mathematics

MDPI

Article

An Improved Variable Kernel Density Estimator Based on L_2 Regularization

Yi Jin [1,2], Yulin He [3,4,*] and Defa Huang [3]

[1] Department of Trace Inspection Technology, Criminal Investigation Police University of China, Shenyang 110854, China; jinyi@cipuc.edu.cn
[2] Key Laboratory of Impression Evidence Examination and Identification Technology, The Ministry of Public Security of the People's Republic of China, Shenyang 110854, China
[3] Big Data Institute, College of Computer Science & Software Engineering, Shenzhen University, Shenzhen 518060, China; huangdefa2017@email.szu.edu.cn
[4] National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen 518060, China
* Correspondence: yulinhe@szu.edu.cn; Tel.: +86-18531315747

Abstract: The nature of the kernel density estimator (KDE) is to find the underlying probability density function (*p.d.f*) for a given dataset. The key to training the KDE is to determine the optimal bandwidth or Parzen window. All the data points share a fixed bandwidth (scalar for univariate KDE and vector for multivariate KDE) in the fixed KDE (FKDE). In this paper, we propose an improved variable KDE (IVKDE) which determines the optimal bandwidth for each data point in the given dataset based on the integrated squared error (ISE) criterion with the L_2 regularization term. An effective optimization algorithm is developed to solve the improved objective function. We compare the estimation performance of IVKDE with FKDE and VKDE based on ISE criterion without L_2 regularization on four univariate and four multivariate probability distributions. The experimental results show that IVKDE obtains lower estimation errors and thus demonstrate the effectiveness of IVKDE.

Keywords: probability density function; kernel density estimation; Parzen window; bandwidth; kernel function

Citation: Jin, Y.; He, Y.; Huang, D. An Improved Variable Kernel Density Estimator Based on L_2 Regularization. *Mathematics* **2021**, *9*, 2004. https://doi.org/10.3390/math9162004

Academic Editor: Filipe J. Marques

Received: 25 June 2021
Accepted: 17 August 2021
Published: 21 August 2021

1. Introduction

It is very important for many machine learning algorithms to estimate the unknown probability density functions (*p.d.f*.s) of given datasets, e.g., Bayesian classifiers [1,2], density-based clustering algorithms [3,4], and mutual information-based feature selection algorithms [5,6]. In order to obtain the unknown *p.d.f.*, an effective kernel density estimator (KDE) should be thoroughly constructed in advance. The classical KDE training method is the Parzen window method [7], which uses the superposition of multiple kernel functions with a fixed Parzen window (i.e., bandwidth) to fit the unknown *p.d.f.* The most used kernels [8] include uniform, triangular, Epanechnikov, biwieght, triweight, cosine, and Gaussian kernels. Compared with the kernels, the bandwidth plays a more important role in *p.d.f.* estimation: a large bandwidth will result in an over-smoothed estimation, while a small bandwidth will lead to an under-smoothed estimation.

How to determine an optimal bandwidth is a key point for training a KDE. In order to select an appropriate bandwidth, the effective error criterion should firstly be designed [9]. Commonly used error criteria include the integrated squared error (ISE) and the mean integrated squared error (MISE). Currently, there are two main ways to design KDE, i.e, the classical Parzen window method with the fixed bandwidth parameter named the fixed kernel density estimator (FKDE) and the modified Parzen window method with the variable bandwidth parameter named the variable kernel density estimator (VKDE). The representative studies corresponding to FKDE and VKDE are summarized as follows.

- Fixed kernel density estimator. The rule-of-thumb-based KDE (RoT-KDE) [10] was designed based on the asymptotic MISE (AMISE) criterion by assuming the unknown *p.d.f.* as normal *p.d.f.* Due to the inappropriate assumption of the true *p.d.f.*, RoT-KDE is a naive KDE and inclined to select the over-smoothed bandwidth [8]. Apart from the sample and direct RoT-KDE, there are three other sophisticated KDEs, i.e., bootstrap-based KDE (BS-KDE) [11] , biased cross-validation-based KDE (BCV-KDE) [12], and unbiased cross-validation-based KDE (UCV-KDE) [13]. BS-KDE determined the optimal bandwidth based on the MISE criterion by using the bootstrap technology to estimate the true *p.d.f.* BCV-KDE was also designed based on the MISE criterion, which calulated the optimal bandwidth by establishing the relationship between the true *p.d.f.* and the derivative of the estimated *p.d.f.* UCV-KDE used the ISE criterion to optimize the bandwidth by representing the true *p.d.f.* with the estimated leave-one-out *p.d.f.* In RoT-KDE, BS-KDE, BCV-KDE, and UCV-KDE, all samples in the given dataset enjoy a fixed bandwidth and do not use the bandwidth to adjust the roles of data points for *p.d.f.* estimation.
- Variable kernel density estimator. The model of VKDE was firstly proposed by Breiman et al. [14], who introduced the variable bandwidths for each data point in the given dataset and represented the bandwidth with distance from the data point to its k-th nearest neighbor. Jones [15] clarified the difference between VKDE employing a different bandwidth for each data point and VKDE with bandwidth as a function of estimation location. Terrell and Scott [16] derived the optimization rule for variable bandwidths based on the asymptotic mean squared error (AMSE) criterion. Hall et al. [17] improved the VKDE proposed in [16] by further analyzing the rates of VKDE convergence. Wu et al. [18] proposed a strategy to express the variable bandwidth in VKDE as the product of a local bandwidth factor and a global smoothing parameter. Suaray [19] proposed a VKDE for the *p.d.f.* estimation of censored data. Klebanov [20] proposed an axiomatic approach to construct a VKDE which guaranteed the density estimation invariance under linear transformations of original density as well as under splitting of density into several well-separated parts.

Compared with FKDEs, the main merit of VKDEs is that the variable bandwidths can flexibly adjust the importance of data points during the *p.d.f* estimation. This paper focuses on the improvement of VKDE. Jones [21] discussed the roles of ISE and MISE criteria in *p.d.f.* estimation. We consider using the ISE criterion to calculate the optimal bandwidths for the VKDE. The mathematical analysis indicates that the ISE criterion usually leads to an over-smoothed *p.d.f.* estimation. Inspired by the integration of empirical and structural risks, we propose an improved variable KDE (IVKDE) which determines the optimal bandwidth for each data point based on the ISE criterion with an L_2 regularization term in this paper. The ISE and L_2 regularization represent the empirical and structural risks for constructing VKDE, respectively. In order to obtain the optimally variable bandwidths, an effective optimization scheme is developed to solve the improved objective function. We conduct the exhaustive experiments to validate the rationality, feasibility, and effectiveness of IVKDE. The experimental results show that IVKDE is convergent and able to obtain the desirable *p.d.f.* estimation. In comparison with FKDE and VKDE based on the ISE criterion without L_2 regularization on four univariate and four multivariate probability distributions, IVKDE obtains lower estimation errors and thus demonstrate the effectiveness of IVKDE.

The remainder of this paper is organized as follows. In Section 2, we describe the basic principles of the variable kernel density estimator. In Section 3, we introduce the improved variable kernel density estimator. In Section 4, we provide experimental results and analysis. Finally, in Section 5, we conclude this paper and discuss future works.

2. Basic Principle of VKDE

For the given dataset $\mathbb{X} = \{\tilde{\mathrm{x}}_n | \tilde{\mathrm{x}}_n = (x_{n1}, x_{n2}, \cdots, x_{n\mathcal{D}}), x_{nd} \in \Re, n = 1, 2, \cdots, \mathcal{N}, d = 1, 2, \cdots, \mathcal{D}\}$, the classical fixed KDE (FKDE), i.e., Parzen window method [7], is constructed as

$$\begin{aligned}\hat{f}_{\mathrm{FKDE}}(\tilde{\mathrm{x}}) &= \hat{f}_{\mathrm{FKDE}}(x_1, x_2, \cdots, x_{\mathcal{D}}) \\ &= \frac{1}{\mathcal{N}\prod_{d=1}^{\mathcal{D}} h_d}\sum_{n=1}^{\mathcal{N}} \kappa\left(\frac{x_1 - x_{n1}}{h_1}, \frac{x_2 - x_{n2}}{h_2}, \cdots, \frac{x_{\mathcal{D}} - x_{n\mathcal{D}}}{h_{\mathcal{D}}}\right),\end{aligned} \tag{1}$$

where

$$\begin{aligned}\kappa(\tilde{\mathrm{u}}) &= \frac{1}{\left(\sqrt{2\pi}\right)^{\mathcal{D}}}\exp\left(-\frac{1}{2}\tilde{\mathrm{u}}^{\mathrm{T}}\tilde{\mathrm{u}}\right) \\ &= \frac{1}{\left(\sqrt{2\pi}\right)^{\mathcal{D}}}\exp\left(-\frac{1}{2}\sum_{d=1}^{\mathcal{D}} u_d^2\right),\end{aligned} \tag{2}$$

$\tilde{\mathrm{u}} = (u_1, u_2, \cdots, u_{\mathcal{D}}) \in \Re^{\mathcal{D}}$ is the $\mathcal{D}$-variate Gaussian kernel, and $\tilde{\mathrm{h}} = (h_1, h_2, \cdots, h_{\mathcal{D}})$, $h_d > 0, d = 1, 2, \cdots, \mathcal{D}$ is the bandwidth. Substituting Equation (2) into Equation (1) yields the estimated *p.d.f.* of dataset $\mathbb{X}$ as

$$\hat{f}_{\mathrm{FKDE}}(\tilde{\mathrm{x}}) = \frac{1}{\mathcal{N}}\sum_{n=1}^{\mathcal{N}} \mathbb{N}(\tilde{\mathrm{x}}_n, \Sigma), \tag{3}$$

where

$$\mathbb{N}(\tilde{\mathrm{x}}_n, \Sigma) = \frac{1}{\left(\sqrt{2\pi}\right)^{\mathcal{D}}|\Sigma|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(\tilde{\mathrm{x}} - \tilde{\mathrm{x}}_n)\Sigma^{-1}(\tilde{\mathrm{x}} - \tilde{\mathrm{x}}_n)^{\mathrm{T}}\right] \tag{4}$$

is the $\mathcal{D}$-dimensional Gaussian distribution with mean vector $\tilde{\mathrm{x}}_n = (x_{n1}, x_{n2}, \cdots, x_{n\mathcal{D}})$ and covariance matrix $\Sigma = \begin{bmatrix} h_1^2 & 0 & \cdots & 0 \\ 0 & h_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h_{\mathcal{D}}^2 \end{bmatrix}$. Equation (3) reflects that the estimated *p.d.f.* is the superposition of $\mathcal{N}$ Gaussian *p.d.f.*s.

The *p.d.f.* of dataset $\mathbb{X}$ estimated by VKDE is

$$\hat{f}_{\mathrm{VKDE}}(\tilde{\mathrm{x}}) = \frac{1}{\mathcal{N}}\sum_{n=1}^{\mathcal{N}} \mathbb{N}(\tilde{\mathrm{x}}_n, \Sigma_n), \tag{5}$$

where the covariance matrix of $\mathbb{N}(\tilde{\mathrm{x}}_n, \Sigma_n)$ is $\Sigma_n = \begin{bmatrix} h_{n1}^2 & 0 & \cdots & 0 \\ 0 & h_{n2}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h_{n\mathcal{D}}^2 \end{bmatrix}$, $h_{nd} > 0$, $n = 1, 2, \cdots, \mathcal{N}, d = 1, 2, \cdots, \mathcal{D}$. Equation (5) can be further transformed into the following Equation (6):

$$\begin{aligned}\hat{f}_{\mathrm{VKDE}}(\tilde{\mathrm{x}}) &= \frac{1}{\mathcal{N}}\sum_{n=1}^{\mathcal{N}} \frac{1}{\left(\sqrt{2\pi}\right)^{\mathcal{D}}|\Sigma_n|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(\tilde{\mathrm{x}} - \tilde{\mathrm{x}}_n)\Sigma_n^{-1}(\tilde{\mathrm{x}} - \tilde{\mathrm{x}}_n)^{\mathrm{T}}\right] \\ &= \frac{1}{\mathcal{N}}\sum_{n=1}^{\mathcal{N}}\prod_{d=1}^{\mathcal{D}} \frac{1}{\sqrt{2\pi}h_{nd}}\exp\left[-\frac{1}{2}\left(\frac{x_d - x_{nd}}{h_{nd}}\right)^2\right],\end{aligned} \tag{6}$$

where $\tilde{\mathrm{h}}_n = (h_{n1}, h_{n2}, \cdots, h_{n\mathcal{D}})$, $n = 1, 2, \cdots, \mathcal{D}$ is the variable bandwidth vector corresponding to the n-th data point. There are $\mathcal{N}\mathcal{D}$ bandwidth parameters which need to be determined in VKDE.

3. Proposed IVKDE

In this section, we firstly provide an improved VKDE which uses an L_2 regularization term-based objective function to evaluate the efficiency of variable bandwidths. Then, a bandwidth optimization algorithm is developed to solve the optimal variable bandwidths based on the above-mentioned objective function.

The purpose of VKDE training is to make the estimated *p.d.f.* $\hat{f}_{\mathrm{VKDE}}(\tilde{\mathrm{x}})$ as close to the true *p.d.f.* $f(\tilde{\mathrm{x}})$ as possible. In Equation (6), we can find that the performance of VKDE is only related to the selection of bandwidth vectors. We want to select the bandwidth vectors which can minimize the error between *p.d.f.* $\hat{f}_{\mathrm{VKDE}}(\tilde{\mathrm{x}})$ and $f(\tilde{\mathrm{x}})$. In order to measure the estimated error, an effective error criterion should firstly be designed. The integrated squared error (ISE)

$$\begin{aligned}\mathrm{ISE}(\tilde{\mathrm{h}}_1,\tilde{\mathrm{h}}_2,\cdots,\tilde{\mathrm{h}}_{\mathcal{N}}) &= \int_{-\infty}^{+\infty}\left[\hat{f}_{\mathrm{VKDE}}(\tilde{\mathrm{x}})-f(\tilde{\mathrm{x}})\right]^2\mathrm{d}\tilde{\mathrm{x}}\\ &= \int_{-\infty}^{+\infty}\left[\hat{f}_{\mathrm{VKDE}}(\tilde{\mathrm{x}})\right]^2\mathrm{d}\tilde{\mathrm{x}}-2\int_{-\infty}^{+\infty}\left[\hat{f}_{\mathrm{VKDE}}(\tilde{\mathrm{x}})f(\tilde{\mathrm{x}})\right]\mathrm{d}\tilde{\mathrm{x}}\\ &\quad+\int_{-\infty}^{+\infty}[f(\tilde{\mathrm{x}})]^2\mathrm{d}\tilde{\mathrm{x}}\end{aligned} \tag{7}$$

is used in our proposed IVKDE to measure the estimated error.

In Equation (7), we can see that the third term $\int_{-\infty}^{+\infty}[f(\tilde{\mathrm{x}})]^2\mathrm{d}\tilde{\mathrm{x}}$ is unrelated to the unknown bandwidth vectors. Thus, the optimal variable bandwidth vectors can be obtained by minimizing the simplified ISE criterion:

$$\mathrm{ISE}^*(\tilde{\mathrm{h}}_1,\tilde{\mathrm{h}}_2,\cdots,\tilde{\mathrm{h}}_{\mathcal{N}}) = \int_{-\infty}^{+\infty}\left[\hat{f}_{\mathrm{VKDE}}(\tilde{\mathrm{x}})\right]^2\mathrm{d}\tilde{\mathrm{x}}-2\int_{-\infty}^{+\infty}\left[\hat{f}_{\mathrm{VKDE}}(\tilde{\mathrm{x}})f(\tilde{\mathrm{x}})\right]^2\mathrm{d}\tilde{\mathrm{x}}. \tag{8}$$

Equation (8) is a data-driven error measurement which easily leads to a data-adaptive KDE and further makes the estimated *p.d.f.* more inclined to fit the given dataset $\mathbb{X}$. In order to guarantee the good generalization capability of KDE, we give the following objective function to select the bandwidth vectors for our proposed IVKDE:

$$\mathrm{L}(\tilde{\mathrm{h}}_1,\tilde{\mathrm{h}}_2,\cdots,\tilde{\mathrm{h}}_{\mathcal{N}}) = \mathrm{ISE}^*(\tilde{\mathrm{h}}_1,\tilde{\mathrm{h}}_2,\cdots,\tilde{\mathrm{h}}_{\mathcal{N}}) + \frac{\xi}{\mathcal{N}}\sum_{n=1}^{\mathcal{N}}\left\|\tilde{\mathrm{h}}_n\right\|_2^2, \tag{9}$$

where the second term is the L_2 regulation term, $\left\|\tilde{\mathrm{h}}_n\right\|_2$ is the L_2 norm of bandwidth vector $\tilde{\mathrm{h}}_n$, $n=1,2,\cdots,\mathcal{N}$, and $\xi>0$ is the regulation factor.

Substituting Equation (6) into $\int_{-\infty}^{+\infty}\left[\hat{f}_{\mathrm{VKDE}}(\tilde{\mathrm{x}})\right]^2\mathrm{d}\tilde{\mathrm{x}}$ and $\int_{-\infty}^{+\infty}\left[\hat{f}_{\mathrm{VKDE}}(\tilde{\mathrm{x}})f(\tilde{\mathrm{x}})\right]\mathrm{d}\tilde{\mathrm{x}}$ terms yields

$$\begin{aligned}&\int_{-\infty}^{+\infty}\left[\hat{f}_{\mathrm{VKDE}}(\tilde{\mathrm{x}})\right]^2\mathrm{d}\tilde{\mathrm{x}}\\ &\quad=\int_{-\infty}^{+\infty}\left[\frac{1}{\mathcal{N}}\sum_{n=1}^{\mathcal{N}}\prod_{d=1}^{\mathcal{D}}\frac{1}{\sqrt{2\pi}h_{nd}}\exp\left[-\frac{1}{2}\left(\frac{x_d-x_{nd}}{h_{nd}}\right)^2\right]\right]^2\mathrm{d}\tilde{\mathrm{x}}\\ &\quad=\frac{1}{\left(2\sqrt{\pi}\right)^{\mathcal{D}}\mathcal{N}^2}\sum_{n=1}^{\mathcal{N}}\frac{1}{h_{n1}h_{n2}\cdots h_{n\mathcal{D}}}\\ &\quad+\frac{1}{\left(\sqrt{2\pi}\right)^{\mathcal{D}}\mathcal{N}^2}\sum_{n=1}^{\mathcal{N}}\sum_{\substack{m=1\\m\neq n}}^{\mathcal{N}}\frac{1}{\prod_{d=1}^{\mathcal{D}}\sqrt{h_{nd}^2+h_{md}^2}}\exp\left[-\frac{1}{2}\sum_{d=1}^{\mathcal{D}}\left(\frac{x_{nd}-x_{md}}{\sqrt{h_{nd}^2+h_{md}^2}}\right)^2\right]\end{aligned} \tag{10}$$

and

$$\begin{aligned}\int_{-\infty}^{+\infty}\left[\hat{f}_{\mathrm{VKDE}}(\tilde{\mathrm{x}})f(\tilde{\mathrm{x}})\right]\mathrm{d}\tilde{\mathrm{x}} &= \mathrm{E}\left[\hat{f}_{\mathrm{VKDE}-n}(\tilde{\mathrm{x}}_n)\right]\\ &= \frac{1}{\mathcal{N}}\sum_{n=1}^{\mathcal{N}}\hat{f}_{\mathrm{VKDE}-n}(\tilde{\mathrm{x}}_n)\\ &= \frac{1}{\mathcal{N}}\sum_{n=1}^{\mathcal{N}}\left[\frac{1}{\mathcal{N}-1}\sum_{\substack{m=1\\m\neq n}}^{\mathcal{N}}\prod_{d=1}^{\mathcal{D}}\frac{1}{\sqrt{2\pi}h_{md}}\exp\left[-\frac{1}{2}\left(\frac{x_{nd}-x_{md}}{h_{md}}\right)^2\right]\right]\\ &= \frac{1}{\left(\sqrt{2\pi}\right)^{\mathcal{D}}\mathcal{N}(\mathcal{N}-1)}\sum_{n=1}^{\mathcal{N}}\sum_{\substack{m=1\\m\neq n}}^{\mathcal{N}}\frac{1}{h_{m1}h_{m2}\cdots h_{m\mathcal{D}}}\exp\left[-\frac{1}{2}\sum_{d=1}^{\mathcal{D}}\left(\frac{x_{nd}-x_{md}}{h_{md}}\right)^2\right],\end{aligned} \tag{11}$$

respectively, where $\hat{f}_{\mathrm{VKDE}-n}(\tilde{\mathrm{x}}_n)$, $n = 1, 2, \cdots, \mathcal{N}$ is a leave-one-out estimator trained through an unbiased cross-validation (UCV) method.

IVKDE needs to use the optimal bandwidth vectors that can minimize the objective function with the L_2 regulation term. In order to solve the optimal bandwidths, we should firstly calculate the partial derivative of $\mathrm{L}(\tilde{\mathrm{h}}_1, \tilde{\mathrm{h}}_2, \cdots, \tilde{\mathrm{h}}_{\mathcal{N}})$ with respect to h_{nd}, $n = 1, 2, \cdots, \mathcal{N}$, $d = 1, 2, \cdots, \mathcal{D}$. Let

$$\begin{aligned}\Delta h_{nd} &= \frac{\partial \mathrm{L}(\tilde{\mathrm{h}}_1, \tilde{\mathrm{h}}_2, \cdots, \tilde{\mathrm{h}}_{\mathcal{N}})}{\partial h_{nd}}\\ &= -\frac{1}{\left(2\sqrt{\pi}\right)^{\mathcal{D}}\mathcal{N}^2 h_{nd}\prod_{k=1}^{\mathcal{D}}h_{nk}} + \frac{2h_{nd}\Delta_1}{\left(\sqrt{2\pi}\right)^{\mathcal{D}}\mathcal{N}^2} - \frac{2\Delta_2}{\left(\sqrt{2\pi}\right)^{\mathcal{D}}\mathcal{N}(\mathcal{N}-1)h_{nd}\prod_{k=1}^{\mathcal{D}}h_{nk}},\end{aligned} \tag{12}$$

where

$$\Delta_1 = \sum_{\substack{m=1\\m\neq n}}^{\mathcal{N}}\left[\left[\frac{\left(\frac{x_{md}-x_{nd}}{\sqrt{h_{nd}^2+h_{md}^2}}\right)^2-1}{\left(h_{nd}^2+h_{md}^2\right)\prod_{k=1}^{\mathcal{D}}\sqrt{h_{nk}^2+h_{mk}^2}}\right]\exp\left[-\frac{1}{2}\sum_{k=1}^{\mathcal{D}}\left(\frac{x_{mk}-x_{nk}}{\sqrt{h_{nk}^2+h_{mk}^2}}\right)^2\right]\right] \tag{13}$$

and

$$\Delta_2 = \sum_{\substack{m=1\\m\neq n}}^{\mathcal{N}}\left[\left[-1+\frac{(x_{nd}-x_{md})^2}{h_{nd}^2}\right]\exp\left[-\frac{1}{2}\sum_{k=1}^{\mathcal{D}}\left(\frac{x_{mk}-x_{nk}}{h_{nk}}\right)^2\right]\right]. \tag{14}$$

We can find that it is very difficult to calculate the analytic solution of h_{nd}, $n = 1, 2, \cdots, \mathcal{N}$, $d = 1, 2, \cdots, \mathcal{D}$ from $\Delta h_{nd} = 0$. Here, we design the following Algorithm 1 which uses the gradient descent method to solve the optimal bandwidths for IVKDE based on the objective function as shown in Equation (9). Algorithm 1 iteratively determines the optimal bandwidths based on the decaying learning rate adjustment. Because the minimization of $\mathrm{L}(\tilde{\mathrm{h}}_1, \tilde{\mathrm{h}}_2, \cdots, \tilde{\mathrm{h}}_{\mathcal{N}})$ is required, the negative gradient is used in Algorithm 1.

Algorithm 1 Solving the optimal bandwidths for IVKDE.

Input: The given dataset $\mathbb{X}$, the regulation factor $\xi > 0$, the maximum value of learning rate α_{Max}, the minimum value of learning rate α_{Min}, the maximum number of iterations $\mathcal{T}_{\text{Max}}$, the stopping threshold $\delta > 0$, and the initial bandwidth $h_{nd}^{(0)}$, $n = 1, 2, \cdots, \mathcal{N}$, $d = 1, 2, \cdots, \mathcal{D}$.

Output: The optimal bandwidth h_{nd}, $n = 1, 2, \cdots, \mathcal{N}$, $d = 1, 2, \cdots, \mathcal{D}$.

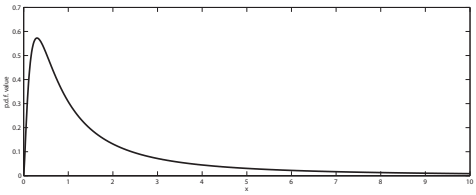
$t = 1$; // t is the number of iterations.
repeat
  for $n = 1; n <= \mathcal{N}; n++$ **do**
    for $d = 1; d <= \mathcal{D}; d++$ **do**
      $h_{nd}^{(t)} = h_{nd}^{(t-1)} - \left(\alpha_{\text{Max}} - \frac{\alpha_{\text{Max}} - \alpha_{\text{Min}}}{\mathcal{T}_{\text{Max}}} t\right) \Delta h_{nd}^{(t-1)}$;
    end for
  end for
  $t = t + 1$;
until $\left| \mathrm{L}\left(\tilde{\mathrm{h}}_1^{(t)}, \tilde{\mathrm{h}}_2^{(t)}, \cdots, \tilde{\mathrm{h}}_{\mathcal{N}}^{(t)}\right) - \mathrm{L}\left(\tilde{\mathrm{h}}_1^{(t-1)}, \tilde{\mathrm{h}}_2^{(t-1)}, \cdots, \tilde{\mathrm{h}}_{\mathcal{N}}^{(t-1)}\right) \right| < \delta$ or $t > \mathcal{T}_{\text{Max}}$
$h_{nd} = h_{nd}^{(t)}, n = 1, 2, \cdots, \mathcal{N}, d = 1, 2, \cdots, \mathcal{D}$.

4. Experimental Results and Analysis

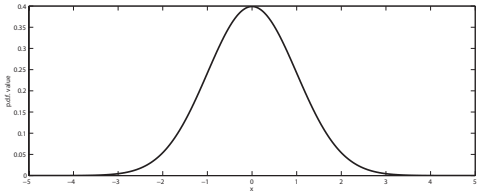
We conduct three experiments based on eight different probability distributions as shown in Table 1 to validate the rationality, feasibility, and effectiveness of the proposed IVKDE. The graphics of these eight *p.d.f.*s for the given parameters are presented in Figure 1.

Table 1. Four univariate and four multivariate probability distributions ($f^{(i)}(\tilde{\mathrm{x}})$ in bimodal, trimodal, and quadrimodal normal distributions is the two-dimensional normal distribution with mean vector $\vec{\mu}^{(i)}$ and covariance matrix $\Sigma^{(i)}$).

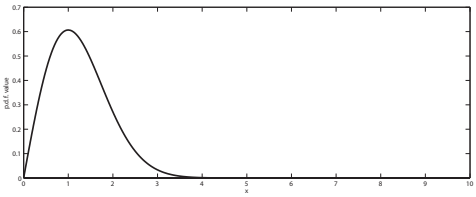
	Probability Distribution	Probability Density Function
Univariate	F	$f(x) = \frac{n_1^{\frac{n_1}{2}} n_2^{\frac{n_2}{2}} x^{\frac{n_1}{2}-1}}{\left[\int_0^1 x^{\frac{n_1}{2}-1}(1-x)^{\frac{n_2}{2}-1} dx\right](n_1 x + n_2)^{\frac{n_1+n_2}{2}}}, n_1, n_2 = 1, 2, 3, \cdots; x \in [0, +\infty)$
	Normal	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \mu \in (-\infty, +\infty), \sigma > 0; x \in (-\infty, +\infty)$
	Rayleigh	$f(x) = \frac{x}{\sigma^2} e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}, \sigma > 0; x \in [0, +\infty)$
	Student's T	$f(x) = \frac{\int_0^{+\infty} x^{\frac{v+1}{2}-1} e^{-x} dx}{\sqrt{\pi v} \int_0^{+\infty} x^{\frac{v}{2}-1} e^{-x} dx} \left(1 + \frac{x^2}{v}\right)^{-\frac{v+1}{2}}, v > 0; x \in (-\infty, +\infty)$
Multivariate	Two-dimensional normal	$f(\tilde{\mathrm{x}}) = (2\pi)^{-\frac{\mathcal{M}}{2}} \lvert\Sigma\rvert^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathrm{x} - \vec{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathrm{x} - \vec{\mu})\right], \tilde{\mathrm{x}} = (x_1, x_2)$, $\vec{\mu}$ is the mean vector and Σ is the covariance matrix.
	Bimodal normal	$f(\tilde{\mathrm{x}}) = \sum_{i=1}^{2} \varepsilon_i f^{(i)}(\tilde{\mathrm{x}}), \tilde{\mathrm{x}} = (x_1, x_2), \sum_{i=1}^{2} \varepsilon_i = 1, \varepsilon_i \geq 0, i = 1, 2$
	Trimodal normal	$f(\tilde{\mathrm{x}}) = \sum_{i=1}^{3} \varepsilon_i f^{(i)}(\tilde{\mathrm{x}}), \tilde{\mathrm{x}} = (x_1, x_2), \sum_{i=1}^{3} \varepsilon_i = 1, \varepsilon_i \geq 0, i = 1, 2, 3$
	Quadrimodal normal	$f(\tilde{\mathrm{x}}) = \sum_{i=1}^{4} \varepsilon_i f^{(i)}(\tilde{\mathrm{x}}), \tilde{\mathrm{x}} = (x_1, x_2), \sum_{i=1}^{4} \varepsilon_i = 1, \varepsilon_i \geq 0, i = 1, 2, 3, 4$



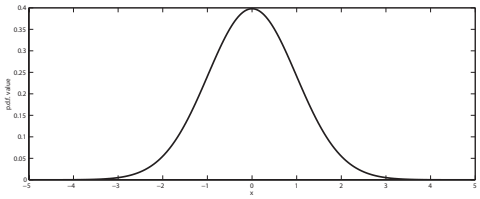
(**a**) F distribution ($n_1 = n_2 = 20$)



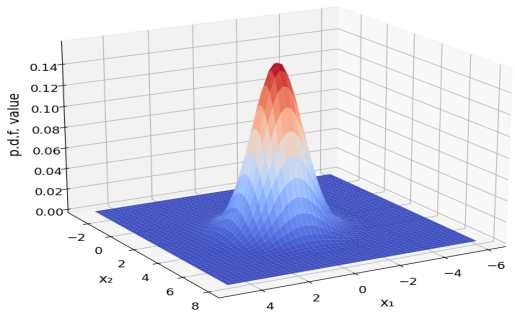
(**b**) Normal distribution ($\mu = 0, \sigma = 1$)



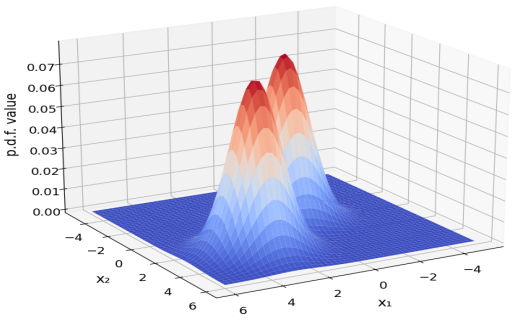
(**c**) Rayleigh distribution ($\sigma = 1$)



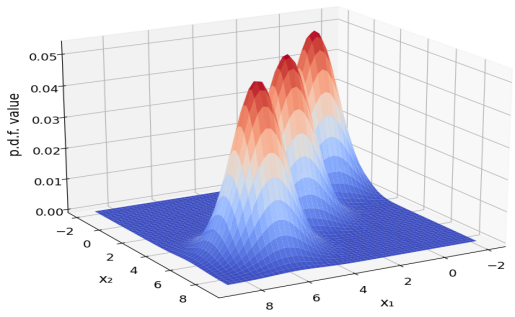
(**d**) Student's T distribution ($v = 10$)



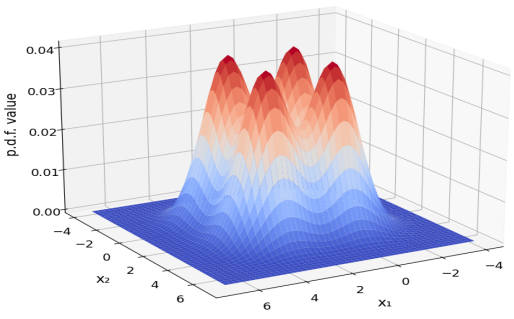
(**e**) Two-dimensional normal distribution ($\bar{\mu} = (0,3)$, $\Sigma_{11} = 1$, $\Sigma_{12} = 0$, $\Sigma_{21} = 0$, $\Sigma_{22} = 1$)



(**f**) Bimodal normal distribution ($\bar{\mu}^{(1)} = (0,0), \bar{\mu}^{(2)} = (3,3)$, $\Sigma^{(1)}$ and $\Sigma^{(2)}$ are identify matrices, $\varepsilon_1 = \varepsilon_2 = \frac{1}{2}$)



(**g**) Trimodal normal distribution ($\bar{\mu}^{(1)} = (0,0), \bar{\mu}^{(2)} = (3,3), \bar{\mu}^{(3)} = (6,6)$, $\Sigma^{(1)}$, $\Sigma^{(2)}$, and $\Sigma^{(2)}$ are identify matrices, $\varepsilon_1 = \varepsilon_2 = \varepsilon_3 = \frac{1}{3}$)



(**h**) Quadrimodal normal distribution ($\bar{\mu}^{(1)} = (0,0), \bar{\mu}^{(2)} = (3,3), \bar{\mu}^{(3)} = (0,3), \bar{\mu}^{(3)} = (3,0)$, $\Sigma^{(1)}$, $\Sigma^{(2)}$, $\Sigma^{(2)}$, and $\Sigma^{(4)}$ are identify matrices, $\varepsilon_1 = \varepsilon_2 = \varepsilon_3 = \varepsilon_4 = \frac{1}{4}$)

Figure 1. Graphics of eight *p.d.f.*s.

4.1. Experiential Setup

The rationality is to check the convergence of Algorithm 1, the feasibility is to show the estimation capability of IVKDE to the given *p.d.f.*s, and the effectiveness is demonstrated by comparing the estimation performances of IVKDE with FKDE and VKDE. For FKDE and VKDE, the optimal bandwidths are also determined with the gradient descent method. The synthetic datasets obeying the above-mentioned distributions can be accessible in any country accessed via our BaiduPan (https://pan.baidu.com/s/1YhkkrckQA_e2GNd8haLE1g, accessed on 25 June 2021) with extraction code vn6j. All the estimators are implemented with the Python programming language and run on a PC with an Intel(R) Quad-core 3.00 GHz i5-7400 CPU and 16 GB memory.

4.2. Rationality of IVKDE

We test the convergence of Algorithm 1 based on the random data points obeying F, normal, two-dimensional normal, and bimodal normal distributions with the following parameters:

- F: $\mathcal{N} = 1000$ and $n_1 = n_2 = 20$;
- Normal: $\mathcal{N} = 1000$, $\mu = 0$, and $\sigma = 1$;
- Two-dimensional normal: $\mathcal{N} = 1000$, $\bar{\mu} = (0, 3)$, and $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$;
- Bimodal normal: $\mathcal{N} = 1000$, $\bar{\mu}^{(1)} = (0, 0)$, $\bar{\mu}^{(2)} = (3, 3)$, $\Sigma^{(1)} = \Sigma^{(2)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, and $\varepsilon_1 = \varepsilon_2 = \frac{1}{2}$.

For each distribution, we repeat the running of Algorithm 1 10 times with the following parameters: $\mathcal{T}_{\text{Max}} = 2500$, $\alpha_{\text{Max}} = 1$, $\alpha_{\text{Min}} = 0.001$, $\delta = 0$, and $h_{nd}^{(0)} = 0.5$. We check the variation of the bandwidth sum with an increase in iteration numbers, where the bandwidth sum is calculated as

$$\text{sum}^{(t)}\left(\tilde{h}_1, \tilde{h}_2, \cdots, \tilde{h}_{\mathcal{N}}\right) = \frac{1}{\mathcal{N}\mathcal{D}} \sum_{n=1}^{\mathcal{N}} \sum_{d=1}^{\mathcal{D}} h_{nd}^{(t)}, t = 1, 2, \cdots, \mathcal{T}_{\text{Max}}. \tag{15}$$

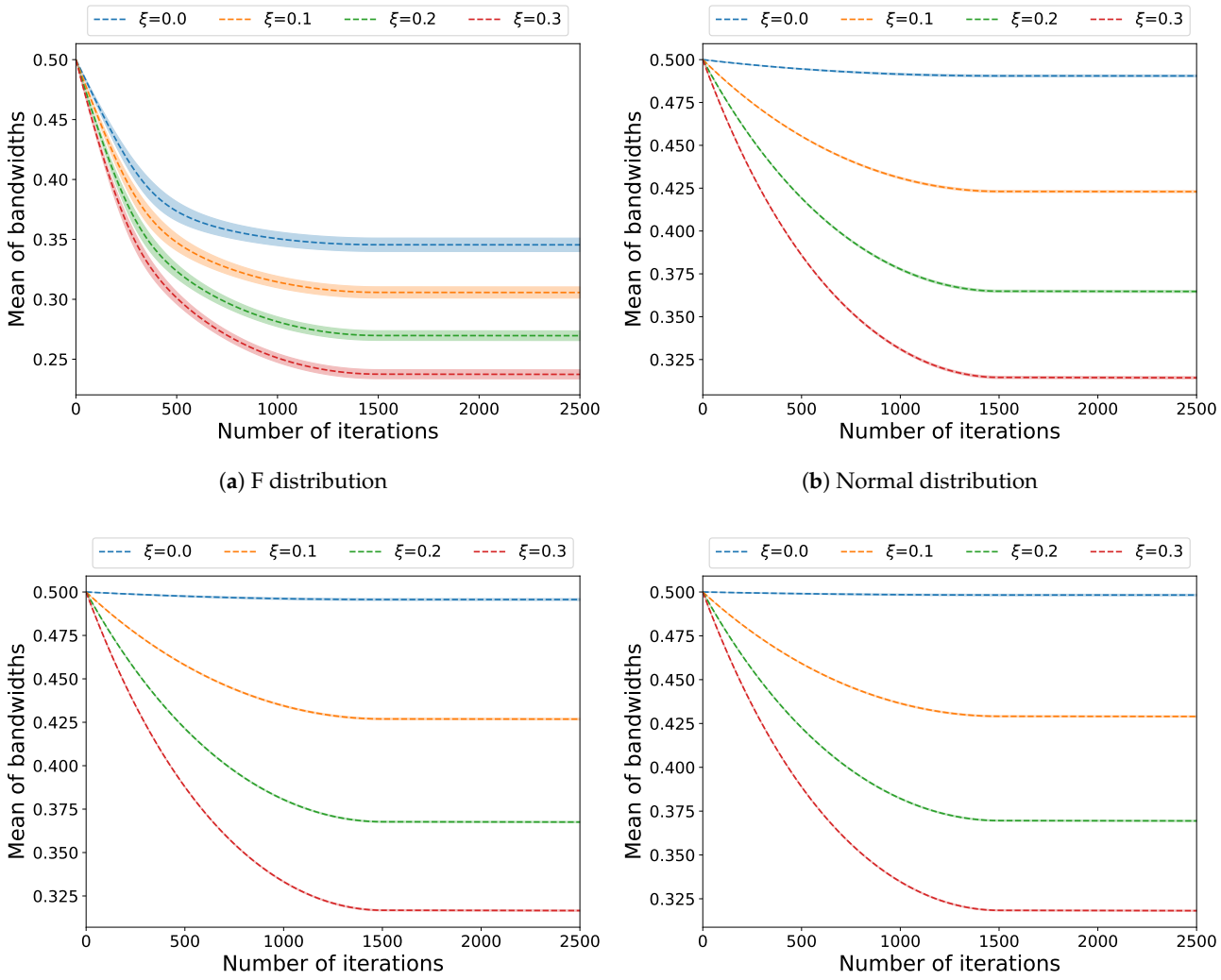
In Figure 2, we can see that Algorithm 1 is convergent for the different regulation factor ξs on the given *p.d.f.* The curves of bandwidth sums firstly decrease and then keep stable with the increase in iteration numbers. This indicates that Algorithm 1 is convergent and can find the optimal bandwidths for IVKDE.

4.3. Feasibility of IVKDE

We check the *p.d.f.* estimation capability of IVKDE based on F and two-dimensional normal distributions with the following parameters:

- F: $\mathcal{N} = 1000$ and $n_1 = n_2 = 20$;
- Two-dimensional normal: $\mathcal{N} = 1000$, $\bar{\mu} = (0, 3)$, and $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

We use Algorithm 1 to determine the optimal bandwidths for each distribution based on the random data points, where the parameters of Algorithm 1 are set as $\mathcal{T}_{\text{Max}} = 1500$, $\alpha_{\text{Max}} = 0.4$, $\alpha_{\text{Min}} = 0.01$, $\delta = 10^{-8}$, $h_{nd}^{(0)} = 0.3$, $\xi = 0.3$ for F distribution and $\mathcal{T}_{\text{Max}} = 500$, $\alpha_{\text{Max}} = 0.2$, $\alpha_{\text{Min}} = 0.1$, $\delta = 10^{-8}$, $h_{nd}^{(0)} = 0.36$, and $\xi = 0.3$ for two-dimensional normal distribution. The estimated *p.d.f.*s are presented in Figures 3 and 4. In these two figures, we can intuitively find that IVKDE can estimate the underlying *p.d.f.*s based on the given data points. The estimated *p.d.f.*s are very close to the true *p.d.f.*s. The experimental results show that IVKDE is feasible to estimate the unknown *p.d.f.*



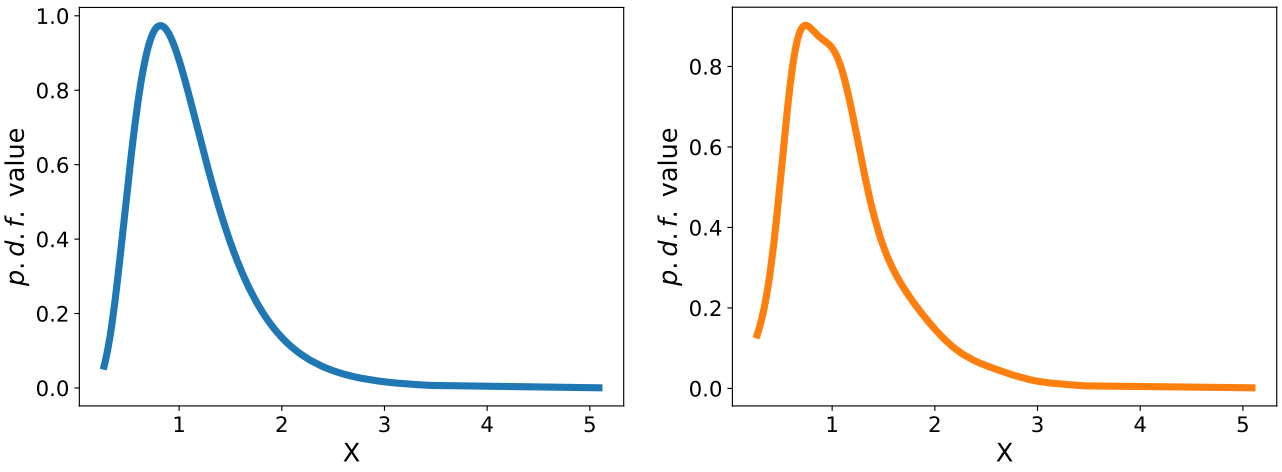
(**a**) F distribution

(**b**) Normal distribution

(**c**) Two-demonsional normal distribution

(**d**) Bimodal normal distribution

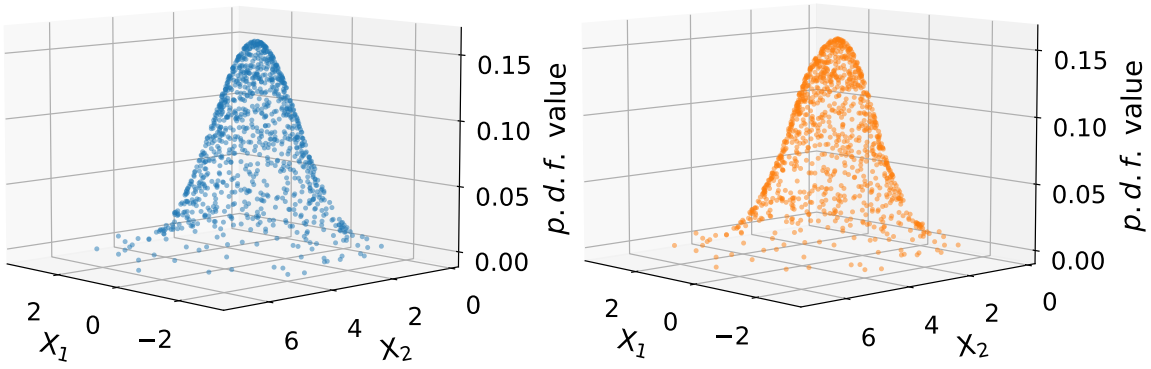
Figure 2. Convergences of Algorithm 1 on 4 given *p.d.f.*s.



(**a**) True *p.d.f.*

(**b**) Estimated *p.d.f.* with IVKDE

Figure 3. Estimation capability of IVKDE on F distribution.



(**a**) True *p.d.f.* (**b**) Estimated *p.d.f.* with IVKDE

Figure 4. Estimation capability of IVKDE on two-dimensional normal distribution.

4.4. Effectiveness of IVKDE

On eight probability distributions, as shown in Table 1, we compare the *p.d.f.* estimation performance of IVKDE with FKDE and VKDE. The parameters of these three kernel density estimators are summarized in Table 2. The comparative results among FKDE, VKDE, and IVKDE are listed in Table 3. We use the mean absolute error (MAE) to evaluate the training and testing performances of these three kernel density estimators. Assume the true and estimated *p.d.f.* values for the given dataset $\mathbb{X}$ are $y_1, y_2, \cdots, y_{\mathcal{N}}$ and $\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_{\mathcal{N}}$, respectively. Then, the MAE on dataset $\mathbb{X}$ is calculated as

$$\mathrm{MAE}(\mathbb{X}) = \frac{1}{\mathcal{N}} \sum_{n=1}^{\mathcal{N}} |y_n - \hat{y}_n|. \tag{16}$$

Table 2. Parameter settings of FKDE, VKDE, and IVKDE.

No.	Probability Distribution	FKDE					VKDE					IVKDE					
		$\mathcal{T}_{\mathrm{Max}}$	α_{Max}	α_{Min}	δ	$h_{nd}^{(0)}$	$\mathcal{T}_{\mathrm{Max}}$	α_{Max}	α_{Min}	δ	$h_{nd}^{(0)}$	$\mathcal{T}_{\mathrm{Max}}$	α_{Max}	α_{Min}	δ	$h_{nd}^{(0)}$	ξ
1	F	1000	1	10^{-5}	10^{-8}	1	1500	0.4	0.01	10^{-8}	0.3	1500	0.4	0.01	10^{-8}	0.3	0.3
2	Normal						1200	0.5	0.1	10^{-8}	0.34	1200	0.5	0.1	10^{-8}	0.34	0.13
3	Rayleigh						1000	1	0.1	10^{-8}	0.5	1000	1	0.1	10^{-8}	0.5	0.45
4	Student's T						1500	1	0.5	10^{-8}	0.5	1500	1	0.5	10^{-8}	0.5	0.15
5	Two-dimensional normal						500	0.2	0.1	10^{-8}	0.36	500	0.2	0.1	10^{-8}	0.36	0.3
6	Bimodal normal						500	0.2	0.1	10^{-8}	0.36	500	0.2	0.1	10^{-8}	0.36	0.3
7	Trimodal normal						500	0.2	0.01	10^{-8}	0.38	500	0.2	0.01	10^{-8}	0.38	0.01
8	Quadrimodal normal						500	0.5	0.1	10^{-8}	0.5	500	0.5	0.1	10^{-8}	0.5	0.3

ξ is the regulation factor; $\mathcal{T}_{\mathrm{Max}}$ is the maximum number of iterations; α_{Max} is the maximum value of learning rate; α_{Min} is the maximum value of learning rate; δ is the stopping threshold; $h_{nd}^{(0)}, n = 1, 2, \cdots, \mathcal{N}, d = 1, 2, \cdots, \mathcal{D}$ are the initial bandwidths.

In Table 3, we can find that IVKDE obtains the significantly better *p.d.f.* estimation performances on training and testing datasets than FKDE and VKDE. We carry out the statistical test on the comparative results based on the sign test method [22]. For the pairwise comparison between methods A and B, A is significantly better than B under the given significance level if the number of A's wins reaches the critical number. There are eight different probability distributions which are used to compare the estimation performances of FKDE, VKDE, and IVKDE. The critical win number is $\frac{8}{2} + 1.96 \times \frac{\sqrt{8}}{2} \approx 7$ in our comparison for the given significance level 0.05. The win numbers of IVKDE vs. FKDE and VKDE on training datasets are 7 and 8, respectively. This indicates that IVKDE obtains significantly better *p.d.f.* estimation performances than FKDE and VKDE on training datasets. The win numbers of IVKDE vs. FKDE and VKDE on testing datasets are 6 and 8, respectively. This indicates that IVKDE obtains significantly better *p.d.f.* estimation

performances than VKDE on testing datasets. The experimental and statistical results show that IVKDE can improve the *p.d.f.* estimation performance of VKDE and thus demonstrate the effectiveness of IVKDE.

Table 3. Competitive results among FKDE, VKDE, and IVKDE on 8 different probability distributions.

No.	Probability Distribution	MAE on Training Set			MAE on Testing Set		
		FKDE	VKDE	IVKDE	FKDE	VKDE	IVKDE
1	F	0.02921	0.03965	0.02891	0.02964	0.04111	0.03171
2	Normal	0.01416	0.01511	0.01389	0.01376	0.01489	0.0137
3	Rayleigh	0.02259	0.05127	0.02797	0.02222	0.05137	0.02859
4	Student's T	0.00999	0.01607	0.00959	0.00980	0.01583	0.00970
5	Two-dimensional normal	0.00486	0.00502	0.00485	0.00500	0.00509	0.00498
6	Bimodal normal	0.00518	0.00463	0.00456	0.00530	0.00465	0.00455
7	Trimodal normal	0.00364	0.00363	0.00363	0.00359	0.00359	0.00359
8	Quadrimodal normal	0.00232	0.00235	0.00228	0.00247	0.00248	0.00241

5. Conclusions and Future Works

This paper presented an improved variable kernel density estimator (IVKDE) by using both integrated squared error (ISE) and L_2 regularization to determine the optimal bandwidths. The L_2 regularization can effectively avoid the over-smoothed bandwidth selection. The experimental results demonstrated the rationality, feasibility, and effectiveness of the proposed IVKDE. Future works will be carried out according to the following research directions: (1) using IVKDE to estimate the unknown *p.d.f.* for a large-scale dataset [23] and (2) finding the practical applications for IVKDE in data mining and machine learning fields.

Author Contributions: Methodology, Y.J.; Writing—Original Draft Preparation, Writing—Review and Editing, Y.H.; Validation, D.H. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was supported by the Basic Research Foundation of Strengthening Police with Science and Technology of the Ministry of Public Security (2017GABJC09), Open Foundation of Key Laboratory of Impression Evidence Examination and Identification Technology, The Ministry of Public Security of the People's Republic of China (HJKF201901), Basic Research Foundation of Shenzhen (20210312191246002), and the Scientific Research Foundation of Shenzhen University for Newly-Introduced Teachers (2018060).

Data Availability Statement: The data presented in this study are available in BaiduPan https://pan.baidu.com/s/1YhkkrckQA_e2GNd8haLE1g (accessed on 25 June 2021) with extraction code vn6j.

Acknowledgments: We would like to thank the editors and two anonymous reviewers whose meticulous readings and valuable suggestions helped us to improve this paper significantly after two rounds of review.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

p.d.f.	Probability Density Function
KDE	Kernel Density Estimator
ISE	Integrated Squared Error
MISE	Mean Integrated Squared Error
FKDE	Fixed Kernel Density Estimator
VKDE	Variable Kernel Density Estimator
RoT	Rule-of-Thumb
BS	Bootstrap

BCV	Biased Cross-Validation
UCV	Unbiased Cross-Validation
IVKDE	Improved Variable Kernel Density Estimator
MAE	Mean Absolute Error

References

1. John, G.H.; Langley, P. Estimating continuous distributions in Bayesian classifiers. In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, San Francisco, CA, USA, 18–20 August 1995; pp. 338–345.
2. Wang, X.Z.; He, Y.L.; Wang, D.D. Non-naive Bayesian classifiers for classification problems with continuous attributes. *IEEE Trans. Cybern.* **2014**, *44*, 21–39. [CrossRef] [PubMed]
3. Azzalini, A.; Menardi, G. Clustering via nonparametric density estimation: The R package pdfCluster. *J. Stat. Softw.* **2014**, *57*, 1–26. [CrossRef]
4. Cuevas, A.; Febrero, M.; Fraiman, R. Cluster analysis: A further approach based on density estimation. *Comput. Stat. Data Anal.* **2001**, *36*, 441–459. [CrossRef]
5. Kwak, N.; Choi, C.H. Input feature selection by mutual information based on Parzen window. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 1667–1671. [CrossRef]
6. Peng, H.C.; Long, F.H.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [CrossRef] [PubMed]
7. Parzen, E. On estimation of a probability density function and mode. *Ann. Math. Stat.* **1962**, *33*, 1065–1076. [CrossRef]
8. Wand, M.P.; Jones, M.C. *Kernel Smoothing*; Chapman and Hall: London, UK, 1994.
9. Silverman, B.W. *Density Estimation for Statistics and Data Analysis*; Routledge: Abingdon, UK, 2018.
10. Chen, S. Optimal bandwidth selection for kernel density functionals estimation. *J. Probab. Stat.* **2015**, *2015*, 242683. [CrossRef]
11. Taylor, C.C. Bootstrap choice of the smoothing parameter in kernel density estimation. *Biometrika* **1989**, *76*, 705–712. [CrossRef]
12. Scott, D.W.; Terrell, G.R. Biased and unbiased cross-validation in density estimation. *J. Am. Stat. Assoc.* **1987**, *82*, 1131–1146. [CrossRef]
13. Bowman, A.W. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **1984**, *71*, 353–360. [CrossRef]
14. Breiman, L.; Meisel, W.; Purcell, E. Variable kernel estimates of multivariate densities. *Technometrics* **1977**, *19*, 135–144. [CrossRef]
15. Jones, M.C. Variable kernel density estimates and variable kernel density estimates. *Aust. J. Stat.* **1990**, *32*, 361–371. [CrossRef]
16. Terrell, G.R.; Scott, D.W. Variable kernel density estimation. *Ann. Stat.* **1992**, *20*, 1236–1265. [CrossRef]
17. Hall, P.; Hu, T.C.; Marron, J.S. Improved variable window kernel estimates of probability densities. *Ann. Stat.* **1995**, *23*, 1–10. [CrossRef]
18. Wu, T.J.; Chen, C.F.; Chen, H.Y. A variable bandwidth selector in multivariate kernel density estimation. *Stat. Probab. Lett.* **2007**, *77*, 462–467. [CrossRef]
19. Suaray, K. Variable bandwidth kernel density estimation for censored data. *J. Stat. Theory Pract.* **2011**, *5*, 221–229. [CrossRef]
20. Klebanov, I. Axiomatic Approach to Variable Kernel Density Estimation. *arXiv* **2018**, arXiv:1805.01729. Available online: https://arxiv.org/abs/1805.01729 (accessed on 4 May 2018).
21. Jones, M.C. The roles of ISE and MISE in density estimation. *Stat. Probab. Lett.* **1991**, *12*, 51–56. [CrossRef]
22. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
23. ur Rehman, M.H., Liew, C.S., Abbas, A., Jayaraman, P.P, Wah, T.Y., Khan, S.U. Big data reduction methods: A survey. *Data Sci. Eng.* **2016**, *4*, 265–284. [CrossRef]