Теорія ймовірностей
та математична статистика
Вип. 1(100)/2019, с. 60–74

Teoriya Ĭmovirnosteĭ
ta Matematychna Statystyka
No. 1(100)/2019, pp. 60–74

УДК 519.21

УЗАГАЛЬНЕНІ ОЦІНКИ НАДАРАЯ – ВАТСОНА ДЛЯ СПОСТЕРЕЖЕНЬ ІЗ СУМІШІ

Г. М. ДИЧКО, Р. Є. МАЙБОРОДА

Анотація. Розглядається узагальнення оцінок Надарая – Ватсона для спостережень із суміші зі змінними концентраціями. Доведено консистентність та асимптотичну нормальність отриманих оцінок.

Ключові слова і фрази. Модель суміші зі змінними концентраціями, непараметрична регресія, асимптотична нормальність, оцінка Надарая – Ватсона.

2010 *Mathematics Subject Classification.* Primary 62G08; Secondary 62G20.

1. Вступ

Моделі суміші кількох компонентів із різними розподілами спостережуваних змінних природно виникають у статистичному аналізі даних медико-біологічних, соціологічних та економічних досліджень. Теорія параметричного оцінювання у таких моделях викладена, наприклад, у [11], приклади застосування до медико-біологічних даних — у [13]. Комп'ютерні засоби, що реалізують алгоритми аналізу сумішей, описано у [1, 7].

При дослідженні сумішей, як і однорідних даних, для аналізу залежностей між змінними, що спостерігаються, зручно використовувати ті чи інші моделі регресійного аналізу. Наприклад, роботи [3, 4] присвячено побудові оцінок параметрів у моделях суміші лінійних регресій для випадку, коли розподіли змінних описуються параметричними моделями. У роботах [6, 8] розглянуто оцінювання коефіцієнтів лінійних регресій у випадку непараметричних моделей для розподілів похибок регресії та регресорів.

У цій роботі розглядається суміш кількох регресій, у якій для функцій регресії не задані параметричні моделі. Таким чином, модель стає непараметричною не тільки відносно таких «заважаючих» характеристик, як розподіл похибки, а й щодо основної характеристики, що вимагає оцінювання — функції регресії. При цьому ми припускаємо, що ймовірності змішування (концентрації компонентів у суміші) є різними для різних спостережень і відомими. Таким чином, досліджувана модель укладається у загальну схему моделей сумішей зі змінними концентраціями [9, 10].

Однією з найбільш поширених непараметричних оцінок функції регресії є оцінка Надарая – Ватсона [5, 12, 15]. Ми розглянемо модифікацію цієї оцінки для випадку спостережень із суміші зі змінними концентраціями й отримаємо умови її консистентності та асимптотичної нормальності. Це, зокрема, дозволяє визначити теоретично оптимальне значення параметра згладжування (bandwidth) для таких оцінок.

Далі у п. 2 формально описана непараметрична регресійна модель для сумішей зі змінними концентраціями та вводяться модифіковані оцінки Надарая – Ватсона. У п. 3 сформульовані теореми про консистентність та асимптотичну нормальність цих оцінок. Результати п. 3 використовуються у п. 4 для визначення теоретично оптимальних параметрів налаштування оцінок (ядра та параметра згладжування). Доведення теорем вміщено у п. 5, а результати перевірки якості оцінок на модельованих даних — у п. 6. Висновкам присвячено п. 7.

2. Постановка задачі і побудова оцінки

У моделі суміші зі змінними концентраціями розглядаються об'єкти O, кожен із яких належить одній з M популяцій (компонентів суміші). Номер компонента, якому належить O, будемо позначати $\kappa(O)$. Вибірка містить n об'єктів $O_1, \dots, O_n$. Для кожного із цих об'єктів спостерігається вектор змінних $\xi_{j;n} = \xi(O_j)$. Значення[1] $\kappa_{j;n} = \kappa(O_j)$ не спостерігаються, але відомі концентрації компонентів у суміші під час j-го спостереження, тобто

$$p^m_{j;n} = \mathsf{P}\{\kappa_{j;n} = m\},\ j = 1, \dots, n;\ m = 1, \dots, M.$$

Розподіл спостережуваних характеристик об'єкта залежить від того, якому компоненту суміші він належить:

$$F_m(A) = \mathsf{P}\{\xi(O) \in A \mid \kappa(O) = m\},\ m = 1, \dots, M,$$

для будь-якої борелевої множини A.

Розподіли компонентів F_m вважаються невідомими.

Таким чином, розподіли спостережуваних змінних задаються моделлю суміші:

$$\mathsf{P}\{\xi_{j;n} \in A\} = \sum_{m=1}^{M} p^m_{j;n} F_m(A).$$

Вектори $(\xi_{j;n}, \kappa_{j;n})$ $j = 1, \dots, n$, вважаються незалежними в сукупності.

(Приклади застосування таких моделей до аналізу реальних даних див. [8, 10].)

Ми розглядаємо випадок, коли вектори спостережуваних характеристик складаються із двох числових змінних X (регресор) та Y (відгук): $\xi(O_j) = (X_j, Y_j)$. При цьому залежність між відгуком і регресором описується непараметричною регресійною моделлю, у якій функція регресії може бути різною для різних компонентів суміші:

$$Y_j = g^{(\kappa_j)}(X_j) + \varepsilon_j,\ j = 1, \dots, n. \tag{1}$$

Тут

$g^{(m)}$ — невипадкова функція регресії для m-го компонента суміші;

ε_j — випадкова похибка регресії.

Розподіл похибок може залежати від того, якому компоненту належить об'єкт, але вважається, що

$$\mathsf{E}[\varepsilon_j \mid \kappa_j = m] = 0,\ \sigma^2_{(m)} = \mathrm{Var}[\varepsilon_j \mid \kappa_j = m] < \infty,\ m = 1, \dots, M,$$

(дисперсії похибок $\sigma^2_{(m)}$ вважаються невідомими). Крім того, ми будемо вважати, що при фіксованому κ_j похибки регресії ε_j і регресори X_j є незалежними між собою.

Також ми будемо припускати, що розподіли регресорів X для всіх компонентів є абсолютно неперервними відносно міри Лебега. Позначимо $f^{(m)}$ — (невідому) щільність розподілу X у об'єктів, що належать m-му компоненту.

Задача полягає в тому, щоб оцінити функцію $g^{(k)}(x)$ за спостереженнями (X_j, Y_j), $j = 1, \dots, n$. (При відомих концентраціях p^m_j.)

Якби функції регресії всіх компонентів були однаковими, тобто $g^{(m)}(x) = g(x)$, $m = 1, \dots, M$, то для оцінювання $g(x)$ у точці $x = x_0$ можна було б використати

[1] Тут і далі нижній індекс $;n$ позначає, що спостереження належить вибірці обсягу n. В асимптотичній теорії досліджується поведінка оцінок при $n \to \infty$, причому ці вибірки при різних n не вважаються якось пов'язаними. Для спрощення позначень індекс $;n$ опускаємо там, де це не створює двозначності, тобто $\kappa_{j;n} = \kappa_j$ і т. д.

звичайну оцінку Надарая–Ватсона:

$$\hat{g}_n(x_0) = \frac{\sum_{j=1}^n Y_j K\Big(\frac{x_0 - X_j}{h}\Big)}{\sum_{j=1}^n K\Big(\frac{x_0 - X_j}{h}\Big)}. \tag{2}$$

Тут

$K : \mathbb{R} \to \mathbb{R}$ — функція, яку звуть ядром оцінки;

h — число, що зветься параметром згладжування (bandwidth) оцінки.

Ядро K і число h є «параметрами налаштування» оцінки Надарая–Ватсона, їх можна вибирати у досить широкому діапазоні, намагаючись отримати оптимальну, у певному розумінні, оцінку.

За досить широких умов ця оцінка буде консистентною.

Але зрозуміло, що коли функції регресії є різними, $\hat{g}_n(x_0)$ не буде оцінювати жодну з них. Щоб оцінити $g^{(k)}(x_0)$ потрібно виділити вплив m-го компонента на оцінку і пригасити впливи всіх інших компонентів. Це можна зробити, використовуючи підсумовування з навантаженням у формулі (2):

$$\hat{g}_n^{(k)}(x_0) = \frac{\sum_{j=1}^n a_{j;n}^k Y_j K\Big(\frac{x_0 - X_j}{h}\Big)}{\sum_{j=1}^n a_{j;n}^k K\Big(\frac{x_0 - X_j}{h}\Big)}. \tag{3}$$

Тут $a_{j;n}^k$ — невипадкові вагові коефіцієнти (навантаження), що підбираються за концентраціями компонентів так, щоб отримати консистентну оцінку $g^{(k)}(x_0)$.

(Схожий підхід був використаний у роботі [14] для отримання модифікованих ядерних оцінок щільностей розподілу компонентів суміші.)

Перейдемо до опису можливого вибору вагових коефіцієнтів $a_{j;n}^k$. Спочатку введемо ряд позначень.

Для масивів концентрацій $\mathbf{p} = (p_{j;n}^m,\ j = 1, \dots, n,\ m = 1, \dots, M,\ n = 1, 2, \dots)$ та вагових коефіцієнтів $\mathbf{a} = (a_{j;n}^m,\ j = 1, \dots, n,\ m = 1, \dots, M,\ n = 1, 2, \dots)$ будемо позначати $\mathbf{p}^m = (p_1^m, \dots, p_n^m)^T$ — вектор-стовпець концентрацій m-го компонента для всіх спостережень і, аналогічно, $\mathbf{a}^m = (a_1^m, \dots, a_n^m)^T$.

Операцію усереднення по всій вибірці позначимо кутовими дужками:

$$\langle \mathbf{p}^m \rangle_n \stackrel{\text{def}}{=} \frac{1}{n} \sum_{j=1}^n p_{j;n}^m.$$

При записі операцій із масивами всередині кутових дужок додавання, множення і піднесення до степеня трактуються як поелементні дії:

$$\langle \mathbf{a}^i \mathbf{p}^m \rangle_n = \frac{1}{n} \sum_{j=1}^n a_{j;n}^i p_{j;n}^m, \quad \big\langle (\mathbf{a}^i)^2 \big\rangle_n = \frac{1}{n} \sum_{j=1}^n \big(a_{j;n}^i\big)^2.$$

Будемо позначати $\langle \mathbf{p}^m \rangle = \lim_{n\to\infty} \langle \mathbf{p}^m \rangle_n$, якщо ця границя існує (аналогічно для інших векторів і їх комбінацій). Помітимо, що $\langle \mathbf{a}^i, \mathbf{p}^m \rangle_n = \langle \mathbf{a}^i \mathbf{p}^m \rangle_n$ можна трактувати як скалярний добуток $\mathbf{a}^i$ і $\mathbf{p}^m$ у $\mathbb{R}^n$.

Нехай Γ_n — матриця Грама векторів концентрацій $\mathbf{p}^m$, $m = 1, \dots, M$: $\Gamma_n = \big(\langle \mathbf{p}^i, \mathbf{p}^m \rangle_n\big)_{i,m=1}^M$. Далі ми припускаємо, що вектори $\mathbf{p}^i$, $i = 1, \dots, M$, є лінійно незалежними, отже, $\det \Gamma_n \neq 0$.

Вагові коефіцієнти a_j^i визначимо як

$$a_j^k = \frac{1}{\det \Gamma_n} \sum_{m=1}^M (-1)^{m+k} \gamma_{km} p_j^m, \tag{4}$$

де γ_{im} — i, m-й мінор матриці Γ_n.

Коефіцієнти $\mathbf{a}$, визначені (4), називають мінімаксними (див. п. 2.1 у [9] або [10] про властивості цих коефіцієнтів).

Помітимо, що для мінімаксних $a^i_{j;n}$ виконується тотожність

$$\langle \mathbf{a}^i, \mathbf{p}^m \rangle = \frac{1}{n}\sum_{j=1}^{n} a^i_j p^m_j = \mathbf{1}\{i = m\}, \quad m = 1, \dots, M, \tag{5}$$

де $\mathbf{1}\{A\}$ — індикатор події A.

Якщо для деякої борелевої функції $r(x)$ розглядати навантажені функціональні моменти вибірки

$$\hat{r}^{(k)}_n = \frac{1}{n}\sum_{j=1}^{n} a^k_{j;n} r(\xi_{j;n})$$

як оцінки для відповідних теоретичних моментів k-го компонента:

$$\bar{r}^{(k)} = \mathsf{E}[r(\xi(O)) \mid \kappa(O) = k],$$

то умова (5) забезпечує незміщенність $\hat{r}^{(k)}_n$. Коефіцієнти $\mathbf{a}$, що задовольняють умову (5), називають незміщенними коефіцієнтами. При використанні у (3) незміщенних коефіцієнтів, математичні сподівання чисельника і знаменника у правій частині (3) будуть залежати лише від розподілу k-го компонента суміші. Це і дає можливість застосовувати таку модифікацію оцінок Надарая–Ватсона для оцінювання $g^{(k)}$.

Надалі ми будемо використовувати для побудови модифікованих оцінок Надарая–Ватсона (3) лише мінімаксні коефіцієнти, визначені (4). Але основні результати наступного розділу можна узагальнити на випадок будь-яких незміщенних коефіцієнтів.

3. Асимптотична поведінка оцінок

У цьому розділі досліджується поведінка модифікованих оцінок Надарая–Ватсона $\hat{g}^{(k)}_n(x_0)$, визначених (3), при зростанні обсягу вибірки n до нескінченності. При цьому ми вважаємо, що вагові коефіцієнти $a^k_{j;n}$ визначені (4).

Будемо казати, що вимірна за Лебегом функція $K : \mathbb{R} \to \mathbb{R}$ є фінітним обмеженим ядром, якщо існують такі $c > 0$ і $C < \infty$, що $K(x) = 0$ для всіх x, таких, що $|x| > c$ та $|K(x)| < C$ для всіх $x \in \mathbb{R}$.

Символом $\xrightarrow{\mathrm{P}}$ будемо позначати збіжність за ймовірністю.

Наступна теорема встановлює умови поточкової консистентності модифікованих оцінок Надарая–Ватсона.

Теорема 3.1. *Нехай для деякого фіксованого $k \in 1, \dots, M$ виконуються такі умови.*

1. K — фінітне обмежене ядро.

2. $x_0 \in \mathbb{R}$ є точкою неперервності функцій $g^{(k)}$ та $f^{(k)}$.

3. $f^{(k)}(x_0) > 0$.

4. Функції $g^{(m)}$ і $f^{(m)}$ обмежені в деякому околі точки x_0 для всіх $m = 1, \dots, M$.

5. Існує $c_0 > 0$, таке, що $\det \Gamma_n > c_0$ для всіх n.

6. $h = h_n \to 0$, $h_n n \to \infty$ при $n \to \infty$.

Тоді

$$\hat{g}^{(k)}_n(x_0) \xrightarrow{\mathrm{P}} g^{(k)}(x_0).$$

Для того, щоб отримати умови асимптотичної нормальності модифікованих оцінок Надарая–Ватсона, нам будуть потрібні деякі позначення.

Позначимо

$$D = \int_{-\infty}^{\infty} z^2 K(z) dz,$$

$$d_-^2 = \int_{-\infty}^{0} (K(z))^2 dz, \quad d_+^2 = \int_0^{\infty} (K(z))^2 dz, \; d^2 = d_-^2 + d_+^2.$$

Точками над символом функції будемо позначати похідні по dx:

$$\dot{f}^{(k)}(x) = \frac{d}{dx} f^{(k)}(x), \quad \ddot{g}^{(k)}(x) = \frac{d^2}{dx^2} g^{(k)}(x).$$

Теорема 3.2. *Нехай для деякого фіксованого $k \in 1, \dots, M$ виконуються такі умови.*

1. Функція $g^{(k)}$ має неперервну другу похідну $\ddot{g}^{(k)}(x)$ в околі точки x_0.

2. Функції $g^{(m)}$, $m = 1, \dots, M$ мають границі

$$g^{(m)}(x_0-) = \lim_{x \uparrow x_0} g^{(m)}(x), \quad g^{(m)}(x_0+) = \lim_{x \downarrow x_0} g^{(m)}(x).$$

3. Існує

$$\Gamma = \lim_{n \to \infty} \Gamma_n$$

і $\det \Gamma \neq 0$.

4. Для всіх $m = 1, \dots, M$ існують границі

$$\big\langle (\mathbf{a}^k)^2 \mathbf{p}^m \big\rangle = \lim_{n \to \infty} \big\langle (\mathbf{a}^k)^2 \mathbf{p}^m \big\rangle_n.$$

5. У деякому околі точки x_0 функції $f^{(m)}$ неперервні для всіх $m = 1, \dots, M$, $f^{(k)}$ — неперервно диференційовна і $f^{(k)}(x_0) > 0$.

6. K — фінітне обмежене ядро і $\int_{-\infty}^{\infty} zK(z)dz = 0$.

7. $h = h_n = Hn^{-1/5}$, де $H > 0$ — деяке фіксоване число.

Тоді

$$n^{2/5}\Big(\hat{g}_n^{(k)}(x_0) - g^{(k)}(x_0)\Big) \xrightarrow{\mathrm{W}} \mathcal{N}\Big(\mu^{(k)}(x_0), S_{(k)}^2(x_0)\Big),$$

де

$$\mu^{(k)}(x_0) = H^2 D\left(\frac{\dot{g}^{(k)}(x_0)\dot{f}^{(k)}(x_0)}{f^{(k)}(x_0)} + \frac{1}{2}\ddot{g}^{(k)}(x_0)\right),$$

$$S_{(k)}^2(x_0) = \frac{1}{H(f^{(k)}(x_0))^2} \sum_{m=1}^{M} f^{(m)}(x_0) \big\langle (\mathbf{a}^k)^2 \mathbf{p}^m \big\rangle R_m,$$

$$R_m = \sigma_{(m)}^2 d^2 + \Big(g^{(k)}(x_0) - g^{(m)}(x_0-)\Big)^2 d_-^2 + \Big(g^{(k)}(x_0) - g^{(m)}(x_0+)\Big)^2 d_+^2.$$

Зауваження 3.1. 1. У цих теоремах вимога, щоб K було обмеженим фінітним ядром, не є обов'язковою. Можна розглядати також і необмежені ядра, що набувають ненульових значень на всій прямій. Але тоді потрібно буде накладати додаткові обмеження на функції регресії $g^{(m)}$ та щільності $f^{(m)}$, не тільки в околі тієї точки x_0, де оцінюється $g^{(k)}$, а й на всій прямій.

2. Обмеження швидкості прямування параметра згладжування до 0, тобто умова $h_n n \to \infty$, у теоремі 3.1 є принциповою. Таке саме обмеження вимагається і для консистентності звичайних оцінок Надарая – Ватсона за однорідною вибіркою.

3. Швидкість збіжності оцінок порядку $n^{-2/5}$, яку забезпечує теорема 3.2, є непокращуваною навіть для оцінювання по однорідних вибірках, якщо розглядаються двічі неперервно диференційовні функції регресії. Саме для досягнення такої швидкості збіжності оцінок потрібно обрати параметр згладжування $h_n \sim n^{-1/5}$.

4. Умова $f^{(k)}(x_0) > 0$ в обох теоремах — принципова. Якщо $f^{(k)}(x) = 0$ у деякому околі точки x_0, то непараметрично оцінити $g^{(k)}(x_0)$ неможливо. Дійсно, у цьому випадку не існує об'єктів із k-го компонента, у яких значення регресора могло б потрапити в окіл точки x_0. Якщо, наприклад, $f^{(k)}(x) > 0$ для $x < x_0$, але $f^{(k)}(x) = 0$ для $x > x_0$, то консистентне оцінювання $g^{(k)}(x_0)$ можливе, але використовувати для

цього оцінки Надарая–Ватсона недоцільно через так званий крайовий ефект (див. [5, приклад 4.3]).

5. Вимоги $\det\Gamma_n > c_0 > 0$ у теоремі 3.1 і $\det\Gamma \neq 0$ у теоремі 3.2 є «асимптотичними аналогами» умови $\det\Gamma_n \neq 0$, тобто незалежності векторів концентрацій різних компонентів. Їх можна дещо послабити (вимагати, наприклад, щоб послідовність $\det\Gamma_n$ не прямувала до 0 занадто швидко), але повністю відмовитись від них не можна. Наприклад, неможливо проводити консистентне непараметричне оцінювання функцій регресії, коли концентрації у суміші є сталими.

4. Оптимальний вибір параметрів налаштування оцінки

Модифікована оцінка Надарая–Ватсона, визначена (3), має два параметри налаштування, які вибирає сам статистик, що нею користується. Це параметр згладжування h і ядро оцінки K. Природно обирати їх так, щоб точність оцінки була найкращою. У теоремі 3.2 параметр згладжування задається як $h = Hn^{-1/5}$, тому його вибір еквівалентний вибору значення H.

Із точки зору асимптотики, отриманої у теоремі 3.2, розкид відхилень оцінки $\hat g_n^{(k)}(x_0)$ від $g^{(k)}(x_0)$ характеризується коливанням випадкової величини η з розподілом $\mathcal{N}(\mu^{(k)}(x_0), S^2_{(k)}(x_0))$ навколо 0. Ці коливання природно характеризувати другим моментом η:

$$\mathrm{aMSE}\Big(\hat g_n^{(k)}(x_0)\Big) \stackrel{\mathrm{def}}{=} \mathsf{E}\,\eta^2 = S^2_{(k)}(x_0) + (\mu^{(k)}(x_0))^2.$$

Цю величину будемо називати асимптотичною середньоквадратичною похибкою (asymptotic mean squared error, aMSE) оцінки.

Обмежимось розглядом моделей, у яких виконуються умови теореми 3.2 і, крім того, функції регресії всіх компонентів є неперервними в околі точки x_0. Тоді

$$\mathrm{aMSE}\Big(\hat g_n^{(k)}(x_0)\Big) = \frac{d^2L_2}{H} + H^4D^2L_1,$$

де

$$L_1 = \left(\frac{\dot g^{(k)}(x_0)\dot f^{(k)}(x_0)}{f^{(k)}(x_0)} + \frac{1}{2}\ddot g^{(k)}(x_0)\right)^2,$$

$$L_2 = \frac{1}{(f^{(k)}(x_0))^2}\sum_{m=1}^{M} f^{(m)}(x_0)\big\langle(\mathbf{a}^k)^2\mathbf{p}^m\big\rangle\left(\sigma^2_{(m)} + \Big(g^{(k)}(x_0) - g^{(m)}(x_0)\Big)^2\right).$$

Легко бачити, що мінімум $\mathrm{aMSE}\Big(\hat g_n^{(k)}(x_0)\Big)$ по H досягається на

$$H_{\mathrm{opt}} = \sqrt[5]{\frac{d^2L_2}{4D^2L_1}}.$$

Величину H_{opt} будемо називати теоретично оптимальним значенням константи H, а

$$h_{\mathrm{opt}} = H_{\mathrm{opt}}n^{-1/5}$$

— теоретично оптимальним значенням параметра згладжування.

Підставляючи H_{opt} в означення aMSE, отримуємо найменше можливе значення aMSE для модифікованої оцінки Надарая–Ватсона:

$$\mathrm{aMSE}_{\mathrm{opt}} = \frac{5}{4}\sqrt[5]{4d^8D^2L_2^4L_1}.$$

У цьому виразі від ядра K залежить лише добуток d^8D^2. Як відомо [5, п. 3.4.3], найменше значення d^8D^2 досягається на ядрі Єпанєчнікова:

$$K_{\mathrm{Epn}}(x) = \frac{3}{4}(1-x^2)\mathbf{1}\{|x|\le 1\}.$$

Таким чином, в умовах теореми 3.2, ядро Єпанєчнікова є оптимальним для оцінювання, якщо функції регресії всіх компонентів є неперервними у точці оцінювання.

На жаль, надати таку ж однозначну рекомендацію для вибору параметра згладжування h не можна. Теоретично оптимальне значення h_{opt} залежить від невідомих параметрів моделі. Тому його неможливо безпосередньо використати для оцінювання. Можливі різні підходи до вибору значення h, близького до оптимального, але їх придатність для оцінки функції регресії за спостереженнями із суміші потребує додаткового дослідження.

5. Доведення теорем

Для запису різних математичних сподівань і ймовірностей нам буде зручно ввести у розгляд випадкові вектори $(X^{(m)}, Y^{(m)})$, розподіл яких такий самий, як і умовний розподіл $\xi(O)$ за умови, що $\kappa(O) = m$. Відповідно, $\varepsilon^{(m)} = Y^{(m)} - g^{(m)}(X^{(m)})$ — випадкова величина, розподілена як похибка регресії для m-го компонента суміші та незалежна від $X^{(m)}$.

Доведення теореми 3.1.

Позначимо

$$\hat r_n^{(k)}(x_0) = \hat r_n(x_0) = \frac{1}{nh}\sum_{j=1}^{n} a_{j;n}^k Y_j K\left(\frac{x_0 - X_j}{h}\right),$$

$$\hat f_n^{(k)}(x_0) = \hat f_n(x_0) = \frac{1}{nh}\sum_{j=1}^{n} a_{j;n}^k K\left(\frac{x_0 - X_j}{h}\right). \tag{6}$$

Помітимо, що

$$\hat g_n^{(k)}(x_0) = \frac{\hat r_n(x_0)}{\hat f_n(x_0)}. \tag{7}$$

Ми покажемо, що, в умовах теореми,

$$\hat r_n(x_0) \xrightarrow{\text{P}} g^{(k)}(x_0) f^{(k)}(x_0) \text{ при } n \to \infty \tag{8}$$

і

$$\hat f_n(x_0) \xrightarrow{\text{P}} f^{(k)}(x_0) \text{ при } n \to \infty. \tag{9}$$

Звідси випливає твердження теореми.

Для доведення (8), покажемо, що $\mathsf{E}\,\hat r_n(x_0) \to g^{(k)}(x_0) f^{(k)}(x_0)$ і $\operatorname{Var} \hat r_n(x_0) \to 0$.

Почнемо з математичного сподівання:

$$\begin{aligned}
\mathsf{E}\,\hat r_n(x_0) &= \frac{1}{nh}\sum_{j=1}^{n} a_{j;n}^k \,\mathsf{E}\Big(g^{(\kappa_j)}(X_j) + \varepsilon_j\Big) K\left(\frac{x_0 - X_j}{h}\right) = \\
&= \sum_{m=1}^{M} \big\langle \mathbf{a}^k \mathbf{p}^m \big\rangle_n \frac{1}{h}\,\mathsf{E}\Big(g^{(m)}\big(X^{(m)}\big) + \varepsilon^{(m)}\Big) K\left(\frac{x_0 - X^{(m)}}{h}\right) = \\
&= \frac{1}{h}\,\mathsf{E}\, g^{(k)}\big(X^{(k)}\big) K\left(\frac{x_0 - X^{(k)}}{h}\right),
\end{aligned} \tag{10}$$

оскільки $\langle \mathbf{a}^k \mathbf{p}^m \rangle_n = 0$ при $m \neq k$, $\mathsf{E}\,\varepsilon^{(k)} = 0$ і $\varepsilon^{(k)}$ не залежить від $X^{(k)}$.

Отже,

$$\begin{aligned}
\mathsf{E}\,\hat r_n(x_0) &= \frac{1}{h}\int_{-\infty}^{\infty} g^{(k)}(x) K\left(\frac{x_0 - x}{h}\right) f^{(k)}(x)\,dx = \\
&= \int_{-\infty}^{\infty} g^{(k)}(x_0 - zh) K(z) f^{(k)}(x_0 - zh)\,dz.
\end{aligned} \tag{11}$$

(Ми зробили заміну змінної $z = (x_0 - x)/h$.)

Унаслідок умов 1 і 2 теореми 3.1, із (11) випливає, що при $h \to 0$,

$$\mathsf{E}\,\hat{r}_n(x_0) \to g^{(k)}(x_0) f^{(k)}(x_0). \tag{12}$$

Тепер оцінимо дисперсію.

$$\operatorname{Var}\hat{r}_n(x_0) = \frac{1}{n^2h^2}\sum_{j=1}^{n}(a_{j;n}^k)^2 \operatorname{Var}\left[\left(g^{(\kappa_j)}(X_j)+\varepsilon_j\right)K\left(\frac{x_0-X_j}{h}\right)\right] \le$$

$$\le \frac{1}{nh^2}\sum_{m=1}^{M}\left\langle (\mathbf{a}^k)^2\mathbf{p}^m\right\rangle \mathsf{E}\left[\left(g^{(m)}\big(X^{(m)}\big)+\varepsilon^{(m)}\right)K\left(\frac{x_0-X^{(m)}}{h}\right)\right]^2. \tag{13}$$

Враховуючи незалежність $\varepsilon^{(m)}$ та $X^{(m)}$ отримуємо, що

$$J_m \stackrel{\text{def}}{=} \frac{1}{h}\,\mathsf{E}\left[\left(g^{(m)}\big(X^{(m)}\big)+\varepsilon^{(m)}\right)K\left(\frac{x_0-X^{(m)}}{h}\right)\right]^2 =$$

$$= \frac{1}{h}\int_{-\infty}^{\infty}\left(\big(g^{(m)}(x)\big)^2+\sigma_{(m)}^2\right)\left(K\left(\frac{x_0-x}{h}\right)\right)^2 f^{(m)}(x)dx =$$

$$= \int_{-c}^{c}\left(\big(g^{(m)}(x_0-zh)\big)^2+\sigma_{(m)}^2\right)(K(z))^2 f^{(m)}(x_0-zh)dz.$$

(Ми скористались фінітністю ядра K.)

Враховуючи умову 4 теореми, отримуємо, що, при достатньо малих h, $J_m \le C_1$, де $C_1 < \infty$ — деяка константа. З умови 5 випливає

$$\sup_{m=1,\dots,M;j=1,\dots,n;n\ge 1}\left|a_{j:n}^m\right| \le C_2 < \infty.$$

Тому, продовжуючи (13), отримуємо

$$\operatorname{Var}\hat{r}_n(x_0) \le \frac{MC_1C_2}{nh} \to 0 \text{ при } n\to\infty$$

внаслідок умови 6 теореми.

Із цієї збіжності з урахуванням (12) отримуємо збіжність (8).

Збіжність (9) можна довести так само, якщо у попередніх міркуваннях покласти $Y_j = 1$.

Із (8) і (9) випливає твердження теореми.

Теорему доведено.

Доведення асимптотичної нормальності спирається на центральну граничну теорему у схемі серій. Ми скористаємось тим її варіантом, який запропоновано у роботі [2, теорема 5, п. 4, гл. 8].

Твердження 5.1. *Нехай $\eta_{j;n}$, $j = 1,\dots,n$, $n = 1,2,\dots$ — послідовність серій випадкових величин, таких, що:*

1. $\eta_{j;n}$, $j = 1,\dots,n$ незалежні в сукупності при кожному фіксованому n.

2. $\mathsf{E}\,\eta_{j,n} = 0$ для всіх $j = 1,\dots,n$, $n = 1,2,\dots$

3. $B_n = \sum_{j=1}^{n}\operatorname{Var}\eta_{j;n} \to B_\infty$, $n\to\infty$; $0 < B_\infty < \infty$.

4. При деякому $s > 2$

$$D_2(n) \stackrel{\text{def}}{=} \sum_{j=1}^{n}\mathsf{E}\min((\eta_{j;n})^2, |\eta_{j;n}|^s) \to 0, \text{ при } n\to\infty. \tag{14}$$

Тоді

$$\sum_{j=1}^{n}\eta_{j;n} \xrightarrow{\text{W}} \mathcal{N}(0, B_\infty).$$

Доведення теореми 3.2. Позначимо $\beta_n = n^{2/5}$,

$$Z_n = \frac{\beta_n}{nh}\sum_{j=1}^{n} a_{j;n}^k\Big(Y_j - g^{(k)}(x_0)\Big)K\left(\frac{x_0 - X_j}{h}\right).$$

Тоді

$$\beta_n\Big(\hat{g}_n^{(k)}(x_0) - g^{(k)}(x_0)\Big) = \frac{Z_n}{\hat{f}_n^{(k)}(x_0)}, \tag{15}$$

де $\hat{f}_n^{(k)}(x_0)$ визначено (6).

Ми покажемо, що

$$Z_n \xrightarrow{\mathrm{W}} \mathcal{N}\Big(\mu^{(k)}(x_0)f^{(k)}(x_0), S_{(k)}^2(x_0)\Big) \text{ при } n\to\infty. \tag{16}$$

Звідси, з урахуванням (9), випливатиме твердження теореми.

Для доведення (16) ми покажемо, що

$$Z_n - \mathsf{E}\, Z_n \xrightarrow{\mathrm{W}} \mathcal{N}\Big(0, S_{(k)}^2(x_0)\Big) \text{ при } n\to\infty \tag{17}$$

і

$$\mathsf{E}\, Z_n \to \mu^{(k)}(x_0)f^{(k)}(x_0). \tag{18}$$

Почнемо з доведення (18).

Так само, як у (10)–(11), отримуємо

$$\begin{aligned}\mathsf{E}\, Z_n &= \frac{\beta_n}{h}\,\mathsf{E}\Big(g^{(k)}(X^{(k)}) - g^{(k)}(x_0)\Big)K\left(\frac{x_0 - X^{(k)}}{h}\right) = \\ &= \beta_n\int_{-\infty}^{\infty}\Big(g^{(k)}(x_0 - zh) - g^{(k)}(x_0)\Big)f^{(k)}(x_0 - zh)K(z)dz.\end{aligned} \tag{19}$$

Запишемо розвинення Тейлора для функцій $g^{(k)}$ і $f^{(k)}$ в околі точки x_0:

$$\begin{aligned}g^{(k)}(x_0 - zh) - g^{(k)}(x_0) &= -zh\dot{g}^{(k)}(x_0) + \frac{(zh)^2}{2}\ddot{g}^{(k)}(x^*),\\ f^{(k)}(x_0 - zh) &= f^{(k)}(x_0) - zh\dot{f}^{(k)}(x^{**}),\end{aligned}$$

де x^*,x^{**} — проміжні точки між x_0 і $x_0 - hz$.

Підставивши ці розвинення у (19), з урахуванням умов 1, 2 і 6 теореми, отримуємо

$$\mathsf{E}\, Z_n = \beta_n\left(\frac{h^2}{2}\Big(2\dot{g}^{(k)}(x_0)\dot{f}^{(k)}(x_0) + \ddot{g}^{(k)}(x_0)f^{(k)}(x_0)\Big)D + o(h^2)\right).$$

Звідси, враховуючи що $\beta_n = n^{2/5}$, а $h = Hn^{-1/5}$, отримуємо

$$\mathsf{E}\, Z_n \to \frac{H^2}{2}D\Big(2\dot{g}^{(k)}(x_0)\dot{f}^{(k)}(x_0) + \ddot{g}^{(k)}(x_0)f^{(k)}(x_0)\Big) \text{ при } n\to\infty,$$

тобто (18).

Перейдемо до доведення (17). Для цього запишемо

$$Z_n - \mathsf{E}\, Z_n = \sum_{j=1}^{n}\eta_{j;n},$$

де $\eta_{j;n} = (\tilde{\eta}_{j;n} - \mathsf{E}\,\tilde{\eta}_{j;n})$,

$$\tilde{\eta}_{j;n} = \frac{\beta_n}{nh}a_{j;n}^k\Big(Y_j - g^{(k)}(x_0)\Big)K\left(\frac{x_0 - X_j}{h}\right), \tag{20}$$

і застосуємо твердження 5.1. Перша і друга умови твердження, вочевидь, виконані. Перевіримо третю і знайдемо B_∞.

$$\operatorname{Var}\tilde{\eta}_{j;n} = \frac{(a^k_{j;n})^2\beta_n^2}{(nh)^2}(S_{j,1} - S_{j,2}),$$

де

$$S_{j,1} \stackrel{\text{def}}{=} \mathsf{E}(\tilde{\eta}_{j;n})^2 = \sum_{m=1}^{M} p^m_{j;n}\,\mathsf{E}\left[\left(\varepsilon^{(m)} + g^{(m)}(X^{(m)}) - g^{(k)}(x_0)\right)K\left(\frac{x_0 - X^{(m)}}{h}\right)\right]^2,$$

$$S_{j,2} \stackrel{\text{def}}{=} (\mathsf{E}\,\tilde{\eta}_{j;n})^2 = \left[\sum_{m=1}^{M} p^m_{j;n}\,\mathsf{E}\left(g^{(m)}(X^{(m)}) - g^{(k)}(x_0)\right)K\left(\frac{x_0 - X^{(m)}}{h}\right)\right]^2.$$

Так само, як при доведенні теореми 3.1, отримуємо

$$\sum_{j=1}^{n}(a^k_{j;n})^2 S_{j,1} = nh\sum_{m=1}^{M}\left\langle\left(\mathbf{a}^{(k)}\right)^2\mathbf{p}^m\right\rangle_n \times$$
$$\times\int_{-\infty}^{\infty}\left[\sigma^2_{(m)} + \left(g^{(m)}(x_0 - zh) - g^{(k)}(x_0)\right)^2\right](K(z))^2 f^{(k)}(x_0 - zh)dz$$

і

$$\sum_{j=1}^{n}(a^k_{j;n})^2 S_{j,2} = nh^2\sum_{m_1,m_2=1}^{M}\left\langle\left(\mathbf{a}^{(k)}\right)^2\mathbf{p}^{m_1}\mathbf{p}^{m_2}\right\rangle_n Q_n(m_1)Q_n(m_2),$$

де

$$Q_n(m) = \int_{-\infty}^{\infty}\left(g^{(m)}(x_0 - zh) - g^{(k)}(x_0)\right)K(z)f^{(m)}(x_0 - zh)dz.$$

Внаслідок умов 5–7 теореми, $Q_n(m)$ — обмежені, а $\beta_n^2/nh \to 1/H$. Тому

$$\frac{\beta_n^2}{(nh)^2}\sum_{j=1}^{n}\left(a^k_{j;n}\right)^2 S_{j,2} \to 0, \text{ при } n\to\infty,$$

і

$$B_\infty = \lim_{n\to\infty}\operatorname{Var} Z_n = \lim_{n\to\infty}\sum_{j=1}^{n}\left(a^k_{j;n}\right)^2 S_{j,1} = S^2_{(k)}(x_0). \tag{21}$$

Третя умова твердження 5.1 виконується.

Перевіримо четверту умову. Позначимо

$$\eta^{(m)}_n = \frac{\beta_n}{nh}\left(Y^{(m)} - g^{(k)}(x_0)\right)K\left(\frac{x_0 - X^{(m)}}{h}\right).$$

Тоді

$$D_2(n) = \sum_{j=1}^{n}\mathsf{E}\min\left((\eta_{j;n})^2, |\eta_{j;n}|^s\right) \le$$
$$\le \max\left(1, \sup_{j=1,\dots,n}|a^{(k)}_{j;n}|^s\right)\sum_{m=1}^{M}\mathsf{E}\min\left(\left(\eta^{(m)}_n - \mathsf{E}\,\eta^{(m)}_n\right)^2, \left(\eta^{(m)}_n - \mathsf{E}\,\eta^{(m)}_n\right)^s\right).$$

Внаслідок умови 3 теореми, $\max(1, \sup_{j=1,\dots,n}|a^{(k)}_{j;n}|^s) < C < \infty$ для всіх n.

Далі, оскільки для будь-якого $s > 1$ і довільних $a, b \in \mathbb{R}$,

$$|a+b|^s \le 2^{s-1}(|a|^s + |b|^s),$$

то

$$\min\Big(\big(\eta_n^{(m)}-\mathsf{E}\,\eta_n^{(m)}\big)^2,\big(\eta_n^{(m)}-\mathsf{E}\,\eta_n^{(m)}\big)^s\Big)\le$$
$$\le 2^{s-1}\Big[\min((\eta_n^{(m)})^2,(\eta_n^{(m)})^s)+\max((\mathsf{E}\,\eta_n^{(m)})^2,(\mathsf{E}\,\eta_n^{(m)})^s)\Big].$$

Так само, як у доведенні теореми 3.1, отримуємо, що $n\max\big(\big(\mathsf{E}\,\eta_n^{(m)}\big)^2,\big(\mathsf{E}\,\eta_n^{(m)}\big)^s\big)\to 0$ при $n\to\infty$.

Отже, для доведення (14) досить показати, що

$$n\,\mathsf{E}\min\Big(\big(\eta_n^{(m)}\big)^2,\big(\eta_n^{(m)}\big)^s\Big)\to 0,\ \text{при } n\to\infty \tag{22}$$

при деякому $s>2$ для всіх $m=1,\dots,M$.

Помітимо, що для будь-якого $\tau>0$,

$$n\,\mathsf{E}\min\Big(\big(\eta_n^{(m)}\big)^2,\big(\eta_n^{(m)}\big)^s\Big)=n\,\mathsf{E}\big(\eta_n^{(m)}\big)^2\mathbf{1}\Big\{\big|\eta_n^{(m)}\big|>\tau\Big\}+n\,\mathsf{E}\big(\eta_n^{(m)}\big)^s\mathbf{1}\Big\{\big|\eta_n^{(m)}\big|<\tau\Big\}\le$$
$$\le n\,\mathsf{E}\big(\eta_n^{(m)}\big)^2\mathbf{1}\Big\{\big|\eta_n^{(m)}\big|>\tau\Big\}+n\tau^{s-2}\,\mathsf{E}\big(\eta_n^{(m)}\big)^2.$$

Так само, як при оцінці J_m у доведенні теореми 3.1, можна показати, що

$$\sup_n n\,\mathsf{E}\big(\eta_n^{(m)}\big)^2<\infty.$$

Тому для доведення (22) досить переконатись, що для всіх $\tau>0$,

$$n\,\mathsf{E}\big(\eta_n^{(m)}\big)^2\mathbf{1}\Big\{\big|\eta_n^{(m)}\big|>\tau\Big\}\to 0 \text{ при } n\to\infty. \tag{23}$$

Позначимо

$$V_h(z,u)=\Big(g^{(m)}(x_0-zh)+u-g^{(k)}(x_0)\Big)K(z),$$

$F_\varepsilon^{(m)}(u)$ — розподіл $\varepsilon^{(m)}$.

Записавши математичне сподівання у вигляді інтеграла та використавши заміну змінних, аналогічну зробленій в (11), отримуємо

$$n\,\mathsf{E}\big(\eta_n^{(m)}\big)^2\mathbf{1}\Big\{\big|\eta_n^{(m)}\big|>\tau\Big\}=\frac{(\beta_n)^2}{nh}J_n,$$

де

$$J_n=\int_{-\infty}^{\infty}\int_{-c}^{c}(V_h(z,u))^2\mathbf{1}\Big\{|V_h(z,u)|>\tau\frac{nh}{\beta_n}\Big\}f^{(m)}(x_0-zh)dzF_\varepsilon^{(m)}(du). \tag{24}$$

Помітимо, що $(\beta_n)^2/(nh)\to 1/H$ при $n\to\infty$.

Покажемо, що $J_n\to 0$. Для цього скористаємось теоремою Лебега про мажоровану збіжність.

Дійсно, при $n\to\infty$, $\frac{nh}{\beta_n}\to\infty$, отже,

$$(V_h(z,u))^2\mathbf{1}\Big\{|V_h(z,u)|>\tau\frac{nh}{\beta_n}\Big\}f^{(m)}(x_0-zh)\to 0,\ \text{при } n\to\infty,$$

для всіх можливих z і u. Крім того, унаслідок умов 2, 5 і 6 теореми, ця функція, при достатньо малих h мажорується функцією $C_1u^2+C_2$, де C_1 і C_2 — деякі скінченні константи. Унаслідок скінченності другого моменту $\varepsilon^{(m)}$,

$$\int_{-\infty}^{\infty}\int_{-c}^{c}\big(C_1u^2+C_2\big)dzF_\varepsilon^{(m)}(du)<\infty,$$

отже, за теоремою Лебега, отримуємо $J_n\to 0$. Звідси випливає (23), а отже, і виконання четвертої умови твердження 5.1.

Застосовуючи це твердження, отримуємо (17), що, разом із (18) дає твердження теореми.

Теорему доведено. □

6. Дослідження точності оцінок на модельованих даних

Поведінка оцінок для вибірок скінченного обсягу була досліджена у п'яти імітаційних експериментах на модельованих даних.

У всіх експериментах розглядалась двокомпонентна суміш із концентраціями, визначеними як

$$p^1_{j:n} = \frac{j}{n}; \quad p^2_{j:n} = 1 - \frac{j}{n}; \quad j = 1, \dots, n.$$

Обсяг вибірки n у кожному експерименті змінювався від 100 до 10000, із генерацією $B = 1000$ вибірок для кожного значення розміру. За отриманими вибірками з оцінок $\hat{g}^{(m)}_n(x_0)$ розраховувались вибіркові середні $\mathsf{E}_* \hat{g}^{(m)}_n(x_0)$ і вибіркові стандартні відхилення $\sqrt{\mathrm{Var}_* \hat{g}^{(m)}_n(x_0)}$.

Далі для кожного експерименту наведено таблиці значень нормованих зміщень

$$\mathrm{Bias} = n^{2/5}\Big(\mathsf{E}_* \hat{g}^{(m)}_n(x_0) - g^{(m)}(x_0)\Big)$$

і нормованих середньоквадратичних відхилень:

$$\mathrm{STD} = n^{2/5}\sqrt{\mathrm{Var}_* \hat{g}^{(m)}_n(x_0)},$$

окремо для кожного компонента $m = 1, 2$.

Теоретичним граничним значенням нормованих зміщення та стандартного відхилення відповідає рядок таблиці, що позначено символом ∞.

Регресор X для обох компонентів у всіх експериментах моделювався з рівномірним розподілом на $[0, 1]$.

В усіх експериментах у оцінках використано ядро Єпанєчнікова. В експериментах 1–3 застосовано теоретично оптимальне значення параметра $H = H_{\mathrm{opt}}$, визначене у п. 4.

Експеримент 1. Функції регресії були задані як

$$g^{(m)}(x) = (-1)^m x(1 - x), \ m = 1, 2. \tag{25}$$

Оцінювання проводилися у точці $x_0 = 0{,}5$. Розподіл похибок ε_j був обраний $\mathcal{N}(0; 0{,}0025)$ для обох компонентів. Теоретично оптимальне значення параметра згладжування $H = H_{\mathrm{opt}} = 0{,}995$ для обох компонентів. Результати обчислень, наведено в табл. 1.

Таблиця 1. Результати моделювання експерименту 1

	Bias		STD	
n	$\kappa = 1$	$\kappa = 2$	$\kappa = 1$	$\kappa = 2$
100	0,182	−0,174	0,307	0,302
500	0,193	−0,205	0,330	0,336
1000	0,195	−0,163	0,339	0,340
2500	0,183	−0,191	0,361	0,358
5000	0,195	−0,190	0,343	0,362
10000	0,186	−0,194	0,366	0,356
∞	0,198	−0,198	0,396	0,396

Результати моделювання в цьому випадку добре узгоджуються зі значеннями, отриманими за асимптотичними формулами при достатньо великих обсягах вибірки.

Експеримент 2. У цьому експерименті для дослідження впливу важких хвостів розподілу похибок, ε_j моделювались із розподілом Стьюдента з 10-ма ступенями

вільності. Теоретично оптимальне значення $H = H_{\text{opt}} = 1{,}816$ для двох компонентів. Результати представлені в табл. 2.

Функції регресії визначались за (25).

Таблиця 2. Результати моделювання експерименту 2

	Bias		STD	
n	$\kappa = 1$	$\kappa = 2$	$\kappa = 1$	$\kappa = 2$
100	0,463	−0,477	1,418	1,432
500	0,618	−0,566	1,311	1,282
1000	0,715	−0,674	1,323	1,342
2500	0,626	−0,674	1,283	1,287
5000	0,673	−0,668	1,331	1,281
10000	0,746	−0,679	1,298	1,286
∞	0,660	-0,660	1,319	1,319

Порівняно з результатами першого експерименту, дисперсія оцінок зросла, а точність наближення за асимптотичними формулами трохи знизилась.

Експеримент 3. У цьому експерименті похибки мають розподіл $\mathcal{N}(0; 1{,}25)$ для обох компонентів. Отже, їх дисперсії та теоретично оптимальні параметри H для двох компонентів такі самі, як у експерименті 2. Результати моделювання, наведено в табл. 3.

Таблиця 3. Результати моделювання експерименту 3

	Bias		STD	
n	$\kappa = 1$	$\kappa = 2$	$\kappa = 1$	$\kappa = 2$
100	0,436	−0,433	1,418	1,520
500	0,639	−0,652	1,293	1,286
1000	0,660	−0,636	1,364	1,310
2500	0,688	−0,651	1,319	1,276
5000	0,617	−0,662	1,256	1,294
10000	0,624	−0,684	1,318	1,288
∞	0,660	−0,660	1,319	1,319

З отриманих результатів видно, що точність наближення асимптотичними формулами дещо покращилася порівняно з експериментом із важкими хвостами похибок регресій.

Експеримент 4. У цьому і наступному експериментах використовуються функції регресії вигляду:

$$g^{(1)}(x) = \begin{cases} -3x^2 + 3x, & \text{при } x < 0{,}7, \\ 3x^2 - 3x, & \text{при } x \geq 0{,}7, \end{cases}$$

$$g^{(2)}(x) = -x + 1,$$

тобто функція регресії першого компонента мала розрив у точці $x = 0{,}7$.

Похибки регресії ϵ_j мали розподіл $\mathcal{N}(0; 0{,}0025)$.

Експеримент 4 полягає в тому, щоб проаналізувати асимптотичну поведінку в точці $x_0 = 0{,}5$ (у точці неперервності обох компонентів). Параметр згладжування емпірично було вибрано $H = 1$ для обох компонентів. Результати наведені в табл. 4.

Точність наближення асимптотичними формулами тепер менша, ніж у попередніх експериментах, але достатня для характеризації якості оцінок при великих обсягах даних.

Таблиця 4. Результати моделювання експерименту 4

	Bias		STD	
n	$\kappa = 1$	$\kappa = 2$	$\kappa = 1$	$\kappa = 2$
100	−1,671	−0,014	0,558	0,376
500	−1,486	−0,009	0,353	0,271
1000	−1,144	−0,017	0,241	0,230
2500	−0,632	−0,003	0,186	0,206
5000	−0,601	−0,008	0,189	0,213
10000	−0,601	0,003	0,198	0,208
infinity	−0,600	0,000	0,209	0,209

Експеримент 5. Тепер для моделі експерименту 4, проведемо оцінювання функції регресії другого компонента (неперервної) в $x_0 = 0{,}7$ (у точці розриву першого компонента) із параметром $H = 1$. Результати вміщені в табл. 5.

Таблиця 5. Результати моделювання експерименту 5

	Bias	STD
n	$\kappa = 1$	$\kappa = 2$
100	0,119	0,564
500	0,010	0,491
1000	0,022	0,517
2500	0,015	0,535
5000	−0,002	0,529
10000	−0,039	0,532
∞	0,000	0,725

Тепер асимптотичні формули дають лише порядок величини STD навіть для $n = 10000$. Отже, для випадку можливих розривів у функціях регресії деяких компонентів, асимптотичними наближеннями слід користуватись з обережністю.

7. Висновки

Таким чином, ми розглянули модифікацію оцінок Надарая–Ватсона для оцінювання функцій регресії компонентів суміші зі змінними концентраціями, отримали умови їх консистентності й асимптотичної нормальності. Асимптотичні результати дозволили визначити оптимальне ядро для оцінки та вказати теоретично оптимальне значення параметра згладжування. Результати моделювання підтверджують можливості використання асимптотичних формул для наближеного опису поведінки оцінок на даних скінченного обсягу.

References

1. T. Benaglia, D. Chauveau, D. Hunter, D. Young, `mixtools`: *An R Package for Analyzing Finite Mixture Models*, Journal of Statistical Software, **32** (2009), no. 6, 1–29.
2. A. A. Borovkov, *Probability theory*, Springer, New York, 2013.
3. S. Faria, G. Soromenhob, *Fitting mixtures of linear regressions*, Journal of Statistical Computation and Simulation, **80** (2010), no. 2, 201–225.
4. B. Grün, F. Leisch, *Fitting finite mixtures of linear regression models with varying & fixed effects in R*, Compstat 2006 — Proceedings in Computational Statistics (A. Rizzi, M. Vichi), Physica Verlag, Heidelberg, 2006, 853–860.
5. W. Hardle, M. Muller, S. Sperlich, A. Werwatz, *Nonparametric and Semiparametric Models*, Springer-Verlag, Berlin, 2004.

6. D. Liubashenko, R. Maiboroda, *Linear regression by observations from mixtures with varying concentrations*, Modern Stochastics: Theory and Applications, **2** (2015), 343–353.
7. P. Macdonald, J. Du, *mixdist: Finite Mixture Distribution Models. R package version 0.5-2.* Online publication, 2012. — http://www.math.mcmaster.ca/peter/mix/mix.html
8. R. Maiboroda, V. Miroshnichenko, *Confidence ellipsoids for regression coefficients by observations from a mixture*, Modern Stochastics: Theory and Applications, **5** (2018), no. 2, 225–245.
9. R. Maiboroda, O. Sugakova, *Estimation and classification by observations form mixture*, Kyiv University Publishers, Kyiv, 2008. (Ukrainian)
10. R. Maiboroda, O. Sugakova, *Statistics of mixtures with varying concentrations with application to DNA microarray data analysis*, Journal of Nonparametric Statistics, **24** (2012), no. 1, 201–205.
11. G. J. McLachlan, D. Peel, *Finite mixture models*, Wiley-Interscience, New York, 2000.
12. E. A. Nadaraya, *On Estimating Regression*, Theory of Probability and its Applications, **9** (1964), no. 1, 141–142.
13. P. Schlattmann, *Medical Applications of Finite Mixture Models*, Springer-Verlag, New York, 2009.
14. O. V. Sugakova, *Asymptotics of a kernel estimate for the density of a distribution constructed from observations of a mixture with varying concentration*, Theory Probab. Math. Statist., **59** (1999), 161–171.
15. G. S. Watson, *Smooth regression analysis*, Sankhyā: The Indian Journal of Statistics, Series A, **26** (1964), no. 4, 359–372.

Кафедра теорії ймовірностей, статистики та актуарної математики, механіко-математичний факультет, Київський національний університет імені Тараса Шевченка, проспект Глушкова, 6, Київ, Україна, 03127
Адреса електронної пошти: galia.dychko@gmail.com

Кафедра теорії ймовірностей, статистики та актуарної математики, механіко-математичний факультет, Київський національний університет імені Тараса Шевченка, проспект Глушкова, 6, Київ, Україна, 03127
Адреса електронної пошти: mre@univ.kiev.ua

Стаття надійшла до редколегії 13.12.2018

GENERALIZED NADARAYA–WATSON ESTIMATOR FOR OBSERVATIONS FROM MIXTURE

H. M. DYCHKO, R. E. MAIBORODA

Abstract. A generalization of Nadaraya–Watson kernel regression estimators is considered for estimation by observations from a mixture with varying concentrations. Consistency and asymptotic normality of the estimators are shown.

ОБОБЩЕННЫЕ ОЦЕНКИ НАДАРАЯ – ВАТСОНА ДЛЯ НАБЛЮДЕНИЙ ИЗ СМЕСИ

Г. М. ДЫЧКО, Р. Е. МАЙБОРОДА

Аннотация. Рассматривается обобщение оценок Надарая – Ватсона по наблюдениям из смеси с переменными концентрациями. Доказаны состоятельность и асимптотическая нормальность полученных оценок.