Communications for Statistical Applications and Methods
2024, Vol. 31, No. 3, 337–347

https://doi.org/10.29220/CSAM.2024.31.3.337
Print ISSN 2287-7843 / Online ISSN 2383-4757

The shifted Chebyshev series-based plug-in for bandwidth selection in kernel density estimation

Soratja Klaichim[ab], Juthaphorn Sinsomboonthong[a], Thidaporn Supapakorn[1,a]

[a]Department of Statistics, Faculty of Science, Kasetsart University, Thailand;
[b]Faculty of Liberal Arts, Rajamangala University of Technology Rattanakosin, Thailand

Abstract

Kernel density estimation is a prevalent technique employed for nonparametric density estimation, enabling direct estimation from the data itself. This estimation involves two crucial elements: selection of the kernel function and the determination of the appropriate bandwidth. The selection of the bandwidth plays an important role in kernel density estimation, which has been developed over the past decade. A range of methods is available for selecting the bandwidth, including the plug-in bandwidth. In this article, the proposed plug-in bandwidth is introduced, which leverages shifted Chebyshev series-based approximation to determine the optimal bandwidth. Through a simulation study, the performance of the suggested bandwidth is analyzed to reveal its favorable performance across a wide range of distributions and sample sizes compared to alternative bandwidths. The proposed bandwidth is also applied for kernel density estimation on real dataset. The outcomes obtained from the proposed bandwidth indicate a favorable selection. Hence, this article serves as motivation to explore additional plug-in bandwidths that rely on function approximations utilizing alternative series expansions.

Keywords: kernel density estimation, Chebyshev, shifted Chebyshev, plug-in, bandwidth

1. Introduction

Density estimation is indeed the process of constructing an estimate of the probability density function (pdf) from an available data. This estimation not only represents the data distribution but also provides summary statistics such as the mean, median, variance, moments and quantiles. Furthermore, density estimates provide information about distribution characteristics, including skewness, kurtosis, and multimodality within the data. The estimation of pdf is a fundamental concept in statistics and a widely researched topic. There are two commonly methods for density estimation: parametric and nonparametric methods. The parametric method assumes that the data is drawn from a known distribution, whereas the nonparametric method aims to estimate the density function directly from the data. Several nonparametric density estimation techniques commonly include the histogram, naïve density estimator, nearest neighbor method, and orthogonal series estimator. Kernel density estimation (KDE) is a widely used nonparametric density estimation. The KDE relies on the kernel function which determines the weight assigned to each data point, and the bandwidth which controls the smoothness of the estimate. Hence, the selection of the bandwidth is the most crucial in the context of kernel density estimation.

[1] Corresponding author: Department of Statistics, Faculty of Science, Kasetsart University, Bangkok, Thailand 10903.
E-mail: fscitdps@ku.ac.th

Published 31 May 2024 / journal homepage: http://csam.or.kr

The selection of the bandwidth is a critical issue that arises in the context of KDE, as the performance of the KDE depends on the chosen bandwidth. A small bandwidth value results in an undersmoothed density, while a large bandwidth value leads to an oversmoothed density (Gramacki, 2018). There are various methods to determine a bandwidth for KDE. The primary categories of bandwidths for KDE include rules-of-thumb (ROT), cross-validation (CV), and plug-in (PI). The plug-in method has been demonstrated to offer excellent performance in many cases. Due to its demonstrated excellent performance, the plug-in method is the common first choice in practical applications. However, there is still room for further improvement in its implementation (Wand and Jones, 1995). Plug-in bandwidths are operated on the straightforward concept of substituting estimated values of the unknown quantities into formulas for achieving the asymptotically optimal bandwidth.

In order to overcome optimal bandwidth, this study introduces the proposed plug-in bandwidth. This bandwidth leverages the first kind shifted Chebyshev polynomials, providing a solution to the problem. The effectiveness of the methods relies on the estimation of integrated squared density derivative functionals, a subject that has been explored by many researchers. Silverman (1986) provided a comprehensive overview of density estimation techniques, including discussions on bandwidth selection and the role of integrated squared density derivative functionals. Sheather and Jones (1991) discussed a data-driven method for bandwidth selection in kernel density estimation, which related to integrated squared density derivative functionals. Raykar and Duraiswami (2006) developed the algorithms for estimating density derivatives using the univariate Gaussian kernel. These algorithms are utilized to calculate the optimal bandwidth for kernel density estimation. Tenreiro (2011, 2020) proposed direct plug-in bandwidth for the KDE based on the Fourier series and the Hermite series. In a recent study, Dharmani (2022) introduced a bandwidth selection by employing the near Gaussian assumption. This assumption enables the use of the Gram-Charlier A series as an approximation to the function for the purpose of estimating its density derivative. The objective of this paper is to derive a bandwidth by using the first kind shifted Chebyshev polynomials as an approximation to the density function. This is aimed at estimating the integrated squared density derivative functionals.

The remaining sections of this article are organized as follows. Section 2 provides an overview of the fundamental properties of kernel density estimation. In Section 3, various methods for bandwidth selection are discussed. These methods include least squares cross-validation bandwidth, an improved version of rules of thumb bandwidth, and the Sheather and Jones plug-in bandwidth. Section 4 offers a brief definition of the first kind shifted Chebyshev polynomials, then utilizes them to approximate the underlying density function, and finally presents the proposed plug-in bandwidth based on this estimator. Section 5 presents a simulation study of the proposed bandwidth, examining its performance under different distributions and sample sizes using the R programming language. Additionally, Section 6 applies the proposed bandwidth to real dataset. Finally, Section 7 concludes the article with a summary of the findings.

2. Kernel density estimation

The kernel density estimator for a random sample $X_1, X_2, \ldots, X_n$ drawn from a common and typically unknown density $f(x)$, as defined by Rosenblatt (1956) and Parzen (1962), is expressed as

$$\hat{f}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{2.1}$$

where $K(x)$ is the kernel function and h is the bandwidth with positive value.

The kernel function $K(x)$ plays as the weight function and satisfies the following properties: $K(x) \geq 0, K(x) = K(-x), \int K(x)\,dx = 1, \int xK(x)\,dx = 0$ and $k_2 = \int x^2 K(x)\,dx \neq 0$. The bandwidth h determines the level of smoothness of the density estimate.

In practice, it is common to consider a global error criterion that measures the distance between the estimated density function $\hat{f}(x;h)$ and the true density function $f(x)$. One such error criterion is the integrated squared error (ISE) given by ISE $(\hat{f}(x;h)) = \int_{-\infty}^{\infty}[\hat{f}(x;h) - f(x)]^2 dx$. A more appropriate approach would involve analyzing the expected value of this quantity, known as the mean integrated square error (MISE), which is defined as

$$\begin{aligned}\text{MISE}\left(\hat{f}(x;h)\right) &= \text{E}\int_{-\infty}^{\infty}\left[\hat{f}(x;h) - f(x)\right]^2 dx\\ &= \int_{-\infty}^{\infty}\text{Bias}^2\left(\hat{f}(x;h)\right)dx + \int_{-\infty}^{\infty}\text{Var}\left(\hat{f}(x;h)\right)dx\\ &= \frac{1}{4}h^4 k_2^2\int\left(f''(x)\right)^2 dx + \frac{1}{nh}k_0 + o\left\{(nh)^{-1} + h^4\right\}, \end{aligned} \tag{2.2}$$

where $k_0 = \int (K(x))^2\,dx$ and $k_2 = \int x^2 K(x)\,dx$ (Gramacki, 2018).

The assumptions are $f(x)$ is assumed to be sufficiently smooth: Its second derivative $f''(x)$ is bounded, continuous and square integrable. Also, if $(nh)^{-1} \to 0$ and $h \to 0$ as $n \to \infty$ then MISE$\hat{f}(x;h) \to 0$. It obtains the asymptotic mean integrated square error (AMISE) as follows:

$$\begin{aligned}\text{AMISE}\,\hat{f}(x;h) &= \frac{1}{4}h^4 k_2^2\int\left(f''(x)\right)^2 dx + \frac{1}{nh}k_0\\ &= \frac{1}{4}h^4 k_2^2\theta_2 + \frac{1}{nh}k_0\,, \end{aligned} \tag{2.3}$$

where $\theta_2 = \int (f''(x))^2\,dx$, $k_0 = \int (K(x))^2\,dx$ and $k_2 = \int x^2 K(x)\,dx$.

3. Bandwidth selection

Several methods are available for determining appropriate bandwidth for KDE. The three main types of bandwidths are as follows: Cross-validation (CV), rules-of-thumb (ROT) and plug-in (PI) (Gramacki, 2018). Cross-validation involves techniques like least squares cross-validation (LSCV), biased cross-validation (BCV), and smoothed cross-validation (SCV). Rules-of-thumb includes approaches such as Silverman's rule of thumb and its improved version. Plug-in methods have been explored by various authors, including Park and Marron (1990), Sheather and Jones (1991), and Hall *et al.* (1991), among others. This article provides concise explanations of the following methods chosen for study: least squares cross-validation, the improved version of Silverman's rule of thumb, and Sheather and Jones plug-in bandwidth.

3.1. Least squares cross validation

A well-known method for selecting the bandwidth is least squares cross-validation (LSCV), proposed by Rudemo (1982) and Bowman (1984). The main objective is to find the optimal bandwidth h that minimizes the ISE using the estimator $\hat{f}(x;h)$ for density $f(x)$. The integrated squared error of $\hat{f}(x;h)$

is represented as

$$\begin{aligned}\text{ISE}\,\hat{f}(x;h) &= \int \left[\hat{f}(x;h) - f(x)\right]^2 dx \\ &= \int \hat{f}(x;h)^2\,dx - 2\int \hat{f}(x;h)\,f(x)\,dx + \int f(x)^2\,dx. \end{aligned} \tag{3.1}$$

(Silverman, 1986; Wand and Jones, 1995).

The first term $\int \hat{f}(x;h)^2\,dx$ of (3.1) can be calculated from the data which was proved by Härdle (1991) as

$$\int \hat{f}(x;h)^2\,dx = \frac{1}{n^2 h}\sum_{i=1}^{n}\sum_{j=1}^{n} K * K\left(\frac{X_j - X_i}{h}\right), \tag{3.2}$$

where $K * K(x)$ is the convolution of $K(x)$.

The second term $\int \hat{f}(x;h)\,f(x)\,dx$ of (3.1), which depends on h and involves the unknown density $f(x)$, has to be estimated. Notice that, $\int \hat{f}(x;h)\,f(x)\,dx$ is the expected value of $\hat{f}(x;h)$ which can be estimated by

$$\text{E}\left[\hat{f}(x;h)\right] = \frac{1}{n}\sum_{i=1}^{n}\hat{f}_{-i}(X_i;h), \tag{3.3}$$

where $\hat{f}_{-i}(X_i;h) = (1/((n-1)h))\sum_{j\neq i}^{n} K((x - X_j)/h)$ is the estimator based on the sample with X_i deleted.

Finally, the last term $\int f(x)^2\,dx$ of (3.1) is independent of the bandwidth h. Therefore, the last term can be moved to the left side of the equation and can be written as

$$\text{ISE}\,\hat{f}(x;h) - \int f(x)^2\,dx = \int \hat{f}(x;h)^2\,dx - 2\int \hat{f}(x;h)\,f(x)\,dx. \tag{3.4}$$

As a result, Equations (3.2) and (3.3) are inserted into the Equation (3.4), leading to the least squares cross-validation function as

$$\text{LSCV}(h) = \frac{1}{n^2 h}\sum_{i=1}^{n}\sum_{j=1}^{n} K * K\left(\frac{X_j - X_i}{h}\right) - \frac{2}{n(n-1)h}\sum_{i=1}^{n}\sum_{j\neq i}^{n} K\left(\frac{X_i - X_j}{h}\right) \tag{3.5}$$

(Härdle *et al.*, 2004). The bandwidth that minimizes the function LSCV (h) is denoted by h_{LSCV}.

3.2. Rules of thumb

The optimal bandwidth minimizes AMISE $\hat{f}(x;h)$ with respect to h, and solving the first partial derivative with respect to h yields

$$h_{\text{AMISE}} = \left[\frac{k_0}{k_2^2\theta_2 n}\right]^{\frac{1}{5}}, \tag{3.6}$$

where $\theta_2 = \int (f''(x))^2\,dx$, $k_0 = \int (K(x))^2\,dx$ and $k_2 = \int x^2 K(x)\,dx$.

The rule-of-thumb bandwidth is determined by replacing the density function $f(x)$ with the normal distribution having zero mean and variance σ^2 and using the Gaussian kernel function in Equation (3.6). The rule-of-thumb bandwidth, denoted by h_{ROT1}, is calculated using the formula

$$h_{\mathrm{ROT1}} = 1.06\sigma n^{-\frac{1}{5}}. \tag{3.7}$$

The rule-of-thumb bandwidth is sensitive to outliers, which cause an overestimation of σ and lead to a larger bandwidth. To make the estimator more robust, the interquartile range (IQR) is used. The improved version of the rule-of-thumb bandwidth, denoted as h_{ROT2}, is defined as

$$h_{\mathrm{ROT2}} = 1.06 n^{-\frac{1}{5}} \min\left(\sigma, \frac{\mathrm{IQR}}{1.34}\right) \tag{3.8}$$

(Härdle *et al.*, 2004).

3.3. Plug-in

The concept of plug-in bandwidth was originally introduced by Woodroofe (1970). This concept is based on the idea of using an optimal bandwidth that minimizes $\mathrm{AMISE}(\hat{f}(x; h_{\mathrm{SCBS}}))$. Plug-in bandwidth is based on the substitution of the unknown quantity $\theta_2 = \int (f''(x))^2\,dx$. Sheather and Jones (1991) also proposed the plug-in bandwidth, denoted as h_{SJDP}. They provided a solution for this bandwidth selection method as

$$h_{\mathrm{SJDP}} = \left[\frac{k_0}{k_2^2 \hat{\psi}_4(\gamma(h))\, n}\right]^{\frac{1}{5}}. \tag{3.9}$$

The pilot bandwidth for the estimation of ψ_4 is a function γ of h. The choice of γ is defined by

$$\gamma(h) = \left(\frac{2K^{(4)}(0)\, k_2}{k_0}\right)^{\frac{1}{7}} \left(-\frac{\hat{\psi}_4(g_1)}{\hat{\psi}_6(g_2)}\right)^{\frac{1}{7}} h^{\frac{5}{7}}, \tag{3.10}$$

where $\hat{\psi}_4(g_1)$ and $\hat{\psi}_6(g_2)$ are kernel estimates of ψ_4 and ψ_6. The choice of g_1 and g_2 are formulated by

$$g_1 = \left(\frac{-2K^{(4)}(0)}{\hat{\psi}_6 k_2 n}\right)^{\frac{1}{7}} \quad \text{and} \quad g_2 = \left(\frac{-2K^{(6)}(0)}{\hat{\psi}_8 k_2 n}\right)^{\frac{1}{9}}, \tag{3.11}$$

where $\hat{\psi}_6 = -15/(16\sqrt{\pi}\hat{\sigma}^7)$, $\hat{\psi}_8 = 105/(32\sqrt{\pi}\hat{\sigma}^9)$, $K^{(4)}(0) = 3/\sqrt{2\pi}$ and $K^{(6)}(0) = -15/\sqrt{2\pi}$ (Wand and Jones, 1995).

4. The shifted Chebyshev series based plug-in bandwidth

Bandwidth selection in KDE using the AMISE criteria involves estimating the second-order derivative of the unknown density being estimated. The first kind shifted Chebyshev series expansion can be used as an approximation method for an unknown density function. This section will cover the necessary background on the first kind shifted Chebyshev polynomials and derive the bandwidth.

4.1. The shifted Chebyshev polynomials

The first kind Chebyshev polynomials of degree m, where $m \in \{0, 1, 2, \ldots\}$, are denoted as $T_m(x)$ and defined on the interval $[-1, 1]$. A more general recurrence relation is

$$T_{m+1}(x) = 2xT_m(x) - T_{m-1}(x) \quad \text{with} \quad T_0(x) = 1 \quad \text{and} \quad T_1(x) = x. \tag{4.1}$$

In order to use the first kind Chebyshev polynomials on a finite range $[a, b]$, a transformation can be applied to generate the so-called the first kind shifted Chebyshev polynomials. This transformation involves using the equation

$$y = \left(\frac{b-a}{2}\right)x + \left(\frac{b+a}{2}\right), \text{ then } x = \left(\frac{2}{b-a}\right)y + \left(\frac{b+a}{b-a}\right). \tag{4.2}$$

Afterward, the first kind shifted Chebyshev polynomials are generated by

$$T_m^*(x) = T_m\left(\frac{2}{b-a}y - \frac{b+a}{b-a}\right), \text{ for } x \in [a, b]. \tag{4.3}$$

In the context of the interval $[a, b]$, the approximating function can be approximated using the first kind shifted Chebyshev series as

$$f(y) = \sum_{i=0}^{m-1} c_i T_i^*(y), \tag{4.4}$$

where the coefficients are defined via the formula

$$c_i = \frac{2}{m}\sum_{i=0}^{m} f\left(\frac{b-a}{2}\tilde{x_k} + \frac{b+a}{2}\right) T_i(\tilde{x_k}), \tag{4.5}$$

where

$$\tilde{x_k} = \cos\left(\frac{2k-1}{2m}\right)\pi, \quad k = 0, 1, \ldots, m-1 \tag{4.6}$$

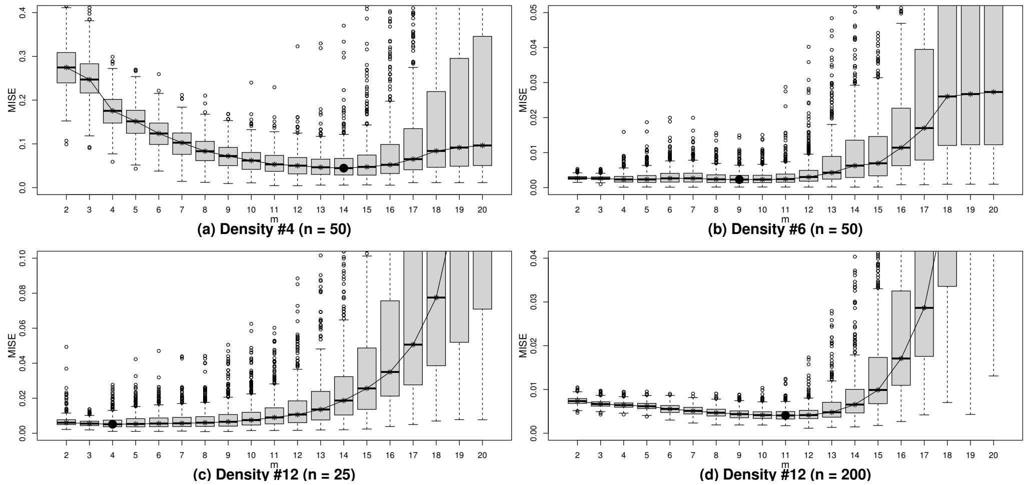
is the Chebyshev zero nodes.

The integration of the squared function of the second-order derivative of the first kind shifted Chebyshev series expansion $\theta_2 = \int (f''(y))^2\, dy$ can be expressed as

$$\begin{aligned} \theta_{2,m} &= \int (f''(y))^2\, dy \\ &= \frac{4}{m^2}\int \left[\sum_{i=0}^{m-3}\left(\sum_{i=1}^{m} f\left(\frac{b-a}{2}\tilde{x_k} + \frac{b+a}{2}\right) T_i(\tilde{x_k})\right) T_i^{*\prime\prime}(x)\right]^2 dy. \end{aligned} \tag{4.7}$$

By substituting the values of θ_2 from (4.7) into the expressions for the optimal bandwidth derived in (3.6), the resulting bandwidth can be described as the first kind shifted Chebyshev series-based bandwidth:

$$h_{\text{SCBS}} = \left[\frac{k_0}{k_2^2\theta_{2,m}n}\right]^{\frac{1}{5}}. \tag{4.8}$$



(a) Density #4 (n = 50)

(b) Density #6 (n = 50)

(c) Density #12 (n = 25)

(d) Density #12 (n = 200)

Figure 1: *Boxplot of MISE* $(\hat{f}(x; h_{SCBS}))$ *with the number of term* m.

4.2. The optimal value of m

The bandwidth h_{SCBS} is influenced by the number of terms m in the first kind shifted Chebyshev expansion used in estimating $\theta_{2,m}$. The performance and smoothness of this expansion are also affected by m, acting as a smoothing parameter and representing the number of terms in the series expansion. The best choice for the number of terms is the smallest m that results in the lowest ISE. Then, the error in function expansion using the first kind shifted Chebyshev expansion is determined by MISE $(\hat{f}(x; h_{\text{SCBS}})) = \text{E} \int [\hat{f}(x; h_{\text{SCBS}}) - f(x)]^2 dx$.

5. Simulation study

In this section, the aim is to evaluate the performance of the proposed bandwidth h_{SCBS} and compare with three other bandwidths used for density estimation: least squares cross-validation bandwidth (h_{LSCV}), the improved version of the rules of thumb bandwidth (h_{ROT2}), and the Sheather and Jones plug-in bandwidth (h_{SJDP}). This study compares different bandwidths for density estimation using fifteen normal mixture densities constructed by Marron and Wand (1992). These densities include various shapes, such as unimodal, bimodal, trimodal, and multimodal, each defined and visualized in Marron and Wand's work. For each distribution, sample sizes of n = 25, 50, 100, 150, and 200 are considered, and the MISE of the estimator is computed over 500 replications. The Gaussian kernel function is used in all cases.

The main idea is to find the optimal number of terms in the expansion (m) that allows a good approximation of $f(x)$ using the first kind shifted Chebyshev series expansion, in the sense of the mean integrated squared error (MISE). The results have been presented through box plots showing the estimated MISE as a function of the number of terms in the expansion, as shown in Figure 1. This graph illustrates the influence of the terms m on MISE across three density distributions (#4, #6, and #12). The x-axis labels represent the sequential terms in the expansion, and the medians of MISE are displayed in the box plots. Furthermore, a solid circle is used to indicate the optimal number of terms in the expansion, which corresponds to the smallest MISE. Figure 1(a) displays the MISE of Density #4 with sample size of n = 50, while Figure 1(b) shows the MISE of Density #6 with sample size n =

Table 1: MISE $(\hat{f}(x; h_{\text{SCBS}})) \times 10^{-3}$ bases on the bandwidths h_{LSCV}, h_{ROT2}, h_{SJDP} and h_{SCBS} with 500 replications for each case

	n	h_{LSCV}	h_{ROT2}	h_{SJDP}	h_{SCBS}
Density #1	25	10.7819	4.8937	7.7200	**2.9581**
	50	4.4220	2.7017	3.7667	**2.0824**
	100	2.5376	1.5025	1.8990	**1.2831**
	150	1.8510	1.1071	1.3038	**0.9956**
	200	1.2925	0.9060	1.0161	**0.8406**
Density #2	25	15.8366	9.6809	14.2788	**6.8853**
	50	8.0684	5.2273	6.4727	**4.4258**
	100	4.3664	2.9407	3.3472	**2.6654**
	150	3.2559	2.2902	2.5283	**2.1106**
	200	2.7399	2.0048	2.1457	**1.8822**
Density #3	25	129.2468	154.7279	90.0448	**72.9098**
	50	62.6229	143.2790	64.9015	**44.6778**
	100	36.2592	131.1196	46.5380	**22.7486**
	150	24.5300	125.3598	35.7687	**19.9659**
	200	19.9220	117.4577	30.3761	**16.7587**
Density #4	25	219.2996	172.4827	136.0272	**84.3708**
	50	89.2615	136.3549	74.9258	**51.4540**
	100	39.6160	114.3767	41.9968	**33.0942**
	150	27.6475	98.8623	28.8491	**26.2868**
	200	22.6444	90.6911	23.0213	**22.9113**
Density #5	25	827.9669	471.6216	753.9018	**292.4579**
	50	366.6813	213.6145	276.7979	**191.1883**
	100	197.7864	120.3047	143.0313	**117.9369**
	150	139.2668	91.8233	103.4045	**90.9849**
	200	110.0786	76.4904	84.0808	**75.2508**
Density #6	25	7.9374	3.9941	6.1431	**3.7115**
	50	4.6343	**2.4640**	3.0836	2.6055
	100	2.3224	1.7155	1.7884	**1.6465**
	150	1.7449	1.3714	1.3327	**1.2983**
	200	1.4163	1.1692	1.1060	**1.1312**
Density #7	25	16.2093	18.0853	8.5167	**7.8264**
	50	8.0472	14.4206	4.7440	**4.6070**
	100	3.8401	11.2955	2.8610	**2.7684**
	150	2.9414	9.5903	2.1900	**2.1349**
	200	2.3237	8.4671	1.7893	**1.7445**
Density #8	25	12.2326	6.2189	8.5853	**5.4875**
	50	6.9255	4.0860	4.6909	**3.9579**
	100	3.5754	2.9847	2.7308	**2.6340**
	150	2.6956	2.4547	2.1257	**1.9760**
	200	2.2223	2.1839	1.7520	**1.6741**
Density #9	25	10.4642	4.1865	5.6758	**4.2808**
	50	4.2284	2.9857	3.1319	**2.8806**
	100	2.6069	2.1917	2.0087	**1.8892**
	150	1.8859	1.8096	1.5154	**1.4570**
	200	1.5083	1.5946	1.2736	**1.2083**
Density #10	25	36.8391	23.5191	26.6129	**22.3816**
	50	26.1360	21.0004	21.4497	**20.6412**
	100	19.5172	20.0515	19.5826	**14.7539**
	150	14.3548	19.3202	18.6060	**11.9755**
	200	10.8535	18.9143	17.9090	**10.5816**
Density #11	25	8.0731	4.5490	6.4799	**4.1913**
	50	4.9207	**3.0088**	3.4638	3.0759
	100	2.7689	2.2636	2.2999	**2.1852**
	150	2.2054	1.8821	1.8253	**1.8198**
	200	1.8125	1.6684	1.5879	**1.6174**
Density #12	25	16.7257	9.6699	12.1060	**8.6525**
	50	10.4177	8.1942	8.6793	**7.9743**
	100	7.8214	7.5209	7.3742	**6.0648**
	150	6.0406	7.1254	6.6118	**4.9554**
	200	4.8533	6.7446	5.9887	**4.2651**
Density #13	25	10.6791	5.8031	7.6241	**5.7830**
	50	6.4439	4.3061	4.7116	**4.3038**
	100	4.0942	3.4850	3.3782	**3.2327**
	150	3.3745	3.0413	2.8025	**2.7256**
	200	3.0361	2.8499	2.6000	**2.5457**
Density #14	25	32.0359	24.3535	15.4503	**13.6719**
	50	15.4430	22.3324	11.9413	**10.5232**
	100	10.0967	20.7340	9.5673	**8.0145**
	150	7.9991	19.4951	8.1350	**6.7710**
	200	6.4908	18.5153	7.1848	**5.9495**
Density #15	25	20.8581	28.7160	19.9445	**15.7466**
	50	14.3004	27.7959	12.0152	**10.7535**
	100	10.2774	26.6659	8.6503	**8.1826**
	150	8.5594	25.5584	7.5136	**7.1010**
	200	7.1760	24.7048	6.7325	**6.1503**

50. Figure 1(c) and Figure 1(d) show MISE for Density #12 with two different sample sizes n = 25 and n = 200, respectively. The performance is influenced by the combination of density and sample size, and specific combinations yield better results.

The simulation results evaluate the performance of the plug-in bandwidth by finding the bandwidth that minimizes the mean integrated squared error, MISE $(\hat{f}(x; h_{\text{SCBS}}))$. Table 1 shows the MISE values for different bandwidths, and the bold text value indicates the smallest MISE associated with the bandwidths h_{LSCV}, h_{ROT2}, h_{SJDP} and h_{SCBS}. Overall, the proposed bandwidth h_{SCBS} demonstrates good performance compared to the other bandwidths, except for Density #6 (n = 50) and #11 (n = 50), where the improved version of rules of thumb bandwidth (h_{ROT2}) shows better performance. The simulation results indicate that the suggested bandwidth is a good choice for various scenarios.

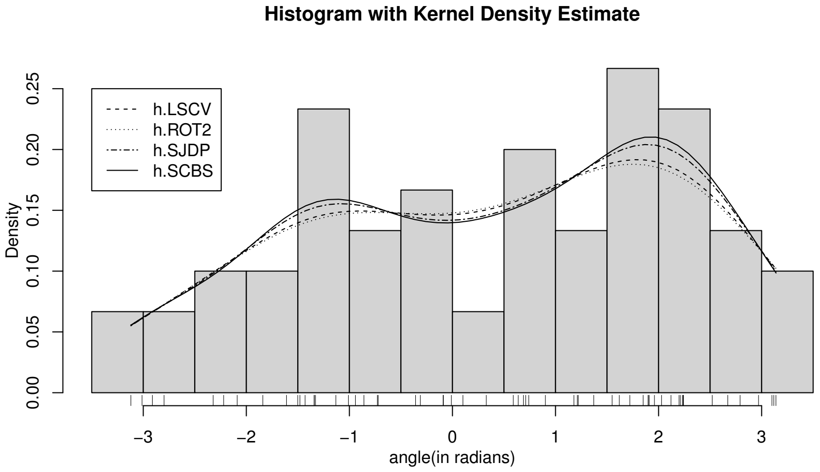


Figure 2: *Histogram and kernel density estimates for the "flywheels" dataset.*

Table 2: MSE $(\hat{f}(x;h_{\text{SCBS}})) \times 10^{-3}$ for kernel density estimates with varying bandwidths

Bandwidth	MSE
h_{LSCV}	1.9152
h_{ROT2}	2.0158
h_{SJDP}	1.6829
h_{SCBS}	**1.6156**

It provides excellent performance compared to all the other bandwidths under consideration, even though the calculation is quite complex.

6. Real data analysis

In this section, kernel density estimation is applied to real datasets. The performance of the proposed bandwidth h_{SCBS} is verified against other bandwidths. All calculations are performed using the R programming language.

A real dataset named "flywheels" from Anderson-Cook (1999) and comprising 60 observations on flywheel imbalance angles, will be utilized. This analysis focuses on how different bandwidth choices influence kernel density estimates and histograms, serving as methods to understand data distribution. Figure 2 displays a histogram with 14 bins for this dataset. The density seems to exhibit asymmetric bimodal behavior. The kernel density estimate, using different bandwidth options such as h_{LSCV}, h_{ROT2}, h_{SJDP} and the proposed h_{SCBS}, is also overlaid on Figure 2. Different bandwidth options will be compared to find the best approach. The kernel density estimate using the suggested bandwidth h_{SCBS} fits well across the dataset, as confirmed in Table 2 by the lowest mean square error (MSE) value for h_{SCBS}.

7. Conclusion

When selecting bandwidth for kernel density estimation, the direct plug-in method is the common initial approach, but there is room for enhancement. Estimating the bandwidth involves finding the integration of the squared function of the second-order derivative of the unknown density to be estimated. This article introduces a bandwidth selection technique by incorporating the estimation of $\theta_{2,m} = \int (f''(x))^2\, dx$ through the first kind shifted Chebyshev polynomials as an approximation to the function $f(x)$.

The simulation studies revealed that the first kind shifted Chebyshev series-based plug-in bandwidth (h_{SCBS}) performs well among other bandwidths, such as least squares cross-validation bandwidth, improved rule of thumb bandwidth, and Sheather and Jones plug-in bandwidth. When applying kernel density estimation to estimate the density of the “flywheels” dataset, the results indicate that the proposed bandwidth (h_{SCBS}), with the lowest MSE, offers the best performance compared to other bandwidth methods. Even though obtaining the proposed bandwidth might be complex, it is still a favorable choice for practical application.

Acknowledgement

The authors would like to thank Kasetsart University and Rajamangala University of Technology Rattanakosin for the support.

References

Anderson-Cook CM (1999). A tutorial on one-way analysis of circular-linear data, *Journal of Quality Technology*, **31**, 109–119.

Bowman AW (1984). An alternative method of cross-validation for the smoothing of density estimates, *Biometrika*, **71**, 353–360.

Dharmani B (2022). Gram-charlier a series based extended rule-of-thumb for bandwidth selection in univariate kernel density estimation, *Austrian Journal of Statistics*, **51**, 141–163.

Gramacki A (2018). *Nonparametric Kernel Density Estimation and Its Computational Aspects*, Spring er, Switzerland.

Hall P, Sheather SJ, Jones MC, and Marron JS (1991). On optimal data-based bandwidth selection in kernel density estimation, *Biometrika*, **78**, 263–269.

Härdle W (1991). *Smoothing Techniques: With Implementation in S*, Springer, New York.

Härdle W, Müller M, Sperlich S, and Werwatz A (2004). *Nonparametric and Semiparametric Models*, Springer, Berlin.

Marron JS and Wand MP (1992). Exact mean integrated squared error, *The Annals of Statistics*, **20**, 712–736.

Park BU and Marron JS (1990). Comparison of data-driven bandwidth selectors, *Journal of the American Statistical Association*, **85**, 66–72.

Parzen E (1962). On estimation of a probability density function and mode, *The Annals of Mathematical Statistics*, **33**, 1065–1076.

Raykar VC and Duraiswami R (2006). Fast optimal bandwidth selection for kernel density estimation, In *Proceedings of the 2006 SIAM International Conference on Data Mining*, Bethesda, MD, 524–528.

Rosenblatt M (1956). A central limit theorem and a strong mixing condition, *Proceedings of the National Academy of Sciences of the United States of America*, **42**, 43–47.

Rudemo M (1982). Empirical choice of histograms and kernel density estimators, *Scandinavian Journal of Statistics*, **9**, 65–78.

Silverman BW (1986). *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London.

Sheather SJ and Jones MC (1991). A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society: Series B (Methodological)*, **53**, 683–690.

Tenreiro C (2011). Fourier series-based direct plug-in bandwidth selectors for kernel density estimation, *Journal of Nonparametric Statistics*, **23**, 533–545.

Tenreiro C (2020). Bandwidth selection for kernel density estimation: A Hermite series-based direct plug-in approach, *Journal of Statistical Computation and Simulation*, **90**, 3433–3453.

Wand MP and Jones MC (1995). *Kernel Smoothing*, Chapman & Hall/CRC, New York.

Woodroofe M (1970). On choosing a delta-sequence, *The Annals of Mathematical Statistics*, **41**, 1665–1671.

Received September 17, 2023; Revised January 26, 2024; Accepted February 12, 2024