УДК 519.21 https://doi.org/10.17721/1812-5409.2023/1.5

Д. Д. Горбунов[1], *студ. магістратури*
Р. Є. Майборода[2], *д-р. наук, проф.*

Крос-валідація у локально-лінійній регресії для спостережень з суміші

[1]Кафедра теорії ймовірностей, статистики та актуарної математики, Київський національний університет імені Тараса Шевченка, 64/13, вул.Володимирська, 01601 Київ, Україна
e-mail: danielhorbunov@knu.ua
[2]Кафедра теорії ймовірностей, статистики та актуарної математики, Київський національний університет імені Тараса Шевченка, 64/13, вул.Володимирська, 01601 Київ, Україна
e-mail: rostmaiboroda@gmail.com

D. D. Horbunov[1], *M. Sc. Student*
R. E. Maïboroda[2], *Dr of Sci., Prof.*

Cross-validation for local-linear regression by observations from mixture

[1]Department of Probability Theory, Statistics and Actuarial Mathematics, Taras Shevchenko National University of Kyiv, 64/13,Volodymyrska Street, 01601 Kyiv, Ukraine
e-mail: danielhorbunov@knu.ua
[2]Department of Probability Theory, Statistics and Actuarial Mathematics, Taras Shevchenko National University of Kyiv, 64/13,Volodymyrska Street, 01601 Kyiv, Ukraine
e-mail: rostmaiboroda@gmail.com

Запропоновано модифікацію оцінок локально-лінійної регресії для оцінювання невідомих функцій регресії компонент суміші зі змінними концентраціями. Досліджуються можливості техніки крос-валідації для вибору параметра згладжування оцінки. Якість отриманих оцінок порівнюється у імітаційних експериментах з якістю модифікованих оцінок Надарая-Ватсона.

Ключові слова: Модель суміші зі змінними концентраціями, непараметрична регресія, техніка крос-валідації, локально-лінійна регресія.

We consider a generalization of local-linear regression for estimation of components' regression functions by observations from mixture with varying concentrations. A cross-validation technique is developed for the bandwidth selection, based on minimization of conditional integrated mean squared error. For the simulations, two approaches on bandwidth selection are used: the naive method, based on the optimal choice for the modified Nadaraya-Watson estimator, and cross-validation technique, mentioned earlier. Performance of the obtained estimator is compared with the modified Nadaraya-Watson estimator performance by simulations. Simulations show that the modified local-linear estimator overcomes boundary effect inherent in the Nadaraya-Watson estimator and its modification for the mixture with varying concentrations.

Key Words: Mixture with varying concentrations, nonparametric regression, cross-validation technique, local-linear regression.

Communicated by Rozora I.V.

1 Вступ

Моделі сумішей зі змінними концентраціями природно виникають в задачах аналізу даних медико-біологічних та соціологічних досліджень [1]. Про можливості застосування цієї моделі у аналізі нейробіологічних даних див. [2].

Для багатовимірних даних залежність між змінними для різних компонент суміші часто буває природно описувати різними регресійними моделями. У роботі [3], запропоновано модифікацію класичної непараметричної оцінки Надарая-Ватсона [4],[5] на випадок спостережень з суміші. Але ці оцінки, як і оцінки Надарая-Ватсона для однорідних вибірок, виявляють так званий крайовий ефект: надзвичайно велике зміщення на кінцях інтервалу зміни незалежної змінної. Одним з можливих способів усунення цього ефекту є використання техніки локально-лінійної регресії [6].

У даній роботі запропонована модифікація локально лінійного оцінювання, яка дозволяє будувати оцінки функції регресії для кожної компоненти суміші окремо. При цьому виникає проблема вибору параметра згладжування

оцінки. Для її розв'язання пропонується модифікація техніки крос-валідації. Якість роботи отриманої оцінки при фікованих значеннях залежної змінної порівнюється із якістю оцінок Надарая-Ватсона у імітаційних експериментах.

2 Постановка задачі

Розглянемо дані, що описуються моделлю суміші зі змінними концентраціями. Кожен спостережуваний об'єкт належить до однієї з M популяцій (компонент суміші). Розподіл спостережуваної характеристики $\xi_j = (X_j, Y_j)$ j-го об'єкта визначається непараметричною регресійною моделлю:

$$Y_j = g^{(\kappa_j)}(X_j) + \varepsilon_j,\ j = \overline{1,n}, \qquad (1)$$

де κ_j є номером компоненти, до якої належить j-ий об'єкт, $g^{(m)}$ є невідомою функцією регресії для m-ої компоненти, ε_j є центрованою похибкою з скінченною дисперсією $\sigma^2_{(m)}$ для m-ої компоненти. Справжні значення номерів κ_j є невідомими, але відомі ймовірності змішування (концентрації) $p^{(m)}_{j:n} = \mathbf{P}\{\kappa_j = m\}$, $j = \overline{1,n}$, $m = \overline{1,M}$. Розподіли регресорів X_j є абсолютно неперервними відносно міри Лебега, при цьому існують щільності розподілу регресора для всіх компонент суміші $f^{(m)}$, $m = 1,\dots,M$. Ці щільності є невідомими. Випадкові величини X_j, ε_j $j = 1,\dots,n$ вважаються незалежими в сукупності при фіксованій послідовності κ_j.

Потрібно оцінити невідомі функції регресії $g^{(m)}$ для кожної компоненти суміші, $m = \overline{1,M}$.

3 Побудова оцінки

Нагадаємо схему побудови локально лінійних оцінок у випадку однорідної вибірки, тобто коли у моделі (1) є лише одна компонента ($M = 1$) і функція регресії $g^{(m)}(x) = g(x)$ є спільною для всіх спостережень. Для того, щоб оцінити $g(x)$ в фіксованій точці $x = x_0$, розглядають лінійне наближення $g(x) \approx a + b(x - x_0)$ і підганяють невідомі коефіцієнти a і b, мінімізуючи локальний функціонал методу найменших квадратів (2.1) [6]:

$$J(a,b) = \sum_{j=1}^{n} K\left(\frac{x_0 - X_j}{h}\right)(Y_j - a - b(X_j - x_0))^2,$$

де K — ядро (інтегровна функція з невід'ємними значеннями), $h > 0$ — параметр згладжування. Чим меншим обрано h, тим ближче до x_0 повинно потрапити X_j для того, щоб вплив j-того спостереження на підігнані значення був помітним. Локально-лінійною оцінкою $\hat{g}(x_0)$ для $g(x_0)$ є координата $\hat{a}$ точки мінімуму $(\hat{a}, \hat{b})$ функціоналу J. Формули (2.1)-(2.4) [6] дають явний вигляд цієї оцінки.

Для узагальнення цих формул ми введемо додаткове навантаження мінімаксними ваговими коефіцієнтами щоб виділити певну компоненту. Ці коефіцієнти описані у [1]. У [3] вони використані для модифікації оцінок Надарая-Ватсона.

Введемо спочатку деякі позначення. Для масивів концентрацій $\mathbf{p} = (p^m_{j:n},\ j = \overline{1,n},\ m = \overline{1,M},\ n \geqslant 1)$ та вагових коефіцієнтів $\mathbf{a} = (a^m_{j:n},\ j = \overline{1,n},\ m = \overline{1,M},\ n \geqslant 1)$ будемо позначати $p^m = (p^m_{1:n},\dots,p^m_{n:n})^T$ вектор-стовпець концентрацій m-ої компоненти для всіх спостережень і, аналогічно, $a^m = (a^m_{1:n},\dots,a^m_{n:n})^T$.

Операцію усереднення по всій вибірці позначимо кутовими дужками:

$$\langle p^m \rangle_n = \frac{1}{n}\sum_{j=1}^{n} p^m_{j:n}.$$

Всі арифметичні операції над векторами, які вказані в кутових дужках, визначаються поелементно. Будемо позначати $\langle p^m \rangle = \lim\limits_{n\to+\infty} \langle p^m \rangle_n$, якщо така границя існує. Зауважимо, що операція усереднення $\langle a^k p^m \rangle_n$ задає скалярний добуток a^k та p^m у $\mathbb{R}^n$.

Розглянемо матрицю Грама на основі операції усереднення по добуткам покомпонентних векторів концентрацій: $\Gamma_n = (\langle p^k p^l \rangle_n)^M_{k,l=1}$. Надалі будемо припускати, що $\{p^m\}^M_{m=1}$ є лінійно незалежними, а тому $\det\Gamma_n \neq 0$. Вагові коефіцієнти $a^m_{j:n}$, визначені за правилом

$$a^m_{j:n} = \frac{1}{\det\Gamma_n}\sum_{m=1}^{M}(-1)^{m+k}\gamma_{km}p^m_{j:n},$$

де γ_{km} є k,m-им мінором матриці Γ_n, називають мінімаксними ваговими коефіцієнтами. З властивостями отриманих коефіцієнтів можна ознайомитися у [1].

Оцінка локально-лінійної регресії із наван-

таженням на m-ту компоненту має вигляд:

$$\hat{g}^{(m)}(x_0) = \frac{\sum_{j=1}^n \tilde{w}_{j:n}^m Y_j}{\sum_{j=1}^n \tilde{w}_{j:n}^m}, \quad (2)$$

де

$$\tilde{w}_{j:n}^m = (\hat{s}_2^m(x_0) - \hat{s}_1^m(x_0)(X_j - x_0))a_{j:n}^m w_{j:n}^{x_0}, \quad (3)$$

$$\hat{s}_l^m(x_0) = \sum_{j=1}^n a_{j:n}^m w_{j:n}^{x_0}(X_j - x_0)^l, \quad (4)$$

$l = 1, 2$, та $w_{j:n}^{x_0} = K((X_j - x_0)/h)$.

Якщо в формулах (2)-(4) покласти $a_j^m = 1$, отримуємо звичайні формули локально-лінійної оцінки для однорідної вибірки (2.1)-(2.4) [6]. Якщо у (2) вагові коефіцієнти $\tilde{w}_{j:n}^m$ з (3) замінити на $a_{j:n}^m w_{j:n}^{x_0}$, отримаємо модифіковані оцінки Надарая-Ватсона, що розглядались у [3].

4 Вибір параметра згладжування

Наївний вибір параметра згладжування. У роботі [3] показано, що узагальнена оцінка Надарая-Ватсона є конзистентною та асимптотично нормальною, коли параметр згладжування вибирається в залежності від обсягу вибірки як $h = h_n = Hn^{-1/5}$, при $n \to +\infty$. З точки зору мінімізації асимптотичної проінтегрованої середньоквадратичної похибки, теоретично оптимальним значенням H є H_{opt}, що визначається наступним чином:

$$H_{opt} = \left(\frac{d^2 L_2}{4D^2 L_1}\right)^{1/5}, \text{ де}$$

$$D = \int_{-\infty}^{+\infty} u^2 K(u)du,\ d^2 = \int_{-\infty}^{+\infty} (K(u))^2 du,$$

$$L_1 = \left(\frac{(g^{(m)})'(x_0)(f^{(m)})'(x_0)}{f^{(m)}(x_0)} + \frac{(g^{(m)})''(x_0)}{2}\right)^2,$$

$$L_2 = \frac{1}{(f^{(m)}(x_0))^2}\sum_{k=1}^M f^{(k)}(x_0)\langle(a^m)^2 p^k\rangle \Delta_{m,k},$$

$$\Delta_{m,k} = (\sigma_{(k)}^2 + (g^{(m)}(x_0) - g^{(k)}(x_0))^2).$$

Зрозуміло, що це значення не можна безпосередньо використовувати для оцінювання за реальними даними, оскільки для його обчислення потрібно знати невідомі параметри моделі. У даній роботі воно використовується для порівняння можливостей локально-лінійних оцінок та оцінок Надарая-Ватсона. Надалі будемо називати наївним підходом вибір для локально лінійних оцінок параметра згладжування $h = H_{opt}n^{-1/5}$, тобто теоретично-оптимальний для оцінок Надарая-Ватсона.

Вибір параметра згладжування на основі крос-валідації. Визначимо теоретичну проінтегровану середньоквадратичну похибку із навантаженням на m-ту компоненту як:

$$\mathrm{ISE}(h; m) = \int_{-\infty}^{+\infty} (\hat{g}_n^{(m)}(x) - g^{(m)}(x))^2 f^{(m)}(x)dx$$

Будемо шукати значення $h > 0$, яке забезпечує (наближену) мінімізацію $\mathrm{ISE}(h; m)$. Оскільки $g^{(m)}$ та $f^{(m)}$ невідомі, побудуємо оцінку для $\mathrm{ISE}(h; m)$. Запишемо $\mathrm{ISE}(h; m)$:

$$\mathrm{ISE}(h; m) = I_1^{(m)} - 2 \cdot I_2^{(m)} + I_3^{(m)},$$

$$I_1^{(m)} = \int_{-\infty}^{+\infty} (\hat{g}_n^{(m)}(x))^2 f^{(m)}(x)dx,$$

$$I_2^{(m)} = \int_{-\infty}^{+\infty} \hat{g}_n^{(m)}(x) g^{(m)}(x) f^{(m)}(x)dx,$$

$$I_3^{(m)} = \int_{-\infty}^{+\infty} (g^{(m)}(x))^2 f^{(m)}(x)dx.$$

Інтеграл $I_3^{(m)}$ не залежить від h, а тому не впливає на мінімізацію. Відкинувши його, маємо функціонал, еквівалентний до $\mathrm{ISE}(h; m)$:

$$\mathrm{CV}^*(h; m) = I_1 - 2 \cdot I_2$$

Оцінимо інтеграли в $\mathrm{CV}^*(h; m)$ наступним чином. Якщо ввести додаткове спостереження $(X^{(m)}, Y^{(m)})$, розподіл першої координати якого має щільність $f^{(m)}$, незалежне від вибірки $\mathbb{X} = \{(X_j, Y_j)\}_{j=1}^n$, то:

$$\mathbb{E}[(\hat{g}_n^{(m)}(X^{(m)}))^2 \mid \mathbb{X}] = I_1,$$
$$\mathbb{E}[g^{(m)}(X^{(m)})\hat{g}_n^{(m)}(X^{(m)}) \mid \mathbb{X}] = I_2$$

Тобто інтеграли $I_1^{(m)}$ та $I_2^{(m)}$ можна було б наблизити зваженими вибірковими середніми, якби у нас були додаткові спостереження незалежні від тих, по яких будувалась оцінка. В таких випадках для оцінювання використовують техніку крос-валідації.

Введемо наступні оцінки для інтегралів:

$$I_1^{(m)} \approx \hat{I}_1^{(m)} = \sum_{j=1}^{n} a_{j:n}^m (\hat{g}_{j-}^{(m)}(X_j))^2,$$

$$I_2^{(m)} \approx \hat{I}_2^{(m)} = \sum_{j=1}^{n} a_{j:n}^m g^{(m)}(X_j)\hat{g}_{j-}^{(m)}(X_j),$$

де $\hat{g}_{j-}^{(m)}$ є оцінкою для $g^{(m)}$ на основі всіх спостережень, окрім j-го. Проблема виникає в оцінюванні $g^{(m)}$ в $\hat{I}_2^{(m)}$. Якщо використовувати для оцінювання всю вибірку, то $g^{(m)}$ і $\hat{g}_{j-}^{(m)}(X_j)$ будуть залежними між собою. Щоб усунути цю залежність, розіб'ємо вибірку на дві половини з приблизно однаковою кількістю спостережень: перша частина вибірки іде на обчислення оцінок крос-валідації, а друга використовується для оцінювання функції регресії $g^{(m)}$.

Підставляючи замість інтегралів побудовані оцінки, отримуємо оцінку для $\mathrm{CV}^*(h;m)$:

$$\widehat{\mathrm{CV}}(h;m) = \sum_{j=1}^{[n/2]} a_{j:[n/2]}^m (\hat{g}_{j-}^{(m)}(X_j))^2 - \\ - 2 \cdot \sum_{j=1}^{[n/2]} a_{j:[n/2]}^m \hat{g}_{n-[n/2]}^{(m)}(X_j)\hat{g}_{j-}^{(m)}(X_j),$$

де $\hat{g}_{j-}^{(m)}(x)$ підраховується за першими $[n/2]$ спостереженнями з вибірки (за виключенням j-го), а $\hat{g}_{n-[n/2]}^{(m)}(x)$ на основі тих спостережень, що залишилися. Оцінка параметра згладжування на основі техніки крос-валідації визначається як точка мінімуму оціненого функціоналу:

$$h_{\widehat{\mathrm{CV}}}^{(m)} = \arg\min_{h>0} \widehat{\mathrm{CV}}(h;m)$$

5 Імітаційний експеримент

Проаналізуємо поведінку оцінки, порівнюючи запропоновані оцінки з узагальненою оцінкою Надарая-Ватсона з [3] для одного модельного розподілу даних. Імітаційні експерименти було проведено на вибірках обсягу $n = 100, 250, 500, 750, 1000$. Для кожного обсягу вибірки було згенеровано по $B = 1000$ незалежних копій вибірок, по яких будувались оцінки.

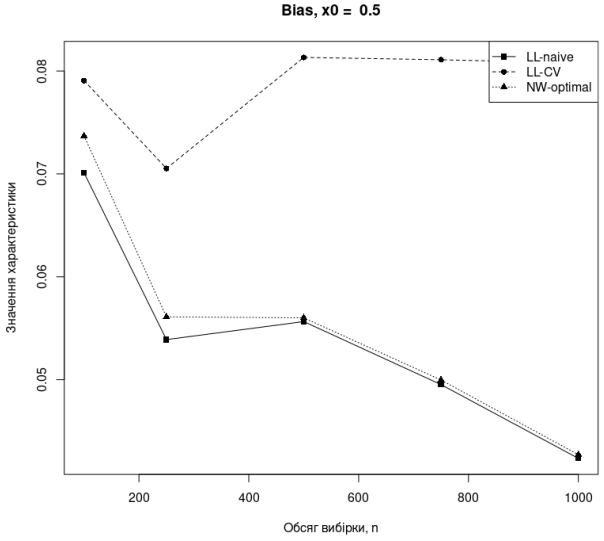
Розглядалась двокомпонентна суміш, тобто $M = 2$. Ймовірності змішування визначені як

$$\mathbf{P}\{\kappa_j = 1\} = \frac{j}{n},\ \mathbf{P}\{\kappa_j = 2\} = 1 - \frac{j}{n},\ j = \overline{1,n}.$$

Розподіл регресора є рівномірним на $[0,1]$. Розподіл похибок є центрованим гауссовим з дисперсією 1.25, тобто $\varepsilon_j \sim N(0, 1.25)$. Функції регресії визначені як

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

В якості ядра оцінки береться ядро Єпанєчнікова $K(x) = 3/4(1-x^2)\mathbf{1}_{|x|\leqslant 1}$. Якість поточкової збіжності оцінок характеризується зміщенням $\mathrm{Bias}(\hat{g}^{(m)}(x_0)) = \mathbb{E}[\hat{g}^{(m)}(x_0) - g^{(m)}(x_0)]$ та дисперсією $\mathbb{D}[\hat{g}^{(m)}(x_0)]$, котрі оцінюються за вибірковим середнім та вибірковою дисперсією по значеннях оцінок на модельованих вибірках. Зобразимо значення поточкових характеристик відносно збільшення обсягу вибірки:



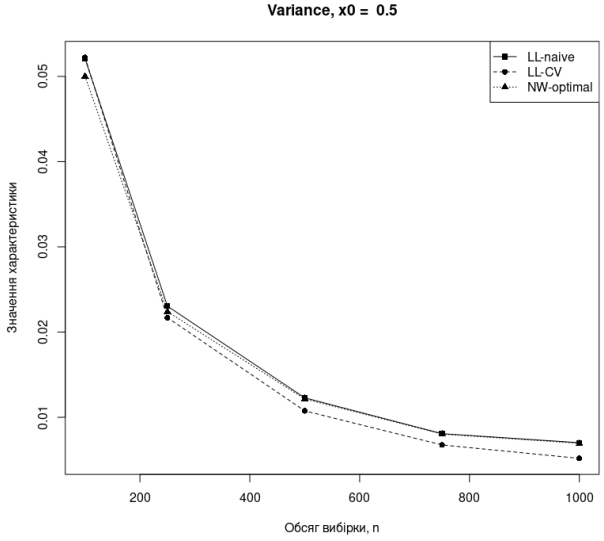
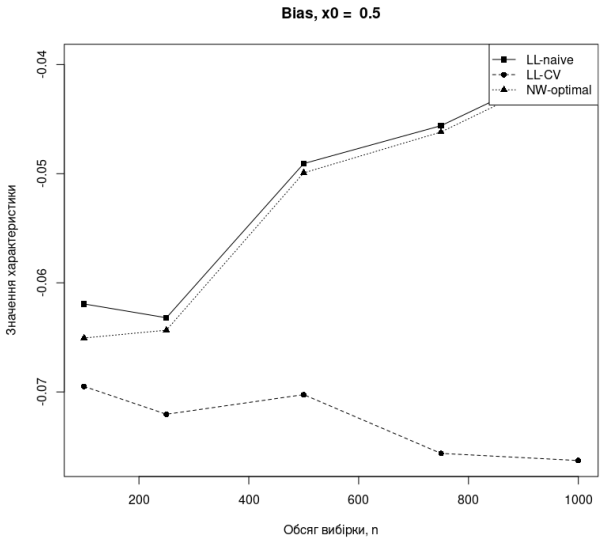
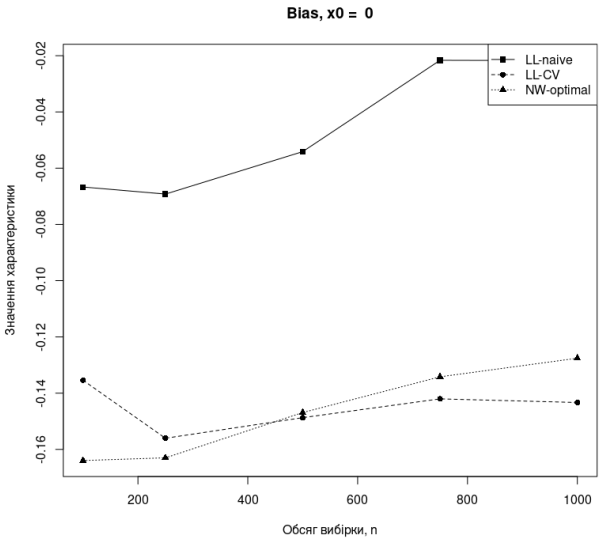


Рис. 1: Зміщення та дисперсія оцінок для $g^{(1)}(x_0)$, $x_0 = 0.5$.





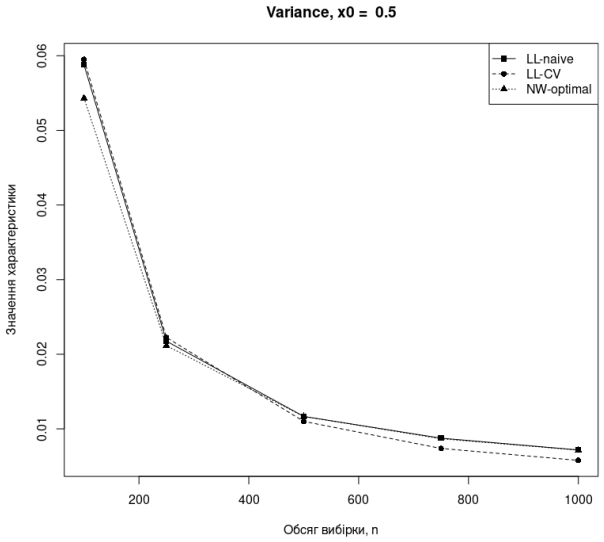


Рис. 2: Зміщення та дисперсія оцінок для $g^{(2)}(x_0)$, $x_0 = 0.5$.

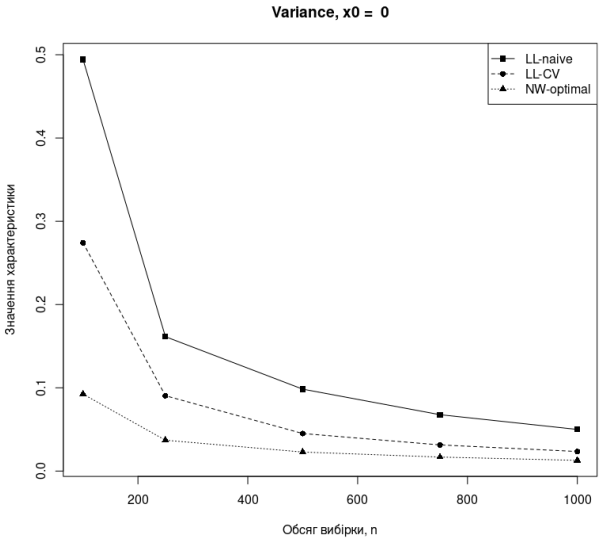


Рис. 3: Зміщення та дисперсія оцінок для $g^{(1)}(x_0)$, $x_0 = 0$.

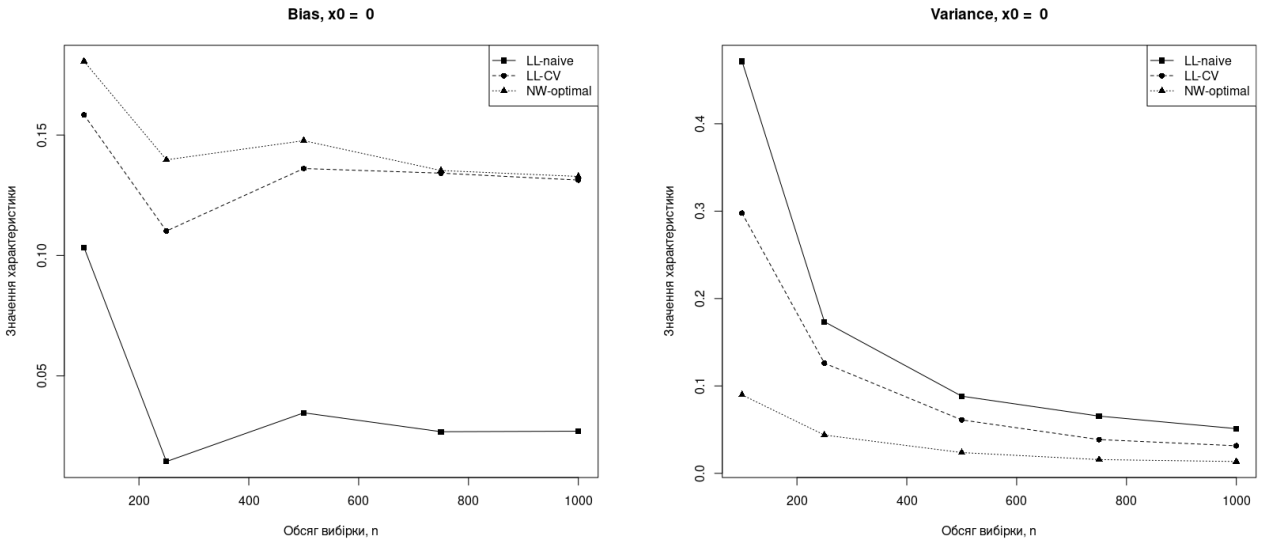


Рис. 4: Зміщення та дисперсія оцінок для $g^{(2)}(x_0)$, $x_0 = 0$.

6 Висновки

У проведених експериментах побудована модифікація оцінки локально-лінійної регресії для моделі суміші з декількома регресіями демонструє кращу поведінку в граничній точці носія регресора ($x_0 = 0$). З наведених вище рисунків це добре видно порівняно меншим зміщенням та дисперсією відносно оцінки Надарая-Ватсона. В інших імітаційних експериментах, які використовують інші розподіли регресорів та функції регресії, поведінка може бути гіршою.

Теоретичне дослідження асимптотичної поведінки запропонованих оцінок має бути предметом подальшої роботи.

Список використаних джерел

1. Майборода Р.Є., Сугакова О.В. "Оцінювання та класифікація за спостереженнями із суміші". К.: ВПЦ "Київський університет", 213 р. - 2008

2. A. Pidnebesna, I. Fajnerová, J. Horáček, J. Hlinka. Mixture Components Inference for Sparse Regression: Introduction and Application for Estimation of Neuronal Signal from fMRI BOLD. *Applied Mathematical Modelling*, Vol. 116 (2023), p. 735-748.

3. H. Dychko, R. Maiboroda. A generalized Nadaraya–Watson estimator for observations obtained from a mixture. *Theory of Probability and Mathematical Statistics*, Vol. 100 (2020), p. 61-76, DOI: 10.1090/tpms/1098.

4. E. A. Nadaraya. On Estimating Regression. *Theory of Probability and its Applications*, Vol. 9 (1964), No. 1, p. 141-142.

5. G. S. Watson. Smooth regression analysis. *Sankhya: The Indian Journal of Statistics*, Series A, Vol. 26 (1964), No. 4, p. 359-372.

6. J. Fan. Local Linear Regression Smoothers and their minimax efficiencies. *The Annals of Statistics*, Vol. 21 (1993), No. 1, p. 196-216.

References

1. MAIBORODA R., SUGAKOVA O. (2008) *Otsiniuvannia ta klasyfikatsiia za sposterezhenniamy iz sumishi.* Kyiv: *Kyivskyi universytet*, 213 p.

2. A. PIDNEBESNA, I. FAJNEROVÁ, J. HORÁČEK, J. HLINKA. (2023) *Mixture Components Inference for Sparse Regression: Introduction and Application for Estimation of Neuronal Signal from fMRI BOLD.* In Applied Mathematical Modelling, Vol. 116, p. 735-748.

3. DYCHKO H., MAIBORODA R. (2020) *A generalized Nadaraya–Watson estimator for observations obtained from a mixture.* In Theory of Probability and Mathematical Statistics, Vol. 100, p. 61-76, DOI: 10.1090/tpms/1098.

4. NADARAYA E. (1964) *On Estimating Regression.* In Theory of Probability and its Applications, Vol. 9, No. 1, p. 141-142.

5. WATSON G. (1964) *Smooth regression analysis.* In Sankhya: The Indian Journal of Statistics, Series A, Vol. 26, No. 4, p. 359-372.

6. FAN J. (1993) *Local Linear Regression Smoothers and their minimax efficiencies.* In The Annals of Statistics, Vol. 21, No. 1, p. 196-216.

Надійшла до редколегії 01.06.2023