УДК 519.254:519.652
В 1747+Б 192.14

П.О. Приставка, канд. техн. наук

НЕПАРАМЕТРИЧНА ОЦІНКА ЗАЛЕЖНОСТЕЙ МЕТОДОМ НАЙМЕНШИХ КВАДРАТІВ

Національний авіаційний університет, obaybuz@inbox.ru

Розглянуто можливість відтворення одновимірних регресійних залежностей з використанням оцінки двовимірної функції щільності. Поширено зазначений підхід для оцінки двовимірної регресії з використанням наближення тривимірних функцій щільності.

Постановка проблеми

Відтворення регресії у випадку лінійної форми зв'язку між ознаками об'єкта спостережень є класичною задачею та не викликає труднощів.

У разі відтворення нелінійних залежностей виникає проблема ідентифікації типу регресійної моделі.

При оцінці двовимірних та багатовимірних регресій обчислювальна складність процедур ідентифікації та параметричного відтворення багатократно зростає.

Одним із напрямів при вирішенні задачі оцінки функції регресії є непараметричне відтворення.

Аналіз досліджень

Проблемі непараметричного відтворення залежностей присвячено багато публікацій, серед яких окремо слід відзначити роботи В.Н. Вапніка та В.Я. Катковніка [1; 2].

У роботі [2] розглянуто підхід до відтворення одновимірних регресій на підставі оцінок функцій щільності розподілу ймовірностей за використанням ядерних оцінок.

Альтернативою зазначеному підходу є використання при відтворенні регресій оцінок функцій щільності на основі поліноміальних сплайнів на підставі *B*-сплайнів, близьких до інтерполяційних у середньому [3–6].

Постановка задачі

Нехай маємо об'єкт спостереження, який характеризується ознаками T, Q реалізації об'єкта, що спостерігаються під час експерименту, – двійки дійсних чисел (t_r, q_r), $r = \overline{1,N}$,

$t \in [t_{\min}, t_{\max}]$,

$q \in [q_{\min}, q_{\max}]$.

Припустимо існування залежності (регресійної або функціональної) ознаки Q від T у вигляді

$q(t) \in C^k$.

Задамося рівномірним розбиттям осей спостереження ознак T і Q:

$t_i = t_{\min} + \delta t(i - 0{,}5)$, $i = \overline{1, n_t}$;

$q_j = q_{\min} + \delta q(j - 0{,}5)$, $j = \overline{1, n_q}$;

$$\delta t = \frac{t_{\max} - t_{\min}}{n_t};$$

$$\delta q = \frac{q_{\max} - q_{\min}}{n_q},$$

де n_t, n_q – довільні цілі додатні числа.

Будемо вважати, що задано рівномірні розбиття Δ_t, Δ_q осей спостереження T і Q, які, у свою чергу, задають рівномірне розбиття Δ_{tq} прямокутної площини спостереження одночасної реалізації ознак об'єкта дослідження.

Мета – знайти оцінку наближення $\hat{q}(t)$ залежності $q(t)$. Для цього, як завжди, вимагатимемо виконання умови досягнення мінімуму залишкової дисперсії, а саме:

$$\min_{[t_{\min}, t_{\max}]} S^2_{зал} = \min_{[t_{\min}, t_{\max}]} \left(\frac{1}{n_t - 1} \sum_{i=1}^{n_t} \left(q(t_i) - \hat{q}(t_i) \right)^2 \right).$$

Виклад основного матеріалу

Неважко показати, що, коли $n_t \to \infty$ (а отже і $\delta t \to 0$), для

$q(t) \in C^k$

виконується

$$\lim_{n_t \to \infty} S^2_{зал} = \lim_{n_t \to \infty} \left(\frac{1}{n_t - 1} \sum_{i=1}^{n_t} \left(q(t_i) - \bar{q}_i \right)^2 \right) = 0,$$

де

$$\bar{q}_i = \frac{1}{\delta t} \int_{t_i - \delta t/2}^{t_i + \delta t/2} q(t)\, dt,$$

або, коли $n_t, n_q \to \infty$ ($\delta t, \delta q \to 0$), для

$q(t) \in C^k$

маємо

$$\lim_{n_t, n_q \to \infty} S^2_{зал} = \lim_{n_t \to \infty} \left(\frac{1}{n_t - 1} \sum_{i=1}^{n_t} \left(q(t_i) - \frac{1}{n_q} \sum_{j=1}^{n_q} q_{ij} \right)^2 \right) = 0, \quad (1)$$

де q_{ij} (t_{ij}) – значення реалізації ознаки Q (T) на (i, j)-му елементі розбиття Δ_{tq}.

Подамо інше зображення $S^2_{зал}$ із правої частини виразу (1).

Для цього розглянемо величину

$$\bar{q}_i = \frac{1}{n_q}\sum_{j=1}^{n_q} q_{ij} = \frac{\sum_{j=1}^{n_q} q_{ij} n_{ij}}{\sum_{j=1}^{n_q} n_{ij}} = \frac{\sum_{j=1}^{n_q} q_{ij} f_{ij}}{\sum_{j=1}^{n_q} f_{ij}},$$

де n_{ij}. – кількість реалізацій об'єкта спостережень на (i,j)-му елементі розбиття Δ_{tq}; f_{ij} – випадковості реалізації двійок (t_i, q_j), $i=\overline{1,n_t}$, $j=\overline{1,n_q}$;

$$f_{ij} = \frac{n_{ij}}{n_t n_q}$$

або

$$f_{ij} = \delta t\, \delta q\, \bar{p}_{ij};$$

$$\bar{p}_{ij} = \frac{1}{\delta t\, \delta q}\int_{t_i-\delta t/2}^{t_i+\delta t/2}\int_{q_j-\delta q/2}^{q_j+\delta q/2} p(t,q)\,dtdq,$$

де $p(t,q)$ – функція щільності розподілу ймовірностей реалізації ознак T і Q.

Для неперервної $p(t,q)$ маємо

$$\lim_{n_q\to\infty}\bar{q}_i = \lim_{n_q\to\infty}\frac{\sum_{j=1}^{n_q} q_{ij}\int_{t_i-\delta t/2}^{t_i+\delta t/2}\int_{q_j-\delta q/2}^{q_j+\delta q/2} p(t,q)\,dtdq}{\sum_{j=1}^{n_q}\int_{t_i-\delta t/2}^{t_i+\delta t/2}\int_{q_j-\delta q/2}^{q_j+\delta q/2} p(t,q)\,dtdq} =$$

$$= \frac{\int_{q_{\min}}^{q_{\max}} q\; p(t_i,q)\,dq}{\int_{q_{\min}}^{q_{\max}} p(t_i,q)\,dq}. \qquad (2)$$

Отже, при побудові обчислювальних схем як дискретизованої оцінки $\hat{q}(t)$ залежності $q(t)$ можемо взяти величину, близьку за значенням до правої частини (2):

$$\hat{q}(t_i) = \frac{\int_{q_{\min}}^{q_{\max}} q\; p(t_i,q)\,dq}{\int_{q_{\min}}^{q_{\max}} p(t_i,q)\,dq} \approx \frac{\sum_{j=1}^{n_q} q_{ij}\hat{p}_{ij}}{\sum_{j=1}^{n_q}\hat{p}_{ij}}, \qquad (3)$$

де $\hat{p}_{ij}$, $i=\overline{1,n_t}$, $j=\overline{1,n_q}$ – оцінка функції $p(t,q)$ у точці (t_i,q_j), за яку беруть значення локального поліноміального сплайна від двох змінних на основі B-сплайнів другого порядку, близького до інтерполяційного у середньому [3; 4].

Обчислювальна схема застосування подібного сплайна не викликає труднощів, більш того, дозволяє отримати неперервне наближення функції щільності за розбиттям Δ_{tq}, що, у свою чергу, дає можливість отримати неперервну оцінку залежності $\hat{q}(t)$.

Згідно з роботою [3], наведемо розгорнутий вигляд зазначеного сплайна:

$$S_{2,0}(p,t,q) = \frac{1}{64}\Big(\big(p_{i-1,j-1}+6p_{i-1,j}+p_{i-1,j+1}+6p_{i,j-1}+$$
$$+36p_{i,j}+6p_{i,j+1}+p_{i+1,j-1}+6p_{i+1,j}+p_{i+1,j+1}\big)+$$
$$+\big(-2p_{i-1,j-1}-12p_{i-1,j}-2p_{i-1,j+1}+2p_{i+1,j-1}+12p_{i+1,j}+$$
$$+2p_{i+1,j+1}\big)x+\big(-2p_{i-1,j-1}+2p_{i-1,j+1}-12p_{i,j-1}+$$
$$+12p_{i,j+1}-2p_{i+1,j-1}+2p_{i+1,j+1}\big)y+\big(4p_{i-1,j-1}-$$
$$-4p_{i-1,j+1}-4p_{i+1,j-1}+4p_{i+1,j+1}\big)xy+\big(p_{i-1,j-1}+$$
$$+6p_{i-1,j}+p_{i-1,j+1}-2p_{i,j-1}-12p_{i,j}-2p_{i,j+1}+$$
$$+p_{i+1,j-1}+6p_{i+1,j}+p_{i+1,j+1}\big)x^2+\big(p_{i-1,j-1}-2p_{i-1,j}+$$
$$+p_{i-1,j+1}+6p_{i,j-1}-12p_{i,j}+6p_{i,j+1}+p_{i+1,j-1}-2p_{i+1,j}+$$
$$+p_{i+1,j+1}\big)y^2+\big(-2p_{i-1,j-1}+2p_{i-1,j+1}+4p_{i,j-1}-4p_{i,j+1}-$$
$$-2p_{i+1,j-1}+2p_{i+1,j+1}\big)x^2y+\big(-2p_{i-1,j-1}+4p_{i-1,j}-$$
$$-2p_{i-1,j+1}+2p_{i+1,j-1}-4p_{i+1,j}+2p_{i+1,j+1}\big)xy^2+$$
$$+\big(p_{i-1,j-1}-2p_{i-1,j}+p_{i-1,j+1}-2p_{i,j-1}+4p_{i,j}-$$
$$-2p_{i,j+1}+p_{i+1,j-1}-2p_{i+1,j}+p_{i+1,j+1}\big)x^2y^2\Big),$$

де

$$x = \frac{2}{\delta t}(t-i\delta t),\ |x|\le 1;$$

$$y = \frac{2}{\delta q}(q-j\delta q),\ |y|\le 1.$$

Зазначений підхід неважко узагальнити, коли маємо об'єкт спостереження, який характеризується трьома ознаками T, Q, G реалізації об'єкта, – трійки дійсних чисел

(t_r, q_r, g_r), $r=\overline{1,N}$,

$t\in[t_{\min}, t_{\max}]$,

$q\in[q_{\min}, q_{\max}]$,

$g\in[g_{\min}, g_{\max}]$.

Нехай потрібно знайти оцінку $\hat{g}(t,q)$ існуючої залежності (регресійної або функціональної)

$g(t,q)\in C^{k,l}$.

Якщо задано рівномірні розбиття Δ_t, Δ_q, Δ_g осей спостереження ознак T, Q, G:

$t_i = t_{\min} + \delta t\,(i-0{,}5)$, $i = \overline{1, n_t}$;

$q_j = q_{\min} + \delta q\,(j-0{,}5)$, $j = \overline{1, n_q}$;

$g_\nu = g_{\min} + \delta g\,(\nu-0{,}5)$, $\nu = \overline{1, n_g}$,

де

$$\delta t = \frac{t_{\max} - t_{\min}}{n_t};$$

$$\delta q = \frac{q_{\max} - q_{\min}}{n_q};$$

$$\delta g = \frac{g_{\max} - g_{\min}}{n_g},$$

то наведені розбиття визначають рівномірне розбиття Δ_{tqg} паралелепіпедного простору спостереження одночасної реалізації ознак об'єкта спостереження.

Поставимо за вимогу виконання умови

$$\min_{\left[(t_{\min}, q_{\min}),(t_{\max}, q_{\max})\right]} S^2_{зал} =$$

$$= \min_{\substack{(t_{\min}, q_{\min}),\\ (t_{\max}, q_{\max})}} \left(\frac{1}{n_t \cdot n_q - 1} \sum_{i=1}^{n_t} \sum_{j=1}^{n_q} \left(g(t_i, q_j) - \hat{g}(t_i, q_j) \right)^2 \right).$$

Тоді при $n_t, n_q \to \infty$ $(\delta t, \delta q \to 0)$ для $q(t,q) \in C^{k,l}$

$$\lim_{n_t, n_q \to \infty} S^2_{зал} =$$

$$= \lim_{n_t, n_q \to \infty} \left(\frac{1}{n_t \;\; n_q - 1} \sum_{i=1}^{n_t} \sum_{j=1}^{n_q} \left(g(t_i, q_j) - \bar{g}_{ij} \right)^2 \right) = 0,$$

де

$$\bar{g}_{ij} = \frac{1}{\delta t \;\; \delta q} \int_{t_i - \delta t/2}^{t_i + \delta t/2} \int_{q_j - \delta q/2}^{q_j + \delta q/2} g(t,q)\, dt dq,$$

або при $n_t, n_q, n_g \to \infty$ $(\delta t, \delta q, \delta g \to 0)$ для $g(t,q) \in C^{k,l}$

маємо:

$$\lim_{n_t, n_q, n_g \to \infty} S^2_{зал} =$$

$$= \lim_{n_t, n_q \to \infty} \left(\frac{1}{n_t \;\; n_q - 1} \sum_{i=1}^{n_t} \sum_{j=1}^{n_q} \left(g(t_i, q_j) - \frac{1}{n_g} \sum_{\nu=1}^{n_g} g_{ij\nu} \right)^2 \right) = 0,$$

де $g_{ij\nu}$ $(t_{ij\nu},\ q_{ij\nu})$ – значення реалізації ознаки G $(T,\ Q)$ на (i,j,ν)-му елементі розбиття Δ_{iqg}.

Величину $\bar{g}_{ij}$ можемо подати у вигляді

$$\bar{g}_{ij} = \frac{1}{n_g} \sum_{\nu=1}^{n_g} g_{ij\nu} = \frac{\sum_{\nu=1}^{n_g} g_{ij\nu} n_{ij\nu}}{\sum_{\nu=1}^{n_g} n_{ij\nu}} = \frac{\sum_{\nu=1}^{n_g} g_{ij\nu} f_{ij\nu}}{\sum_{\nu=1}^{n_g} f_{ij\nu}},$$

де $n_{ij\nu}$ – кількість спостережень на (i,j,ν)-му елементі розбиття Δ_{iqg};

$$f_{ij\nu} = \frac{n_{ij\nu}}{n_t \;\; n_q \;\; n_g} = \delta t \;\; \delta q \;\; \delta g \;\; \bar{p}_{ij\nu};$$

$$\bar{p}_{ij\nu} = \frac{1}{\delta t \;\; \delta q \;\; \delta g} \int_{t_i - \delta t/2}^{t_i + \delta t/2} \int_{q_j - \delta q/2}^{q_j + \delta q/2} \int_{g_\nu - \delta g/2}^{g_\nu + \delta g/2} p(t,q,g)\, dt dq dg;$$

$p(t,q,g)$ – функція щільності розподілу ймовірностей ознак T, Q, G.

Тоді для неперервної $p(t,q,g)$ маємо

$$\lim_{n_g \to \infty} \bar{g}_{ij} =$$

$$= \lim_{n_g \to \infty} \frac{\sum_{\nu=1}^{n_g} g_{ij\nu} \int_{t_i - \delta t/2}^{t_i + \delta t/2} \int_{q_j - \delta q/2}^{q_j + \delta q/2} \int_{g_\nu - \delta g/2}^{g_\nu + \delta g/2} p(t,q,g)\, dt dq dg}{\sum_{\nu=1}^{n_g} \int_{t_i - \delta t/2}^{t_i + \delta t/2} \int_{q_j - \delta q/2}^{q_j + \delta q/2} \int_{g_\nu - \delta g/2}^{g_\nu + \delta g/2} p(t,q,g)\, dt dq dg} =$$

$$= \frac{\int_{g_{\min}}^{g_{\max}} g \;\; p(t_i, q_j, g)\, dq}{\int_{g_{\min}}^{g_{\max}} p(t_i, q_j, g)\, dq}.$$

Отже, за дискретизовану оцінку $\hat{g}(t,q)$ залежності $g(t,q)$ можемо взяти величину

$$\hat{q}(t_i, q_j) = \frac{\int_{g_{\min}}^{g_{\max}} g \;\; p(t_i, q_j, g)\, dq}{\int_{g_{\min}}^{g_{\max}} p(t_i, q_j, g)\, dq} \approx \frac{\sum_{\nu=1}^{n_g} g_{ij\nu} \hat{p}_{ij\nu}}{\sum_{\nu=1}^{n_g} \hat{p}_{ij\nu}}, \qquad (4)$$

де $\hat{p}_{ij\nu}$, $i = \overline{1, n_t}$, $j = \overline{1, n_q}$, $\nu = \overline{1, n_g}$ – оцінка функції $p(t,q,g)$ у точці (t_i, q_j, g_ν), за яку можна запропонувати використання значення локального поліноміального сплайна від трьох змінних на основі B-сплайнів [5]:

$$S_{2,0}(p,t,q,g) =$$

$$= \sum_{i=1}^{n_t} \sum_{j=1}^{n_q} \sum_{\nu=1}^{n_g} B_{2,h_t}(t - ih_t)\, B_{2,h_q}(q - jh_q)\, B_{2,h_g}(g - \nu h_g)\, p_{i,j,\nu},$$

$$B_{2,h_w}(w) = \begin{cases} 0, & |w| \ge 3h_w/2, \\ (3 + 2w/h_w)^2/8, & w \in [-3h_w/2; -h_w/2], \\ 3/4 - (2w/h_w)^2/4, & w \in [-h_w/2; h_w/2], \\ (3 - 2w/h_w)^2/8, & w \in [h_w/2; 3h_w/2], \end{cases}$$

$w = t, q, g$.

Викладений підхід до відтворення одновимірних та двовимірних регресій має обмеження за використанням. Це пов'язано з тим, що не завжди в розпорядженні дослідника є дані, за якими можна отримати адекватну оцінку функції щільності розподілу ймовірностей (обмежений обсяг, приналежність реалізацій різним генеральним сукупностям, суттєва неоднорідність). Застосування сплайн-оцінок на основі *B*-сплайнів з урахуванням заходів щодо підвищення адекватності оцінки [3] дозволяє поліпшити процес обробки "неякісних" даних, проте, зазначені обмеження мають бути прийняті до уваги.

Одержаний результат забезпечує оцінку функції регресії як значення, яке є найбільш вірогідним при заданому аргументі (аргументах), що є важливим в умовах інформативної невизначеності. Щодо обчислювальної складності процесу відтворення, то згідно з формулами (3), (4) вона не є високою, а саму обчислювальну схему можна легко модифікувати залежно від математичного апарату оцінки функції щільності та вимог до обробки даних. Так, якщо у формулах (3), (4) використовувати лише ті значення аргументів, для яких випадковості більші деякого межового рівня, буде забезпечено автоматичне ігнорування малоймовірних (аномальних) реалізацій.

Наведені дослідження зв'язку регресій із щільностями реалізацій випадкових величин дозволяють аналізувати застосування функцій щільності для візуалізації просторових об'єктів, а саме: визначення найбільш ймовірних координат простору реалізацій ознак, що є визначенням координат локалізації об'єктів. У такому разі відтворення регресії є частковим випадком, а загальним розв'язком є області, які визначено ізолініями функції щільності при заданому рівні випадковості. Останні зауваження виводять на низку прикладних задач: класифікацію та кластеризацію спостережень, візуалізацію рудних тіл в георозвідці, виявлення областей з підвищеним рівнем забруднення при екологічному моніторингу тощо.

Висновки

Виходячи з методу найменших квадратів, показано можливість непараметричної оцінки одновимірних та двовимірних регресій на підставі сплайн-оцінок функцій щільності. Наведені необхідні функціонали можливо реалізувати в прикладному програмному забезпеченні задач відтворення залежностей. Зазначений підхід може бути поширений на випадок відтворення багатовимірних регресійних залежностей.

Список літератури

1. *Вапник В.Н.* Восстановление зависимостей по эмпирическим данным. –М.: Наука, 1979. – 448 с.
2. *Катковник В.Я.* Непараметрическая идентификация и сглаживание данных: метод локальной аппроксимации. – М.: Наука, 1985. – 336 с.
3. *Приставка П.О.* Оцінка функції щільності розподілу двох змінних поліноміальним сплайном на основі *B*-сплайнів // Актуальні проблеми автоматизації та інформаційних технологій. – Д.: Навч. кн. – 2001. – Т. 5. – С. 3–12.
4. *Білецький А.Я., Приставка П.О., Фоменко Г.В.* Оцінка вірогідності непараметричного відтворення функції щільності розподілу ймовірностей двох змінних // Вісн. НАУ. – 2001. – №4(11). – С. 121–126.
5. *Білецький А.Я., Приставка П.О.* Застосування поліноміального сплайна трьох змінних на основі *B*-сплайнів при опрацюванні результатів спостережень // Вісн. НАУ. – 2001. – №3(10). – С. 153–155.
6. *Приставка П.О.* Обчислювальна технологія обробки масивів реалізацій двовимірної випадкової величини // Вісн. НАУ. – 2002. – №3(14). – С. 187–193.

Стаття надійшла до редакції 23.05.03

Ф.А. Приставка

Непараметрическая оценка зависимостей методом наименьших квадратов

Рассмотрена возможность восстановления одномерных регрессионных зависимостей с использованием оценки двухмерной функции плотности. Предложенный подход обобщён для оценки двухмерной регрессии с использованием приближений трехмерных функций плотности.

Ph.O.Pristavka

Nonparametryc restoration of mean square regression

The question of restoration of mean square non linear regression of one and two variables based on the density distribution functions of two- and three-dimensional random variable was considered in the article.