UDC 519.22
DOI: https://doi.org/10.17721/1812-5409.2025/2.6

Rostyslav MAIBORODA, DSc (Phys. & Math.), Prof.
ORCID ID: 0000-0002-3899-558X
e-mail: rostmaiboroda@knu.ua
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

Olena SUGAKOVA, DSc (Phys. & Math.), Assoc. Prof.
ORCID ID: 0000-0002-6529-0788
e-mail: olenasugakova@knu.ua
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

# BANDWIDTH SELECTION FOR DENSITY ESTIMATION BY MIXTURE WITH VARYING CONCENTRATIONS

***Finite mixture models arise in statistics of biological and medical data when the investigated subjects belong to sub-populations with different distributions of observed variable. In the model of mixture with varying concentrations (MVC) the concentrations of the mixture components can vary from observation to observation. We consider estimation of a probability density for a mixture component in MVC by a modification of the kernel density estimator (KDE). To apply KDE one needs to select a tuning parameter named the bandwidth. Two approaches to the bandwidth selection are considered. The first one is a modification of the Silverman's rule of thumb. The second one is a version of the leave-one-out cross-validation algorithm. We present results of simulation which show that both algorithms demonstrate similar behavior for nearly Gaussian densities and the cross-validation outperforms the Silverman's rule of thumb on highly non-Gaussian densities.***

**K e y w o r d s:** ***finite mixture model, varying concentrations, kernel density estimator, bandwidth selection, Silwerman's rule of thumb, leave-one-out cross-validation.***

**AMS 2020 classification: 62G07, 62G09.**

## Introduction

Nonparametric density estimators are widely used in applied statistics. One of the most popular such estimators is the kernel density estimator (KDE) (Hastie, Tibshirani, & Friedman, 2009, Section 6.6). To apply KDE one needs to select a tuning parameter called the bandwidth. Among different methods of bandwidth selection the Silverman's rule of thumb is maybe the simplest one, and the cross-validation technique is one of most advanced ones (Silverman, 2018, Section 3.4), (Stone, 1984). Standard versions of these techniques were developed for homogeneous samples of independent identically distributed observations.

In this paper we discuss KDE and bandwidth selection rules in the case when observed subjects belong to $M$ different sub-populations with different distributions of the observed variable $\xi$. If the sub-population which the subject belongs to is unknown, then the observed distribution of $\xi$ is a mixture of its distributions for different components. In classical finite mixture models the mixing probabilities are the same for all observed subjects (Titterington, Smith, & Makov, 1985), (McLachlan, & Peel, 2000). We consider more flexible model of mixture with varying concentrations (MVC) at which the mixing probabilities are different for different subjects (see (Maiboroda, & Sugakova, 2012) for theory and applications to genetical data analysis, (Pidnebesna et al., 2023) for applications in neuroscience). A modification of KDE and Silverman's rule of thumb for MVC is considered in (Sugakova, 1999).

The main goal of this paper is to introduce a modification of the cross-validation technique for MVC KDE and to compare its performance with the performance of KDE with the bandwidth selected by the Silverman's rule of thumb. We performed a simulation study to analyze accuracy of these estimators. Note that the jackknife technique modification for MVC was considered in (Maiboroda, Miroshnichenko, & Sugakova, 2022). In the case of independent identically distributed observations the jackknife is rather similar to the cross-validation. But their modifications for MVC data differ significantly.

## 1. Mixtures with varying concentrations and components estimation

Assume that the observed subjects $O_1,\dots, O_n$ are obtained from $M$ different sub-populations (components of the mixture). The true number $\kappa(O)$ of the sub-population which the subject $O$ belongs to is unknown, but the probabilities $p_j^i = \mathsf{P}\{\kappa(O) = i\}$ are given. These probabilities are called the concentrations of components or the mixing probabilities.

For each subject we observe a numerical random variable $\xi_j = \xi(O_j)$. The distribution of $\xi(O)$ depends on $\kappa(O)$:

$$F_i(x) = \mathsf{Pr}\{\xi(O) < x \mid \kappa(O) = i\}.$$

So, the cumulative distribution function of $\xi_j$ is

$$\mathsf{P}\{\xi_j < x\} = \sum_{i=1}^{M} p_j^i F_i(x). \tag{1}$$

Formula (1) is called the model of mixture with varying concentrations (MVC). It is assumed that $\xi_j$ are independent. The CDFs $F_i$ are unknown. We suppose that they have probability densities $f_i$:

$$F_i(x) = \int_{-\infty}^{x} f_i(t)dt.$$

 

Then the probability density of $\xi_j$ is $\sum_{i=1}^m p_j^i f_i(x)$. Our aim is to estimate $f_k$ by the data $\Xi = (\xi_1, \dots, \xi_n)$. Let us start from estimation of $F_k$. The usual way to estimate CDF by a homogeneous sample is to use the empirical CDF

$$\hat{F}_n(x) = \frac{1}{n}\sum_{j=1}^n 1\{\xi_j < x\}.$$

But in the MVC case the empirical CDF would estimate some mixture of different $F_i$, $i = 1, \dots, M$, not the CDF $F_k$ of a chosen mixture component. To estimate $F_k$ one can use some weighted version of empirical CDF with weights which suppress the influence of nuisance components. In (Maiboroda, & Sugakova, 2012) weighted CDF with minimax weights is proposed for this purpose:

$$\hat{F}_{k;n}(x) = \frac{1}{n}\sum_{j=1}^n a_j^k 1\{\xi_j < x\}.$$

Here the matrix of minimax weights $\mathbf{a} = (a_j^i)_{j=1,\dots n, i=1,\dots M} \in \mathbb{R}^{n\times M}$ is defined as $\mathbf{a} = \mathbf{p}\mathbf{\Gamma}_n^{-1}$, where

$$\mathbf{p} = (p_j^i)_{j=1,\dots n, i=1,\dots M} \in \mathbb{R}^{n\times M},\ \mathbf{\Gamma}_n = \frac{1}{n}\mathbf{p}^T\mathbf{p}.$$

As discussed in (Maiboroda, & Sugakova, 2012), if $\det\mathbf{\Gamma}_n \neq 0$, then $\hat{F}_{k;n}(x)$ is a minimax estimator to $F_k(x)$ with respect to the quadratic loss function. With the weighted CDF $\hat{F}_{k;n}$ at hand, one can readily construct estimators for numerical characteristics of the $k$-th component distribution. Say, to estimate the $m$-th moment of $F_k$,

$$\mu_k^m = \int_{-\infty}^{\infty} x^m F_k(dx),$$

one can use

$$\hat{\mu}_{k;n}^m = \int_{-\infty}^{\infty} x^m \hat{F}_{k;n}(dx) = \frac{1}{n}\sum_{j=1}^n a_j^k (\xi_j)^m.$$

The variance

$$\sigma^2(F_k) = \int_{-\infty}^{\infty} (x - \mu_k^1)^2 F_k(dx)$$

can be estimated by

$$\hat{\sigma}_k^2 = \int_{-\infty}^{\infty} (x - \hat{\mu}_{k;n}^1)^m \hat{F}_{k;n}(dx) = \frac{1}{n}\sum_{j=1}^n a_j^k (\xi_j - \hat{\mu}_{k;n}^1)^2. \quad (2)$$

Consistency and asymptotic normality of these estimators is discussed in (Maiboroda, & Sugakova, 2012).

In what follows we will also need an estimator for the interquartile range (IQR) of $F_k$. Recall that IQR of a CDF $F_k$ can be defined as

$$\mathrm{IQR}(F_k) = \sup\{x:\ F_k(x) < 3/4\} - \inf\{x:\ F_k(x) > 1/4\}.$$

(If the function $F_k$ has an inverse $F_k^{-1}$, then $\mathrm{IQR}(F_k) = F_k^{-1}(3/4) - F_k^{-1}(1/4)$). So, we define the estimator for $\mathrm{IQR}(F_k)$ as

$$\widehat{\mathrm{IQR}}_k = \sup\{x:\ \hat{F}_{k;n}(x) < 3/4\} - \inf\{x:\ \hat{F}_{k;n}(x) > 1/4\}. \quad (3)$$

Asymptotic behavior of such estimators is discussed in (Maiboroda, & Sugakova, 2020), (Maiboroda, Miroshnichenko, & Sugakova, 2024).

**2. Modified kernel density estimator for mixture components**

Let us now consider estimation of the $k$th component density $f_k$. In (Sugakova, 1999) a modified kernel density estimator (MKDE) for $f_k(x)$ is considered of the form

$$\hat{f}_{k,n}(x) = \frac{1}{n}\sum_{j=1}^n a_j^k K\left(\frac{x - \xi_j}{h}\right), \quad (4)$$

where $K : \mathbb{R} \to \mathbb{R}$ is a kernel, i.e. a function satisfying $\int f(z)dz = 1$, $h >$ is the bandwidth. The bandwidth is a tuning parameter of the estimator which must be selected carefully to ensure a good accuracy. The integrated squared error

$$\mathrm{ISE}(h) = \mathrm{ISE}(\hat{f}_{k;n}) = \int_{-\infty}^{\infty} (\hat{f}_{k;n}(x) - f_k(x))^2 dx, \quad (5)$$

and the mean integrated squared error

$$\mathrm{MISE}(h) = \mathrm{MISE}(\hat{f}_{k;n}) = \int_{-\infty}^{\infty} \mathrm{E}(\hat{f}_{k;n}(x) - f_k(x))^2 dx = \mathrm{E}\left[\mathrm{ISE}(\hat{f}_{k;n})\right] \quad (6)$$

are the most useful measures of estimator's accuracy.

The following theorem on the lower bound of asymptotic MISE of kernel density estimators is shown in (Sugakova, 1999).

**Theorem 1.** *Assume that*

*(i) There exists a finite interval* $[b_1, b_2] \subset \mathbb{R}$, *such that* $f_k(x) = 0$ *for all* $x \notin [b_1, b_2]$.

*(ii)* $f_k''(x) = d^2 f_k(x)/dx^2$ *exists for all* $x \in [b_1, b_2]$ *and*

$$\phi_k = \int_{-\infty}^{\infty} (f_k''(x))^2 dx < \infty.$$

*(iii)*

$$\int_{-\infty}^{\infty} zK(z)dz = 0,\ D = \int_{-\infty}^{\infty} z^2 K(z)dz < \infty,\ d^2 = \int_{-\infty}^{\infty} K^2(z)dz < \infty.$$

*(iv)* $A_n = \frac{1}{n}\sum_{j=1}^{n}(a_j^k)^2 \to A$ *as* $n \to \infty$, *where* $0 < A < \infty$.

*(v) There exists* $c > 0$, *such that* $\det \mathbf{\Gamma}_n > c$ *for all* $n$.

*Then*

$$\liminf_{n\to\infty} n^{4/5}\,\mathsf{MISE}(\hat{f}_{k;n}) \geqslant \frac{5}{4}(DAd^2)^{4/5} \tag{7}$$

*and the equality in (7) is attained if*

$$h = h_n^{opt} = \left(\frac{Ad^2}{D^4\phi_k}\right)^{1/5}. \tag{8}$$

The bandwidth $h_n^{\mathsf{opt}}$ defined by (8) is called the theoretically optimal bandwidth. It corresponds to asymptotically least possible value of MISE as $n \to \infty$. But it is impossible to use $h_n^{\mathsf{opt}}$ in the density estimation by real data, since the value of $\phi_k$ depends from the unknown density. So one needs some practical sub-optimal bandwidth selection algorithm.

**3. Modified Silverman's rule of thumb**

The simplest bandwidth selection algorithm is the Silverman's rule of thumb. It is based on the idea to approximate $\phi_k$ by its estimator by the data. This needs some additional assumptions on $f_k$. If $f_k$ was a Gaussian density with variance $\sigma_k^2$, then its $\phi_k$ would be

$$\phi_k = \frac{3}{8\sqrt{\pi}\sigma^5}$$

and the optimal bandwidth would be

$$h_n^{\mathsf{opt}} = \left(\frac{8\sqrt{\pi}Ad^2}{3D^4}\right)^{1/5}\sigma_k.$$

Replacing here $\sigma_k$ by its estimator $\hat{\sigma}_k$ defined by (2), we obtain a simple bandwidth selection rule. To make this rule more robust, one can utilize the IQR for estimation of $\sigma_k$. Namely, since $\sigma_k = \mathsf{IQR}(F_k)/(2\Phi^{-1}(0.75))$ for Gaussian $F_k$, we obtain the following modification of the Silverman's rule of thumb for MVC data:

$$h_n^{\mathsf{Silv}} = \left(\frac{8\sqrt{\pi}Ad^2}{3D^4}\right)^{1/5}\min(\hat{\sigma}_k, \widehat{\mathsf{IQR}}_k/(2\Phi^{-1}(0.75))).$$

Surely, the true $f_k$ to be estimated, is not Gaussian. So this rule of thumb will not be optimal even asymptotically. But one expects that it should lead to appropriate estimators if $f_k$ is a Gaussian-like probability density.

**4. Modified cross-validation bandwidth selector**

Another way to chose the bandwidth is based on the leave-one-out cross-validation technique (LOO-CV). It is based on the approximate minimization of ISE given by (5). Observe that

$$\mathsf{ISE}(h) = J_1(h) - 2J_2(h) + J_3,$$

where

$$J_1(h) = \int_{-\infty}^{\infty} (\hat{f}_{k;n}(x))^2 dx,\ J_2(h) = \int_{-\infty}^{\infty} \hat{f}_{k;n}(x) f_k(x) dx,$$

$$J_3 = \int_{-\infty}^{\infty} (f_k(x))^2 dx.$$

The term $J_3$ does not depend on $h$ and can be dropped. The term $J_1(h)$ can be calculated by the data directly. But $J_2(h)$ depends on $h$ and on the unknown $f_k$. To obtain a feasible bandwidth selector we need to estimate $J_2(h)$. Let us do it by a LOO-CV type estimator. Observe that

$$J_2(h) = \mathsf{E}[\hat{f}_{k;n}(\eta) \mid \Xi],$$

where $\eta$ is a random variable with CDF $F_k$ independent of the data $\Xi$ by which $\hat{f}_{k;n}$ is calculated. To estimate $J_2(h)$ we use the weighted mean

$$\hat{J}_2(h) = \frac{1}{n}\sum_{j=1}^{n} a_j^k \hat{f}_{k;n}^{(j-)}(\xi_j),$$

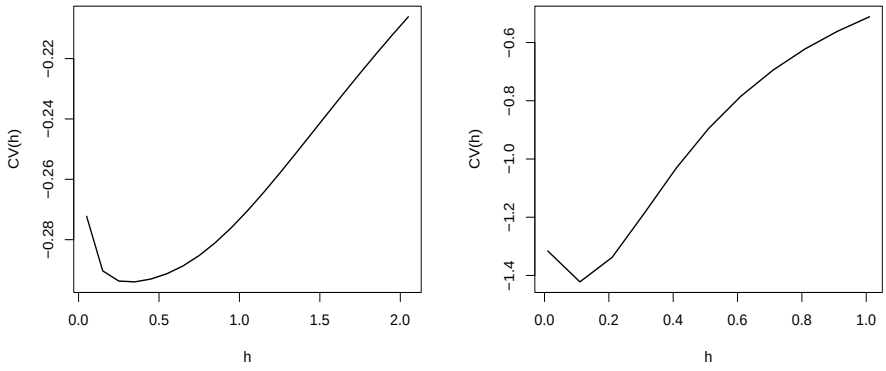


**Fig. 1.** Graph of $CV(h)$ in the Experiment 1. Component 1 on the left panel, Component 2 on the right pannel

where $\hat{f}_{k;n}^{(j-)}(x)$ is the MKDE, calculated by all observations except the $j$th one. More precisely, $\hat{f}_{k;n}^{(j-)}(x)$ is defined as follows. Let

$$p_j^{k,i-} = \begin{cases} p_j^k & \text{if } j \neq i, \\ 0 & \text{if } j = k, \end{cases}$$

$$\mathbf{p}^{(i-)} = (p_j^{l,i-})_{j=1,\dots n, l=1,\dots M} \in \mathbb{R}^{n\times M},\ \mathbf{\Gamma}_n^{(i-)} = \frac{1}{n}(\mathbf{p}^{(i-)})^T\mathbf{p}^{(i-)}.$$

$$\mathbf{a}^{(i-)} = \mathbf{p}^{(i-)}(\mathbf{\Gamma}_n^{(i-)})^{-1}.$$

Then

$$\hat{f}_{k,n}^{(i-)}(x) = \frac{1}{n}\sum_{j=1}^{n} a_j^{k,i-} K\left(\frac{x-\xi_j}{h}\right).$$

Observe that $a_i^{k,i-} = 0$, so $\xi_i$ and $\hat{f}_{k,n}^{(i-)}(x)$ are independent.

Now the cross-validation functional is defined as

$$\mathrm{CV}(h) = J_1(h) - 2\hat{J}_2(h).$$

It is an estimator of $\mathrm{ISE}(h) - J_3$. The cross-validation bandwidth is

$$h_n^{\mathrm{CV}} = \mathrm{argmin}_{h>0}\,\mathrm{CV}(h). \tag{9}$$

In the simulation experiments we considered two-component mixtures ($M = 2$) with the concentrations defined as

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

The Gaussian kernel $K(t) = \exp(-t^2/2)/\sqrt{2\pi}$ was used in the density estimator.

**Experiment 1**. In the first experiment we simulate the sample of size $n = 1000$. Both components are normal: the first one has distribution $N(10,1)$ and the second one $N(0,0.2)$. Plot of $\mathrm{CV}(h_n)$ for the first and second components (Fig. 1) show presence of the global minimum. So, calculation the optimal $h_n$ by formula (9) is reasonable.

**5. Simulation results**

In all the following experiments we generated samples of the size $n = 100, 250, 500, 750, 1000$. For each $n$, $T = 1000$ samples were generated. Firstly, we calculate for every case $h_n^{Silv}$ and $h_n^{CV}$. Than we derive means and variations for $\mathrm{ISE}(h_n)$ defined by (5) for $h_n = h_n^{Silv}$ and $h_n = h_n^{CV}$.

**Experiment 2**. In second experiment both components are normal too, but with another parameters, then in the Experiment 1: the first component $\sim N(-1,1)$, the second one $\sim N(1,1)$. The results are presented in the Tab. 1 and 2.

The results are practically identical for the Silverman's and CV bandwidths, but the estimators based on the Silverman's rule perform slightly better. In the next experiment we work with heavy-tailed distributions.

***Table 1***

**Experiment 2, Component 1**

| $n$ | **mean**$(\mathrm{ISE}(h_n^{CV}))$ | **mean**$(\mathrm{ISE}(h_n^{Silv}))$ | **var**$(\mathrm{ISE}(h_n^{CV}))$ | **var**$(\mathrm{ISE}(h_n^{Silv}))$ |
|---|---|---|---|---|
| 100 | 0.024065 | 0.018527 | 7.23E-04 | 1.93E-04 |
| 250 | 0.011508 | 0.008928 | 9.95E-05 | 3.48E-05 |
| 500 | 0.006502 | 0.005176 | 2.41E-05 | 1.08E-05 |
| 750 | 0.004907 | 0.003936 | 1.50E-05 | 5.63E-06 |
| 1000 | 0.003885 | 0.003137 | 8.67E-06 | 3.69E-06 |

***Table 2***

**Experiment 2, Component 1**

| $n$ | **mean**(ISE($h_n^{CV}$)) | **mean**(ISE($h_n^{Silv}$)) | **var**(ISE($h_n^{CV}$)) | **var**(ISE($h_n^{Silv}$)) |
|---|---|---|---|---|
| 100 | 0.025176 | 0.018637 | 6.59E-04 | 1.82E-04 |
| 250 | 0.011968 | 0.009080 | 1.09E-04 | 3.38E-05 |
| 500 | 0.006694 | 0.005290 | 3.42E-05 | 1.16E-05 |
| 750 | 0.005156 | 0.004025 | 1.71E-05 | 6.66E-06 |
| 1000 | 0.003967 | 0.003087 | 9.07E-06 | 3.01E-06 |

**Experiment 3**. In this experiment both components have Student's t distribution: the first component with $df = 5$ degrees of freedom, the second one with $df = 20$ degrees of freedom.

The results of this experiment presented in Tab. 3, 4 are rather similar to Experiment 1. The Silverman's rule performs a bit better. So the heavy tails of the estimated density does not deteriorate the estimators performance.

***Table 3***

**Experiment 3, Component 1**

| $n$ | **mean**(ISE($h_n^{CV}$)) | **mean**(ISE($h_n^{Silv}$)) | **var**(ISE($h_n^{CV}$)) | **var**(ISE($h_n^{Silv}$)) |
|---|---|---|---|---|
| 100 | 0.024022 | 0.016803 | 6.46E-04 | 1.70E-04 |
| 250 | 0.011216 | 0.007846 | 1.47E-04 | 3.35E-05 |
| 500 | 0.007206 | 0.005233 | 4.34E-05 | 1.21E-05 |
| 750 | 0.005117 | 0.003769 | 2.48E-05 | 6.25E-06 |
| 1000 | 0.003963 | 0.003115 | 9.52E-06 | 4.30E-06 |

***Table 4***

**Experiment 3, Component 2**

| $n$ | **mean**(ISE($h_n^{CV}$)) | **mean**(ISE($h_n^{Silv}$)) | **var**(ISE($h_n^{CV}$)) | **var**(ISE($h_n^{Silv}$)) |
|---|---|---|---|---|
| 100 | 0.025452 | 0.017431 | 9.55E-04 | 1.76E-04 |
| 250 | 0.011807 | 0.008621 | 1.10E-04 | 3.60E-05 |
| 500 | 0.007048 | 0.005089 | 4.78E-05 | 1.19E-05 |
| 750 | 0.004811 | 0.003675 | 1.57E-05 | 6.12E-06 |
| 1000 | 0.004145 | 0.003010 | 1.86E-05 | 3.68E-06 |

**Experiment 4**. In this experiment the first component is normal $\sim N(0,1)$ but the second one is a mixture of two normal distributions $N(3,1)$ and $N(-3,1)$ with concentrations equal 1/2. The results are presented in Tab. 5 and 6.

Here the distribution of the second component significantly differs from the normal one, so one expects that the Silverman's rule would fall. Accordingly, in this case the CV-based estimator outperform the Silverman's one.

***Table 5***

**Experiment 4, Component 1**

| $n$ | **mean**(ISE($h_n^{CV}$)) | **mean**(ISE($h_n^{Silv}$)) | **var**(ISE($h_n^{CV}$)) | **var**(ISE($h_n^{Silv}$)) |
|---|---|---|---|---|
| 100 | 0.024363 | 0.019493 | 4.24E-04 | 1.76E-04 |
| 250 | 0.011588 | 0.009500 | 8.18E-05 | 3.24E-05 |
| 500 | 0.006695 | 0.005624 | 2.37E-05 | 1.11E-05 |
| 750 | 0.004867 | 0.004023 | 1.14E-05 | 5.13E-06 |
| 1000 | 0.003905 | 0.003246 | 6.97E-06 | 3.26E-06 |

***Table 6***

**Experiment 4, Component 2**

| $n$ | **mean**(ISE($h_n^{CV}$)) | **mean**(ISE($h_n^{Silv}$)) | **var**(ISE($h_n^{CV}$)) | **var**(ISE($h_n^{Silv}$)) |
|---|---|---|---|---|
| 100 | 0.020952 | 0.038626 | 2.07E-04 | 2.17E-05 |
| 250 | 0.009808 | 0.027813 | 3.65E-05 | 1.19E-05 |
| 500 | 0.005848 | 0.020735 | 1.20E-05 | 6.20E-06 |
| 750 | 0.004155 | 0.017020 | 6.17E-06 | 4.13E-06 |
| 1000 | 0.003320 | 0.014971 | 2.87E-06 | 3.40E-06 |

## Discussion and conclusion

We have introduced a new modification of CV-approach to the bandwidth selection for the kernel density estimators. This approach allows to chose the bandwidth for the estimation by data from a mixture with varying concentrations. Results of simulations demonstrate that the CV-based estimators perform nearly as well as the modified Silverman's rule of thumb based estimators when the estimated density is similar to the Gaussian one. But in the highly non-Gaussian case the CV bandwidth selection outperforms the Silverman's rule significantly even for small sample sizes. Further experiments are needed for investigation of small sample behavior of CV-based estimators.

**Authors' contribution:** Rostyslav Maiboroda — conceptualization; methodology; Olena Sugakova — methodology; software.

**Sources of funding.** Funding is partially provided by Taras Shevchenko National University of Kyiv.

**References**

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning data mining, inference, and prediction*. Springer.

Maiboroda, R., Miroshnichenko, V., & Sugakova, O. (2022). Jackknife for nonlinear estimating equations. *Modern Stochastics: Theory and Applications*, 9(4), 377–399. https://doi.org/10.15559/22-VMSTA208

Maiboroda, R., Miroshnichenko, V., & Sugakova, O. (2024). Quantile estimators for regression errors in mixture models with varying concentrations. *Bulletin of Taras Shevchenko National University of Kyiv. Physical and Mathematical Sciences*, 1(78), 45–50. https://doi.org/10.17721/1812-5409.2024/1.8

Maiboroda, R., & Sugakova, O. (2012). Statistics of mixtures with varying concentrations with application to dna microarray data analysis. *Nonparametric statistics*, 24(1), 201–215. https://doi.org/10.1080/10485252.2011.630076

Maiboroda, R., & Sugakova, O. (2020). Tests of hypotheses on quantiles of distributions of components in a mixture. *Theory of Probability and Mathematical Statistics*, 101, 179–191. https://doi.org/10.1090/tpms/1120

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. Wiley. https://doi.org/10.1002/0471721182

Pidnebesna, A., Fajnerová, I., Horáček, J., & Hlinka, J. (2023). Mixture components inference for sparse regression: Introduction and application for estimation of neuronal signal from fmri bold. *Applied Mathematical Modelling*, 116, 735–748. https://doi.org/10.1016/j.apm.2022.11.034

Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge.

Stone, C. (1984). An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, 12(4), 1285–-1297. https://doi.org/10.1002/047172118210.1214/aos/1176346792

Sugakova, O. (1999). Asymptotics of a kernel estimate for the density of a distribution constructed from observations of a mixture with varying concentration. *Theory of Probability and Mathematic Statistics*, 59, 161–171.

Titterington, D., Smith, A., & Makov, O. (1985). *Analysis of finite mixture distributions*. Wiley.

**Отримано редакцією журналу / Received: 30.04.25**
**Прорецензовано / Revised: 02.09.25**
**Схвалено до друку / Accepted: 10.10.25**

**Ростислав МАЙБОРОДА, д-р фіз.-мат. наук, проф.**
**ORCID ID: 0000-0002-3899-558X**
**e-mail: rostmaiboroda@knu.ua**
**Київський національний університет імені Тараса Шевченка, Київ, Україна**

**Олена СУГАКОВА, д-р фіз.-мат. наук, доц.**
**ORCID ID: 0000-0002-6529-0788**
**e-mail: olenasugakova@knu.ua**
**Київський національний університет імені Тараса Шевченка, Київ, Україна**

## ВИБІР ПАРАМЕТРА ЗГЛАДЖУВАННЯ ДЛЯ ЯДЕРНОЇ ОЦІНКИ ЗА СПОСТЕРЕЖЕННЯМИ ІЗ СУМІШІ

***Статистичні дані біологічних і медичних досліджень часто являють собою суміш спостережень об'єктів, що належать різним підпопуляціям з різними статистичними властивостями спостережуваних характеристик. Такі дані зручно описувати математичними моделями скінченних сумішей. У пропонованій роботі для опису даних використано модель суміші зі змінними концентраціями (СЗК), у якій концентрації компонентів можуть змінюватись від спостереження до спостереження. Розглянуто задачу непараметричного оцінювання щільності розподілу окремої компоненти суміші в межах моделі СЗК. Для цього використовується модифікована ядерна оцінка щільності (ЯОЩ). Практичне застосування ЯОЩ потребує вибору параметра згладжування. У статті запропоновано два підходи до такого вибору: на основі модифікованого правила Сілвермана і за допомогою алгоритму кросвалідації, адаптованого до СЗК. Якість отриманих оцінок порівнюється за допомогою імітаційного моделювання. Результати моделювання показують, що у випадку щільностей, близьких до гауссових, обидва підходи дають майже однакові результати. Для щільностей, які відрізняються від гауссових, вибір параметра згладжування на основі модифікованої кросвалідації дає значно кращі результати, ніж на основі правила Сілвермана.***

**Ключові слова:** ***модель скінченної суміші, змінні концентрації, ядерна оцінка щільності, вибір параметра згладжування, правило Сілвермана, кросвалідація.***

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження (у зборі, аналізі чи інтерпретації даних, якщо це мало місце), у написанні рукопису та в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study (in the collection, analyses or interpretation of data if applicable), in the writing of the manuscript as well as in the decision to publish the results.