аналитическое исследование этого метода прогнозирования затруднено тем, что оценка $\tilde{\theta}^{(l)}$ является решением задачи линейного программирования [2]. Отметим, что при $n = m$ оценки $\tilde{\theta}^{(l)}$ и $\hat{\theta}^{(l)}$ будут равны, и, следовательно, прогнозы (5) и (6) будут совпадать.

3. Пороговые точки и распределение вероятностей прогнозов

Оценим пороговые точки для прогноза (5) в смысле Хампеля [3], используя методику из [1].

Теорема 1. Если $L = L_+ = C_T^n$, тогда пороговая точка для прогноза (4) регрессионнной модели с выбросами (1) есть единственный корень алгебраического уравнения n-ого порядка из отрезка $[0,1-nT^{-1}]$ $\prod_{t=0}^{n-1}(1-\varepsilon-\frac{t}{T}) = (1-\alpha)\prod_{t=0}^{n-1}(1-\frac{t}{T})$, где $\alpha = \frac{[\delta L]}{L}$.

Следствие. Если n фиксировано и $T \to \infty$, то $\varepsilon^* \to 1-2^{-\frac{1}{n}}$, и оптимальный размер подвыборки $n^* = m$.

Для определения распределения прогноза (4) необходимо знать распределение локальных прогнозов (3) и совместное распределение порядковых статистик. Точное распределение локальных прогнозов было выведено в [1], а в [6] приведены функции совместного распределения двух и более порядковых статистик, но их выражения слишком сложны для анализа. Поэтому рассмотрим случай, когда локальные прогнозы одина-

ково распределены. Пусть локальные прогнозы независимые и одинаково распределенные случайные величины с плотностью $p_\zeta(z)$.

В этом случае порядковые статистики выборки локальных прогнозов можно рассматривать как квантили соответствующего уровня распределения с плотностью $p_\zeta(z)$. Используя асимптотическое совместное распределение квантилей [5] и теорему о линейном преобразовании гауссовской случайной величины [4], можно вывести распределение вероятностей прогноза (5).

Теорема 2. Пусть локальные прогнозы $\hat{x}_{T+\tau}^{(l)}$, $l=1,...,L$ – независимые и одинаково распределенные случайные величины с плотностью $p_\zeta(z)$, тогда при $L\to\infty$ распределение случайной величины

$$\omega=(L-2s)\sqrt{L}\left(\sum_{j=1}^{L-2s}\left(\frac{\lambda_j(1-\lambda_j)}{p_\zeta^2(\zeta_{\lambda_j})}+2\sum\nolimits_{i=j+1}^{L-2s}\frac{\lambda_j(1-\lambda_i)}{p_\zeta(\zeta_{\lambda_j})p_\zeta(\zeta_{\lambda_i})}\right)\right)^{-\frac{1}{2}}\times$$

$$\times\left(\hat{x}_{T+\tau}-\frac{1}{L-2s}\sum_{j=1}^{L-2s}\zeta_{\lambda_j}\right), \qquad (7)$$

где ζ_{λ_i} – квантиль уровня λ_i распределения случайной величины $\hat{x}_{T+\tau}^{(l)}$ (5), $\lambda_i=\delta+(i-1)/L$, $i=1,...,L-2s$, сходится к стандартному гауссовскому распределению с плотностью $p_\omega(z)=(2\pi)^{-\frac{1}{2}}\exp\left(-\frac{z^2}{2}\right)$.

По итогам численных экспериментов был сделан вывод о том, что наибольшая устойчивость и минимальный риск прогнозирования достигаются при $\delta=\frac{1}{2}$ и $n=m$ как в случае прогноза (5), так и в случае прогноза (6).

Литература

1. *Kharin Yu S., Maevsk V.* Local-median method of forecasting for regression time series under outliers / Automation and Remote Control, 2002, Vol. 11.
2. *Portnoy S. ,Koenker R.* The Gaussian Hare and the Laplasian Tortoise: Computability of Squared-Error versus Absolute-Error Estimators / Statistical Science, 1997, Vol. 12.
3. *Хампель Ф., Рочетти Э., Рауссеу П., Штаэль В*. Робастность в статистике / М.: Мир, 1989.
4. *Ивченко Г. И., Медведев Ю. И.* Математическая статистика / М.: Высшая школа, 1984.
5. *Дэвид Г.* Введение в теорию порядковых статистик / М.: Статистика, 1970.
6. *Кендалл М., Стюарт А.* Теория распределений / М.: Наука, 1966.