Journal of Mathematical Problems, Equations and Statistics

E-ISSN: 2709-9407
P-ISSN: 2709-9393
JMPES 2024; 5(1): 143-149

www.mathematicaljournal.com
Received: 25-01-2024
Accepted: 06-03-2024

Corresponding Author:

A modified rule-of-thumb method for kernel density estimation

Abstract
Kernel density estimation (KDE) is a widely used nonparametric technique for estimating the probability density function (PDF) of a random variable. However, the performance of KDE depends largely on the choice of the bandwidth parameter, which controls the trade-off between bias and variance in the estimation. In this study, we investigate the Silverman method for selecting the bandwidth for univariate continuous PDFs, and propose a modified method that improves the accuracy and efficiency of the estimation. We use simulation to compare the two methods, and show that the modified method achieves lower mean squared error (MSE) and mean integrated squared error (MISE) than the Silverman method.

Keywords: Kernel density, silverman rule of thumb, bandwidth selection

Introduction
Kernel density estimation (KDE) is a powerful tool in statistics and data analysis. It is a nonparametric method that estimates the probability density function (PDF) of a random variable without making any assumptions about its underlying distribution. It does so by using a kernel function, which is a smooth and symmetric function that assigns weights to nearby observations, and then taking their weighted average. KDE can be seen as a generalization of the histogram, where the bins are replaced by kernel functions centered at each observation.
The idea of using kernel functions to estimate PDFs dates back to Rosenblatt (1956) [12], who proposed the kernel estimator for univariate continuous PDFs and derived its asymptotic mean squared error (MSE) based on symmetric kernels. Parzen (1962) [11] extended the kernel estimator to multivariate PDFs and gave a general formula for the optimal bandwidth, which is the smoothing parameter that controls the trade-off between bias and variance in the estimation. Since then, many researchers have developed various methods for selecting the bandwidth, such as plug-in methods (Scott *et al*., 1977; Sheather and Jones, 1991; Wand and Jones, 1994; Tenreiro, C. 2020) [16, 17, 21, 20], cross-validation methods (Rudemo, 1982; Bowman and Azzalini, 1997; Stone, 1984; Scott and Terrell, 1987; Savchuk *et al*., 2010) [13, 3, 19, 15, 14], and rule-of-thumb methods (Silverman, 1986) [18].
KDE has been widely applied in various fields of science and engineering because of its ability to describe the shape and features of the data without imposing any parametric assumptions. For example, KDE has been used in medicine to analyze the distribution of blood pressure (Jankowska *et al*., 2017) [7], in machine learning to perform classification and clustering tasks (Lahane and Sangaiah, 2015) [9], in genetics to detect differential expression of genes and cells (Alquicira-Hernandez and Powell, 2021) [1], in petroleum engineering to model the porosity and permeability of reservoir rocks (Corina and Hovda, 2018) [4], in climatology to study the wind speed and direction (Hu *et al*., 2017) [6], in energy economics to forecast the electricity consumption and price (Arora and Taylor, 2016) [2], and in ecology to estimate the home range and habitat selection of wildlife (Fleming and Calabrese, 2017) [5].
In this study, we focus on the rule-of-thumb method by Silverman (1986) [18], which is one of the simplest and computationally efficient methods for choosing the bandwidth for univariate continuous PDFs. We investigate its performance and limitations, and propose a modified method that improves the accuracy and efficiency of the estimation. We use simulation to compare the two methods, and show that the modified method achieves lower MSE and mean integrated squared error (MISE) than the Silverman method.

Kernel estimation
Suppose that $x_1, x_2, \ldots, x_n$ is an independent and identically distributed (i.i.d) random samples

from a PDF f. The kernel density estimator is defined as follows (Rosenblatt, 1956) [12]

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \quad (1)$$

where K represents a non-negative function called the kernel, and $h > 0$ represents the smoothing parameter or bandwidth. The kernel function K is often chosen to be a symmetric probability function that satisfies the following conditions:

1. $\int K(u)\, du = 1$
2. $\int uK(u)\, du = 0$
3. $\int u^2 K(u)\, du = \kappa_2 > 0$

The kernel estimation of the probability density function depends on the kernel function K, which determines the shape of the weights assigned to each observation, and the bandwidth h determines the width of the window around each observation. The choice of the kernel function K has a minor effect on the estimation of the function f, as long as the chosen kernel is smooth and symmetric. One of the most commonly used kernel functions is the standard normal distribution function:

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$$

The choice of the bandwidth h, on the other hand, has a major effect on the estimation of f, as it controls the trade-off between bias and variance. Figure 1 shows how different values of h affect the estimation of f. If h is too small, the estimator will be very sensitive to noise and will overfit the data, resulting in a high variance and low bias. If h is too large, the estimator will be very smooth and will underfit the data, resulting in a low variance and high bias. Therefore, it is important to choose an optimal value of h that balances the bias and variance and minimizes the mean squared error (MSE) or mean integrated squared error (MISE) of the estimator.

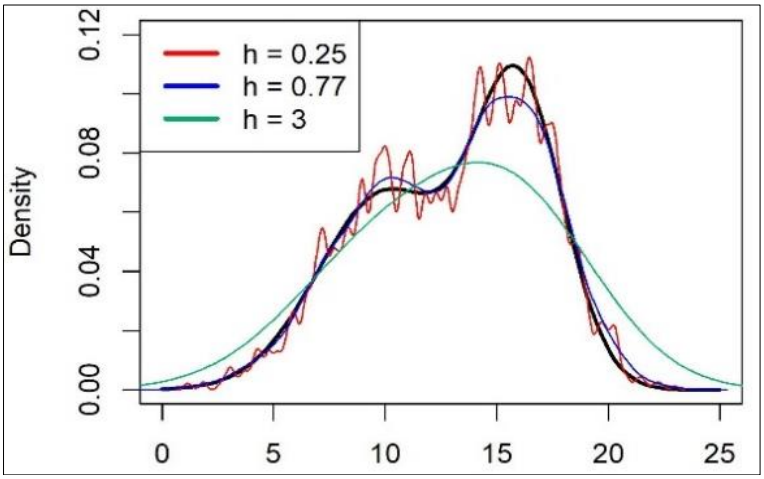

Fig 1: Kernel density estimation of a PDF with different bandwidth (h)

Some properties of kernel estimator

In this section, we will review some of the properties of kernel estimators, such as their approximated bias, variance, MSE and MISE. Since KDE are generally considered biased estimators, and therefore measurements of errors should be used to study the performance of these estimators, and one of the most famous of these measurements is the mean square error (MSE), which is known as follows (Silverman, 1986; Wand and Jones, 1995) [18, 21]:

$$\text{MSE}(\hat{f}) = E\left[\hat{f}(x) - f(x)\right]^2$$

Note that we could write the MSE as the bias squared (i.e., $b^2 = \left[E\left(\hat{f}(x)\right) - f(x)\right]^2$) plus the variance ($Var\left(\hat{f}(x)\right)$). We usually find an approximation to the MSE as follows

$$E\left(\hat{f}(x)\right) = \frac{1}{nh}\sum_{i=1}^{n} E\left[K\left(\frac{x_i - x}{h}\right)\right] = \frac{1}{h} E\left[K\left(\frac{x_i - x}{h}\right)\right] = \frac{1}{h}\int K\left(\frac{x_i - x}{h}\right) f(x)\, \mathrm{d}x$$

let $y = \frac{x_i - x}{h}$, we get

$$E\left(\hat{f}(x)\right) = \int K(y) f(x - yh)\, \mathrm{d}y$$

Using Taylor series expansion of $f(x-yh)$, we have

$$E\left(\hat{f}(x)\right) \approx \int K(y)\left[f(x)+yhf'(x)+\frac{1}{2}y^2h^2f''(x)+o(h^2)\right]\mathrm{d}y$$

By simple algebraic manipulation, the bias can be approximated as follows

$$b \approx \int K(y)\left[f(x)+yhf'(x)+\frac{1}{2}y^2h^2f''(x)+o(h^2)\right]\mathrm{d}y - f(x) = \frac{1}{2}h^2f''(x)\int y^2K(y)\,\mathrm{d}y + o(h^2) = \frac{1}{2}h^2f''(x)\kappa_2 + o(h^2)$$

Therefore, as $h \to 0$, the bias shrinks by $o(h^2)$. In addition, the variance can be approximated as follows,

$$Var\left(\hat{f}(x)\right) = \frac{1}{nh^2}Var\left[K\left(\frac{x_i-x}{h}\right)\right] \le \frac{1}{nh^2}E\left[K^2\left(\frac{x_i-x}{h}\right)\right] = \frac{1}{nh}\int K^2(y)f(x+yh)\,\mathrm{d}y$$

Using Taylor series expansion of $f(x+yh)$, we have

$$Var\left(\hat{f}(x)\right) = \frac{1}{nh}\int K^2(y)[f(x)+yhf'(x)o(h)]\mathrm{d}y = \frac{1}{nh}f(x)\int K^2(y)\,\mathrm{d}y + o\left(\frac{1}{nh}\right) = \frac{1}{nh}f(x)R(k) + o\left(\frac{1}{nh}\right)$$

Where $R(K) = \int K^2(y)\mathrm{d}y$. Note that, as $n \to \infty$ and $h \to 0$, the variance shrinks at rate of $o\left(\frac{1}{nh}\right)$. Hence

$$\mathrm{MSE}\left(\hat{f}\right) \approx \tfrac{1}{4}h^4f''(x)^2\kappa_2^2 + \tfrac{1}{nh}f(x)R(x) + o(h^4) + o\left(\tfrac{1}{nh}\right) \quad (2)$$

The MISE of a kernel estimator is defined as the expected value of its integrated squared error:M

$$\mathrm{ISE}\left(\hat{f}\right) = E\int\left[\hat{f}(x)-f(x)\right]^2\mathrm{d}x = \int \mathrm{MSE}(\hat{f})\,\mathrm{d}x$$

Using the MSE approximation derived in (2), we can obtain an expression for the MISE by applying some algebraic manipulations. The result is:

$$\mathrm{MISE}\left(\hat{f}\right) \approx \frac{1}{4}h^4\int f''(x)^2\mathrm{d}x\,\kappa_2^2K + \frac{1}{nh}\int f(x)\mathrm{d}x\,R(x) + o(h^4) + o\left(\frac{1}{nh}\right)$$

$$= \tfrac{\kappa_2^2K}{4}h^4\int f''(x)^2\mathrm{d}x + \tfrac{R(x)}{nh} + o(h^4) + o\left(\tfrac{1}{nh}\right) \quad (3)$$

Therefore, the MISE depends on both the sample size n and the bandwidth h, and it can be minimized by choosing an optimal value of h that balances the bias and variance terms.
To find the optimal value of the smoothing parameter h for the MISE criteria, we differentiate the previous equation with respect to h and then equate it to zero to obtain

$$h_{opt} = \left[\frac{R(K)}{n\kappa_2^2\int f''(x)^2\mathrm{d}x}\right]^{\frac{1}{5}} \quad (4)$$

Unfortunately, the previous formula (4) cannot be applied in practice, because $\int f''(x)^2\mathrm{d}x$ is unknown and inversely proportional to h_{opt}, so Silverman (1986) suggested using normal distribution density as a reference distribution, and thus

$$\int f''(x)^2\,\mathrm{d}x = \frac{1}{\sigma^5}\int \phi''(x)^2\,\mathrm{d}x = \frac{3}{8\sigma^5\sqrt{\pi}} \approx 0.212\sigma^{-5} \quad (5)$$

where $\phi(x)$ is the standard normal distribution. Substituting back in equation (4) with $R(K) = \frac{1}{2\sqrt{\pi}}$ and $\kappa_2 = 1$, we have

$$h_S = 1.06\sigma n^{-\frac{1}{5}} \quad (6)$$

Where σ is approximated by the standard deviation of the sampled dataset. In order to improve the bandwidth in (6) and make it robust for deviation from normality and less sensitive to outliers and skewness, Silverman (1986) suggests using the interquartile range IQR of the sampled data to estimate the standard deviation, and then apply it as follows:

$$h_{S^*} = 1.06\min\left(\hat{\sigma}, \frac{\mathrm{IQR}}{1.34}\right)n^{-\frac{1}{5}} \quad (7)$$

Note that the interquartile range for the standard normal distribution is equal to 1.34.

In this article, we propose using the standard Cauchy distribution to approximate the standard normal distribution in (5). Both distributions are symmetric and bell-shaped, but the Cauchy distribution has fatter tails. This means that it gives more weight to extreme values, which is a useful feature for our approximation.

Replacing the standard normal distribution with the standard Cauchy distribution, we obtain:

$$\int \phi''(x)^2\,dx \;\approx\; \int \left(\frac{\partial^2}{\partial x^2}\frac{1}{\pi(1+x^2)}\right)^2 dx = \int \left(\frac{6x^2-2}{\pi(1+x^2)^3}\right)^2 dx \;= \frac{3}{4\pi}$$

By substituting the result back in the optimum bandwidth in (4) we obtain

$$h_M = 1.03\sigma n^{-\frac{1}{5}} \tag{8}$$

To make the bandwidth in (8) less sensitive to unusual data values or any deviation from normality, I recommend using the following bandwidth

$$h_{M^*} = 1.03 \min\left(\hat{\sigma}, s^* = \frac{P_{0.977}-P_{0.023}}{4}\right) n^{-\frac{1}{5}} \tag{9}$$

Where P denotes the percentile of the sampled data, and s^* be an estimate of the standard deviation based on the empirical rule of the normal distribution, which states that about 95% of the values are within two standard deviations from the mean.

Simulation

In this section, we compare the modified method of estimating the bandwidth with the Silverman method using Monte Carlo simulation. The models that we used in our simulation study were chosen to represent a variety of shapes and features of the PDFs, such as unimodal, bimodal, symmetric, skewed, and multimodal distributions. Figure 2 displays the graphs of the models, which are described as follows:

Model 1: $N(0,1)$

Model 2: $N(0,3)$

Model 3: $N(0,9)$

Model 4: $\frac{9}{10}N(0,1)+\frac{1}{10}N(0,4)$

Model 5: $\frac{1}{2}N(0,1)+\frac{1}{2}N(6,2)$

Model 6: $\frac{1}{5}N(0,1)+\frac{1}{5}N\left(\frac{1}{2},\left(\frac{2}{3}\right)^2\right)+\frac{3}{5}N\left(\frac{13}{12},\left(\frac{5}{9}\right)^2\right)$

Model 7: $t(5)$

Model 8: $t(10)$

Model 9: $\frac{1}{2}N\left(-1,\left(\frac{2}{3}\right)^2\right)+\frac{1}{2}N\left(1,\left(\frac{2}{3}\right)^2\right)$

Model 10: $\frac{1}{2}N\left(-\frac{3}{2},\left(\frac{1}{2}\right)^2\right)+\frac{1}{2}N\left(\frac{3}{2},\left(\frac{1}{2}\right)^2\right)$

Model 11: $\frac{3}{4}N(0,1)+\frac{1}{4}N\left(\frac{3}{2},\left(\frac{1}{3}\right)^2\right)$

Model 12: $\sum_{i=0}^{7}\frac{1}{8}N\left(3\left[\left(\frac{2}{3}\right)^i-1\right],\left(\frac{2}{3}\right)^{2i}\right)$

Model 13: $\frac{1}{2}N(0,1)+\sum_{i=0}^{4}\frac{1}{10}N\left(\left(\frac{i}{2}-1\right),\left(\frac{1}{10}\right)^2\right)$

Model 14: $\frac{1}{2}N(0,1)+\sum_{i=-2}^{2}\frac{2^{1-i}}{31}N\left(i+\frac{1}{2},\left(\frac{2^{-i}}{10}\right)^2\right)$

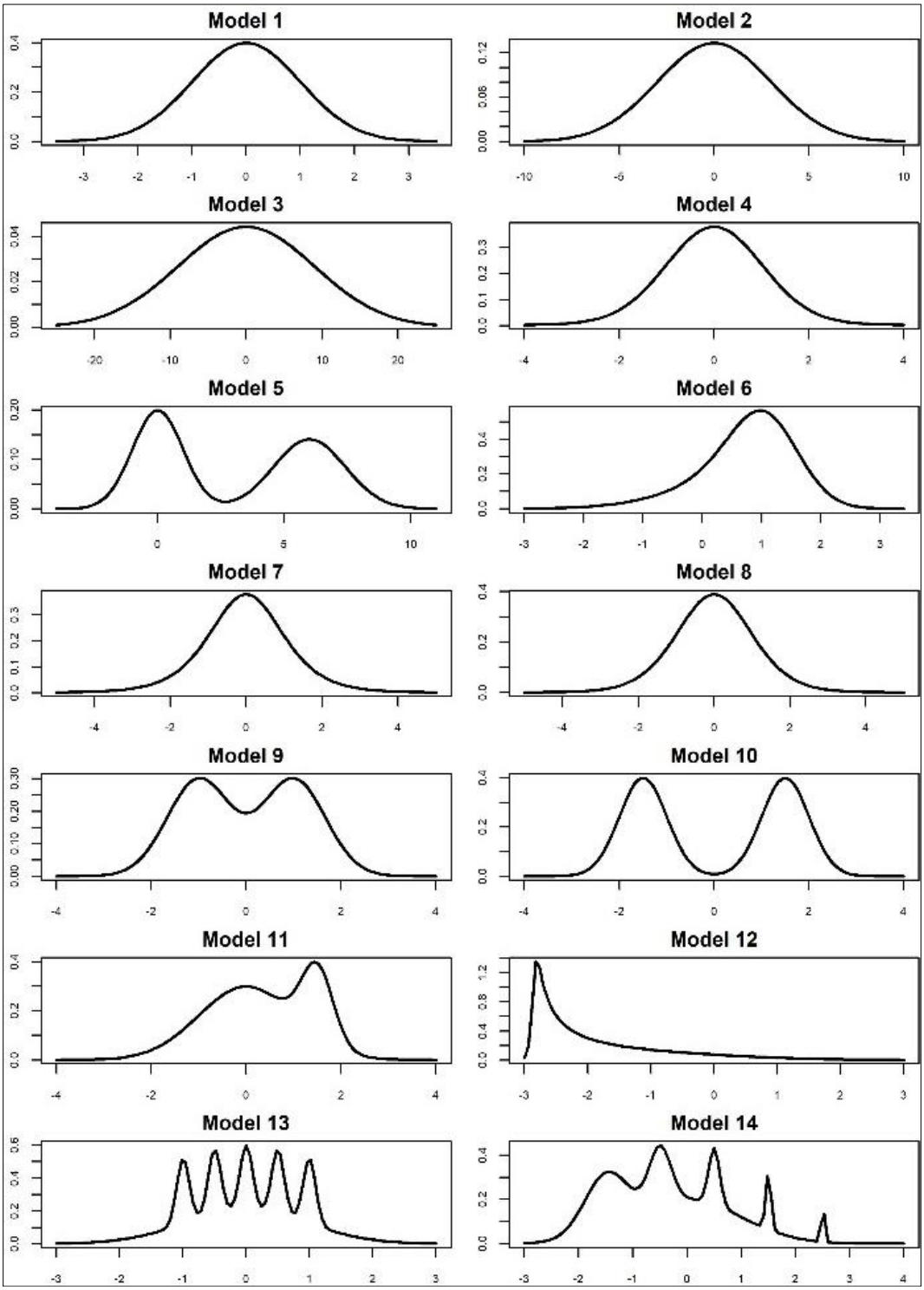

Fig 2: Plots of the 14 PDF's that were used in our simulation

Some of these models were adapted from (Marron and Wand, 1992) [10] and some from Kile (2010) [8]. The simulations were run for 10,000 times for each of the proposed sample sizes (n=30,100,300), and then the performance of the methods was compared using the MSE and MISE, where R 4.2.3 was used to run the simulation.

Results and Discussions

The simulation results are summarized in Tables 1, 2 and 3. Our results showed that the modified method generally achieved lower MSE and MISE than the Silverman method, especially when the PDF was skewed, bimodal, or had multiple modes. For example, for model 5, which is a mixture of two normal distributions with different means and variances, the modified method reduced the MSE by 37.7%, 44.6%, and 49.0% for sample sizes 30, 100, and 300, respectively. Similarly, for model 10, which is a mixture of two normal distributions with the same variance but opposite means, the modified method reduced the MSE by 44.7%, 51.4%, and 56.0% for sample sizes 30, 100, and 300, respectively. For model 13, which is a mixture of a normal distribution and five normal distributions with very small variances, the modified method reduced the MSE by 0.0%, 0.1%, and 0.1% for sample sizes 30, 100, and 300, respectively. These results indicate that the modified method is more adaptive and robust to the shape and features of the PDF than the Silverman method, which assumes a normal distribution.

The simulation results also showed that the performance of both methods improved as the sample size increased, as expected. However, the improvement was more significant for the modified method than for the Silverman method. For example, for model 5, the MSE of the modified method decreased by 75.4%, 41.3%, and 46.2% as the sample size increased from 30 to 100, from 100 to 300, and from 30 to 300, respectively. For the Silverman method, the MSE decreased by 64.8%, 28.9%, and 40.4% for the same sample size changes, respectively. This suggests that the modified method is more efficient and reliable than the Silverman method, especially for large samples.

Table 1: Summary result of the simulation study when the sample size is n=30.

	MSE h_{S^*}	MSE h_{M^*}	MISE h_{S^*}	MISE h_{M^*}
Model 1	0.00227	0.00221	0.01557	0.01518
Model 2	0.00025	0.00024	0.00513	0.00501
Model 3	0.00003	0.00003	0.00171	0.00167
Model 4	0.00187	0.00177	0.01471	0.01396
Model 5	0.00130	0.00081	0.02661	0.01667
Model 6	0.00418	0.00385	0.02388	0.02221
Model 7	0.00174	0.00161	0.01533	0.01453
Model 8	0.00197	0.00184	0.01514	0.01426
Model 9	0.00216	0.00212	0.01747	0.01701
Model 10	0.00763	0.00422	0.07477	0.04139
Model 11	0.00294	0.00286	0.02179	0.02107
Model 12	0.02657	0.02533	0.18517	0.17547
Model 13	0.22065	0.22066	1.33238	1.33228
Model 14	0.00483	0.00482	0.03632	0.03623

Table 2: Summary result of the simulation study when the sample size is n=100.

	MSE h_{S^*}	MSE h_{M^*}	MISE h_{S^*}	MISE h_{M^*}
Model 1	0.00080	0.00079	0.00597	0.00590
Model 2	0.00009	0.00009	0.00199	0.00197
Model 3	0.00001	0.00001	0.00065	0.00065
Model 4	0.00063	0.00062	0.00579	0.00563
Model 5	0.00083	0.00046	0.01642	0.00911
Model 6	0.00149	0.00144	0.00928	0.00901
Model 7	0.00057	0.00056	0.00605	0.00599
Model 8	0.00067	0.00064	0.00590	0.00567
Model 9	0.00107	0.00096	0.00862	0.00773
Model 10	0.00502	0.00244	0.04660	0.02269
Model 11	0.00153	0.00137	0.01169	0.01044
Model 12	0.02116	0.02054	0.14784	0.14320
Model 13	0.20017	0.20007	1.32092	1.32025
Model 14	0.00354	0.00345	0.02749	0.02686

Table 3: Summary result of the simulation study when the sample size is n=300.

	MSE h_{S^*}	MSE h_{M^*}	MISE h_{S^*}	MISE h_{M^*}
Model 1	0.00033	0.00033	0.00256	0.00255
Model 2	0.00004	0.00004	0.00086	0.00085
Model 3	0.00000	0.00000	0.00029	0.00029
Model 4	0.00025	0.00024	0.00254	0.00249
Model 5	0.00051	0.00026	0.00964	0.00486
Model 6	0.00062	0.00062	0.00406	0.00405
Model 7	0.00021	0.00022	0.00268	0.00275
Model 8	0.00027	0.00027	0.00262	0.00256
Model 9	0.00054	0.00046	0.00433	0.00366
Model 10	0.00309	0.00136	0.02748	0.01206
Model 11	0.00085	0.00071	0.00661	0.00558
Model 12	0.01697	0.01663	0.11942	0.11694
Model 13	0.18772	0.18755	1.31503	1.31389
Model 14	0.00295	0.00282	0.02284	0.02185

Conclusion

In conclusion, the simulation study demonstrated that the modified method for selecting the bandwidth for univariate continuous PDFs using KDE outperformed the Silverman method in terms of MSE and MISE. The modified method was more flexible and sensitive to the characteristics of the PDF, and achieved higher accuracy and smoothness. The modified method also showed greater improvement as the sample size increased, indicating its efficiency and reliability. Therefore, we recommend the modified method over the Silverman method for estimating univariate continuous PDFs using KDE.

References

1. Alquicira-Hernandez J, Powell JE. Nebulosa recovers single-cell gene expression signals by kernel density estimation. Bioinformatics. 2021;37(16):2485-7.
2. Arora S, Taylor JW. Forecasting electricity smart meter data using conditional kernel density estimation. Omega. 2016;59:47-59.
3. Bowman AW, Azzalini A. Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations. Vol. 18. Oxford: OUP Oxford; c1997.
4. Corina AN, Hovda S. Automatic lithology prediction from well logging using kernel density estimation. Journal of Petroleum Science and Engineering. 2018;170:664-74.
5. Fleming CH, Calabrese JM. A new kernel density estimator for accurate home-range and species-range area estimation. Methods in Ecology and Evolution. 2017;8(5):571-9.
6. Hu B, Li Y, Yang H, Wang H. Wind speed model based on kernel density estimation and its application in reliability assessment of generating systems. Journal of Modern Power Systems and Clean Energy. 2017;5(2):220-7.
7. Jankowska MM, Natarajan L, Godbole S, Meseck K, Sears DD, Patterson RE, *et al*. Kernel density estimation as a measure of environmental exposure related to insulin resistance in breast cancer survivors. Cancer Epidemiology, Biomarkers & Prevention. 2017;26(7):1078-84.
8. Kile H. Bandwidth selection in Kernel density estimation [master's thesis]. Trondheim: Institutt for matematiske fag; c2010.
9. Lahane P, Sangaiah AK. An approach to EEG based emotion recognition and classification using kernel density estimation. Procedia Computer Science. 2015;48:574-81.
10. Marron JS, Wand MP. Exact mean integrated squared error. The Annals of Statistics. 1992;20(2):712-36.
11. Parzen E. On estimation of a probability density function and mode. The Annals of Mathematical Statistics. 1962;33(3):1065-76.
12. Rosenblatt M. Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics. 1956;27:832-7.
13. Rudemo M. Empirical choice of histograms and kernel density estimators. Scandinavian Journal of Statistics. 1982;9(2):65-78.
14. Savchuk OY, Hart JD, Sheather SJ. Indirect cross-validation for density estimation. Journal of the American Statistical Association. 2010;105(489):415-23.
15. Scott DW, Terrell GR. Biased and unbiased cross-validation in density estimation. Journal of the American Statistical Association. 1987;82(400):1131-46.
16. Scott DW, Tapia RA, Thompson JR. Kernel density estimation revisited. Nonlinear Analysis: Theory, Methods & Applications. 1977;1(4):339-72.
17. Sheather SJ, Jones MC. A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society: Series B (Methodological). 1991;53(3):683-90.
18. Silverman BW. Density estimation for statistics and data analysis. Vol. 26. Boca Raton: CRC Press; c1986.
19. Stone CJ. An asymptotically optimal window selection rule for kernel density estimates. The Annals of Statistics. 1984;12(4):1285-97.
20. Tenreiro C. Bandwidth selection for kernel density estimation: A Hermite series-based direct plug-in approach. Journal of Statistical Computation and Simulation. 2020;90(18):3433-53.
21. Wand MP, Jones MC. Kernel smoothing. Boca Raton: CRC Press; c1994.