АКТУАЛЬНІ ПРОБЛЕМИ АВТОМАТИЗАЦІЇ ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ. Том 13. 2009

УДК 519.254:519.600:519.674

* П.О. Приставка, ** Ю.М. Архангельська

** Національний авіаційний університет*
*** Дніпропетровський національний університет ім. Олеся Гончара*

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ НЕПАРАМЕТРИЧНОЇ ОЦІНКИ ДВОВИМІРНОЇ РЕГРЕСІЇ ЗА ВИКОРИСТАННЯМ СПЛАЙНІВ БЛИЗЬКИХ ДО ІНТЕРПОЛЯЦІЙНИХ У СЕРЕДНЬОМУ

Пропонується інформаційна технологія непараметричної оцінки двовимірної регресії, за використанням лінійних комбінацій *B*-сплайнів близьких до інтерполяційних в середньому.

Ключові слова: *B-сплайн, адекватність, інформаційна технологія, метод локальної апроксимації, метод найменших квадратів, метод «поточкового» відновлення, непараметрична оцінка, регресія, регресограма, сплайн поліноміальна регресія, статистичне оцінювання*

Предлагается информационная технология непараметрической оценки двумерной регрессии с использованием линейных комбинаций B-сплайнов близких к интерполяционным в среднем.

Ключевые слова: *B-сплайн, адекватность, информационная технология, метод локальной аппроксимации, метод наименьших квадратов, метод «поточечного» восстановления, непараметрическая оценка, регрессия, регрессограмма, сплайн-полиномиальная регрессия, статистическое оценивание*

Nonparametric estimation of two-dimensional regression information technology using linear combinations of close to average B-spline interpolation is proposed.

Keywords: *B-spline, adequacy, information technology, the local approximation method, the least-squares method, the "pointwise" reconstruction method, nonparametric estimation, regression, regresogramma, spline-polynomial regression, statistical estimation*

Постановка проблеми в загальному вигляді та її зв'язок з важливими науковими дослідженнями. При обробці великих тривимірних масивів даних, серед інших задач статистичного оцінювання може поставати потреба вибору адекватної моделі двовимірної регресії. Параметричні моделі, відтворення яких не викликає труднощів за використанням методу найменших квадратів (лінійна, поліноміальна, мультиплікативна) не завжди можуть задовольняти потреби дослідника з точки зору адекватності, зокрема, коли є необхідність оцінити локальні особливості регресії. Виходом з такої ситуації може бути застосування складних моделей непараметричного регресійного оцінювання (наприклад, сплайн-поліноміальна регресія), або ж використання фінітних функцій для непараметричного оцінювання. Стосовно відтворення сплайн-поліноміальних регресій відзначимо суттєву обчислювальну складність подібного підходу та складність ідентифікації ступенів поліномів, що входять до складу моделей. Крім того, інтерпретація оцінок параметрів подібних моделей залишається доволі малоінформативною, що схиляє дослідників у практичних цілях віддавати перевагу, хоча й так само мало інформативним, але більш адекватним непараметричним оцінкам. Стосовно останніх – актуальною залишається потреба підвищення адекватності при відносно малій обчислювальній складності оцінювання для забезпечення обробки представницьких вибірок в автоматизованих системах, що функціонують з режимі реального часу.

Аналіз досліджень та постановка задачі. Нехай [1] $X \in \mathbf{R}^d$ – вхідковий вектор, $Y \in \mathbf{R}^1$ – випадкова величина. Традиційною та найбільш вивченою задачею непараметричного регресійного аналізу є оцінювання функції регресії $p(x) = E\{Y | X = x\}$ на деякій підмножині $\mathbf{R}^d$ за незалежними реалізаціями $\{(X_l, Y_l); l = \overline{1, N}\}$ двійки (X, Y) при апріорній інформації про належність p деякому класу функцій. Згідно [1] визначають такі типи непараметричних оцінок регресії: регресограма, ядерна оцінка (оцінка Надарая-Ватсона), оцінка найближчих сусідів, проекційна оцінка та сплайн-оцінка.

На загал, проблемі непараметричного відтворення залежностей, в тому числі регресійних, присвячено багато публікацій [2 – 9]. І Надарая вперше запропонував ядерну оцінку регресії, яка була введена формально, як статистика, та підлягала дослідженню. Подальший розвиток оцінки ядерного типу отримали G. Watson [3] та М Розенбланта [9], в яких містилось багато чисельних прикладів, розширена задача побудови ядра, оптимального у середньому квадратичного для заданої функції регресії. Узагальнення на багатовимірний випадок стало об'єктом досліджень робіт В Жилявського, А. Мєдвєдєва та А. Рубана. Оцінки ядерного типу при достатньо слабких обмеженнях на p і для широкого класу ядер виявляються асимптотично нормальними, спроможними та

асимптотично незсуненими. Проте, незважаючи на високу точність апроксимації за методом ядерних оцінок, вибір як вигляду ядра та ширини вікна і в загальному випадку для вибору цих параметрів необхідна спеціальна структура даних, яка дозволяє знаходити оптимальний вибір.

У [2] розглянуто підходи щодо непараметричної ідентифікації регресії на підставі методу «поточкового» відновлення та методу локальної апроксимації. Застосування методу «поточкового» відновлення робить його практично непридатним у багатовимірному випадку. Забезпечуючи відновлення лише на множині використаних значень X, алгоритм виявляється неекономічним та потребує явно надлишкової сумарної кількості спостережень значень Y.

У будь-якому разі, застосування згаданих методів локальної апроксимації забезпечує значно адекватніше оцінювання регресії у порівнянні із застосуванням параметричних моделей. Однак, у багатовимірному випадку оцінювання ускладнюється отриманням аналітичного виразу вагової функції.

Альтернативою зазначеним підходам може бути інформаційна технологія викладена в даній статті, яка спирається на наступні положення.

Найпростішою та найбільш очевидною непараметричною оцінкою функції регресії є регресограма. Простір $\mathbf{R}^r$ розбивається на підмножини, що не перетинаються. Регресограма визначається [1] як кусково-стала функція, значення котрої на підмножині розбиття дорівнюють середньому спостережень Y_l, для яких прообрази X_l потрапили в цю підмножину. В [6] подано та досліджено нерекурсивну процедуру регуляризації, згідно якої може бути проведена оцінка у вигляді регресограми. Там же показано, що застосування локальних сплайн-операторів, близьких до інтерполяційних у середньому на основі B-сплайнів дозволяє на порядок та більше покращити асимптотичну оцінку функцій $p \in C^r$, визначених з усередненими значеннями на елементах деякого розбиття. При цьому використання таких сплайнів за наявності відомих явних виразів у формі поліномів з визначеними коефіцієнтами при мономах забезпечує високу швидкодію обробки даних у багатовимірному випадку при реалізації обчислювальних схем у відповідному програмному забезпеченні. Проте, інформаційна технологія непараметричної оцінки регресії за використанням процедури регуляризації та згаданих сплайнів може бути розвиненою виходячи з наступного міркування.

Для перевірки адекватності регресійної оцінки широко

застосовують критерій Фішера, оснований на статистиці $f = \hat{\sigma}_\varepsilon^2 / \hat{\sigma}_Y^2$, де $\hat{\sigma}_\varepsilon^2$, $\hat{\sigma}_Y^2$ – відповідно, оцінки залишкової дисперсії та дисперсії Y. При цьому адекватність тим вища, чим меншим є значення f у порівнянні з квантилем розподілу Фішера при заданому рівні похибки першого роду щодо прийняття відповідного рішення. Отже, застосування інваріантного перетворення над $\{(X_l, Y_l); l = \overline{1, N}\}$, яке б приводило до простору з меншою варіабельністю даних, надало б змогу проводити більш якісну оцінку регресії в новій системі координат за використанням локальних сплайн-операторів підвищеної точності. Останнє є справедливим у зв'язку з тим, що в досліджених оцінках якості апроксимації гладких функцій сплайнами про які мова [6], явно присутня норма сплайн-оператора, яка є тим більшою, чим більша варіабельність функції, що наближаємо.

Поставимо за мету в даній статті викласти та дослідити експериментально нову інформаційну технологію непараметричної оцінки двовимірної регресії на основі згаданих вище міркувань.

Виклад основного матеріалу. Подальше викладення матеріалу проведемо з використанням позначень, що мають місце в [6].

Інформаційна технологія непараметричної оцінки двовимірної регресії має наступний вигляд. Нехай задано об'єкт спостереження, який характеризується трьома ознаками T, Q, G, а реалізації об'єкта – трійки дійсних чисел

$$(t_l, q_l, g_l), l = \overline{1, N}, \quad (1)$$
$$t \in [t_{\min}, t_{\max}], \; q \in [q_{\min}, q_{\max}], \; g \in [g_{\min}, g_{\max}].$$

Необхідно знайти оцінку $\hat{\bar{g}}(t,q)$ залежності $\bar{g}(t,q) \in C^{r_1 r_2}$, $r_1, r_2 = 2,3...$ Технологія відтворення непараметричної регресії складається з наступних етапів:

Етап №1. Відтворення площини регресії за рядом реалізації об'єкта спостережень;

Етап №2. Лінійне перетворення реалізації об'єкта відносно площини регресії;

Етап №3. Формування тривимірного ряду за рівномірним розбиттям та новою реалізацією об'єкта (регуляризація даних);

Етап №4. Непараметричне відтворення двовимірної регресії за допомогою поліноміальних сплайнів на основі B-сплайнів, близьких до інтерполяційних у середньому;

Етап №5. Повернення до початкової реалізації об'єкта.

Етап №1. Відтворення двовимірної площини регресії $\overline{g}_0(t,q) \in C^{n/2}$ відбувається за моделлю вигляду

$$\overline{g}_0(t,q) = a_0 + a_1 t + a_2 q,$$

оцінки параметрів якої визначаються так:

$$\hat{a}_0 = \overline{g} - \hat{a}_1 \cdot \overline{t} - \hat{a}_2 \cdot \overline{q}$$

$$\hat{a}_1 = \frac{\sum_{i=1}^{N}(g_i-\overline{g})(t_i-\overline{t})\cdot\sum_{i=1}^{N}(q_i-\overline{q})^2 - \sum_{i=1}^{N}(g_i-\overline{g})(q_i-\overline{q})\cdot\sum_{i=1}^{N}(t_i-\overline{t})\cdot(q_i-\overline{q})}{\sum_{i=1}^{N}(t_i-\overline{t})^2\cdot\sum_{i=1}^{N}(q_i-\overline{q})^2 - \left(\sum_{i=1}^{N}(t_i-\overline{t})\cdot(q_i-\overline{q})\right)^2},$$

$$\hat{a}_2 = \frac{\sum_{i=1}^{N}(g_i-\overline{g})(q_i-\overline{q})\cdot\sum_{i=1}^{N}(t_i-\overline{t})^2 - \sum_{i=1}^{N}(g_i-\overline{g})(t_i-\overline{t})\cdot\sum_{i=1}^{N}(t_i-\overline{t})\cdot(q_i-\overline{q})}{\sum_{i=1}^{N}(t_i-\overline{t})^2\cdot\sum_{i=1}^{N}(q_i-\overline{q})^2 - \left(\sum_{i=1}^{N}(t_i-\overline{t})\cdot(q_i-\overline{q})\right)^2},$$

де $\overline{t} = \frac{1}{N}\sum_{i=1}^{N} t_i$, $\overline{q} = \frac{1}{N}\sum_{i=1}^{N} q_i$, $\overline{g} = \frac{1}{N}\sum_{i=1}^{N} g_i$, та побудовано площину регресії виду

$$\hat{a}_0 + \hat{a}_1 \cdot t + \hat{a}_2 \cdot q - \overline{g}_0(t,q) = 0 \qquad (2)$$

Етап №2. Лінійне (координатне) перетворення реалізацій об'єкта дослідження (1) за допомогою кутів нахилу площини (2).

Обчислено кути обернення α, β відносно ознак T та Q реалізації об'єкта, які дорівнюють:

$$\alpha = \pm \arccos\left(\frac{\sqrt{\hat{a}_1^2+1}}{\sqrt{\hat{a}_1^2+\hat{a}_2^2+1}}\right), \quad \beta = \pm \arccos\left(-\frac{-1}{\sqrt{\hat{a}_1^2+1}}\right).$$

Знак при кутах вибирається відповідно до коефіцієнтів множинної регресії (табл.1).

Таблиця 1

Залежність кутів від коефіцієнтів регресії

	$\hat{a}_2 \geq 0$		$\hat{a}_2 < 0$	
$\hat{a}_1 \geq 0$	α	$-\beta$	$-\alpha$	$-\beta$
$\hat{a}_1 < 0$	α	β	$-\alpha$	β

У результаті проведення зміни реалізації об'єкта, подальшій обробці підлягає масив даних вигляду:

$$(t_l', q_l', g_l'), l = \overline{1, N},$$

$$t' \in [t'_{\min}, t'_{\max}], \; q' \in [q'_{\min}, q'_{\max}], \; g' \in [g'_{\min}, g'_{\max}]$$

обчислення якого відбувається за наступною схемою:

$$(t_l', q_l', g_l') = (t_l, q_l, g_l) \cdot R(\alpha) \cdot R(\beta),$$

де

$$R(\alpha) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & -\sin(\alpha) \\ 0 & \sin(\alpha) & \cos(\alpha) \end{pmatrix}, \; R(\beta) = \begin{pmatrix} \cos(\beta) & 0 & \sin(\beta) \\ 0 & 1 & 0 \\ -\sin(\beta) & 0 & \cos(\beta) \end{pmatrix}.$$

Етапи №3, 4. Формування тривимірного ряду за рівномірним розбиттям та непараметричне відтворення двовимірної регресії за допомогою поліноміальних сплайнів на основі B-сплайнів, близьких до інтерполяційних у середньому проводиться за допомогою процедури регуляризації даних [6, підрозд. 5.4].

Результатом регуляризації є масив точок $\left\{(t_{i,j}'^{(\kappa)}, q_{i,j}'^{(\kappa)}, g_{i,j}'^{(\kappa)}), i,j = \overline{0,2^\kappa - 1}\right\}$, де $\left\{(t_{i,j}'^{(\kappa)}, q_{i,j}'^{(\kappa)}), i,j = \overline{0,2^\kappa - 1}\right\}$ – масив точок, що визначає регулярну сітку вузлів в області визначення незалежних змінних двовимірної регресії; $g_{i,j}'^{(\kappa)}$ – шукана оцінка усередненого значення функції регресії $\overline{g}'(t',q')$ в області $\Delta_{i,j}^{(\kappa)}$ площини аргументів; к – кількість ітерацій процедури регуляризації.

Неперервне наближення $\hat{\overline{g}}'^{(\kappa)}(t'^{(\kappa)}, q'^{(\kappa)})$ функції регресії $\overline{g}'^{(\kappa)}(t'^{(\kappa)}, q'^{(\kappa)})$ визначається у вигляді сплайну, побудованого за регуляризованим масивом [6].

Етап №5. Повернення до початкової області визначення об'єкта спостережень. Зміна реалізації об'єкта полягає у поверненні до координат $\left\{(t_{i,j}^{(\kappa)}, q_{i,j}^{(\kappa)}, g_{i,j}^{(\kappa)}), i,j = \overline{0,2^\kappa - 1}\right\}$, що визначаються:

$$\left(t_{i,j}^{(\kappa)}, q_{i,j}^{(\kappa)}, g_{i,j}^{(\kappa)}\right) = \left(t_{i,j}^{\prime(\kappa)}, q_{i,j}^{\prime(\kappa)}, g_{i,j}^{\prime(\kappa)}\right) \cdot R(-\beta) \cdot R(-\alpha).$$

Пропонуючи той чи інший метод регресійної або функціональної залежності спостережень за масивом її реалізації, впевнено можна зазначити, якщо метод добре працює для поліноміальних функцій, або, наприклад, для лінійних комбінацій тригонометричних функцій, то його використання задовольнить більшість вимог дослідника.

Порівняльний аналіз застосування представленої технології та інформаційної технології відтворення регресійних та інших залежностей [6], проводився на основі статистик: середнього абсолютного $\overline{D}$ та середнього відхилення $\overline{m}$ регресії $\overline{g}(t,q)$ та наближення $\hat{\overline{g}}(t,q)$, а також перевірка статистичної гіпотези про адекватність відтворення регресії (реалізація *F-критерію*, $\overline{f}$). Отримання $\overline{D}$, $\overline{m}$ та $\overline{f}$ було проведено на підставі даних, визначених на нерегулярних сітках вузлів, як реалізацій функцій регресії наступного виду:

$$\overline{g}(t,q) = -t^3 - q^2 + 0{,}5t^2 q + 0{,}5t + 0{,}5q,\ t,q \in [-10{,}10]; \quad (4)$$

$$\overline{g}(t,q) = t^3 + q^2 - 0{,}5t^2 q - 0{,}5t - 0{,}5q,\ t,q \in [-10{,}10]; \quad (5)$$

$$\overline{g}(t,q) = -t^3 - q^2 + 0{,}5t^2 q + 0{,}5t + 0{,}5q,\ t,q \in [-10{,}10]; \quad (6)$$

$$\overline{g}(t,q) = -t^3 - q^2 + 0{,}5t^2 q + 0{,}5t + 0{,}5q,\ t,q \in [-2\pi, 2\pi]. \quad (7)$$

Масиви даних для аналізу отримано за результатами імітаційного моделювання. В якості закону розподілу реалізації об'єкта (1) виступає нормальний закон з параметрами μ та σ. Задаючись обсягом N масиву, що генерується, правило моделювання масиву для подальшої обробки має наступний вигляд:

$$t_l = t_{\min} + randomN(\mu,\sigma)\cdot(t_{\max} - t_{\min}),$$

$$q_l = q_{\min} + randomN(\mu,\sigma)\cdot(q_{\max} - q_{\min}),$$

$$g_l = \overline{g}(t_l, q_l) + \overline{g}(t_l, q_l)\cdot randomN(\mu, 2\sigma),\ l = \overline{1,N},$$

де $randomN(\mu,\sigma)$ – нормально розподілене число з параметрами μ, σ.

Результати аналізу представлено таблицею 2 для випадку двовимірного сплайну $S_{2,1}(g,t,q)$. Кількість ітерацій процедури регуляризації $\kappa = \overline{2,4}$. Моделювання масивів для аналізу здійснювалось з параметрами нормального розподілу $\mu = 1$, $\sigma = 0{,}6$.

Випадки, коли $\kappa = 2$ обсяг даних становив $N < 20$, для $\kappa = 3{,}4$

аналіз проводився на обсязі $N = 1000$. Усі результати наведені для кількості повторень експериментів $M = 100$.

Таблиця 2

Результати порівняння двох технологій відтворення

	$\overline{D}$	$\overline{m}$	$\overline{f}$	$\overline{D}$	$\overline{m}$	$\overline{f}$	$\overline{D}$	$\overline{m}$	$\overline{f}$
	к=2			**к=3**			**к=4**		
Результати за відтворенням функції (4)									
Підхід №1*	5e–5	2e–5	4,32	3e–5	–2e–5	7.96	4e–5	–3e–5	6,4
Підхід №2**	3e–2	–2e–2	5,99	2e–2	–1e–2	8,81	2e–2	–1e–2	9,1
Результати за відтворенням функції (5)									
Підхід №1*	2e–5	5e–6	4.28	3e–5	3e–5	7.47	4e–5	3e–5	8.11
Підхід №2**	2e–2	9e–3	7,34	2e–2	2e–2	8,59	2e–2	2e–2	9,31
Результати за відтворенням функції (6)									
Підхід №1*	7e–4	–6e–6	3.16	6e–4	5e–5	9,96	2e–4	2e–4	15.6
Підхід №2**	5e–2	4e–2	9,21	2e–2	1e–2	6,06	3e–2	2e–2	11,7
Результати за відтворенням функції (7)									
Підхід №1*	6e–2	9e–4	1,99	2e–2	1e–3	1,46	2e–2	1e–4	1,56
Підхід №2**	7e–2	1e–3	2,76	2e–2	1e–3	1.7	2e–2	8e–4	1,65

* – підхід представлений у статті;

** – підхід, опис якого представлено у [4]

Висновки. З аналізу результатів маємо такі висновки:

- усі величини, що є мірою якості оцінки $\hat{\overline{g}}(t,q)$: середнє абсолютне відхилення $\overline{D}$, середнє відхилення $\overline{m}$ та величина $\overline{f}$ за підходом №1 на порядок менші ніж за підходом №2;
- значення середнього абсолютного відхилення $\overline{D}$ за підходом №1 є прийнятною для всіх реалізацій двовимірної функції регресії;
- аналіз середнього відхилення $\overline{m}$ показав, що ця величина близька до нуля, проте, як і у випадку підходу №2, для опуклих вгору двовимірних регресій (4) маємо незначне «відставання» $\hat{\overline{g}}(t,q)$ відносно $\overline{g}(t,q)$; для опуклих вниз двовимірних функцій регресії типу (5) результат відтворення $\hat{\overline{g}}(t,q)$ трохи перевищує істинне значення. Меньшою мірою це

стосується площинних та багатомодальних двовимірних функцій регресій типу (6, 7);

– величина $\tilde{f}$ свідчить про високу адекватність оцінки $\tilde{g}(t,q)$ у вигляді $\hat{\tilde{g}}(t,q)$.

Подальші дослідження можуть мати за мету, наприклад, узагальнення поданої технології на випадок оцінювання регресії трьох та більше аргументів. Крім того, технологія може бути поширена на пошук інших лінійних перетворень, що приводять до нової системи координат: за використанням відстані до площини регресії за рядом реалізації об'єкта спостережень; відхилення від площини регресії за рядом реалізації об'єкта спостережень та методу головних компонент. Запропонована інформаційна технологія непараметричної оцінки функції регресії, за використанням лінійних комбінацій B-сплайнів, близьких до інтерполяційних у середньому може бути використана при розробці автоматизованих систем обробки даних для вирішення задач картографічного моніторингу, моніторингу складних об'єктів та оперативного аналізу показників гідрохімічного та геохімічного моніторингів і т. ін.

Бібліографічні посилання

1. Вероятность и математическая статистика: Энциклопедия / Гл. ред. Ю.В. Прохоров. – М., 1999. – 910 с.

2. **Катовник В.Я.** Непараметрическая идентификация и сглаживание данных. / В.Я. Катовник. – М.,1985. – 336 с.

3. **Живоглядов В.П.** Непараметрические алгоритмы адаптации. / В.П. Живоглядов , А.В. Медведев – Фрунзе, 1974.

4. **Надарая Е.А.** Непараметрическое оценивание плотности вероятности и кривой регрессии. / Е.А. Надарая. – Тбилиси, 1983.

5. **Рубан А.И.** Идентификация стахостических объектов на основе непараметрического подхода. / А.И. Рубан. – Автоматика и телемеханика, 1979, № 11, С. 106–118.

6. **Приставка П.О.** Поліноміальні сплайни при обробці даних: Монографія. / П.О. Приставка. – Д., 2004. – 236 с.

7. **Приставка П.О.** Непараметрична оцінка залежностей методом найменших квадратів . / П.О. Приставка // Вісн. НАУ, 2003. – №1(16). – С. 17–20.

8. **Rosenblatt M.** Conditional probability density and regression estimators. – In: Multivariate Analisis, II, Academic Press, N.T., 1969, p. 25-31.

9. **Watson G.S.** Smooth regression Analisis. – Sakhya, ser. A, 1964, V, 26, part 4, P. 359–372.

Надійшла до редколегії 27.06.09