Applied Mathematics and Computation 219 (2012) 3425–3460

Contents lists available at SciVerse ScienceDirect

Applied Mathematics and Computation

journal homepage: www.elsevier.com/locate/amc

# Optimal bandwidth selection for re-substitution entropy estimation

Yu-Lin He [a,*], James N.K. Liu [b], Xi-Zhao Wang [a], Yan-Xing Hu [b]

[a] College of Mathematics and Computer Science, Hebei University, Baoding, Hebei, China
[b] Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

ARTICLE INFO

*Keywords:*
Information entropy
Re-substitution entropy estimator
Probability density estimation
Optimal bandwidth
Integrated mean square error
Discretization

ABSTRACT

A new fusion approach of selecting an optimal bandwidth for re-substitution entropy estimator (RE) is presented in this study. When approximating the continuous entropy with density estimation, two types of errors will be generated: entropy estimation error (type-I error) and density estimation error (type-II error). These two errors are all strongly dependent on the undetermined bandwidths. Firstly, an experimental conclusion based on 24 typical probability distributions is demonstrated that there is some inconsistency between the optimal bandwidths associated with these two errors. Secondly, two different error measures for type-I and type-II errors are derived. A trade-off between type-I and type-II errors is a fundamental and potential property of our proposed method called $RE_{I+II}$. Thus, the fusion of these two errors is conducted and an optimal bandwidth for $RE_{I+II}$ is solved. Finally, the experimental comparisons are carried out to verify the estimation performance of our proposed strategy. The discretization method is deemed to be the necessary preprocessing technology for the calculation of continuous entropy traditionally. So, the nine mostly used unsupervised discretization methods are introduced to give comparison of their computational performances with that of $RE_{I+II}$. And, five most popular estimators for entropy approximation are also plugged into our comparisons: splitting data estimator (SDE), cross-validation estimator (CVE), $m$-spacing estimator ($m$SE), $m_n$-spacing estimator ($m_n$SE), and nearest neighbor distance estimator (NNDE). The simulation studies on 24 different typical density distributions show that $RE_{I+II}$ can obtain the better estimation performance among the involved methods. Meanwhile, the estimation behaviors of different entropy estimation methods are also revealed based on the comparative results. The empirical analysis demonstrates that $RE_{I+II}$ is more insensitive to data and a better generalizable way for the estimation of continuous entropy. $RE_{I+II}$ makes it possible for a handy optimal bandwidth to be derived from a given dataset.



## 1. Introduction

How to measure the amount of information contained in a certain application domain was a confused and arduous topic until the Shannon entropy [1,2] had been put forward. The Shannon entropy quantifies the amount of information needed to be measured [3]. This landmark breakthrough gives birth to information theory [4] which is developed by Shannon to find the fundamental randomness or uncertainty associated with a random variable. The concept of Shannon entropy is the central role of information theory which is based on probability theory and statistics. Shannon entropy plays the important roles

* Corresponding author.
*E-mail addresses:* csylhe@gmail.com, fuchengrenyulin@126.com (Y.-L. He), csnkliu@inet.polyu.edu.hk (J.N.K. Liu), xizhaowang@ieee.org (X.-Z. Wang), huyanxing@gmail.com (Y.-X. Hu).

0096-3003/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.amc.2012.08.056

over a wide range of machine learning applications, such as, decision tree [5–7], neural language processing [8], text categorization [9], feature selection [10,11], image processing [59], and so on.

In the learning problems, the variables may be discrete and continuous. The "discrete" refers to the variables taking on categorical values. The "continuous" refers to the variables taking on numerical values (integer or real). It is well acknowledged that Shannon entropy can be implemented sophisticatedly and efficiently for the discrete variables as the following Eq. (1):

$$H(X) = -\sum_{i=1}^{n} p(x_i)\mathrm{In}[p(x_i)], \tag{1}$$

where, let $X$ be a discrete random variable taking a finite number of possible values $x_1, x_2, \ldots, x_n$ with probabilities $p(x_1), p(x_2), \ldots, p(x_n)$, respectively such that $p(x_i) \geqslant 0, i = 1, 2, \ldots, n$ and $\sum_{i=1}^{n} p(x_i) = 1$. $\mathrm{In}(u)$, $u \succ 0$ is the natural logarithm which is the logarithm to the base $e$, where $e$ is an irrational constant approximately equal to 2.718281828. Note that $p\mathrm{In}(p) = 0$ when $p = 0$. However, many learning tasks [12–15] are often involved with the continuous variables. The mathematical formula for continuous variables can be summarized as follows by extending the discrete entropy to continuous case:

$$H(X) = -\int_{-\infty}^{+\infty} f(X)\mathrm{In}[f(X)]dx, \tag{2}$$

where, let $X$ be a continuous random variable taking the probability density function $f(X)$ such that $\int_{-\infty}^{+\infty} f(X)dx = 1$. From the Eq. (2), we can find that there are two main handicaps when the entropy computation for continuous variables is implemented: the unknown of probability density function and the evaluation of integral paradigm. In order to bridge over these difficulties, the academic world has made many positive attempts and advanced a number of representative ways.

In many practical applications [7,20,25,31], the main strategy for handling the entropy computation of continuous variables is to discretize the continuous variables into discrete ones and then calculate the discrete entropy according to the Eq. (1). In statistics and machine learning, discretization refers to the process of converting or partitioning continuous attributes, features or variables to discretized or nominal attributes, features, or variables. Over the years, many discretization methods have been proposed and tested to show that discretization helps improve the performance of learning methods and helps understand the learning result. One of taxonomies is to classify primary discretization methods into supervised discretization and unsupervised discretization, where supervised discretization uses the class or label information to select the discretization cut points and unsupervised discretization determines the cut points without the usage of class or label information. In the setting of entropy estimation without class information provided, supervised discretization may not be competent to implement the entropy computation for continuous variables. So, unsupervised discretization is considered as a capable candidate to carry out the continuous entropy computation.

In our study, nine common unsupervised discretization methods are introduced and employed as the competitors. Equal width discretization (EWD) [16] and equal frequency discretization (EFD) [17] are two mostly used and simplest methods. The experimental observations in numerous literatures [18–20] show that the satisfactory performances and reasonable effectiveness of EWD and EFD are not affected by their directness and simplicity. K-means clustering discretization (KMCD) [21] uses k-means clustering [22] to determine intervals for the discrete variables. Ordinal discretization (OD) [23,24] aims at taking advantage of the ordering information implicit in the continuous variables, so the ordering information of continuous variables is preserved when a transformation of discretized data is carried out. Fixed frequency discretization (FFD) [25], non-disjoint discretization (NDD) [26], proportional discretization (PD) [27], and weight proportional discretization (WPD) [28] are designed intentionally for managing the bias and variance generated during the discretization of continuous variables. The gratifying experimental results have been reported when these four discretization methods are applied to naïve Bayesian classifier [29,30]. Mean value and standard deviation discretization (MVSDD) [31] are applied to feature selection and the better experimental results are obtained when continuous variables are discretized by MVSDD.

However, the latest researches [25,57] have demonstrated that the loss of information will be generated as the consequence of discretization. For example, two different continuous values may be represented with the same discrete value. Then, the quantitative and ordinal differences will be lost. If the ignored information is used in the approximating mechanism of continuous entropy, then the renewed approximating mechanism will be distinct and should be more accurate.

In addition to using the unsupervised discretization to implement the continuous entropy computation, some sophisticated entropy estimators for continuous variables are accomplished to overcome the information loss of discretization methods and the handicaps encountered when computing the Eq. (2). The splitting data estimator (SDE) [34] and cross-validation estimator (CVE) [35] are two main implementations of re-substitution estimator model (RE) [32,33] in which the evaluation of integral is excluded and replaced by summation approximation of certain probability density function values. RE is a theoretical nonparametric estimation model with the mean square consistency. Under the given regularity conditions, RE are the first and second consistencies [32,33]. The $m$-spacing estimator ($m$ SE) [36] and $m_n$-spacing estimator ($m_n$ SE) [37] are two branches of estimates of entropy based on the sample-spacing which can be derived as a plug-in integral estimate using a spacing density estimate. Nearest neighbor distance estimator (NNDE) [38] is one of the estimates of entropy based on nearest neighbor distances. All of the above six entropy estimators try to approximate the unknown density function and avoid the troublesome integral evaluation so that the continuous entropy can be calculated easily.

In the mentioned-above sophisticated entropy estimators, RE is most widely used due to its theoretical mean square consistency. And, the estimation performance of RE depends strongly on the selection of bandwidth [32,33]. Thus, motivated by improving the estimation performance of RE via selecting an optimal bandwidth, we propose an entropy estimation strategy in which two different types of estimation errors are fused and a new error measure is derived accordingly. Given 24 typical density distributions, such empirical conclusion is firstly presented: in addition to the entropy estimation error named type-I error, other kind of error called density estimation error (simply type-II error) is also generated in the process of continuous entropy computation and there is always some inconsistency between the optimal bandwidths associated with these two errors. These two errors are all heavily dependent on the bandwidth parameters and react upon each other. So, a trade-off between type-I error and type-II error is considered as a fundamental and potential property of our proposed method called $RE_{I+II}$. Then, based on the theoretical consistencies of RE, a new and practical entropy estimation error measure is designed for $RE_{I+II}$ in this paper. By minimizing the designed estimation error, the optimal bandwidth can be calculated. The bandwidth selected by $RE_{I+II}$ balances the two generated errors. Finally, the simulation comparisons of $SE_{I+II}$ are carried out with 14 different entropy estimation methods based on 24 probability distributions. The experimental results reflect the following conclusions: (1) the optimal bandwidth used in $RE_{I+II}$ indeed performs better than the singled bandwidth by minimizing type-I error. And, compared with the sophisticated SDE and CVE, $SE_{I+II}$ can also obtain a satisfactory estimation. The simulations confirm the validity and effectiveness of the derivation error measure; (2) the discretized estimators are sensitive to the dataset size. It is not feasible to build the entropy estimation when facing with a large-scale dataset. With the increase of dataset size, the estimated error will increase significantly. On the contrary, $RE_{I+II}$ can reduce the estimated error with the augment of dataset size; (3) the overall performance of $RE_{I+II}$ also goes beyond the estimated behaviors of $m$ SE, $m_n$ SE, and NNDE. The application conditions of $m$ SE, $m_n$ SE, and NNDE are also discussed by the experimental comparisons.

The rest of the paper is organized as follows: In Sections 2 and 3, we summarize nine unsupervised discretization methods and 5 common entropy estimators. The Section 4 discusses the error generations in the process of entropy estimation with re-substitution estimator. The proposed entropy estimation method $RE_{I+II}$ is brought forward in Section 5. In Section 6, the experimental simulations are carried out and the corresponding analyses to empirical observations are also presented. Finally, we make a conclusion and outline the main directions for future research.

## 2. The typical discretization methods

To illustrate the method in question, we firstly introduce the number of denotations and explain their meanings:

$X$ is a continuous random variable taking the probability density function $f(X)$. $x_1, x_2, \ldots, x_n$ are $n$ continuous observations which obey the probability density $f(X)$. The goal of discretization is to divide the domain of continuous variable into some disjoint or non-disjoint intervals after sorting data in ascending or descending order so that every continuous observation can keep the information relative to a categorical value [18,42,43]. The endpoints of interval are called cut points or split points in the discretization context. In this section, we will review nine frequently-used unsupervised discretization methods as follows.

### 2.1. Equal width discretization-EWD

When discretizing $n$ continuous observations, EWD [16] divides the number space between $x_{\min}$ and $x_{\max}$ into $k$ intervals with the equal width, where $x_{\min} = \min\{x_1, x_2, \ldots, x_n\}$, $x_{\max} = \max\{x_1, x_2, \ldots, x_n\}$, and $k$ is a user predefined parameter. Thus, the width of every interval is $w = (x_{\max} - x_{\min})/k$ and the corresponding cut points are $x_{\min} + w, x_{\min} + 2w, \ldots, x_{\min} + (k-1)w$. All of the continuous observations in the $i$th interval $[x_{\min} + (i-1)w, x_{\min} + iw), i = 1, 2, \ldots, k$ (Note that $x_{\max}$ is put into the last interval) correspond to the same categorical value $x'_i (i = 1, 2, \ldots, k)$.

### 2.2. Equal frequency discretization-EFD

EFD [17] divides $n$ continuous observations into $k$ intervals so that each interval contains approximately the same number of continuous observations, where $k$ is a user predefined parameter. Thus, each interval contains $\lceil n/k \rceil$ ($\lceil u \rceil$ denotes the rounding of the element $u$ to the nearest integers towards infinity) continuous observations with adjacent (possibly identical) values. Note that the identical continuous observations must be placed in the same interval. In fact it is not always possible to generate $k$ equal frequency intervals and the number of continuous observations in the last interval is not always equal to $\lceil n/k \rceil$.

### 2.3. K-means clustering discretization-KMCD

KMCD [21] uses the k-means clustering technology [22] to determine the intervals for the continuous observations. The parameter $k$ needs to be determined beforehand. The continuous observations assembled into the same cluster correspond to the same categorical identifier.

### 2.4. Ordinal discretization-OD

OD [23,24] aims at taking advantage of the ordering information implicit in the continuous variables, so as not to make values 1 and 2 as dissimilar as values 1 and 1000 [30,44]. In fact, the theme of OD is extending the current one-dimensional variable into the multidimensional case. Let $X = \{x_1, x_2, \ldots, x_n\}$ be the attribute which takes $n$ continuous values. OD firstly discretizes $X$ by using some primary discretization method (e.g., EWD, EFD, or KMCD). Then, the discretized attribute $X^* = \{x_1^*, x_1^*, \ldots, x_k^*\}$ can be obtained. In order to implement the intention of preserving the ordering information, $k-1$ binary attributes are introduced as the extended attributes of original continuous attribute $X$. The $j$-th binary attribute $X_j^*(j = 1, 2, \ldots, k-1)$ denotes the special split to discretized attribute $X^* = \{x_1^*, x_1^*, \ldots, x_k^*\}$ : $X_j^* = \left\{\left\{x_1^*, \ldots, x_j^*\right\}, \left\{x_{j+1}^*, \ldots, x_k^*\right\}\right\}, j = 1, 2, \ldots, k-1$. The following Fig. 1 explains the mechanism of OD in detail.

From Fig. 1 we can see that OD converts one continuous variable into multiple discretized cases. In the classification context [23,24,30,45], the empirical results show that OD can improve the performance compared with the discretization technologies treating the ordered attributes as nominal quantities. However, unlike the classification application, it is really confusing which binary attribute should be used to compute the continuous entropy approximately. So, in this study we use the following formulas (3) to merge the $k-1$ individual entropies for the sake of preserving the useful ordering information:

$$H(X) \approx H(X^*) = \frac{\sum_{j=1}^{k-1} H\left(X_j^*\right)}{k-1}. \tag{3}$$

### 2.5. Fixed frequency discretization-FFD

The basic principle of FFD [25,30] is similar to EFD's. A sufficient interval frequency $m = 30$ is set to discretize the continuous variable. The empirical results show when $m = 30$, it is commonly seemed as the minimum sample size for managing the discretization bias and variance [25], the better performance can be obtained in naïve Bayesian classifier context [25,30]. The amount of intervals in FFD is $\lceil n/m \rceil$, where $n$ is the number of continuous observations. The continuous observations in an interval will be represented with the same qualitative identifier.

### 2.6. Non-disjoint discretization-NDD

The unique feature of NDD [25,26,30] is the non-disjoint (or overlapped) intervals allowing for generation. NDD forms a series of overlapping intervals for the continuous observations and always locates an observation value toward the middle of an interval. Firstly, the $k'$ "atomic intervals" need to be generated and every "atomic interval" contains $m'$ continuous observations, where $k'$ and $m'$ satisfy the following equation:

$$m' = m/3, \quad \text{and} \quad k' = n/m'. \tag{4}$$

In NDD, the value of interval frequency $m$ is also set to 30 [26] and $n$ denotes the number of continuous observations. Actually, a total of $\lceil n/10 \rceil$ "atomic intervals" are constructed, and each one contains 10 continuous observations.

After the "atomic intervals" are formed, a total of $k$ "actual intervals" can be constructed, where every "actual interval" is designed by combining three consecutive atomic intervals. The formation of "actual interval" can be presented in Fig. 2. Let the number of continuous observations be 100, viz. $n = 100$.

As a result, each actual interval has frequency equal to 30 by comprising three consecutive atomic intervals, and except in the case of falling into the first or the last atomic interval, a continuous observation is always towards the middle of its cor-

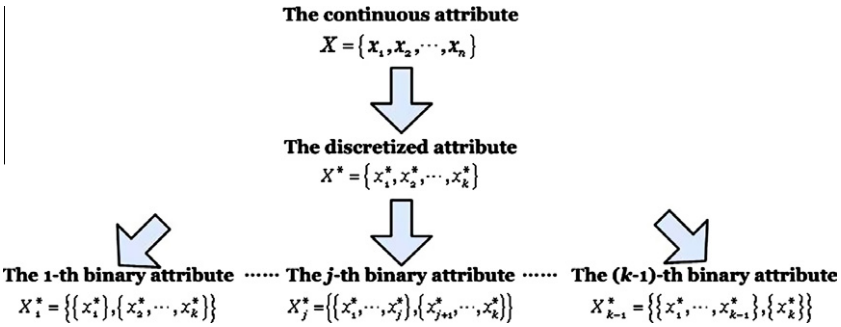



**Fig. 1.** The ordinal discretization.

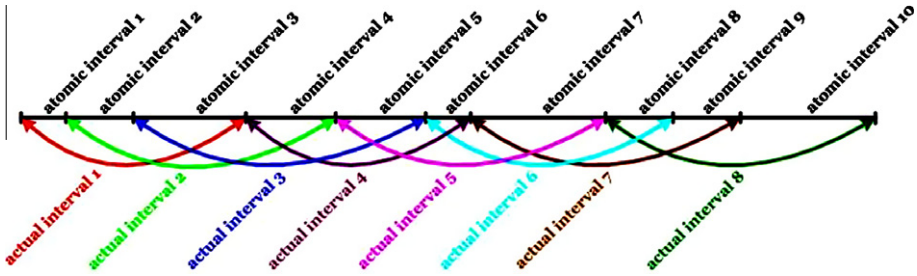



**Fig. 2.** The formation of actual intervals in NDD.

responding interval [25,26,30]. From the perspective of the whole set of continuous observations, the discretized intervals (i.e., actual intervals) formed for two continuous observations might overlap with each other.

### 2.7. Proportional discretization-PD

By tuning the interval size and number, PD [25,27,30] wants to balance the discretization bias and variance [25]. As described in Yang and Webb's works [25,27,30], increasing interval size (decreasing interval number) will decrease variance but increase bias. Conversely, decreasing interval size (increasing interval number) will decrease bias but increase variance. PD aims to resolve this conflict by setting the interval size and number proportional to the number of training instances. With increasing number of training instances, both discretization bias and variance tend to be decreasing. Bias can decrease because the interval number increases. Variance can decrease because the interval size increases. The desired interval size is $m$ and the desired interval number is $k$, PD employs the following Eq. (5) to calculate $m$ and $k$:

$$\begin{cases} m \times k = n \\ m = k \end{cases}, \tag{5}$$

where $n$ is the number of continuous observations. As a result, the interval size $m$ is equal to the interval number $k(m = k \approx \sqrt{n})$.

### 2.8. Weight proportional discretization-WPD

WPD [28,29] is the improved version of PD [25,27,30]. It is credible that variance reduction can contribute more to lower probability estimation error than bias reduction [46]. Thus, fewer intervals each containing more observations would be of greater utility [46]. Accordingly, WPD weighs discretization variance reduction more than bias reduction by setting a minimum interval size $m_{min}$ to make the probability estimation more reliable. Note that the increase of training data allows for the increase of both the interval size $m$ and the interval number $k$. Given the same definitions $m$ and $k$ as in Eq. (5), WPD employs the following Eq. (6) to calculate $m$ and $k$:

$$\begin{cases} m \times k = n \\ m - m_{\min} = k \end{cases}, \tag{6}$$

where, $m_{\min} = 30$ is the minimum interval size. The $m_{min}$ is set to 30 as it can mitigate PD's disadvantage on smaller datasets by establishing a suitable bias and variance trade-off, meanwhile it still retains PD's advantage on larger datasets by allowing additional training data to be used to reduce both bias and variance. As a result, the interval size $m$ can be approximated as $(30 + \sqrt{900 + 4n})/2$.

### 2.9. Mean value and standard deviation discretization-MVSDD

In Peng, Long, and Ding's work [31], they used $\mu \pm \alpha \cdot \sigma$ to discretize the continuous observations in the framework of feature selection, where $\mu$ and $\sigma$ are the mean value and standard deviation of continuous observations respectively, $\alpha = 0$, or 0.5, or 1 [51]. That is to say, the cut points of discretized intervals are $\mu - \alpha \times \sigma$ and $\mu + \alpha \times \sigma$. The empirical results show that the choice of $\alpha$ will have some influences on the ordering of selected features, but the selected features are almost the same [31,47]. MVSDD is actually a very robust discretization way for selecting features. So, MVSDD is also employed as a competitor in our study.

## 3. The existing continuous entropy estimators

In addition to calculate the continuous entropy with the discretization technologies [16,17,21,23,25–28,31] mentioned above, many continuous entropy estimators are also studied widely. Now, five commonly used entropy estimators, splitting data estimator (SDE) [34], cross-validation estimator (CVE) [35], $m$-spacing estimator ($m$ SE) [36], $m_n$-spacing estimator

($m_n$ SE) [37], and nearest neighbor distance estimator (NNDE) [38], will be presented as follows. Unlike the discretization way, these entropy estimators evaluate the continuous entropy from the continuous observations directly.

### 3.1. Splitting data estimator (SDE) and cross-validation estimator (CVE)

The estimators SDE [34] and CVE [35] are the derivatives of RE [32]. RE depends on the following Eq. (7) to evaluate the continuous entropy from the continuous observations $x_1, x_2, \ldots, x_n$:

$$H_{RE} = -\frac{1}{n}\sum_{i=1}^{n} \mathrm{In}[\hat{f}(x_i)], \tag{7}$$

where, $\hat{f}_n(X)$ is the estimated density function based on all $n$ continuous observations. RE provides a theoretical model for entropy estimation with density estimation approach. In Ahmad and Lin's work [32], the consistencies of RE had been proved under the regularity conditions: $E[H_{RE} - H]^2 \to 0$ as $n \to \infty$. Based on this theoretical nonparametric estimation model, SDE [34] and CVE [35] are proposed. SDE calculates the continuous entropy according to the following procedures: Firstly, the original dataset $X = \{x_1, x_2, \ldots, x_n\}$ is partitioned into two parts: $X' = \{x'_i | x'_i = x_{2i-1}, i = 1, 2, \ldots, \lfloor (n+1)/2 \rfloor\}$ and $X'' = \{x''_i | x''_i = x_{2i}, i = 1, 2, \ldots, \lfloor n/2 \rfloor\}$, where $\lfloor u \rfloor$ denotes that rounding the element $u$ to the nearest integers towards zero. Then, SDE employs Eq. (8) to compute the continuous entropy:

$$H_{\mathrm{SDE}} = -\frac{1}{\lfloor n/2 \rfloor}\sum_{i=1}^{\lfloor n/2 \rfloor} \mathrm{In}\left[\hat{f}_{\lfloor (n+1)/2 \rfloor}(x''_i)\right], \tag{8}$$

where, $\hat{f}_{\lfloor (n+1)/2 \rfloor}(X) = \frac{1}{\lfloor (n+1)/2 \rfloor h_{\mathrm{SDE}}}\sum_{i=1}^{\lfloor (n+1)/2 \rfloor} K\left(\frac{x - x'_i}{h_{\mathrm{SDE}}}\right)$ is the estimated density function by using Parzen window method [39] and $h_{\mathrm{SDE}}$ is the bandwidth parameter. In our study, Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$ is employed. The entropy estimation rule of CVE [35] is:

$$H_{\mathrm{CVE}} = -\frac{1}{n}\sum_{i=1}^{n} \mathrm{In}[\hat{f}_{-i}(x_i)], \tag{9}$$

where, $\hat{f}_{-i}(x_i) = \frac{1}{(n-1)h_{\mathrm{CVE}}}\sum_{j \neq i}^{n} K\left(\frac{x_i - x_j}{h_{\mathrm{CVE}}}\right)$ ($h_{\mathrm{CVE}}$ is the bandwidth parameter) denotes the kernel density estimation of $x_i (i = 1, 2, \ldots, n)$ based on the other samples, and this is taken for cross-validation.

For the determination of bandwidth $h_{\mathrm{SDE}}$ and $h_{\mathrm{CVE}}$, a lot of theoretical studies [33–35] had been proposed to define their properties. Due to the complex mathematical terms in the expression of mean square error, it always makes the choice of optimal bandwidth impractical [33,35]. So, the approximation rules for bandwidth selection are given. For example, Joe [33] considered the optimal bandwidth for CVE as:

$$h_{\mathrm{CVE}} = c \cdot n^{-1/4.5} \cdot \mathrm{IQR}(x_1, x_2, \ldots, x_n), c \in [0.75, 1], \tag{10}$$

where, $\mathrm{IQR}(x_1, x_2, \ldots, x_n)$ is the interquartile range [50] of dataset $x_1, x_2, \ldots, x_n$. Györfi and Meulen [34] also put forward the required bandwidth for SDE, viz. $h_{\mathrm{SDE}} \to n^{-\alpha}, \alpha \prec 1$. In this study, we let the optimal bandwidth for SDE be:

$$h_{\mathrm{SDE}} = c \cdot \lfloor (n+1)/2 \rfloor^{-1/4.5} \cdot \mathrm{IQR}(X'), c \in [0.75, 1], X' = \{x'_i | x'_i = x_{2i-1}, i = 1, 2, \ldots, \lfloor (n+1)/2 \rfloor\}. \tag{11}$$

### 3.2. m-spacing estimator (mSE) and $m_n$-spacing estimator ($m_n$ SE)

The spacing estimators [36,37] construct the density estimation based on the notion of "spacing". The so-called "spacing" can be established as the following steps. Firstly, the original dataset $X = \{x_1, x_2, \ldots, x_n\}$ is sorted in the ascending order and the ordering dataset is $\bar{X} = \{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n\}$, where $\bar{x}_1 \leqslant \bar{x}_2 \leqslant \ldots \leqslant \bar{x}_n$. Then, $\bar{x}_{i+m} - \bar{x}_i$ is called a spacing or order $m$, or $m$-spacing ($1 \leqslant i \leqslant i + m \leqslant n$). Based on the $m$-spacing, $m$ SE [36] considers the entropy computation as Eq. (12) for the fixed $m$:

$$H_{mSE} = \frac{1}{n}\sum_{i=1}^{n-m} \mathrm{In}\left[\frac{n}{m}(\bar{x}_{i+m} - \bar{x}_i)\right] - \Psi(m) + \mathrm{In}(m), \tag{12}$$

where, $\Psi(u)$ is the digamma function [48] which is the logarithmic derivative of the gamma function $\Gamma(u)$ [48], $\Psi(u) = d[\mathrm{In}\Gamma(u)]/du$. In the following comparison, without loss of generality, we let $m = 1$ in the Eq. (12).

In order to decrease the asymptotic estimation variance, $m_n$ SE [37] computes the continuous entropy with mathematical expression (13) by modifying Eq. (12) slightly:

$$H_{m_nSE} = \frac{1}{n}\sum_{i=1}^{n-m_n} \mathrm{In}\left[\frac{n}{m_n}(\bar{x}_{i+m_n} - \bar{x}_i)\right], \tag{13}$$

where, $m_n = n^{1/3}$ is designed in our study.

### 3.3. Nearest neighbor distance estimator (NNDE)

NNDE [38] is defined as the following Eq. (14) by considering the nearest neighbor methodology:

$$H_{\mathrm{NNDE}} = \frac{1}{n}\sum_{i=1}^{n}\ln(nd_i) + \ln 2 + \gamma_E, \tag{14}$$

where, $d_i = \min_{\substack{1 \leqslant j \leqslant n \\ j \neq i}} \|x_i - x_j\|$ denotes the nearest neighbor distance between $x_i (1 \leqslant i \leqslant n)$ and other data $x_j (1 \leqslant j \leqslant n, j \neq i)$, $\gamma_E$ is Euler-Mascheroni constant $\gamma_E \approx 0.57722$ [49].

## 4. Different error generations

The research objective of this work is to find a practical bandwidth selection rule for RE directly. As described above, the estimation model of RE is $H_{RE} = -\frac{1}{n}\sum_{i=1}^{n}\ln[\hat{f}(x_i)]$ based on the given dataset $X = \{x_1, x_2, \ldots, x_n\}$, where many technologies, for example, SDE, CVE, $m$SE, $m_n$SE, and NNDE, can be used to estimate the unknown density function. Here, we try to estimate this problematic density with the following mathematical equation [40,41,60]:

$$\hat{f}(x_i) = \frac{1}{nh_{SE}}\sum_{j=1}^{n} K\left(\frac{x_i - x_j}{h_{SE}}\right) = \frac{1}{nh_{SE}}\sum_{j=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x_i - x_j}{h_{SE}}\right)^2\right], \quad (i = 1, 2, \ldots, n). \tag{15}$$

By bringing the Eq. (15) into Eq. (7), the entropy estimator can be summarized as follows:

$$H_{RE} = -\frac{1}{n}\sum_{i=1}^{n}\ln[\hat{f}(x_i)] = -\frac{1}{n}\sum_{i=1}^{n}\ln\left\{\frac{1}{nh_{SE}}\sum_{j=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x_i - x_j}{h_{SE}}\right)^2\right]\right\}. \tag{16}$$

From Eqs. (15) and (16), we can see these two estimations are all dependent on the bandwidth $h_{SE}$ which imposes a direct influence on density estimation and an indirect influence on entropy estimation. In order to quantify these influences, the necessary measures should be introduced. The estimation errors can depict the corresponding estimation performances. In Joe's work [33], the mean square error is employed to measure the entropy estimation error. And, the $L_1$-error criterion is used in SDE by Györfi and Meulen [34]. Using the experiences of these two typical error criteria for reference, we employ the following mean square error to evaluate the performance of entropy estimation:

$$E_{\mathrm{I}} = E[H_{RE}(X) - H(X)]^2. \tag{17}$$

Meanwhile, the integrated squared error, given by

$$E_{\mathrm{II}} = \int [\hat{f}(X) - f(X)]^2 dx, \tag{18}$$

is used as the criterion function to measure the density estimation error. The main reason that $E_{\mathrm{II}}$ is used is that it can estimate $f(X)$ over the whole data space and measure the error between $f(X)$ and $\hat{f}(X)$ globally. For the sake of convenience, we note $E_{\mathrm{I}}$ as type-I error and $E_{\mathrm{II}}$ as type-II error.

In order to display the influences of bandwidths on these two estimation errors more intuitively, an experimental simulation is carried out. In this experiment, 24 typical probability distributions are used as the testing pools. Table 1 itemizes the detailed information of these 24 probability distributions, including the density functions, the continuous entropy values [52], and the corresponding support intervals.

Our experiment is arranged as the following procedures:

Step 1: For every probability density in Table 1, 150 random samples are generated $X^{(p)} = \left\{x_1^{(p)}, x_2^{(p)}, x_{150}^{(p)}\right\}, (p = 1, 2, \ldots, 24)$;

Step 2: Use Eq. (15) to estimate the density value $\hat{f}\left(x_i^{(p)}, h\right)$ for each data point $x_i^{(p)} (i = 1, 2, \ldots, 150)$, and compute the error between the estimated density value $\hat{f}\left(x_i^{(p)}, h\right)$ and the true density value $f\left(x_i^{(p)}\right)$, noted as $\left[\hat{f}\left(x_i^{(p)}, h\right) - f\left(x_i^{(p)}\right)\right]^2$, $(i = 1, 2, \ldots, 150)$;

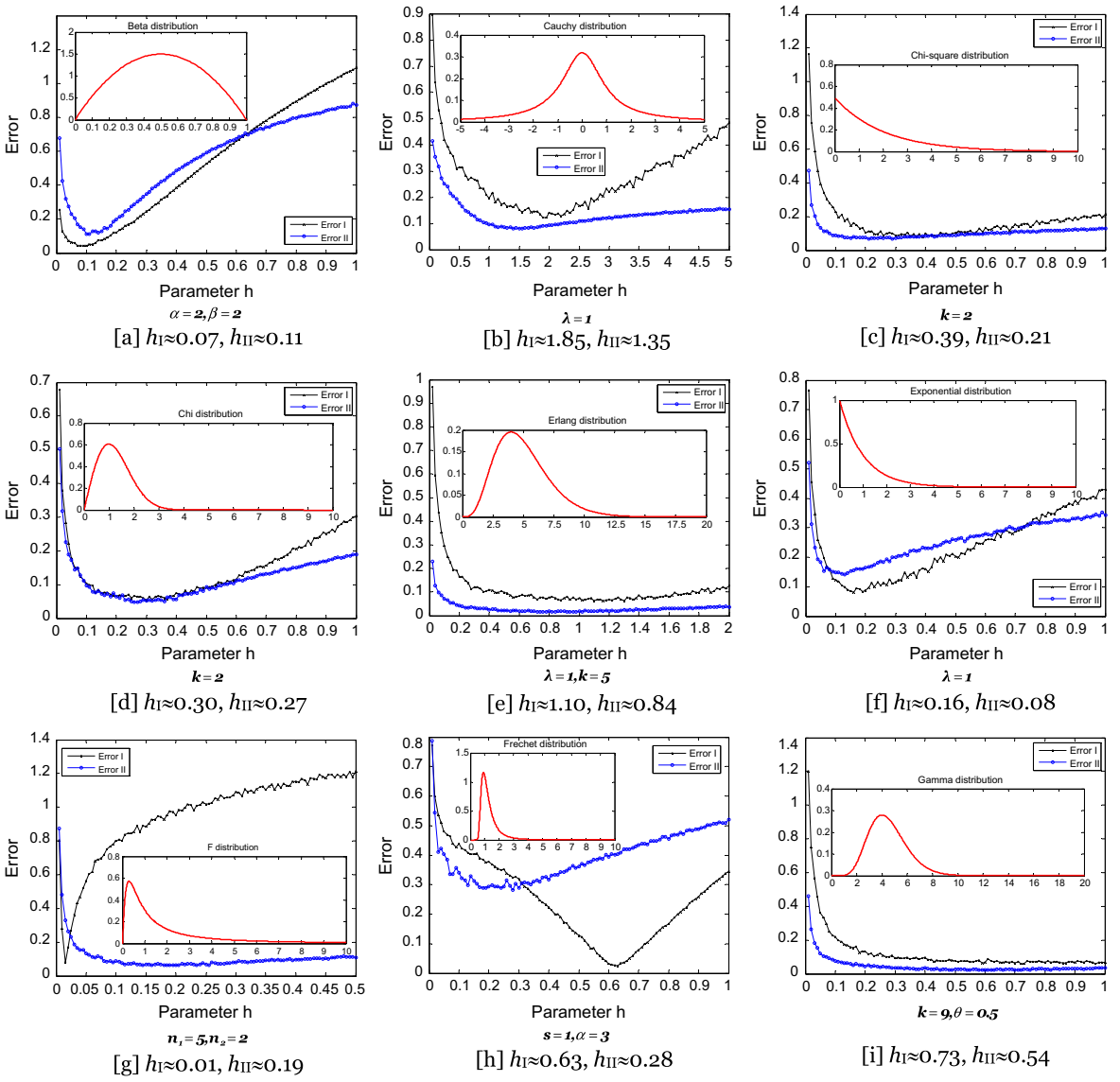
Step 3: The density estimation error on dataset $X^{(p)} = \left\{x_1^{(p)}, x_2^{(p)}, \ldots, x_{150}^{(p)}\right\}, (p = 1, 2, \ldots, 24)$ is $E_{\mathrm{II}}^{(p)} = \frac{1}{150}\sum_{i=1}^{150}\left[\hat{f}\left(x_i^{(p)}, h\right) - f\left(x_i^{(p)}\right)\right]^2, (p = 1, 2, \ldots, 24)$;

Step 4: Use Eq. (16) to estimate the entropy value $H_{RE}(X^{(p)}, h), (p = 1, 2, \ldots, 24)$, and the entropy estimation error between the estimated entropy and the true entropy $H(X^{(p)}), (p = 1, 2, \ldots, 24)$, noted as $E_{\mathrm{I}}^{(p)} = [H_{RE}(X^{(p)}, h) - H(X^{(p)})]^2$, $(p = 1, 2, \ldots, 24)$;

**Table 1**
The typical 24 probability distributions.

| # | Distribution | Density function | Continuous entropy value $H(x)=\int_{-\infty}^{+\infty} f(x)\ln[f(x)]dx$ | Support interval |
|---|---|---|---|---|
| 1 | Beta | $f(x)=\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \alpha\succ 0, \beta\succ 0$ | $\ln[B(\alpha,\beta)]-(\alpha-1)\Psi(\alpha)-(\beta-1)\Psi(\beta)+(\alpha+\beta-2)\Psi(\alpha+\beta)$ [a,b] | $x\in[0,1]$ |
| 2 | Cauchy | $f(x)=\frac{1}{\pi}\left(\frac{\lambda}{x^2+\lambda^2}\right), \lambda\succ 0$ | $\ln(4\pi\lambda)$ | $x\in(-\infty,+\infty)$ |
| 3 | Central Chi-Squared | $f(x)=\frac{2^{-\frac{k}{2}}x^{\frac{k}{2}-1}e^{-\frac{x}{2}}}{\Gamma(k/2)}, k\succ 0$ | $\ln[2\Gamma(k/2)]+(1-\frac{k}{2})\Psi(k/2)+\frac{k}{2}$ [b] | $x\in[0,+\infty)$ |
| 4 | Chi | $f(x)=\frac{2^{1-\frac{k}{2}}x^{k-1}e^{-\frac{x^2}{2}}}{\Gamma(k/2)}, k\succ 0$ | $\ln\left[\frac{\Gamma(k/2)}{\sqrt{2}}\right]-\frac{k-1}{2}\Psi(k/2)+\frac{k}{2}$ | $x\in[0,+\infty)$ |
| 5 | Erlang | $f(x)=\frac{\lambda^k x^{k-1}e^{-\lambda x}}{(k-1)!}, k\succ 0, \lambda\succ 0$ | $(1-k)\Psi(k)+\ln\left[\frac{\Gamma(k)}{\lambda}\right]+k$ | $x\in[0,+\infty)$ |
| 6 | Exponential | $f(x)=\lambda\exp(-\lambda x), \lambda\succ 0$ | $1-\ln(\lambda)$ | $x\in[0,+\infty)$ |
| 7 | F | $f(x)=\frac{n_1^{\frac{n_1}{2}}n_2^{\frac{n_2}{2}}}{B(\frac{n_1}{2},\frac{n_2}{2})}\frac{x^{\frac{n_1}{2}-1}}{(n_2+n_1x)^{\frac{n_1+n_2}{2}}}, n_1\succ 0, n_2\succ 0$ | $\ln\left[\frac{n_1}{n_2}B(\frac{n_1}{2},\frac{n_2}{2})\right]+(1-\frac{n_1}{2})\Psi(\frac{n_1}{2})-(1+\frac{n_2}{2})\Psi(\frac{n_2}{2})+\frac{n_1+n_2}{2}\Psi(\frac{n_1+n_2}{2})$ | $x\in[0,+\infty)$ |
| 8 | Frechet | $f(x)=\frac{\alpha}{s}\left(\frac{x}{s}\right)^{-1-\alpha}e^{(-\frac{x}{s})^{-\alpha}}, \alpha\succ 0, s\succ 0$ | $1+\frac{\gamma_E}{\alpha}+\gamma_E+\ln(\frac{s}{\alpha})$ [c] | $x\in(0,+\infty)$ |
| 9 | Gamma | $f(x)=x^{k-1}\frac{e^{-x/\theta}}{\theta^k\Gamma(k)}, k\succ 0, \theta\succ 0$ | $k+\ln\theta+\ln[\Gamma(k)]+(1-k)\Psi(k)$ | $x\in[0,+\infty)$ |
| 10 | Gumbel | $f(x)=\frac{z\exp(-z)}{\beta}, z=\exp[-(x-\mu)/\beta], \mu\in R, \beta\succ 0$ | $\ln(\beta)+\gamma_E+1$ | $x\in(-\infty,+\infty)$ |
| 11 | Hyperbolic Secant | $f(x)=\frac{1}{2}\mathrm{sech}(\frac{\pi}{2}x)$ | $(4/\pi)K$ [d] | $x\in(-\infty,+\infty)$ |
| 12 | Laplace | $f(x)=\frac{1}{2b}\exp\left(-\frac{\lvert x-\mu\rvert}{b}\right), \mu\in R, b\succ 0$ | $\ln(2b)+1$ | $x\in(-\infty,+\infty)$ |
| 13 | Logistic | $f(x)=\frac{1}{4s}\mathrm{sech}^2(\frac{x-\mu}{2s}), \mu\in R, s\succ 0$ | $\ln(s)+2$ | $x\in(-\infty,+\infty)$ |
| 14 | Lognormal | $f(x)=\frac{1}{x\sigma\sqrt{2\pi}}\exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right], \mu\in R, \sigma^2\succ 0$ | $\frac{1}{2}+\frac{1}{2}\ln(2\pi\sigma^2)+\mu$ | $x\in(0,+\infty)$ |
| 15 | Normal | $f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \mu\in R, \sigma^2\succ 0$ | $\ln(\sigma\sqrt{2\pi e})$ | $x\in(-\infty,+\infty)$ |
| 16 | Pareto | $f(x)=\frac{\alpha x_m^\alpha}{x^{\alpha+1}}, x_m\succ 0, \alpha\succ 0$ | $\ln(\frac{x_m}{\alpha})+\frac{1}{\alpha}+1$ | $x\in[x_m,+\infty)$ |
| 17 | Rayleigh | $f(x)=\frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right), \sigma\succ 0$ | $1+\ln\left(\frac{\sigma}{\sqrt{2}}\right)+\frac{\gamma_E}{2}$ | $x\in[0,+\infty)$ |
| 18 | Semicircle | $f(x)=\frac{2}{\pi R^2}\sqrt{R^2-x^2}, R\succ 0$ | $\ln(\pi R)-\frac{1}{2}$ | $x\in[-R,R]$ |
| 19 | Student's-t | $f(x)=\frac{(1+x^2/v)^{-\frac{v+1}{2}}}{\sqrt{v}B(1/2,v/2)}, v\succ 0$ | $\frac{v+1}{2}\left[\Psi(\frac{v+1}{2})-\Psi(\frac{v}{2})\right]+\ln\left[\sqrt{v}B(\frac{1}{2},\frac{v}{2})\right]$ | $x\in(-\infty,+\infty)$ |
| 20 | Triangular | $f(x)=\begin{cases}\frac{2(x-a)}{(b-a)(c-a)}, & a\leqslant x\leqslant c\\ \frac{2(b-x)}{(b-a)(b-c)}, & c\leqslant x\leqslant b\end{cases}$ | $\ln(\frac{b-a}{2})+\frac{1}{2}$ | $x\in[a,b]$ |
| 21 | Truncated Normal | $f(x)=\frac{\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma})-\Phi(\frac{a-\mu}{\sigma})}, \mu\in R, a\in R, b\in R, \sigma^2\succ 0$ | $\ln\left(\sigma Z\sqrt{2\pi e}\right)+\frac{\alpha\phi(\alpha)-\beta\phi(\beta)}{2Z}-\frac{[\phi(\alpha)-\phi(\beta)]^2}{2Z^2}, \alpha=\frac{a-\mu}{\sigma}, \beta=\frac{b-\mu}{\sigma}, Z=\Phi(\beta)-\Phi(\alpha)$ | $x\in[a,b]$ |
| 22 | Uniform | $f(x)=\frac{1}{b-a}$ | $\ln(b-a)$ | $x\in[a,b]$ |
| 23 | Von Mises | $f(x)=\frac{e^{k\cos(x-\mu)}}{2\pi I_0(k)}, \mu\in R, k\succ 0$ | $-k\frac{I_1(k)}{I_0(k)}+\ln[2\pi I_0(k)]$ [e] | $x\in[-\pi,\pi]$ |
| 24 | Weibull | $f(x)=\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}e^{-(x/\lambda)^k}, \lambda\succ 0, k\succ 0$ | $\gamma_E(1-\frac{1}{k})+\ln(\frac{\lambda}{k})+1$ | $x\in[0,+\infty)$ |

[a] $B(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$ is the Beta function.
[b] $\Psi(X)=d\ln[\Gamma(X)]/dx$ is the Digamma function, where $\Gamma(X)=\int_0^\infty u^{x-1}e^{-u}du$ is the Gamma function.
[c] $\gamma_E$ is Euler-Mascheroni's constant $\gamma_E=-\Psi(1)\approx 0.57722$.
[d] $K$ is Catalan's constant $G=\sum_{n=0}^{\infty}(-1)^n/(2n+1)^2\approx 0.91597$.
[e] $I_0(X)$ is the modified Bessel function of order 0, and $I_1(X)$ is the modified Bessel function of order 1.




[a] $h_{\mathrm{I}}\approx 0.07$, $h_{\mathrm{II}}\approx 0.11$ [b] $h_{\mathrm{I}}\approx 1.85$, $h_{\mathrm{II}}\approx 1.35$ [c] $h_{\mathrm{I}}\approx 0.39$, $h_{\mathrm{II}}\approx 0.21$

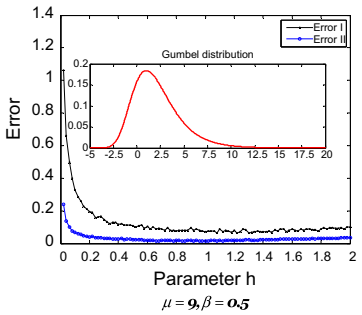
[d] $h_{\mathrm{I}}\approx 0.30$, $h_{\mathrm{II}}\approx 0.27$ [e] $h_{\mathrm{I}}\approx 1.10$, $h_{\mathrm{II}}\approx 0.84$ [f] $h_{\mathrm{I}}\approx 0.16$, $h_{\mathrm{II}}\approx 0.08$

[g] $h_{\mathrm{I}}\approx 0.01$, $h_{\mathrm{II}}\approx 0.19$ [h] $h_{\mathrm{I}}\approx 0.63$, $h_{\mathrm{II}}\approx 0.28$ [i] $h_{\mathrm{I}}\approx 0.73$, $h_{\mathrm{II}}\approx 0.54$

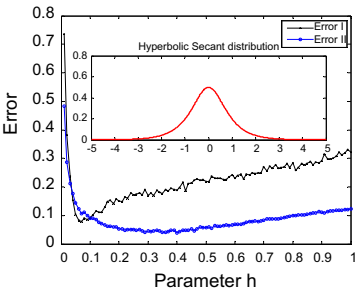
**Fig. 3.** Type-I error and Type-II error along parameter $h$ change.

Step 5: Set $h$ ranging from 0.0001 to 5 in step of 0.0001. Repeat Steps 1–4 for different values of $h$ and 100 times for each value. The average values of type-I errors and type-II errors on the 100 repetitions are recorded respectively. The main purpose of getting the average of 100 results is to reduce the randomness of data. The learning curves are plotted in Fig. 3.

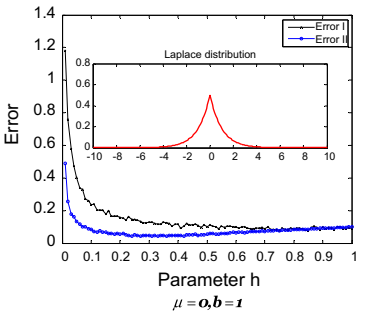
Fig. 3 also gives the parameter(s) used in every distribution. From these comparisons, two experimental observations can be easily summarized: (1) with the increase of parameter $h$, all the type-I errors show the trends of first decrease and then increase; (2) as $h$ keeps increasing, all the type-II errors also hold the trends of first decrease and then increase. These two observations reflect that there are optimal bandwidths that can make the type-I error and type-II error reach the corresponding minimums. The optimal bandwidths $h_{\mathrm{I}}$ and $h_{\mathrm{II}}$ for type-I error and type-II error are selected. However, we can find that the inconsistent optimal bandwidths are conspicuous. Our purpose is to minimize the generated errors, including type-I error and type-I error, in the process of entropy calculation when density estimation technology is used in RE. The reason that optimal bandwidth which minimizes the type-II error is not the best bandwidth selection for the entropy estimation error is as follows.



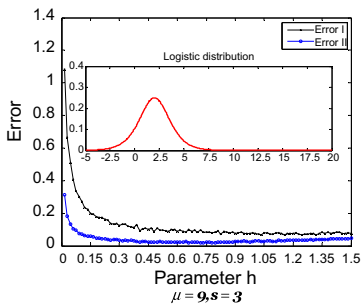

[j] $h_{\mathrm{I}}\approx 1.12$, $h_{\mathrm{II}}\approx 0.90$



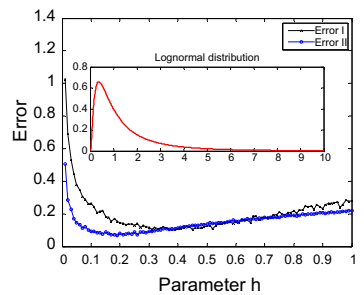

[k] $h_{\mathrm{I}}\approx 0.07$, $h_{\mathrm{II}}\approx 0.40$



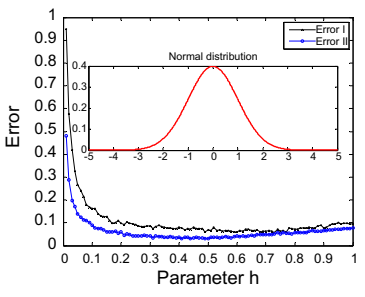

[l] $h_{\mathrm{I}}\approx 0.85$, $h_{\mathrm{II}}\approx 0.33$



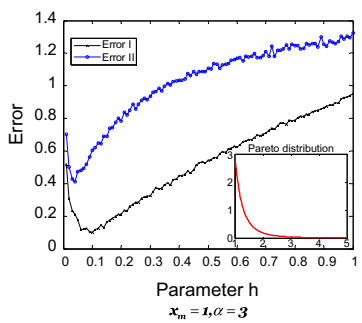

[m] $h_{\mathrm{I}}\approx 1.24$, $h_{\mathrm{II}}\approx 0.66$



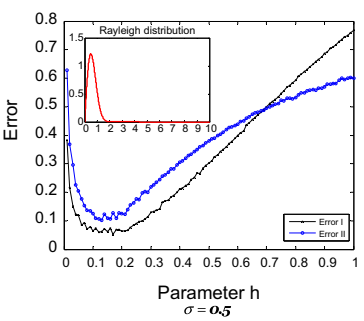

[n] $h_{\mathrm{I}}\approx 0.39$, $h_{\mathrm{II}}\approx 0.19$



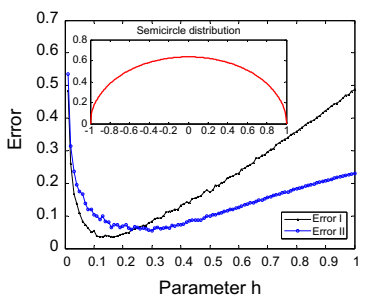

[o] $h_{\mathrm{I}}\approx 0.61$, $h_{\mathrm{II}}\approx 0.37$




[p] $h_{\mathrm{I}}\approx 0.10$, $h_{\mathrm{II}}\approx 0.03$




[q] $h_{\mathrm{I}}\approx 0.17$, $h_{\mathrm{II}}\approx 0.13$




[r] $h_{\mathrm{I}}\approx 0.12$, $h_{\mathrm{II}}\approx 0.30$

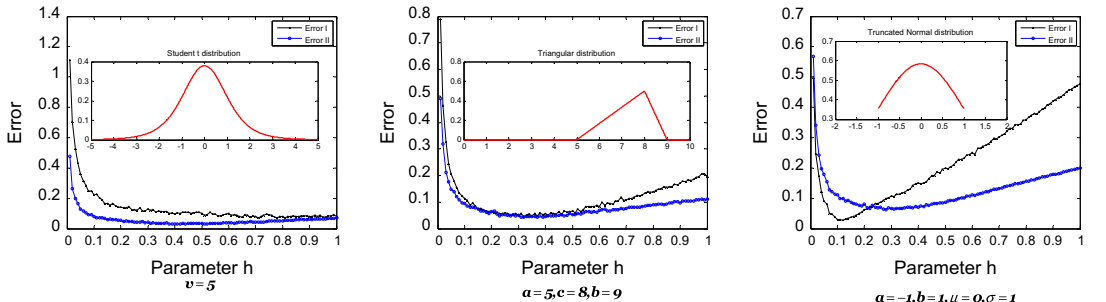
**Fig. 3.** (*continued*)

For the given dataset $X=\{x_1,x_2,\ldots,x_n\}$ obeying the density function $f(X)$, let $h_{\mathrm{I}}=\arg\min_h[(H_{RE}-H)^2]$ and $h_{\mathrm{II}}=\arg\min_h\left\{\frac{1}{n}\sum_{i=1}^{n}[\hat{f}(x_i)-f(x_i)]^2\right\}$. Because when $n\to\infty$, $H=-\frac{1}{n}\sum_{i=1}^{n}\mathrm{ln}[f(x_i)]$ is the unbiased alternative to $H=-\int_{-\infty}^{+\infty}f(X)\mathrm{ln}[f(X)]dx$, and root-$n$ is consistent [35],

$$(H_{RE}-H)^2=\left[\left\{-\frac{1}{n}\sum_{i=1}^{n}\mathrm{ln}[\hat{f}(x_i)]\right\}-\left\{-\frac{1}{n}\sum_{i=1}^{n}\mathrm{ln}[f(x_i)]\right\}\right]^2=\frac{1}{n^2}\left[\sum_{i=1}^{n}\left\{\mathrm{ln}\left[\hat{f}(x_i)\right]-\mathrm{ln}[f(x_i)]\right\}\right]^2. \tag{19}$$

For $\forall i\in\{1,2,\ldots,n\}$, $\hat{f}(x_i)=\frac{1}{nh}\sum_{j=1}^{n}K\left(\frac{x_i-x_j}{h}\right)=\frac{1}{nh}\sum_{j=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x_i-x_j}{h}\right)^2\right]=\frac{1}{n}\sum_{j=1}^{n}\frac{1}{\sqrt{2\pi}h}\exp\left[-\frac{1}{2}\left(\frac{x_i-x_j}{h}\right)^2\right]$. With the increase of parameter $h$, it is easy to verify that the value of the estimated $\hat{f}(x_i)$ decreases monotonously. Then, in the process of $h$ being gradually increasing, we discuss the selections of $h_{\mathrm{I}}$ and $h_{\mathrm{II}}$ according to the following three cases:

**Case I:** Assume that $h_{\mathrm{II}}=\arg\min_h\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}(x_i)-f(x_i)\right]^2\right\}$ such that $\hat{f}(x_i,h_{\mathrm{II}})-f(x_i)\succ 0,\forall i\in\{1,2,\ldots,n\}$. And, $\forall h_{\mathrm{II}}^{*}\in \mathrm{H}_1$, where $H_1$ is the universe of discourse of the discussed parameter h, such that $\hat{f}(x_i,h_{\mathrm{II}}^{*})-f(x_i)\succ 0$, $\forall i\in\{1,2,\ldots,n\}$. If $\frac{1}{n}\sum_{i=1}^{n}[\hat{f}(x_i,h_{\mathrm{II}})-f(x_i)]^2\prec\frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}(x_i,h_{\mathrm{II}}^{*})-f(x_i)\right]^2$, then the inequality $[H_{RE}(h_{\mathrm{II}})-H]^2\prec\left[H_{RE}(h_{\mathrm{II}}^{*})-H\right]^2$ holds.




[s] $h_{\mathrm{I}}\approx 0.81$, $h_{\mathrm{II}}\approx 0.52$ [t] $h_{\mathrm{I}}\approx 0.34$, $h_{\mathrm{II}}\approx 0.29$ [u] $h_{\mathrm{I}}\approx 0.12$, $h_{\mathrm{II}}\approx 0.28$

[v] $h_{\mathrm{I}}\approx 0.25$, $h_{\mathrm{II}}\approx 0.55$ [w] $h_{\mathrm{I}}\approx 0.39$, $h_{\mathrm{II}}\approx 0.24$ [x] $h_{\mathrm{I}}\approx 0.14$, $h_{\mathrm{II}}\approx 0.09$

**Fig. 3.** (*continued*)

Based on $\frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}(x_i,h_{\mathrm{II}})-f(x_i)\right]^2 \prec \frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}(x_i,h_{\mathrm{II}}^*)-f(x_i)\right]^2$, $\hat{f}(x_i,h_{\mathrm{II}})-f(x_i)\succ 0$, and $\hat{f}(x_i,h_{\mathrm{II}}^*)-f(x_i)\succ 0, \forall i\in\{1,2,\ldots,n\}$, we can get the conclusion: $h_{\mathrm{II}}\succ h_{\mathrm{II}}^*$ and $\hat{f}(x_i,h_{\mathrm{II}})\prec\hat{f}(x_i,h_{\mathrm{II}}^*), \forall i\in\{1,2,\ldots,n\}$. Because if $h_{\mathrm{II}}\prec h_{\mathrm{II}}^*$, then, for $\forall i\in\{1,2,\ldots,n\}$,

$$\begin{aligned}
&\left.\begin{array}{l}\hat{f}(x_i,h_{\mathrm{II}})-f(x_i)\succ 0\\ \hat{f}(x_i,h_{\mathrm{II}}^*)-f(x_i)\succ 0\end{array}\right\}\Rightarrow [\hat{f}(x_i,h_{\mathrm{II}})-f(x_i)]^2\succ[\hat{f}(x_i,h_{\mathrm{II}}^*)-f(x_i)]^2\\
&\hat{f}(x_i,h_{\mathrm{II}})\succ\hat{f}(x_i,h_{\mathrm{II}}^*)\Rightarrow\hat{f}(x_i,h_{\mathrm{II}})-f(x_i)\succ\hat{f}(x_i,h_{\mathrm{II}}^*)-f(x_i)\\
&\Rightarrow\frac{1}{n}\sum_{i=1}^{n}[\hat{f}(x_i,h_{\mathrm{II}})-f(x_i)]^2\succ\frac{1}{n}\sum_{i=1}^{n}[\hat{f}(x_i,h_{\mathrm{II}}^*)-f(x_i)]^2.
\end{aligned}$$

This is contradictory to the premise supposition $\frac{1}{n}\sum_{i=1}^{n}[\hat{f}(x_i,h_{\mathrm{II}})-f(x_i)]^2\prec\frac{1}{n}\sum_{i=1}^{n}[\hat{f}(x_i,h_{\mathrm{II}}^*)-f(x_i)]^2$.
Because $h_{\mathrm{II}}\succ h_{\mathrm{II}}^*$, for $\forall i\in\{1,2,\ldots,n\}$,

$$\begin{aligned}
&\hat{f}(x_i,h_{\mathrm{II}})\prec\hat{f}(x_i,h_{\mathrm{II}}^*)\Rightarrow\ln[\hat{f}(x_i,h_{\mathrm{II}})]\prec\ln[\hat{f}(x_i,h_{\mathrm{II}}^*)]\Rightarrow\ln[\hat{f}(x_i,h_{\mathrm{II}})]-\ln[f(x_i)]\prec\ln[\hat{f}(x_i,h_{\mathrm{II}}^*)]-\ln[f(x_i)]\\
&\Rightarrow\sum_{i=1}^{n}\{\ln[\hat{f}(x_i,h_{\mathrm{II}})]-\ln[f(x_i)]\}\\
&\left.\begin{array}{l}\hat{f}(x_i,h_{\mathrm{II}})-f(x_i)\succ 0\Rightarrow\ln[\hat{f}(x_i,h_{\mathrm{II}})]-\ln[f(x_i)]\succ 0\\ \hat{f}(x_i,h_{\mathrm{II}}^*)-f(x_i)\succ 0\Rightarrow\ln[\hat{f}(x_i,h_{\mathrm{II}}^*)]-\ln[f(x_i)]\succ 0\end{array}\right\}\Rightarrow\\
&\prec\sum_{i=1}^{n}\{\ln[\hat{f}(x_i,h_{\mathrm{II}}^*)]-\ln[f(x_i)]\}\\
&\times\frac{1}{n^2}\left[\sum_{i=1}^{n}\{\ln[\hat{f}(x_i,h_{\mathrm{II}})]-\ln[f(x_i)]\}\right]^2\\
&\prec\frac{1}{n^2}\left[\sum_{i=1}^{n}\{\ln[\hat{f}(x_i,h_{\mathrm{II}}^*)]-\ln[f(x_i)]\}\right]^2\Rightarrow[H_{RE}(h_{\mathrm{II}})-H]^2\prec[H_{RE}(h_{\mathrm{II}}^*)-H]^2.
\end{aligned}$$

This result indicates that if $h_{\mathrm{II}}=\arg\min_h\left\{\frac{1}{n}\sum_{i=1}^{n}[\hat{f}(x_i)-f(x_i)]^2\right\}$ such that $\hat{f}(x_i,h_{\mathrm{II}})-f(x_i)\succ 0, \forall i\in\{1,2,\ldots,n\}$, then $h_{\mathrm{I}}=h_{\mathrm{II}}=\arg\min_h[(H_{RE}-H)^2]$. □

**Case II:** Assume that $h_{\mathrm{II}} = \arg\min_h \left\{\frac{1}{n}\sum_{i=1}^{n}[\hat{f}(x_i) - f(x_i)]^2\right\}$ such that $\hat{f}(x_i, h_{\mathrm{II}}) - f(x_i) \prec 0, \forall i \in \{1, 2, \ldots, n\}$. And, $\forall h_{\mathrm{II}}^* \in \mathrm{H}_2$, where $H_2$ is the universe of discourse of the discussed parameter h, such that $\hat{f}(x_i, h_{\mathrm{II}}^*) - f(x_i) \prec 0, \forall i \in \{1, 2, \ldots, n\}$. If $\frac{1}{n}\sum_{i=1}^{n}[\hat{f}(x_i, h_{\mathrm{II}}) - f(x_i)]^2 \prec \frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}(x_i, h_{\mathrm{II}}^*) - f(x_i)\right]^2$, then the inequality $[H_{RE}(h_{\mathrm{II}}) - H]^2 \prec \left[H_{RE}(h_{\mathrm{II}}^*) - H\right]^2$ holds.

Based on $\frac{1}{n}\sum_{i=1}^{n}[\hat{f}(x_i, h_{\mathrm{II}}) - f(x_i)]^2 \prec \frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}(x_i, h_{\mathrm{II}}^*) - f(x_i)\right]^2$, $\hat{f}(x_i, h_{\mathrm{II}}) - f(x_i) \prec 0$, and $\hat{f}(x_i, h_{\mathrm{II}}^*) - f(x_i) \prec 0, \forall i \in \{1, 2, \ldots, n\}$, we can get the conclusion: $h_{\mathrm{II}} \prec h_{\mathrm{II}}^*$ and $\hat{f}(x_i, h_{\mathrm{II}}) \succ \hat{f}(x_i, h_{\mathrm{II}}^*), \forall i \in \{1, 2, \ldots, n\}$. Because if $h_{\mathrm{II}} \succ h_{\mathrm{II}}^*$, then, for $\forall i \in \{1, 2, \ldots, n\}$,

$$\hat{f}(x_i, h_{\mathrm{II}}) \prec \hat{f}(x_i, h_{\mathrm{II}}^*) \Rightarrow \hat{f}(x_i, h_{\mathrm{II}}) - f(x_i) \prec \hat{f}(x_i, h_{\mathrm{II}}^*) - f(x_i) \overset{\substack{\hat{f}(x_i, h_{\mathrm{II}}) - f(x_i) \prec 0 \\ \hat{f}(x_i, h_{\mathrm{II}}^*) - f(x_i) \prec 0}}{\Rightarrow} [\hat{f}(x_i, h_{\mathrm{II}}) - f(x_i)]^2 \succ \left[\hat{f}(x_i, h_{\mathrm{II}}^*) - f(x_i)\right]^2$$
$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n}[\hat{f}(x_i, h_{\mathrm{II}}) - f(x_i)]^2 \succ \frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}(x_i, h_{\mathrm{II}}^*) - f(x_i)\right]^2.$$

This is contradictory to the premise supposition $\frac{1}{n}\sum_{i=1}^{n}[\hat{f}(x_i, h_{\mathrm{II}}) - f(x_i)]^2 \prec \frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}(x_i, h_{\mathrm{II}}^*) - f(x_i)\right]^2$.
Because $h_{\mathrm{II}} \prec h_{\mathrm{II}}^*$, for $\forall i \in \{1, 2, \ldots, n\}$,

$$\hat{f}(x_i, h_{\mathrm{II}}) \succ \hat{f}(x_i, h_{\mathrm{II}}^*) \Rightarrow \mathrm{In}\left[\hat{f}(x_i, h_{\mathrm{II}})\right] \succ \mathrm{In}\left[\hat{f}(x_i, h_{\mathrm{II}}^*)\right] \Rightarrow \mathrm{In}[\hat{f}(x_i, h_{\mathrm{II}})] - \mathrm{In}[f(x_i)] \succ \mathrm{In}\left[\hat{f}(x_i, h_{\mathrm{II}}^*)\right] - \mathrm{In}[f(x_i)]$$
$$\Rightarrow \sum_{i=1}^{n}\{\mathrm{In}[\hat{f}(x_i, h_{\mathrm{II}})] - \mathrm{In}[f(x_i)]\}$$
$$\succ \sum_{i=1}^{n}\left\{\mathrm{In}\left[\hat{f}(x_i, h_{\mathrm{II}}^*)\right] - \mathrm{In}[f(x_i)]\right\} \overset{\substack{\hat{f}(x_i, h_{\mathrm{II}}) - f(x_i) \prec 0 \Rightarrow \mathrm{In}[\hat{f}(x_i, h_{\mathrm{II}})] - \mathrm{In}[f(x_i)] \prec 0 \\ \hat{f}(x_i, h_{\mathrm{II}}^*) - f(x_i) \prec 0 \Rightarrow \mathrm{In}\left[\hat{f}(x_i, h_{\mathrm{II}}^*)\right] - \mathrm{In}[f(x_i)] \prec 0}}{\Rightarrow}$$
$$\times \frac{1}{n^2}\left[\sum_{i=1}^{n}\left\{\mathrm{In}[\hat{f}(x_i, h_{\mathrm{II}})] - \mathrm{In}[f(x_i)]\right\}\right]^2$$
$$\prec \frac{1}{n^2}\left[\sum_{i=1}^{n}\left\{\mathrm{In}\left[\hat{f}(x_i, h_{\mathrm{II}}^*)\right] - \mathrm{In}[f(x_i)]\right\}\right]^2 \Rightarrow [H_{RE}(h_{\mathrm{II}}) - H]^2 \prec \left[H_{RE}(h_{\mathrm{II}}^*) - H\right]^2.$$

This result indicates that if $h_{\mathrm{II}} = \arg\min_h \left\{\frac{1}{n}\sum_{i=1}^{n}[\hat{f}(x_i) - f(x_i)]^2\right\}$ such that $\hat{f}(x_i, h_{\mathrm{II}}) - f(x_i) \prec 0, \forall i \in \{1, 2, \ldots, n\}$, then $h_{\mathrm{I}} = h_{\mathrm{II}} = \arg\min_h [(H_{RE} - H)^2]$. □

The mentioned-above two cases show that for the given dataset $X = \{x_1, x_2, \ldots, x_n\}$ and its density function $f(X)$, if there is a universe H of discourse of the discussed parameter $h$ such that all $\hat{f}(x_i, h) - f(x_i), \forall i \in \{1, 2, \ldots, n\}$ terms keep the same sign for $\forall h \in \mathrm{H}$, then the optimal bandwidth $h_{\mathrm{I}}$ for type-I error can also be selected as the optimal bandwidth for type-II error.

**Table 2**
An example of case III.

| $f(x_1)$ | $f(x_2)$ | $h_{\mathrm{II}}$ | | $h_{\mathrm{II1}}^* \succ h_{\mathrm{II}}$ | |
|---|---|---|---|---|---|
| | | $\hat{f}(x_1, h_{\mathrm{II}})$ | $\hat{f}(x_2, h_{\mathrm{II}})$ | $\hat{f}(x_1, h_{\mathrm{II1}}^*)$ | $\hat{f}(x_2, h_{\mathrm{II1}}^*)$ |
| 0.400 | 0.500 | 0.396 | 0.581 | 0.321 | 0.553 |
| | | $h_{\mathrm{II}}$ | | $h_{\mathrm{II2}}^* \prec h_{\mathrm{II}}$ | |
| | | $\hat{f}(x_1, h_{\mathrm{II}})$ | $\hat{f}(x_2, h_{\mathrm{II}})$ | $\hat{f}(x_1, h_{\mathrm{II2}}^*)$ | $\hat{f}(x_2, h_{\mathrm{II2}}^*)$ |
| 0.500 | 0.600 | 0.435 | 0.616 | 0.437 | 0.651 |

**Case III:** For $h_{II} = \arg\min_h \left\{\frac{1}{n}\sum_{i=1}^{n}[\hat{f}(x_i) - f(x_i)]^2\right\}$, if $\exists\{i_1, i_2, \ldots, i_t\} \subset \{1, 2, \ldots, n\}$ such that $\hat{f}(x_i, h_{II}) - f(x_i) \prec 0, \forall i \in \{i_1, i_2, \ldots, i_t\}$ and $\hat{f}(x_i, h_{II}) - f(x_i) \succ 0, \forall i \in \{1, 2, \ldots, n\} - \{i_1, i_2, \ldots, i_t\} t \prec n$, then for $\forall h_{II}^* \in H_3$, where $H_3$ is the universe of discourse of the discussed parameter h, such that $\frac{1}{n}\sum_{i=1}^{n}[\hat{f}(x_i, h_{II}) - f(x_i)]^2 \prec \frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}(x_i, h_{II}^*) - f(x_i)\right]^2$, then the inequality $[H_{RE}(h_{II}) - H]^2 \prec [H_{RE}(h_{II}^*) - H]^2$ cannot always hold.

We set an example to explain this fact in Table 2:

From Table 2, we can see that when $h_{II1}^* \succ h_{II}$, then $\hat{f}(x_1, h_{II1}^*) \prec \hat{f}(x_1, h_{II})$ and $\hat{f}(x_2, h_{II1}^*) \prec \hat{f}(x_2, h_{II})$. The inequalities $\hat{f}(x_1, h_{II}) - f(x_1) \prec 0$ and $\hat{f}(x_2, h_{II}) - f(x_2) \succ 0$ hold. Then, $\exists h_{II1}^*$ such that

$$[\hat{f}(x_1, h_{II}) - f(x_1)]^2 + [\hat{f}(x_2, h_{II}) - f(x_2)]^2 = 0.007 \prec \left[\hat{f}(x_1, h_{II1}^*) - f(x_1)\right]^2 + \left[\hat{f}(x_2, h_{II1}^*) - f(x_2)\right]^2 = 0.009,$$

But,

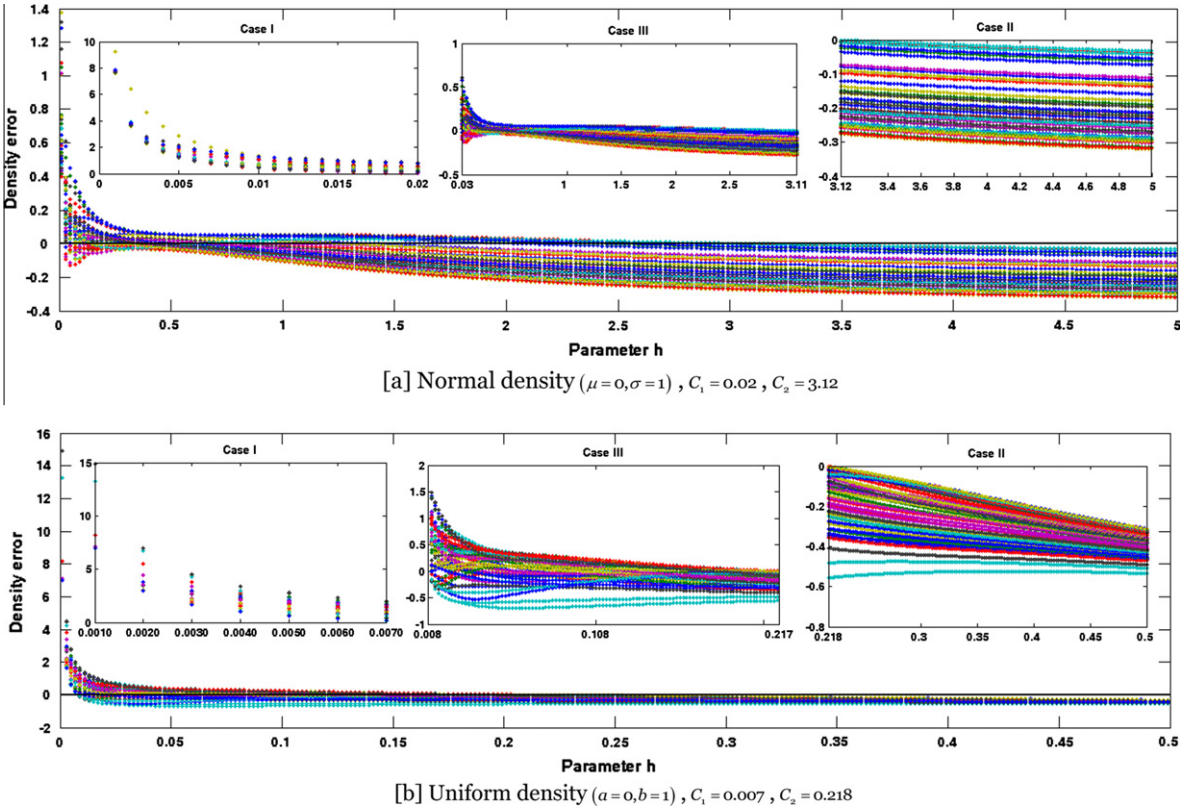
$$\left[\left\{\ln[\hat{f}(x_1, h_{II})] - \ln[f(x_1)]\right\} + \left\{\ln\left[\hat{f}(x_2, h_{II})\right] - \ln[f(x_2)]\right\}\right]^2 = 0.020$$
$$\succ \left[\left\{\ln\left[\hat{f}(x_1, h_{II1}^*)\right] - \ln[f(x_1)]\right\} + \left\{\ln\left[\hat{f}(x_2, h_{II1}^*)\right] - \ln[f(x_2)]\right\}\right]^2 = 0.014.$$

When $h_{II2}^* \prec h_{II}$, then $\hat{f}(x_1, h_{II2}^*) \succ \hat{f}(x_1, h_{II})$ and $\hat{f}(x_2, h_{II2}^*) \succ \hat{f}(x_2, h_{II})$. The inequalities $\hat{f}(x_1, h_{II}) - f(x_1) \prec 0$ and $\hat{f}(x_2, h_{II}) - f(x_2) \succ 0$ also hold. Then, $\exists h_{II2}^*$ such that

$$[\hat{f}(x_1, h_{II}) - f(x_1)]^2 + [\hat{f}(x_2, h_{II}) - f(x_2)]^2 = 0.004 \prec \left[\hat{f}(x_1, h_{II2}^*) - f(x_1)\right]^2 + \left[\hat{f}(x_2, h_{II2}^*) - f(x_2)\right]^2 = 0.007.$$

But,

$$\left[\{\ln[\hat{f}(x_1, h_{II})] - \ln[f(x_1)]\} + \{\ln[\hat{f}(x_2, h_{II})] - \ln[f(x_2)]\}\right]^2 = 0.013$$
$$\succ \left[\left\{\ln\left[\hat{f}(x_1, h_{II2}^*)\right] - \ln[f(x_1)]\right\} + \left\{\ln\left[\hat{f}(x_2, h_{II2}^*)\right] - \ln[f(x_2)]\right\}\right]^2 = 0.003. \quad \square$$




[a] Normal density $(\mu = 0, \sigma = 1)$, $C_1 = 0.02$, $C_2 = 3.12$

[b] Uniform density $(a = 0, b = 1)$, $C_1 = 0.007$, $C_2 = 0.218$

**Fig. 4.** The formations of cut-points on two density distributions.

After we have analyzed the different bandwidth selections for type-I error and type-II error, a graphical illustration is also given to present the impacts of different $h$ on the density error between $\hat{f}(x_i,h)$ and $f(x_i), i=1,2,\ldots,n$. The conclusion have been discussed earlier that with the increase of parameter $h$, the value of the estimated $\hat{f}(x_i,h), i=1,2,\ldots,n$ decreases monotonously. This will lead to the following cases: for the discussed $h\in(0,+\infty)$, there are two cut-points $C_1$ and $C_2$ which divide the interval $(0,+\infty)$ into three parts: $(0,C_1),(C_1,C_2)$, and $[C_2,+\infty)$, such that (1) $\forall h\in(0,C_1]$, $\hat{f}(x_i,h)-f(x_i)\succ 0,\forall i\in\{1,2,\ldots,n\}$ (Case I); (2) $\forall h\in[C_2,+\infty)$, $\hat{f}(x_i,h)-f(x_i)\prec 0$, $\forall i\in\{1,2,\ldots,n\}$ (Case II); (3) $\forall h\in(C_1,C_2)$, $\exists\{i_1,i_2,\ldots,i_t\}\subset\{1,2,\ldots,n\}$ such that $\hat{f}(x_i,h)-f(x_i)\prec 0,\forall i\in\{i_1,i_2,\ldots,i_t\}$ and $\hat{f}(x_i,h_{\mathrm{II}})-f(x_i)\succ 0,\forall i\in\{1,2,\ldots,n\}-\{i_1,i_2,\ldots,i_t\}t\prec n$ (Case III). According to the previous discussions, we can know that if the optimal bandwidth $h_{\mathrm{II}}$ for type-II error falls into the interval $H_1\subset(0,C_1]$ or $H_2\subset[C_2,+\infty)$, then this optimal bandwidth can also be chosen as the necessary parameter for type-I error. However, if $h_{\mathrm{II}}\in H_3\subset(C_1,C_2)$, then the conflict between these two optimal bandwidths will happen. In Fig. 4, two density functions are employed to show the formations of cut-points. For every distribution, 50 data points are generated randomly. For each data point $x_i(i=1,2,\ldots,50)$, the density error between $\hat{f}(x_i,h)$ and $f(x_i),(i=1,2,\ldots,50)$ is computed and plotted in Fig. 4. It shows that with the increase of $h$, the signs of $\hat{f}(x_i,h)-f(x_i)i\in\{1,2,\ldots,n\}$ terms all keep the trends of changing from positive to negative. This reflects the different impacts on error measures will be generated. And, the cut-points to these two distributions are also outlined in the figures.

When $h_{\mathrm{I}}$ and $h_{\mathrm{II}}$ cannot converge to the same approximated value, we try to tune the parameter selection in order to make a trade-off between the generated errors in the process of entropy estimation. From the Eq. (16), we can know that the entropy estimation relies on the density estimation. It makes the type-I error which is related to type-II error to some extent. In view of amending the type-II error with type-I error, a new entropy estimator called $\mathrm{RE}_{\mathrm{I+II}}$ is proposed. Selecting an optimal bandwidth based on the trade-off between these two errors is considered as a fundamental and potential property of $\mathrm{RE}_{\mathrm{I+II}}$.

## 5. The proposed estimator-$\mathrm{RE}_{\mathrm{I+II}}$

In the real application, the underlying density function is always unknown. In order to approximate the optimal bandwidths for the entropy and density estimations, the corresponding error criteria should be designed. We have provided the theoretical error criteria for the type-I error and type-II error in Eqs. (17) and (18) respectively. Now, we will derive the specific mathematical formulations in the following.

### 5.1. Type-I error

From Eq. (17), we know that the measure of type-I error can be described as follows:

$$E_{\mathrm{I}}=E[H_{RE}-H]^2=E[+E(H_{RE})-H]^2=[E(H_{RE})-H]^2+E[H_{RE}-E(H_{RE})]^2=[\mathrm{Bias}(H_{RE})]^2+\mathrm{Var}(H_{RE}).$$

where, $\mathrm{Bias}(H_{RE})$ and $\mathrm{Var}(H_{RE})$ are the estimated bias and variance of $H_{RE}$ respectively. The bias denotes that whether the estimated values center on the real one. And, the variance expresses the measure of dispersion among the estimated values. A good estimator must have both low bias and low variance [25].

Firstly, because the mathematical expression of bias term can be described as follows:

$$\begin{aligned}\mathrm{Bias}(H_{RE})&=E(H_{RE})-H=E\left\{-\frac{1}{n}\sum_{i=1}^{n}\mathrm{In}[\hat{f}(x_i)]\right\}-\left\{-\frac{1}{n}\sum_{i=1}^{n}\mathrm{In}[f(x_i)]\right\}\\&=-\frac{1}{n}\sum_{i=1}^{n}[E\{\mathrm{In}[\hat{f}(x_i)]\}-\mathrm{In}[f(x_i)]]\underline{\mathrm{In}(u)\approx u-1,u\to 1}-\frac{1}{n}\sum_{i=1}^{n}\{E[\hat{f}(x_i)-1]-[f(x_i)-1]\}.\end{aligned}\tag{20}$$

From the Eq. (15) we get $\hat{f}(x_i)=\frac{1}{nh}\sum_{j=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x_i-x_j}{h}\right)^2\right],(i=1,2,\ldots,n)$. Then, the form of Eq. (5.1) can be rewritten as:

$$\mathrm{Bias}(H_{RE})=-\frac{1}{n}\sum_{i=1}^{n}\{E[\hat{f}(x_i)]-f(x_i)\}=-\frac{1}{n}\sum_{i=1}^{n}\left[E\left\{\frac{1}{nh}\sum_{j=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x_i-x_j}{h}\right)^2\right]\right\}-f(x_i)\right],\tag{21}$$

where,

$$\begin{aligned}E[\hat{f}(x_i)]-f(x_i)&=\int\left[\frac{1}{h}K\left(\frac{x_i-y}{h}\right)f(y)\right]dy-f(x_i)\underline{z=\frac{x_i-y}{h}}\int[K(z)f(x_i-hz)]dz-f(x_i)\\&=\int\{K(z)[f(x_i-hz)-f(x_i)]\}dz=\int\left\{K(z)\left[f(x_i)-hzf'(x_i)+\frac{1}{2}h^2z^2f''(x_i)+O(h^2)-f(x_i)\right]\right\}dz\\&=-hf'(x_i)\int zK(z)dz+\frac{1}{2}h^2f''(x_i)\int z^2K(z)dz+O(h^2)\\&\quad\times\int f(z)dz\underline{\int zK(z)dz=0,\int f(z)dz=1}\frac{1}{2}h^2f''(x_i)\int z^2K(z)dz+O(h^2).\end{aligned}\tag{22}$$

By bringing the Eq. (22) into Eq. (21), the other form of bias is obtained:

$$\mathrm{Bias}(H_{RE}) = -\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{2}h^2 f''(x_i)\int z^2K(z)dz\right] + O(h^2) = -\frac{h^2}{2n}\sum_{i=1}^{n}\left[f''(x_i)\int z^2K(z)dz\right] + O(h^2). \tag{23}$$

Then, the expression of variance term can be calculated as follows:

$$\mathrm{Var}(H_{RE}) = E[H_{RE} - E(H_{RE})]^2 = E\left(H_{RE}^2\right) - [E(H_{RE})]^2. \tag{24}$$

We derive $E\left(H_{RE}^2\right)$ and $[E(H_{RE})]^2$ respectively:

$$\begin{aligned} E\left(H_{RE}^2\right) &= E\left\{-\frac{1}{n}\sum_{i=1}^{n}\mathrm{In}[\hat{f}(x_i)]\right\}^2 = \frac{1}{n^2}E\left\{\sum_{i=1}^{n}[\hat{f}(x_i)-1]\right\}^2 = \frac{1}{n^2}E\left\{\sum_{i=1}^{n}\hat{f}(x_i)-n\right\}^2 \\ &= \frac{1}{n^2}E\left\{\left[\sum_{i=1}^{n}\hat{f}(x_i)\right]^2 - 2n\sum_{i=1}^{n}\hat{f}(x_i) + n^2\right\} = \frac{1}{n^2}E\left[\sum_{i=1}^{n}\hat{f}(x_i)\right]^2 - \frac{2}{n}\sum_{i=1}^{n}E[\hat{f}(x_i)] + 1 \\ &= \frac{1}{n^2}\sum_{i=1}^{n}E[\hat{f}^2(x_i)] + \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}E[\hat{f}(x_i)\hat{f}(x_j)] - \frac{2}{n}\sum_{i=1}^{n}E[\hat{f}(x_i)] + 1 \end{aligned} \tag{25}$$

and,

$$\begin{aligned} [E(H_{RE})]^2 &= \left[E\left\{-\frac{1}{n}\sum_{i=1}^{n}\mathrm{In}[\hat{f}(x_i)]\right\}\right]^2 = \frac{1}{n^2}\left[E\left\{\sum_{i=1}^{n}[\hat{f}(x_i)-1]\right\}\right]^2 = \frac{1}{n^2}\left[E\left\{\sum_{i=1}^{n}\hat{f}(x_i)-n\right\}\right]^2 = \frac{1}{n^2}\left\{\sum_{i=1}^{n}E[\hat{f}(x_i)]-n\right\}^2 \\ &= \frac{1}{n^2}\left[\left\{\sum_{i=1}^{n}E[\hat{f}(x_i)]\right\}^2 - 2n\sum_{i=1}^{n}E[\hat{f}(x_i)] + n^2\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\{E[\hat{f}(x_i)]\}^2 + \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}E[\hat{f}(x_i)]E[\hat{f}(x_j)] - \frac{2}{n}\sum_{i=1}^{n}E[\hat{f}(x_i)] + 1. \end{aligned} \tag{26}$$

Above all, through bringing the Eqs. (25) and (26) into equation (24), we can get

$$\mathrm{Var}(H_{RE}) = \frac{1}{n^2}\sum_{i=1}^{n}\left[E\left[\hat{f}^2(x_i)\right] - \{E[\hat{f}(x_i)]\}^2\right]. \tag{27}$$

where,

$$\begin{aligned} E[\hat{f}^2(x_i)] - \{E[\hat{f}(x_i)]\}^2 &= E\left[\frac{1}{nh}\sum_{j=1}^{n}K\left(\frac{x_i-x_j}{h}\right)\right]^2 - \left\{E\left[\frac{1}{nh}\sum_{j=1}^{n}K\left(\frac{x_i-x_j}{h}\right)\right]\right\}^2 \\ &= \frac{1}{n}\int\left[\frac{1}{h^2}K\left(\frac{x_i-y}{h}\right)^2 f(y)\right]dy - \frac{1}{n}\left\{\int\left[\frac{1}{h^2}K\left(\frac{x_i-y}{h}\right)f(y)\right]dy\right\}^2 \underline{\underline{z = \frac{x_i-y}{h}}} \frac{1}{nh}\int[K(z)^2 f(x_i \\ &\quad - hz)]dz - \frac{1}{nh}\left\{\int[K(z)f(x_i-hz)]dz\right\}^2 \\ &= \frac{1}{nh}\int[K^2(z)f(x_i-hz)]dz - \frac{1}{n}\{E[\hat{f}(x_i)]\}^2 \underline{\underline{\frac{1}{n}\{E[\hat{f}(x_i)]\}^2 = O\left(\frac{1}{n}\right)}} \frac{1}{nh} \\ &\quad \times \int\left\{K^2(z)\left[f(x_i) - hzf'(x_i) + \frac{1}{2}h^2z^2f''(x_i) + O(h^2)\right]\right\}dz + O\left(\frac{1}{n}\right) \\ &= \frac{1}{nh}\left\{f(x_i)\int K^2(z)dz - hf'(x_i)\int zK^2(z)dz + \frac{1}{2}h^2f''(x_i)\int z^2K^2(z)dz + O(h^2)\right\} + O\left(\frac{1}{n}\right). \end{aligned} \tag{28}$$

Because there are $\frac{1}{nh}\left[hf'(x_i)\int zK^2(z)dz\right] = O(\frac{1}{n})$ and $\frac{1}{nh}\left[\frac{1}{2}h^2f''(x_i)\int z^2K^2(z)dz\right] = O(\frac{1}{n})$, the final expression of $\mathrm{Var}(H_{RE})$ can be written as:

$$\mathrm{Var}(H_{RE}) = \frac{1}{n^2}\sum_{i=1}^{n}\left[\frac{1}{nh}f(x_i)\int K^2(z)dz\right] + O\left(\frac{1}{nh}\right). \tag{29}$$

Hence, after we have got the formulations of $\mathrm{Bias}(H_{RE})$ and $\mathrm{Var}(H_{RE})$ in Eqs. (23) and (29), the final expression of $E_I$ can be described as:

$$E_{\mathrm{I}} = \left[-\frac{h^2}{2n}\sum_{i=1}^{n}[f''(x_i)\int z^2K(z)dz] + O(h^2)\right]^2 + \frac{1}{n^2}\sum_{i=1}^{n}\left[\frac{1}{nh}f(x_i)\int K^2(z)dz\right] + O\left(\frac{1}{nh}\right)$$
$$= \frac{h^4}{4n^2}\sum_{i=1}^{n}\left[f''(x_i)\int z^2K(z)dz\right] + \frac{1}{n^3h}\sum_{i=1}^{n}\left[f(x_i)\int K^2(z)dz\right] + O\left(\frac{1}{nh}\right) + O(h^2). \tag{30}$$

For Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right)$, $\int z^2K(z)dz = 1$ and $\int K^2(z)dz = \frac{1}{2\sqrt{\pi}}$. So, the Eq. (30) can also be represented as:

$$E_{\mathrm{I}} = \frac{h^4}{4n^2}\left\{\sum_{i=1}^{n}[f''(x_i)]\right\}^2 + \frac{1}{2\sqrt{\pi}n^3h}\sum_{i=1}^{n}[f(x_i)] + O\left(\frac{1}{nh}\right) + O(h^2). \tag{31}$$

### 5.2. Type-II error

By using the Eq. (18) to measure the density estimation error, the type-II error can be expanded as follows:

$$E_{\mathrm{II}} = \int[\hat{f}(X) - f(X)]^2dx = \int[\hat{f}(X)]^2dx - 2\int[\hat{f}(X)f(X)]dx + \int[f(X)]^2dx. \tag{32}$$

From the above equation, we can see that the third term $\int[f(X)]^2dx$ is not related to the unknown parameter $h$. So, the minimization of $E_{\mathrm{II}}$ is same as to minimize $E^*_{\mathrm{II}}$:

$$E^*_{\mathrm{II}} = \int[\hat{f}(X) - f(X)]^2dx = \int[\hat{f}(X)]^2dx - 2\int[\hat{f}(X)f(X)]dx. \tag{33}$$

Noting that the second term in Eq. (33) satisfies the following derivation:

$$2\int[\hat{f}(X)f(X)]dx = 2E[\hat{f}_{-i}(x_i)] = \frac{2}{n}\sum_{i=1}^{n}[\hat{f}_{-i}(x_i)] = \frac{2}{n}\sum_{i=1}^{n}\left\{\frac{1}{(n-1)h}\sum_{j\neq i}^{n}\left[K\left(\frac{x_i - x_j}{h}\right)\right]\right\}$$
$$= \frac{2}{n}\sum_{i=1}^{n}\left\{\frac{1}{(n-1)h}\sum_{j\neq i}^{n}\left[\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x_i - x_j}{h}\right)^2\right\}\right]\right\}$$
$$= \frac{2}{\sqrt{2\pi}n(n-1)h}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left\{\exp\left[-\frac{1}{2}\left(\frac{x_i - x_j}{h}\right)^2\right]\right\} \approx \frac{2}{\sqrt{2\pi}n^2h}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left\{\exp\left[-\frac{1}{2}\left(\frac{x_i - x_j}{h}\right)^2\right]\right\}. \tag{34}$$

and,

$$\int[\hat{f}(X)]^2dx = \int\left\{\frac{1}{nh}\sum_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x - x_i}{h}\right)^2\right\}\right]\right\}^2dx$$
$$= \frac{1}{n^2h^2}\sum_{i=1}^{n}\int\left\{\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x - x_i}{h}\right)^2\right]\right\}^2dx$$
$$+ \frac{2}{n^2h^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\int\left\{\frac{1}{2\pi}\exp\left[-\frac{1}{2}\left(\frac{x - x_i}{h}\right)^2\right]\exp\left[-\frac{1}{2}\left(\frac{x - x_j}{h}\right)^2\right]\right\}dx$$
$$= \frac{1}{2\pi n^2h^2}\sum_{i=1}^{n}\int\left\{\exp\left[-\left(\frac{x - x_i}{h}\right)^2\right]\right\}dx + \frac{1}{\pi n^2h^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\int\left\{\exp\left[-\frac{1}{2}\left\{\left(\frac{x - x_i}{h}\right)^2 + \left(\frac{x - x_j}{h}\right)^2\right\}\right]\right\}dx$$
$$= \frac{1}{2\sqrt{\pi}nh} + \frac{1}{2\sqrt{\pi}n^2h}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left\{\exp\left[-\frac{1}{4}\left(\frac{x_i - x_j}{h}\right)^2\right]\right\}. \tag{35}$$

We can obtain the formulation of type-II error by bringing the equations (34) and (35) into Eq. (33):

$$E^*_{\mathrm{II}} = \frac{1}{2\sqrt{\pi}nh} + \frac{1}{2\sqrt{\pi}n^2h}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left\{\exp\left[-\frac{1}{4}\left(\frac{x_i - x_j}{h}\right)^2\right]\right\} - \frac{2}{\sqrt{2\pi}n^2h}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left\{\exp\left[-\frac{1}{2}\left(\frac{x_i - x_j}{h}\right)^2\right]\right\}$$
$$= \frac{1}{2\sqrt{\pi}nh} + \frac{1}{2\sqrt{\pi}n^2h}\left\{\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left\{\exp\left[-\frac{1}{4}\left(\frac{x_i - x_j}{h}\right)^2\right]\right\} - 2\sqrt{2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left\{\exp\left[-\frac{1}{2}\left(\frac{x_i - x_j}{h}\right)^2\right]\right\}\right\}. \tag{36}$$

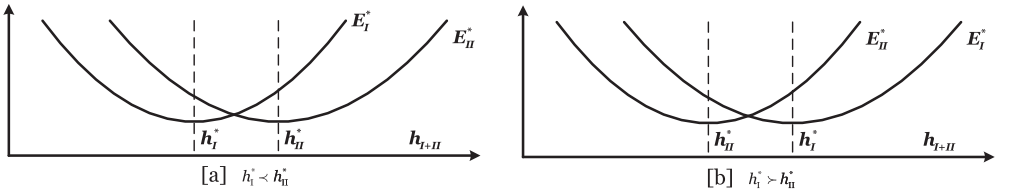



**Fig. 5.** The boundary of optimal parameter $h_{I+II}$.

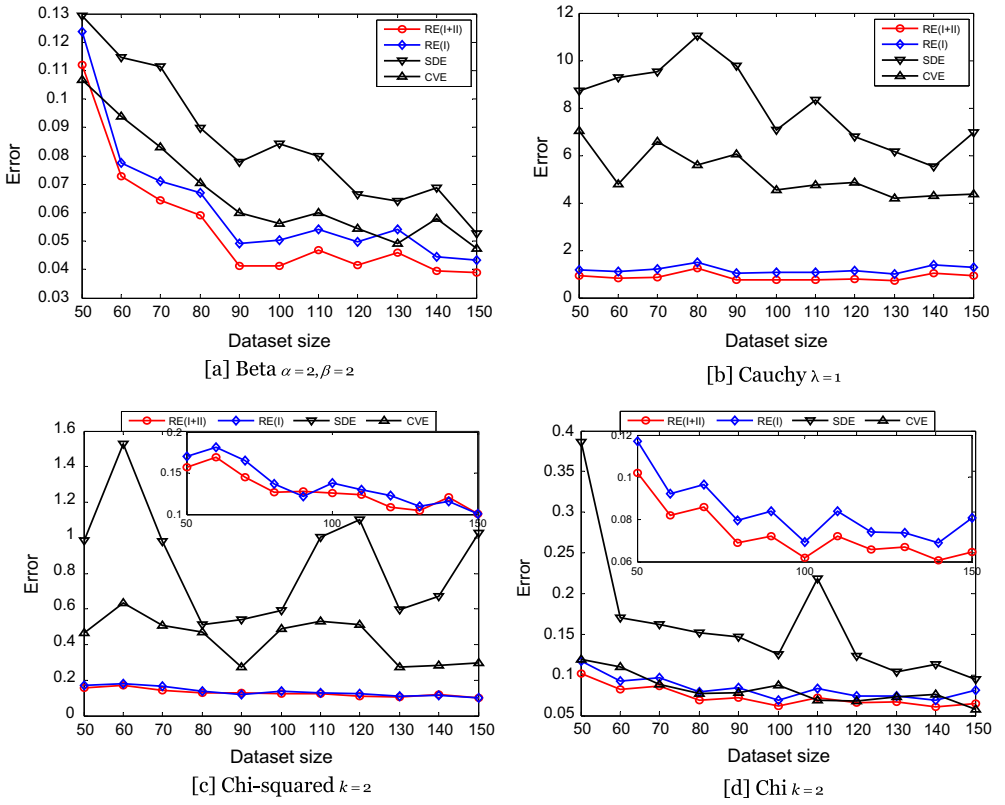



**Fig. 6.** The learning curves of $RE_{I+II}$, $RE_I$, SDE, and CVE on 24 different densities.

### 5.3. $RE_{I+II}$

In subsection 5.1, the unknown density term is considered to be replaced with $\hat{f}(x_i), i = 1, 2, \ldots, n$. So,

$$f''(x_i) = \hat{f}''(x_i) = \frac{d^2\left\{\frac{1}{nh}\sum_{j=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x_i-x_j}{h}\right)^2\right]\right\}}{dx_i^2} = \frac{1}{\sqrt{2\pi}nh^3}\sum_{j=1}^{n}\left\{\exp\left[-\frac{1}{2}\left(\frac{x_i-x_j}{h}\right)^2\right]\times\left[\left(\frac{x_i-x_j}{h}\right)^2-1\right]\right\}. \tag{37}$$

Then, neglecting the terms $O(\frac{1}{nh})$ and $O(h^2)$ as $h \to 0$ and $nh \to \infty$, we can get the type-I error function:

$$\begin{aligned} E_I^* &= \frac{h^4}{4n^2}\left\{\sum_{i=1}^{n}[f''(x_i)]\right\}^2 + \frac{1}{2\sqrt{\pi}n^3h}\sum_{i=1}^{n}[f(x_i)] \\ &= \frac{1}{8\pi n^4h^2}\left\{\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\exp\left[-\frac{1}{2}\left(\frac{x_i-x_j}{h}\right)^2\right]\times\left[\left(\frac{x_i-x_j}{h}\right)^2-1\right]\right\}\right\}^2 + \frac{1}{2\sqrt{2}\pi n^4h^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\exp\left[-\frac{1}{2}\left(\frac{x_i-x_j}{h}\right)^2\right]. \end{aligned} \tag{38}$$

Hence, the estimation model of $RE_{I+II}$ can be described as follows:

$$H_{RE_{I+II}} = -\frac{1}{n}\sum_{i=1}^{n}\ln\left\{\frac{1}{nh_{I+II}}\sum_{j=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x_i-x_j}{h_{I+II}}\right)^2\right]\right\}, \text{ where } h_{I+II} = \arg\ \min_h(E_{I+II}) = \arg\ \min_h(E_I^* + E_{II}^*). \tag{39}$$

The merged error $E_{\mathrm{I+II}} = E_{\mathrm{I}}^* + E_{\mathrm{II}}^*$ can converge to 0, when $h \to 0$ and $nh \to \infty$. Let $h_{\mathrm{I}}^* = \arg\min_h(E_{\mathrm{I}}^*)$ and $h_{\mathrm{II}}^* = \arg\min_h(E_{\mathrm{II}}^*)$. If $h_{\mathrm{I}}^* = h_{\mathrm{II}}^*$, that is to say, the same parameter $h$ makes the errors $E_{\mathrm{I}}^*$ and $E_{\mathrm{II}}^*$ reach the minimizations simultaneously, so, this $h$ can also minimize the error measure $E_{\mathrm{I+II}} = E_{\mathrm{I}}^* + E_{\mathrm{II}}^*$. Under such circumstance, we can get the conclusion: $h_{\mathrm{I+II}} = h_{\mathrm{I}}^* = h_{\mathrm{II}}^*$. If $h_{\mathrm{I}}^* \neq h_{\mathrm{II}}^*$, $\mathrm{RE}_{\mathrm{I+II}}$ intends to minimize the error sum between $E_{\mathrm{I}}^*$ and $E_{\mathrm{II}}^*$. The selected bandwidth $h_{\mathrm{I+II}}$ attempts to minimize the generated errors when the entropy estimation is carried out. From the mathematical formulation of $E_{\mathrm{I+II}} = E_{\mathrm{I}}^* + E_{\mathrm{II}}^*$, we can see that it is not practical to solve the optimal bandwidth by means of the derivative knowledge. Using some optimization algorithm to search the required parameter is considered as a necessary technology. In our study, the extremum of the merged error function was found by using the golden section method (GSM) [56]. In order to overcome the uncertainty of more than one minimum, we check GSM by plotting the error criteria over some disjoint segments of $h$ values. However, in our following experiments, we find that for every dataset that we have tested, there is only a global minimum in $E_{\mathrm{I+II}} = E_{\mathrm{I}}^* + E_{\mathrm{II}}^*$. Based on this empirical observation, we will set the upper and lower bounds for the required bandwidth $h_{\mathrm{I+II}}$. The following derivations are considered:

$$h_{\mathrm{I+II}} = \arg\ \min_h(E_{\mathrm{I}}^* + E_{\mathrm{II}}^*) \Rightarrow \frac{\partial\left[E_{\mathrm{I}}^*(h_{\mathrm{I+II}}) + E_{\mathrm{II}}^*(h_{\mathrm{I+II}})\right]}{\partial h} = 0 \Rightarrow \frac{\partial E_{\mathrm{I}}^*(h_{\mathrm{I+II}})}{\partial h} + \frac{\partial E_{\mathrm{II}}^*(h_{\mathrm{I+II}})}{\partial h} = 0.$$

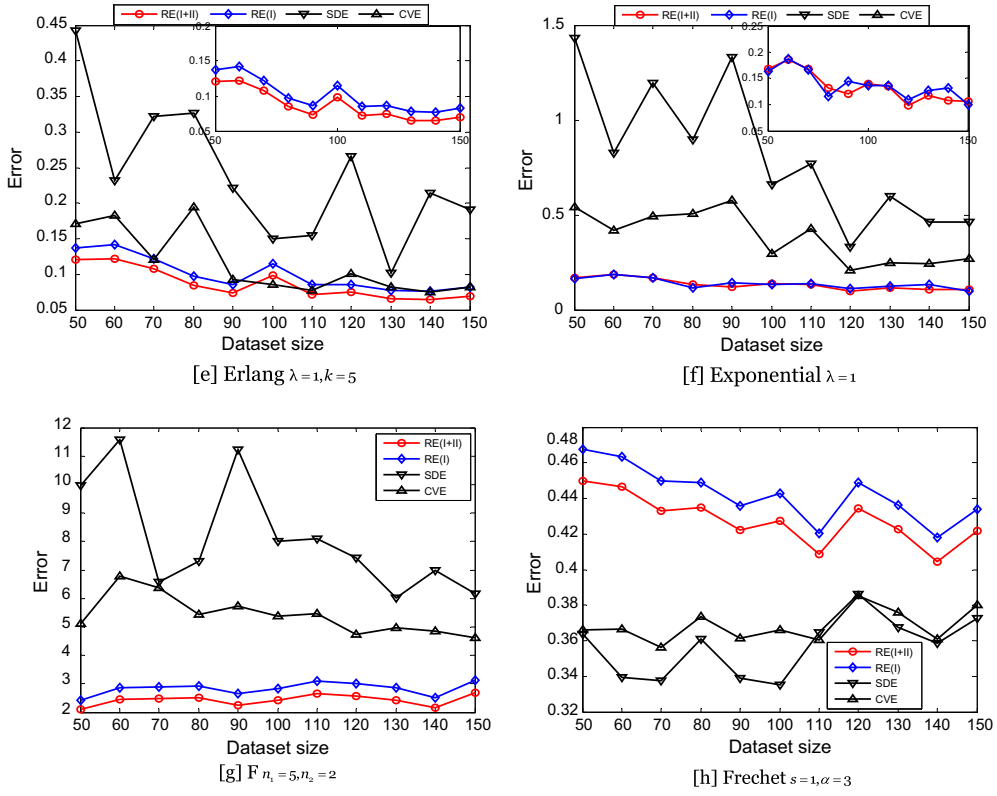
Suppose $h_{\mathrm{I}}^* \prec h_{\mathrm{II}}^*$. Based on $\frac{\partial E_{\mathrm{I}}^*(h_{\mathrm{I+II}})}{\partial h} + \frac{\partial E_{\mathrm{II}}^*(h_{\mathrm{I+II}})}{\partial h} = 0$, we can just derive $\frac{\partial E_{\mathrm{I}}^*(h_{\mathrm{I+II}})}{\partial h} \succ 0$ and $\frac{\partial E_{\mathrm{II}}^*(h_{\mathrm{I+II}})}{\partial h} \prec 0$. According to the geometric interpretation of the derivative, we can see that there is no such $h_{\mathrm{I+II}}$ which satisfies $h_{\mathrm{I+II}} \prec h_{\mathrm{I}}^*$ or $h_{\mathrm{II}}^* \prec h_{\mathrm{I+II}}$ such that $\frac{\partial E_{\mathrm{I}}^*(h_{\mathrm{I+II}})}{\partial h} \prec 0$ and $\frac{\partial E_{\mathrm{II}}^*(h_{\mathrm{I+II}})}{\partial h} \succ 0$. This result can be interpreted in Fig. 5 as follows.

From the first picture in Fig. 5, we can see that if $h_{\mathrm{I+II}} \prec h_{\mathrm{I}}^*$, then we can get $\frac{\partial E_{\mathrm{I}}^*(h_{\mathrm{I+II}})}{\partial h} \prec 0$ and $\frac{\partial E_{\mathrm{II}}^*(h_{\mathrm{I+II}})}{\partial h} \prec 0$. And, if $h_{\mathrm{II}}^* \prec h_{\mathrm{I+II}}$, we can get $\frac{\partial E_{\mathrm{I}}^*(h_{\mathrm{I+II}})}{\partial h} \succ 0$ and $\frac{\partial E_{\mathrm{II}}^*(h_{\mathrm{I+II}})}{\partial h} \succ 0$. The two derived results are all in conflict with $\frac{\partial E_{\mathrm{I}}^*(h_{\mathrm{I+II}})}{\partial h} + \frac{\partial E_{\mathrm{II}}^*(h_{\mathrm{I+II}})}{\partial h} = 0$. So, the conclusion is developed: $h_{\mathrm{I}}^* \prec h_{\mathrm{I+II}} \prec h_{\mathrm{II}}^*$. The same derivation processes can also be obtained by referring the second picture in Fig. 5 when $h_{\mathrm{I}}^* \succ h_{\mathrm{II}}^*$. In this case, we can know $h_{\mathrm{II}}^* \prec h_{\mathrm{I+II}} \prec h_{\mathrm{I}}^*$. In conclusion, the boundary of candidate bandwidth $h_{\mathrm{I+II}}$ can be summarized as:

$$\min\left\{h_{\mathrm{I}}^*, h_{\mathrm{II}}^*\right\} \leqslant h_{\mathrm{I+II}} \leqslant \max\left\{h_{\mathrm{I}}^*, h_{\mathrm{II}}^*\right\}.$$

## 6. The experimental demonstration

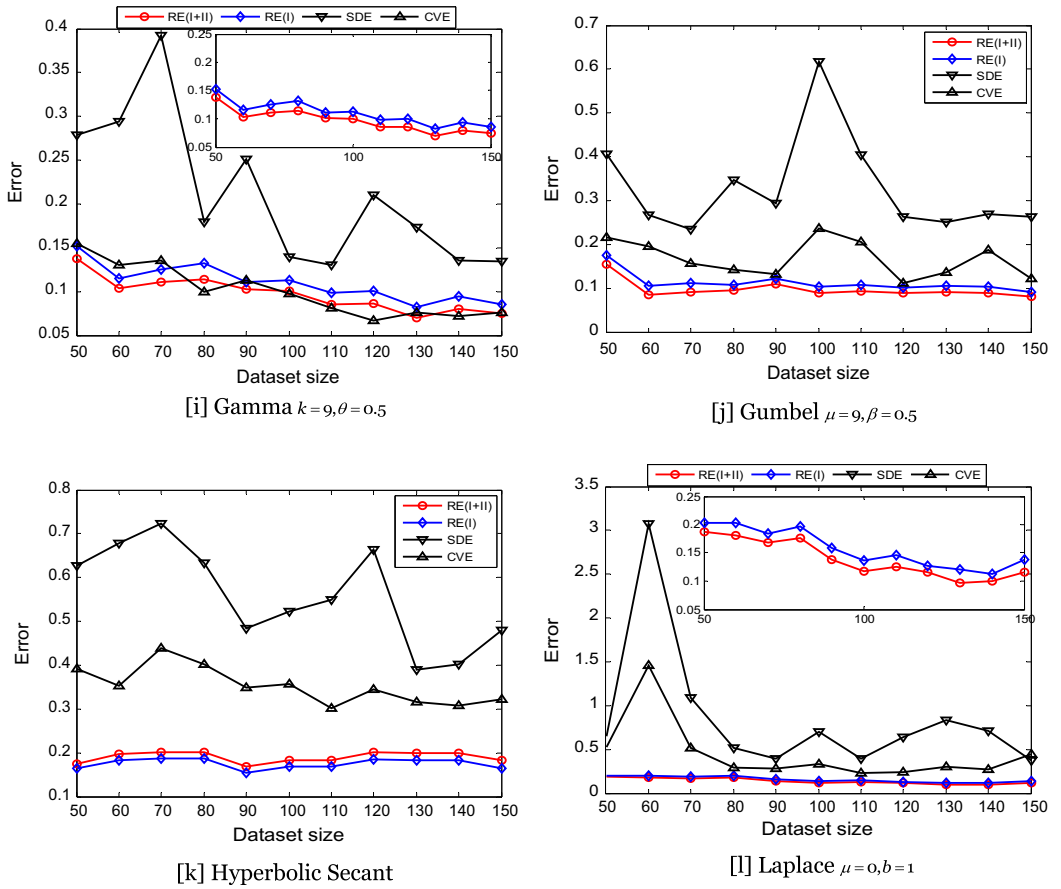
In this section, we want to verify the estimation performance of our proposed estimator-$\mathrm{RE}_{\mathrm{I+II}}$. All our experiments are conducted on a PC having the OS of Windows XP Professional with one Pentium 4 2.8 GHz processor and 1024 MB RAM.




[e] Erlang $\lambda = 1, k = 5$ [f] Exponential $\lambda = 1$

[g] F $n_1 = 5, n_2 = 2$ [h] Frechet $s = 1, \alpha = 3$

**Fig. 6.** (*continued*)




[i] Gamma $k=9,\theta=0.5$

[j] Gumbel $\mu=9,\beta=0.5$

[k] Hyperbolic Secant

[l] Laplace $\mu=0,b=1$

**Fig. 6.** (*continued*)

The related algorithms are implemented with MATLAB 7.1. Our comparisons are divided into three parts: (1) compare $\mathrm{RE}_{\mathrm{I+II}}$ with SDE and CVE which are derived from RE; (2) compare $\mathrm{RE}_{\mathrm{I+II}}$ with 9 typical discretization methods; (3) compare $\mathrm{RE}_{\mathrm{I+II}}$ with three space based estimators: $m$ SE, $m_n$ SE, and NNDE. And, based on the observed experimental results, the theoretical analyses are also conducted.

### 6.1. Compare $\mathrm{RE}_{\mathrm{I+II}}$ with SDE and CVE

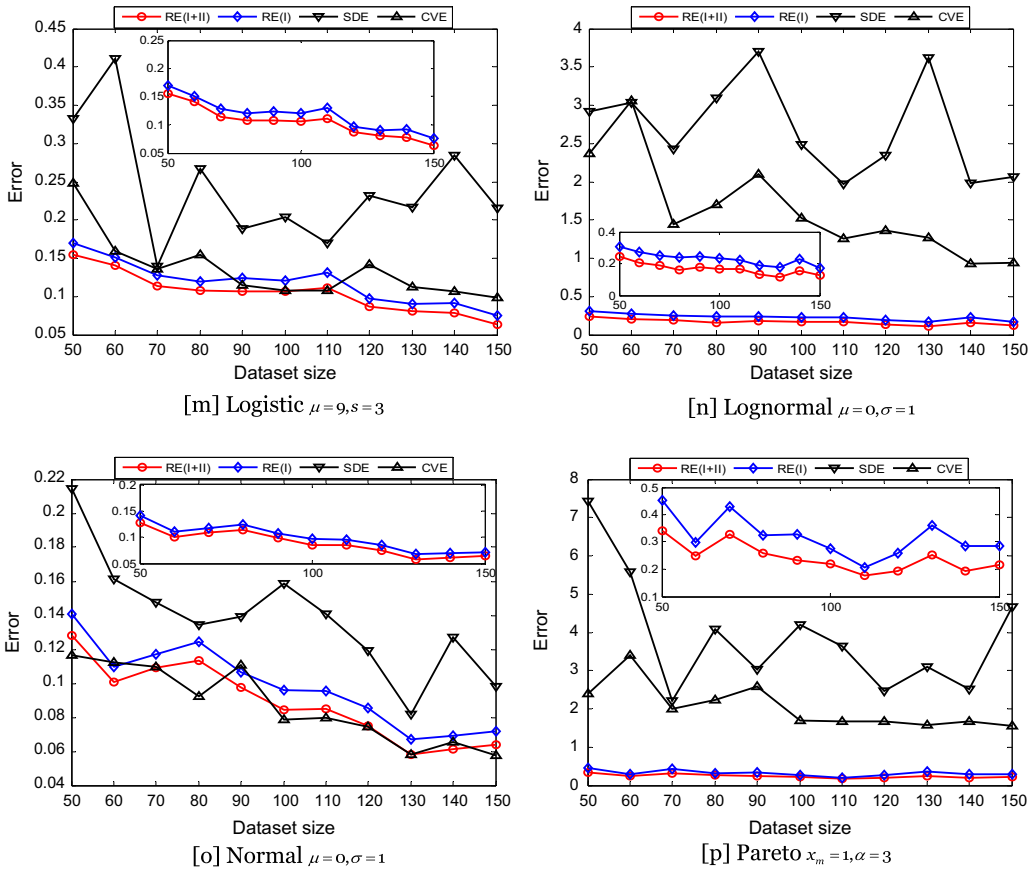
SDE [34] and CVE [35] are two classical derivatives of RE [32]. The computational formulations of SDE and CVE are listed in Eqs. (8) and (9), where the parameters $h_{SDE}$ and $h_{CVE}$ used in this experiment are computed according to the rules shown in the equations (10) and (11) respectively. Let $c = 1$. Meanwhile, in order to demonstrate the effectiveness of the proposed trade-off strategy, the entropy estimator $\mathrm{RE}_{\mathrm{I}}$ is also included. The mathematical expression of $\mathrm{RE}_{\mathrm{I}}$ is:

$$H_{RE_{\mathrm{I}}} = -\frac{1}{n}\sum_{i=1}^{n}\ln\left\{\frac{1}{nh_{\mathrm{I}}}\sum_{j=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x_i-x_j}{h_{\mathrm{I}}}\right)^2\right]\right\},\ \text{where } h_{\mathrm{I}} = \arg\ \min_h\left(E_{\mathrm{I}}^*\right). \tag{40}$$

Unlike $\mathrm{RE}_{\mathrm{I+II}}$, the optimal bandwidth of $\mathrm{RE}_{\mathrm{I}}$ is obtained only by minimizing the entropy estimation error, viz. type-I error. In fact, $\mathrm{RE}_{\mathrm{I}}$ is a more intuitive approach than $\mathrm{RE}_{\mathrm{I+II}}$, because the density estimation error (type-II error) is always neglected when the kernel density estimation is used in the entropy estimation. However, through this experiment, we will get that the adopted trade-off method $\mathrm{RE}_{\mathrm{I+II}}$ is more reasonable. The experimental procedures are arranged as follows:

Step 1: For every probability density listed in Table 1, $n$ random samples are generated $X^{(p)} = \left\{x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)}\right\}, (p = 1, 2, \ldots, 24)$;

Step 2: Compute the estimated entropies $H_{RE_{\mathrm{I+II}}}(X^{(p)}, h_{\mathrm{I+II}}), H_{RE_{\mathrm{I}}}(X^{(p)}, h_{\mathrm{I}}), H_{\mathrm{SDE}}(X^{(p)}, h_{\mathrm{SDE}})$, and $H_{\mathrm{CVE}}(X^{(p)}, h_{\mathrm{CVE}})$ by using four different estimators: $\mathrm{RE}_{\mathrm{I+II}}, \mathrm{RE}_{\mathrm{I}}$, SDE, and CVE;




[m] Logistic $\mu = 9, s = 3$

[n] Lognormal $\mu = 0, \sigma = 1$

[o] Normal $\mu = 0, \sigma = 1$

[p] Pareto $x_m = 1, \alpha = 3$

**Fig. 6.** (*continued*)

Step 3: Calculate the estimated errors:

$$RE(\mathrm{I}+\mathrm{II}) = [H_{RE_{\mathrm{I}+\mathrm{II}}}(X^{(p)}, h_{\mathrm{I}+\mathrm{II}}) - H^{(p)}]^2,\ RE(\mathrm{I}) = [H_{RE_{\mathrm{I}}}(X^{(p)}, h_{\mathrm{I}}) - H^{(p)}]^2,\ \mathrm{SDE} = [H_{\mathrm{SDE}}(X^{(p)}, h) - H^{(p)}]^2,\ \text{and CVE} = [H_{\mathrm{CVE}}(X^{(p)}, h) - H^{(p)}]^2,$$

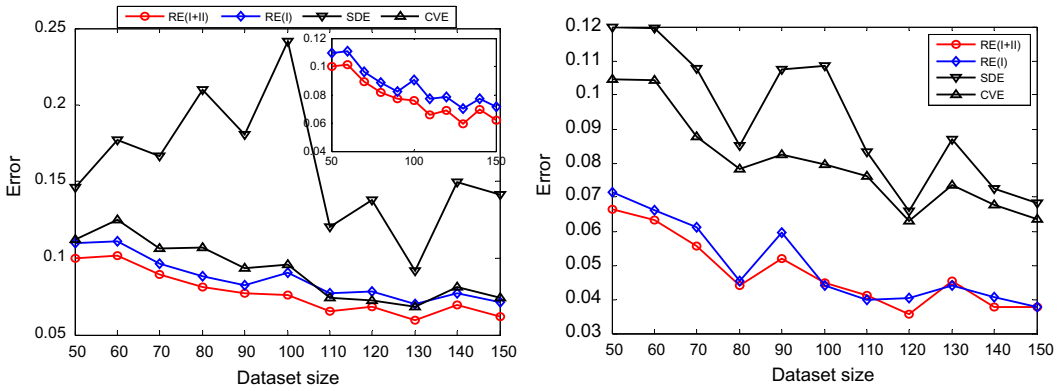
where $H^{(p)}(p = 1, 2, \ldots, 24)$ is the true entropy of the $p$-th density listed in Table 1;

Step 4: Set the size of dataset $n$ ranging from 50 to 150 in step of 10. Repeat the following Steps 1–3 for different values of $n$ and 100 times for each value. The average values of $RE(\mathrm{I}+\mathrm{II}), RE(\mathrm{I}), \mathrm{SDE}$, and CVE terms on the 100 repetitions are recorded respectively. The learning curves are plotted in Fig. 6.

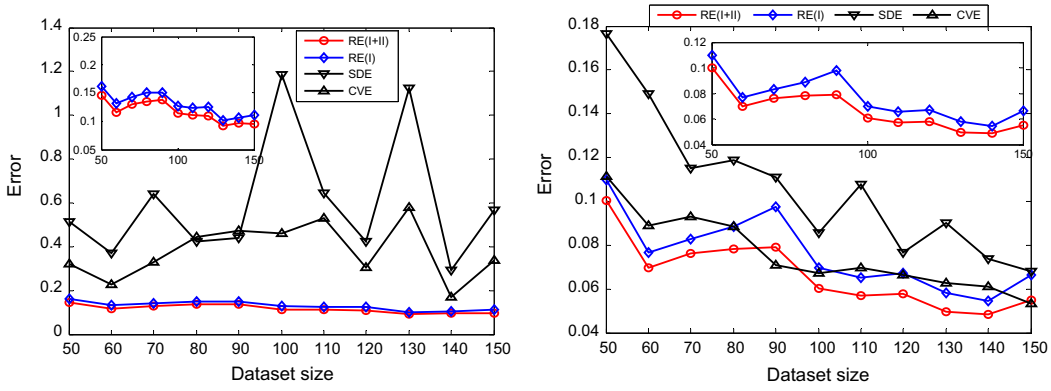
From the experimental results on 24 different density distributions, we can summarize the following three observations:

(1) With the increase of dataset size, the estimated error of $\mathrm{RE}_{\mathrm{I}+\mathrm{II}}$ keeps a more smooth decrease compared with SDE and CVE. Though the learning curves corresponding to SDE and CVE also decline roughly with the increase of number of data points, the variation trends of SDE and CVE are not stable. Even if the same density distribution is studied, the changes of estimated errors are also drastic. This indicates that SDE and CVE are more sensitive to the used dataset;

(2) Compared with the sophisticated estimation methods (SDE and CVE), $\mathrm{RE}_{\mathrm{I}+\mathrm{II}}$ has obtained the better estimation performance. Except that for the Frechet distribution ($s = 1, \alpha = 3$), the estimated performances of $\mathrm{RE}_{\mathrm{I}+\mathrm{II}}$ are all satisfactory. The main reason of the inferior performances of SDE and CVE is due to the fact that these two estimation strategies cannot make use of all the information provided by the current dataset. From the descriptions mentioned above, we can see that all $n$ data are used to estimate the unknown density in RE [32], only $\lfloor (n+1)/2 \rfloor$ data used in SDE [34], and $n-1$ data used in CVE [35]. Theoretically, these three estimators are all convergent when the size of dataset satisfies the condition $n \to \infty$ [32–34]. This means that only if the "very enough" samples are provided, these estimators can obtain the same excellent performances and all make the estimations converge to the real one. However, in the many real applications [7,25,31], this condition is not always held and the size of dataset is limited. For example, in our tests, we set the amount of dataset ranging from 50 to 150. Then, under such dataset sizes, RE can make better use of the available dataset, and provide more precise entropy estimation. Assume there are 5 random samples generated by




[q] Rayleigh $\sigma = 0.5$ [r] Semicircle $R = 1$




[s] Student's-t $\upsilon = 5$ [t] Triangular $a = 5, c = 8, b = 9$

**Fig. 6.** (*continued*)

normal distribution $N(0,1)$: $x_1 = 0.3935$, $x_2 = 3.7930$, $x_3 = 1.2048$, $x_4 = 0.0780$, and $x_5 = -0.6712$. Certainly, 5 samples are far from enough for acceptable entropy estimation. We only want to use this typical example to illustrate the estimation mechanisms of different methods. Besides, SDE can be seen as a special case of CVE: SDE uses $\lfloor (n+1)/2 \rfloor$ samples to estimate the unknown entropy and CVE uses $n-1$ samples. So, only the comparison between RE and CVE is studied. The true entropy of $N(0,1)$ is nearly 1.42 (Computed according to the rule in Table 1). The estimation entropies of RE and CVE are 1.34 and 5.58, respectively. From this computational result, we can see that the estimation of CVE is of so low accuracy ($5.58 \gg 1.42$), however, RE has indeed obtained an ideal estimation ($1.34 \approx 1.42$). We think this result is acceptable because of such a few samples used. With the focus on the fourth sample $x_4 = 0.0780$, we can compute that the density estimations of $x_4$ are 0.37 and 0.19 corresponding to RE and CVE (The true density value of $x_4 = 0.0780$ is nearly 0.39). The density estimation with inadequate information leads to the worse estimation performance of CVE. Thus, the conclusion of this numeric example tells us that the entropy estimation method which cannot make use of all the data information will lead to considerable estimation errors. From the computational rules of SDE and CVE, we can see that the parameters $h$ also play a role on their estimation performances. Whether it is essential that the more appropriate bandwidths want to be chosen for SDE and CVE in this comparison? In order to make clear this doubt, we also compare these three methods from the view of the whole approximation process with the change of parameter $h$. Let $h$ range from 0.0001 to 5 in step of 0.0001. For different values of $h$, the entropy estimations are carried out based on 100 times of random repeated trials. On 24 different density distributions, all the entropy estimations corresponding to different parameter $h$ are averaged. The above processes are conducted repeatedly under different dataset sizes $n$ = 50, 100, and 150. The experimental results are listed in Fig. 7:

From Fig. 7, we can see that the overall performance of RE is also better. Under the different dataset sizes, RE can obtain the minimal estimation errors from these 24 typical density distributions. This observation also demonstrates the effectiveness of entropy estimator listed in Eqs. (15) and (16) from another perspective. It can lead to a more stable and data-insensitive entropy estimator compared with SDE and CVE;

(3) The estimation performance of RE is improved by applying the optimal bandwidth $h_{\mathrm{I+II}}$. On 21 different distributions, the performance of $\mathrm{RE}_{\mathrm{I+II}}$ surpasses $\mathrm{RE}_{\mathrm{I}}$. This indicates that neglecting the density estimation error (type-II error) is not an advisable strategy when the entropy estimation is carried out. We find that $\mathrm{RE}_{\mathrm{I+II}}$ is incapable of estimating the

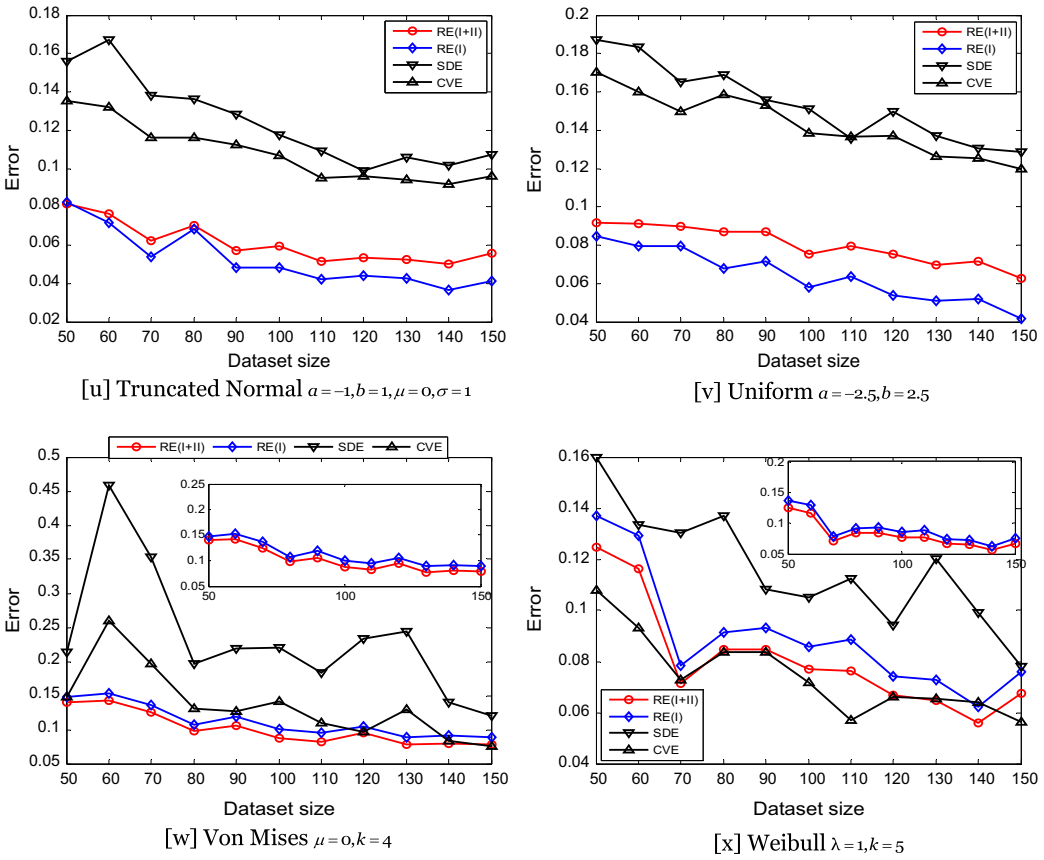



**Fig. 6.** (*continued*)

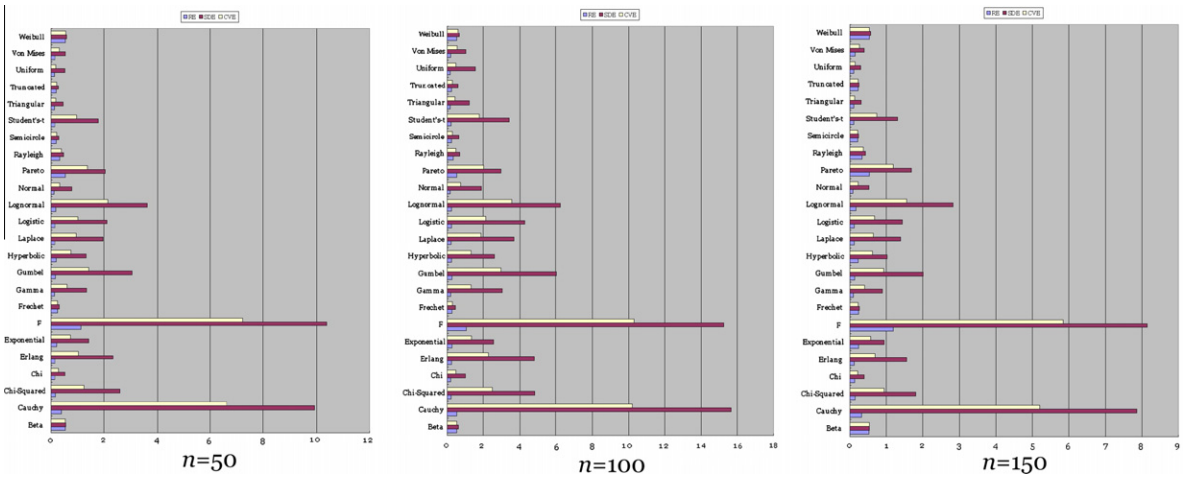



**Fig. 7.** The overall performances of RE, SDE, and CVE.

continuous entropies on Hyperbolic Secant, Truncated Normal, Uniform distributions. The tentative explanations of this observation can be depicted as follows. In fact, the new strategy $RE_{I+II}$ of selecting optimal bandwidth can be recognized as the adjustment to the parameter determined by $RE_I$. Only considering the type-I error is not enough for the determination of optimal bandwidth. So, tuning the inadequate bandwidth based on trading off between type-I error and type-II error seems to be a practicable solution. Through an additional simulation, this tuning can thus be presented and the reason of worse performances of $RE_{I+II}$ on Hyperbolic Secant, Truncated Normal, Uniform distributions

**Table 3**
The summarizations of $h_{min}, h_{max}, h_{ave}, h_{\mathrm{I}}$, and $h_{\mathrm{I+II}}$.

| | | $n = 50$ | | | | | $n = 100$ | | | | | $n = 150$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $h_{min}$ | $h_{max}$ | $h_{ave}$ | $h_{\mathrm{I}}$ | $h_{\mathrm{I+II}}$ | $h_{min}$ | $h_{max}$ | $h_{ave}$ | $h_{\mathrm{I}}$ | $h_{\mathrm{I+II}}$ | $h_{min}$ | $h_{max}$ | $h_{ave}$ | $h_{\mathrm{I}}$ | $h_{\mathrm{I+II}}$ |
| Positive distribution | Beta | 0.068 | 0.116 | 0.092 | 0.085 | 0.100 | 0.050 | 0.099 | 0.074 | 0.064 | 0.092 | 0.046 | 0.090 | 0.068 | 0.052 | 0.080 |
| | Chi | 0.205 | 0.343 | 0.274 | 0.213 | 0.275 | 0.173 | 0.296 | 0.235 | 0.168 | 0.240 | 0.134 | 0.273 | 0.204 | 0.143 | 0.219 |
| | Erlang | 0.629 | 1.128 | 0.879 | 0.727 | 0.916 | 0.569 | 1.040 | 0.804 | 0.519 | 0.795 | 0.448 | 0.947 | 0.698 | 0.446 | 0.748 |
| | Gamma | 0.487 | 0.780 | 0.634 | 0.480 | 0.660 | 0.378 | 0.708 | 0.543 | 0.382 | 0.558 | 0.343 | 0.651 | 0.497 | 0.318 | 0.529 |
| | Gumbel | 0.757 | 1.300 | 1.028 | 0.762 | 1.018 | 0.625 | 1.158 | 0.891 | 0.556 | 0.891 | 0.593 | 1.129 | 0.861 | 0.468 | 0.792 |
| | Logistic | 0.583 | 0.978 | 0.781 | 0.577 | 0.743 | 0.498 | 0.893 | 0.695 | 0.430 | 0.672 | 0.447 | 0.807 | 0.627 | 0.374 | 0.613 |
| | Normal | 0.306 | 0.536 | 0.421 | 0.357 | 0.479 | 0.262 | 0.470 | 0.366 | 0.268 | 0.384 | 0.205 | 0.441 | 0.323 | 0.226 | 0.375 |
| | Student's-$t$ | 0.361 | 0.664 | 0.513 | 0.399 | 0.521 | 0.386 | 0.614 | 0.500 | 0.274 | 0.443 | 0.295 | 0.572 | 0.433 | 0.249 | 0.421 |
| | Von Mises | 0.158 | 0.290 | 0.224 | 0.185 | 0.243 | 0.123 | 0.257 | 0.190 | 0.142 | 0.215 | 0.107 | 0.242 | 0.174 | 0.115 | 0.200 |
| | Weibull | 0.064 | 0.112 | 0.088 | 0.075 | 0.098 | 0.054 | 0.099 | 0.076 | 0.056 | 0.084 | 0.046 | 0.089 | 0.067 | 0.048 | 0.075 |
| Negative distribution | Hyperbolic | 0.063 | 0.269 | 0.166 | 0.301 | 0.438 | 0.026 | 0.146 | 0.086 | 0.226 | 0.352 | 0.024 | 0.087 | 0.055 | 0.197 | 0.339 |
| | Truncated | 0.116 | 0.216 | 0.166 | 0.206 | 0.219 | 0.076 | 0.162 | 0.119 | 0.152 | 0.183 | 0.066 | 0.140 | 0.103 | 0.127 | 0.166 |
| | Uniform | 0.261 | 0.491 | 0.376 | 0.534 | 0.545 | 0.184 | 0.338 | 0.261 | 0.374 | 0.400 | 0.162 | 0.242 | 0.202 | 0.328 | 0.339 |

**Table 4**
The comparison between $RE_{I+II}$ and different discretization methods under dataset size $n = 50$.

| | RE(I+II) | EWD | | | | EFD | | | | KMCD | | | | OD | FFD | NDD | PD | WPD | MVSDD | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Stu | Sco | Squ | Fre | Stu | Sco | Squ | Fre | Stu | Sco | Squ | Fre | | | | | | $\alpha = 0$ | $\alpha = 0.5$ | $\alpha = 1$ |
| Beta | 0.084 | 1.978 | 1.555 | 2.100 | 1.712 | 2.013 | 1.635 | 2.130 | 1.754 | 1.976 | 1.553 | 2.109 | 1.681 | 0.613 | 0.798 | 1.045 | 2.130 | 0.779 | 0.815 | 1.213 | 1.024 |
| Cauchy | 0.544 | 1.774 | 1.711 | 1.716 | 0.486 | 0.643 | 0.511 | 0.526 | 1.087 | 1.022 | 0.773 | 0.809 | 0.573 | 2.338 | 1.858 | 1.612 | 0.526 | 1.878 | 2.025 | 1.998 | 2.189 |
| Chi-Squared | 0.195 | 0.320 | 0.563 | 0.234 | 0.252 | 0.195 | 0.143 | 0.312 | 0.481 | 0.148 | 0.212 | 0.266 | 0.421 | 1.313 | 1.020 | 0.774 | 0.312 | 1.040 | 1.038 | 0.633 | 1.170 |
| Chi | 0.117 | 0.792 | 0.477 | 0.921 | 0.698 | 0.946 | 0.674 | 1.063 | 0.906 | 0.895 | 0.577 | 1.028 | 0.780 | 0.511 | 0.269 | 0.023 | 1.063 | 0.289 | 0.255 | 0.144 | 0.117 |
| Erlang | 0.102 | 0.476 | 0.726 | 0.371 | 0.505 | 0.266 | 0.538 | 0.149 | 0.346 | 0.325 | 0.642 | 0.208 | 0.418 | 1.737 | 1.481 | 1.234 | 0.149 | 1.500 | 1.473 | 1.069 | 1.356 |
| Exponential | 0.175 | 0.455 | 0.202 | 0.562 | 0.497 | 0.888 | 0.668 | 1.005 | 1.107 | 0.832 | 0.496 | 0.958 | 0.965 | 0.624 | 0.327 | 0.081 | 1.005 | 0.347 | 0.340 | 0.067 | 0.524 |
| F | 2.526 | 0.414 | 0.408 | 0.409 | 0.676 | 1.082 | 1.033 | 1.199 | 2.552 | 0.828 | 0.828 | 1.021 | 1.574 | 0.626 | 0.133 | 0.113 | 1.199 | 0.153 | 0.398 | 0.520 | 0.615 |
| Frechet | 0.465 | 1.132 | 0.799 | 1.258 | 0.984 | 1.217 | 0.951 | 1.334 | 1.163 | 1.149 | 0.853 | 1.298 | 1.060 | 0.234 | 0.002 | 0.248 | 1.334 | 0.018 | 0.019 | 0.414 | 0.192 |
| Gamma | 0.133 | 0.145 | 0.336 | 0.124 | 0.158 | 0.102 | 0.151 | 0.219 | 0.203 | 0.074 | 0.257 | 0.175 | 0.165 | 1.361 | 1.113 | 0.867 | 0.219 | 1.133 | 1.101 | 0.705 | 0.974 |
| Gumbel | 0.130 | 0.663 | 0.875 | 0.551 | 0.584 | 0.382 | 0.620 | 0.265 | 0.372 | 0.457 | 0.729 | 0.319 | 0.438 | 1.872 | 1.597 | 1.351 | 0.265 | 1.617 | 1.594 | 1.196 | 1.513 |
| Hyperbolic | 0.183 | 0.503 | 0.303 | 0.620 | 0.654 | 0.722 | 0.596 | 0.839 | 0.951 | 0.622 | 0.465 | 0.757 | 0.952 | 0.792 | 0.493 | 0.247 | 0.839 | 0.513 | 0.478 | 0.116 | 0.445 |
| Laplace | 0.176 | 0.216 | 0.272 | 0.147 | 0.212 | 0.195 | 0.152 | 0.312 | 0.646 | 0.108 | 0.127 | 0.231 | 0.624 | 1.333 | 1.020 | 0.774 | 0.312 | 1.040 | 1.008 | 0.658 | 0.952 |
| Logistic | 0.134 | 0.344 | 0.528 | 0.256 | 0.277 | 0.112 | 0.317 | 0.005 | 0.210 | 0.201 | 0.421 | 0.109 | 0.227 | 1.610 | 1.327 | 1.081 | 0.005 | 1.347 | 1.311 | 0.928 | 1.218 |
| Lognormal | 0.240 | 0.445 | 0.545 | 0.374 | 0.175 | 0.469 | 0.341 | 0.586 | 1.059 | 0.376 | 0.189 | 0.517 | 1.031 | 1.108 | 0.746 | 0.499 | 0.586 | 0.766 | 0.797 | 0.461 | 1.057 |
| Normal | 0.112 | 0.341 | 0.090 | 0.460 | 0.340 | 0.469 | 0.254 | 0.586 | 0.478 | 0.424 | 0.145 | 0.555 | 0.389 | 0.984 | 0.746 | 0.499 | 0.586 | 0.766 | 0.729 | 0.339 | 0.609 |
| Pareto | 0.274 | 0.798 | 0.632 | 0.905 | 1.257 | 1.653 | 1.529 | 1.770 | 2.297 | 1.557 | 1.338 | 1.698 | 2.136 | 0.102 | 0.438 | 0.685 | 1.770 | 0.419 | 0.371 | 0.702 | 0.148 |
| Rayleigh | 0.109 | 1.515 | 1.195 | 1.627 | 1.409 | 1.639 | 1.368 | 1.756 | 1.551 | 1.603 | 1.264 | 1.737 | 1.458 | 0.201 | 0.424 | 0.671 | 1.756 | 0.405 | 0.436 | 0.833 | 0.595 |
| Semicircle | 0.069 | 1.235 | 0.761 | 1.355 | 0.865 | 1.243 | 0.789 | 1.360 | 0.908 | 1.238 | 0.752 | 1.340 | 0.891 | 0.143 | 0.028 | 0.275 | 1.360 | 0.009 | 0.046 | 0.447 | 0.293 |
| Student's-*t* | 0.152 | 0.195 | 0.211 | 0.212 | 0.217 | 0.261 | 0.186 | 0.378 | 0.575 | 0.189 | 0.129 | 0.299 | 0.483 | 1.255 | 0.954 | 0.708 | 0.378 | 0.974 | 0.941 | 0.566 | 0.863 |
| Triangular | 0.085 | 0.644 | 0.236 | 0.777 | 0.383 | 0.695 | 0.336 | 0.812 | 0.471 | 0.671 | 0.304 | 0.778 | 0.447 | 0.717 | 0.520 | 0.274 | 0.812 | 0.540 | 0.509 | 0.107 | 0.309 |
| Truncated | 0.073 | 1.195 | 0.718 | 1.317 | 0.864 | 1.205 | 0.727 | 1.322 | 0.814 | 1.180 | 0.699 | 1.312 | 0.801 | 0.156 | 0.010 | 0.237 | 1.322 | 0.029 | 0.008 | 0.408 | 0.267 |
| Uniform | 0.105 | 0.284 | 0.249 | 0.403 | 0.251 | 0.279 | 0.223 | 0.396 | 0.249 | 0.255 | 0.275 | 0.386 | 0.248 | 1.062 | 0.936 | 0.690 | 0.396 | 0.956 | 0.921 | 0.526 | 0.643 |
| Von Mises | 0.131 | 0.893 | 0.667 | 1.010 | 0.931 | 1.079 | 0.881 | 1.196 | 1.165 | 1.006 | 0.759 | 1.135 | 1.037 | 0.412 | 0.136 | 0.111 | 1.196 | 0.155 | 0.118 | 0.270 | 0.080 |
| Weibull | 0.115 | 1.908 | 1.642 | 2.031 | 1.869 | 2.036 | 1.790 | 2.153 | 2.019 | 1.991 | 1.672 | 2.115 | 1.906 | 0.577 | 0.821 | 1.067 | 2.153 | 0.801 | 0.837 | 1.233 | 1.004 |

**Table 5**
The comparison between $RE_{I+II}$ and different discretization methods under dataset size $n = 100$.

| | RE(I+II) | EWD | | | | EFD | | | | KMCD | | | | OD | FFD | NDD | PD | WPD | MVSDD | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Stu | Sco | Squ | Fre | Stu | Sco | Squ | Fre | Stu | Sco | Squ | Fre | | | | | | $\alpha = 0$ | $\alpha = 0.5$ | $\alpha = 1$ |
| Beta | 0.046 | 2.128 | 1.814 | 2.341 | 1.921 | 2.198 | 1.905 | 2.428 | 2.022 | 2.123 | 1.824 | 2.336 | 1.942 | 0.603 | 1.439 | 3.015 | 2.428 | 1.269 | 0.817 | 1.218 | 1.038 |
| Cauchy | 0.868 | 1.931 | 1.724 | 1.898 | 0.201 | 0.458 | 0.179 | 0.228 | 1.916 | 1.107 | 0.235 | 0.731 | 0.785 | 2.410 | 1.217 | 0.359 | 0.228 | 1.387 | 1.998 | 2.094 | 2.257 |
| Chi-Squared | 0.153 | 0.377 | 0.367 | 0.264 | 0.082 | 0.380 | 0.293 | 0.609 | 0.720 | 0.281 | 0.223 | 0.499 | 0.690 | 1.357 | 0.379 | 1.196 | 0.609 | 0.550 | 1.037 | 0.628 | 1.195 |
| Chi | 0.083 | 0.915 | 0.734 | 1.128 | 0.936 | 1.131 | 0.948 | 1.361 | 1.162 | 1.054 | 0.883 | 1.263 | 1.097 | 0.532 | 0.372 | 1.948 | 1.361 | 0.202 | 0.255 | 0.146 | 0.111 |
| Erlang | 0.094 | 0.376 | 0.488 | 0.174 | 0.231 | 0.080 | 0.220 | 0.149 | 0.198 | 0.190 | 0.321 | 0.064 | 0.188 | 1.777 | 0.840 | 0.736 | 0.149 | 1.010 | 1.470 | 1.071 | 1.328 |
| Exponential | 0.133 | 0.493 | 0.372 | 0.687 | 0.741 | 1.073 | 0.982 | 1.303 | 1.405 | 0.973 | 0.888 | 1.190 | 1.367 | 0.658 | 0.314 | 1.890 | 1.303 | 0.144 | 0.342 | 0.066 | 0.492 |
| F | 2.672 | 0.547 | 0.503 | 0.517 | 0.941 | 1.267 | 1.574 | 1.496 | 3.440 | 0.862 | 1.390 | 1.223 | 2.043 | 0.694 | 0.508 | 2.083 | 1.496 | 0.337 | 0.398 | 0.538 | 0.614 |
| Frechet | 0.411 | 1.203 | 1.048 | 1.417 | 1.257 | 1.402 | 1.213 | 1.632 | 1.408 | 1.317 | 1.145 | 1.523 | 1.379 | 0.264 | 0.643 | 2.219 | 1.632 | 0.473 | 0.021 | 0.419 | 0.197 |
| Gamma | 0.090 | 0.120 | 0.116 | 0.243 | 0.172 | 0.287 | 0.187 | 0.516 | 0.423 | 0.193 | 0.125 | 0.398 | 0.371 | 1.404 | 0.472 | 1.103 | 0.516 | 0.643 | 1.099 | 0.703 | 0.953 |
| Gumbel | 0.105 | 0.594 | 0.631 | 0.404 | 0.328 | 0.197 | 0.310 | 0.032 | 0.176 | 0.321 | 0.425 | 0.092 | 0.197 | 1.914 | 0.957 | 0.619 | 0.032 | 1.127 | 1.589 | 1.193 | 1.506 |
| Hyperbolic | 0.200 | 0.491 | 0.525 | 0.685 | 0.910 | 0.907 | 0.967 | 1.136 | 1.346 | 0.760 | 0.806 | 0.995 | 1.273 | 0.843 | 0.148 | 1.723 | 1.136 | 0.023 | 0.475 | 0.105 | 0.421 |
| Laplace | 0.145 | 0.197 | 0.083 | 0.191 | 0.467 | 0.380 | 0.488 | 0.609 | 1.053 | 0.231 | 0.317 | 0.453 | 0.893 | 1.383 | 0.379 | 1.196 | 0.609 | 0.550 | 1.005 | 0.654 | 0.976 |
| Logistic | 0.089 | 0.339 | 0.298 | 0.163 | 0.107 | 0.073 | 0.137 | 0.303 | 0.429 | 0.089 | 0.112 | 0.181 | 0.342 | 1.657 | 0.686 | 0.890 | 0.303 | 0.856 | 1.309 | 0.928 | 1.204 |
| Lognormal | 0.139 | 0.539 | 0.430 | 0.423 | 0.408 | 0.654 | 0.739 | 0.884 | 1.526 | 0.506 | 0.606 | 0.740 | 1.481 | 1.176 | 0.105 | 1.471 | 0.884 | 0.275 | 0.801 | 0.460 | 1.086 |
| Normal | 0.087 | 0.437 | 0.321 | 0.642 | 0.527 | 0.654 | 0.549 | 0.884 | 0.779 | 0.555 | 0.451 | 0.785 | 0.706 | 1.032 | 0.105 | 1.471 | 0.884 | 0.275 | 0.727 | 0.327 | 0.572 |
| Pareto | 0.188 | 0.737 | 0.753 | 0.902 | 1.509 | 1.839 | 1.948 | 2.068 | 2.758 | 1.644 | 1.818 | 1.910 | 2.766 | 0.076 | 1.079 | 2.655 | 2.068 | 0.909 | 0.376 | 0.706 | 0.099 |
| Rayleigh | 0.075 | 1.595 | 1.431 | 1.806 | 1.641 | 1.824 | 1.661 | 2.054 | 1.871 | 1.733 | 1.590 | 1.951 | 1.832 | 0.161 | 1.065 | 2.641 | 2.054 | 0.895 | 0.440 | 0.841 | 0.612 |
| Semicircle | 0.052 | 1.383 | 1.018 | 1.597 | 1.092 | 1.429 | 1.052 | 1.658 | 1.131 | 1.366 | 1.015 | 1.584 | 1.090 | 0.144 | 0.669 | 2.245 | 1.658 | 0.499 | 0.047 | 0.448 | 0.284 |
| Student's-*t* | 0.114 | 0.196 | 0.081 | 0.224 | 0.433 | 0.446 | 0.520 | 0.675 | 0.909 | 0.294 | 0.381 | 0.519 | 0.801 | 1.330 | 0.314 | 1.262 | 0.675 | 0.484 | 0.936 | 0.566 | 0.896 |
| Triangular | 0.053 | 0.752 | 0.485 | 0.964 | 0.626 | 0.880 | 0.592 | 1.109 | 0.733 | 0.808 | 0.532 | 1.028 | 0.709 | 0.729 | 0.121 | 1.696 | 1.109 | 0.050 | 0.507 | 0.102 | 0.318 |
| Truncated | 0.056 | 1.361 | 0.930 | 1.572 | 1.010 | 1.391 | 0.962 | 1.620 | 1.069 | 1.339 | 0.921 | 1.554 | 1.007 | 0.168 | 0.631 | 2.207 | 1.620 | 0.461 | 0.009 | 0.411 | 0.266 |
| Uniform | 0.083 | 0.431 | 0.045 | 0.647 | 0.103 | 0.464 | 0.036 | 0.693 | 0.094 | 0.416 | 0.044 | 0.616 | 0.097 | 1.063 | 0.296 | 1.280 | 0.693 | 0.466 | 0.918 | 0.519 | 0.640 |
| Von Mises | 0.084 | 0.984 | 0.914 | 1.191 | 1.168 | 1.265 | 1.230 | 1.494 | 1.499 | 1.139 | 1.087 | 1.367 | 1.413 | 0.452 | 0.505 | 2.081 | 1.494 | 0.335 | 0.118 | 0.276 | 0.055 |
| Weibull | 0.069 | 2.002 | 1.864 | 2.216 | 2.095 | 2.221 | 2.073 | 2.450 | 2.293 | 2.123 | 1.986 | 2.353 | 2.210 | 0.537 | 1.461 | 3.037 | 2.450 | 1.291 | 0.839 | 1.238 | 0.997 |

**Table 6**
The comparison between $RE_{I+II}$ and different discretization methods under dataset size $n = 150$.

| | RE(I+II) | EWD | | | | EFD | | | | KMCD | | | | OD | FFD | NDD | PD | WPD | MVSDD | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Stu | Sco | Squ | Fre | Stu | Sco | Squ | Fre | Stu | Sco | Squ | Fre | | | | | | $\alpha = 0$ | $\alpha = 0.5$ | $\alpha = 1$ |
| Beta | 0.032 | 2.246 | 1.961 | 2.598 | 2.074 | 2.321 | 2.033 | 2.679 | 2.153 | 2.261 | 1.983 | 2.596 | 2.087 | 0.594 | 1.735 | 4.310 | 2.679 | 1.692 | 0.817 | 1.221 | 1.039 |
| Cauchy | 1.019 | 2.088 | 1.839 | 1.971 | 0.086 | 0.335 | 0.440 | 0.022 | 2.334 | 0.979 | 0.270 | 0.394 | 1.179 | 2.438 | 0.922 | 1.654 | 0.022 | 0.964 | 2.076 | 2.203 | 2.334 |
| Chi-Squared | 0.104 | 0.268 | 0.241 | 0.207 | 0.212 | 0.502 | 0.474 | 0.860 | 0.877 | 0.399 | 0.400 | 0.778 | 0.858 | 1.381 | 0.084 | 2.491 | 0.860 | 0.126 | 1.037 | 0.628 | 1.191 |
| Chi | 0.056 | 0.992 | 0.875 | 1.338 | 1.098 | 1.254 | 1.138 | 1.611 | 1.389 | 1.163 | 1.056 | 1.528 | 1.294 | 0.550 | 0.667 | 3.243 | 1.611 | 0.625 | 0.254 | 0.148 | 0.095 |
| Erlang | 0.074 | 0.358 | 0.372 | 0.132 | 0.109 | 0.042 | 0.120 | 0.400 | 0.323 | 0.077 | 0.149 | 0.309 | 0.254 | 1.789 | 0.544 | 2.031 | 0.400 | 0.586 | 1.469 | 1.070 | 1.347 |
| Exponential | 0.087 | 0.520 | 0.484 | 0.855 | 0.877 | 1.196 | 1.235 | 1.553 | 1.667 | 1.079 | 1.112 | 1.469 | 1.574 | 0.693 | 0.609 | 3.185 | 1.553 | 0.567 | 0.340 | 0.071 | 0.484 |
| F | 2.429 | 0.582 | 0.497 | 0.531 | 1.092 | 1.389 | 1.901 | 1.747 | 3.857 | 0.926 | 1.763 | 1.532 | 2.050 | 0.722 | 0.803 | 3.378 | 1.747 | 0.761 | 0.438 | 0.587 | 0.643 |
| Frechet | 0.415 | 1.307 | 1.197 | 1.657 | 1.401 | 1.525 | 1.416 | 1.882 | 1.643 | 1.441 | 1.297 | 1.796 | 1.497 | 0.278 | 0.938 | 3.514 | 1.882 | 0.896 | 0.021 | 0.422 | 0.199 |
| Gamma | 0.068 | 0.127 | 0.045 | 0.438 | 0.321 | 0.409 | 0.363 | 0.767 | 0.634 | 0.302 | 0.250 | 0.681 | 0.557 | 1.423 | 0.177 | 2.398 | 0.767 | 0.219 | 1.099 | 0.699 | 0.956 |
| Gumbel | 0.063 | 0.546 | 0.501 | 0.252 | 0.201 | 0.075 | 0.127 | 0.283 | 0.333 | 0.195 | 0.211 | 0.188 | 0.236 | 1.936 | 0.661 | 1.914 | 0.283 | 0.703 | 1.587 | 1.189 | 1.493 |
| Hyperbolic | 0.179 | 0.535 | 0.663 | 0.869 | 1.046 | 1.029 | 1.183 | 1.387 | 1.602 | 0.872 | 1.053 | 1.274 | 1.492 | 0.856 | 0.443 | 3.018 | 1.387 | 0.401 | 0.474 | 0.106 | 0.420 |
| Laplace | 0.098 | 0.146 | 0.107 | 0.306 | 0.592 | 0.502 | 0.681 | 0.860 | 1.204 | 0.303 | 0.550 | 0.733 | 1.129 | 1.409 | 0.084 | 2.491 | 0.860 | 0.126 | 1.003 | 0.655 | 0.955 |
| Logistic | 0.066 | 0.234 | 0.154 | 0.173 | 0.187 | 0.196 | 0.331 | 0.553 | 0.697 | 0.075 | 0.160 | 0.452 | 0.526 | 1.694 | 0.391 | 2.185 | 0.553 | 0.433 | 1.308 | 0.923 | 1.213 |
| Lognormal | 0.096 | 0.526 | 0.335 | 0.378 | 0.532 | 0.777 | 0.998 | 1.135 | 1.845 | 0.615 | 0.920 | 1.025 | 1.826 | 1.206 | 0.190 | 2.766 | 1.135 | 0.148 | 0.802 | 0.490 | 1.107 |
| Normal | 0.055 | 0.492 | 0.439 | 0.835 | 0.704 | 0.777 | 0.766 | 1.135 | 1.027 | 0.671 | 0.605 | 1.047 | 0.875 | 1.036 | 0.190 | 2.766 | 1.135 | 0.148 | 0.727 | 0.328 | 0.591 |
| Pareto | 0.129 | 0.666 | 0.810 | 0.943 | 1.632 | 1.961 | 2.249 | 2.319 | 3.189 | 1.788 | 2.062 | 2.218 | 2.947 | 0.089 | 1.375 | 3.950 | 2.319 | 1.332 | 0.373 | 0.698 | 0.098 |
| Rayleigh | 0.067 | 1.692 | 1.559 | 2.040 | 1.795 | 1.947 | 1.820 | 2.305 | 2.049 | 1.845 | 1.732 | 2.222 | 1.973 | 0.157 | 1.361 | 3.936 | 2.305 | 1.318 | 0.440 | 0.841 | 0.607 |
| Semicircle | 0.034 | 1.503 | 1.150 | 1.857 | 1.247 | 1.551 | 1.185 | 1.909 | 1.278 | 1.497 | 1.154 | 1.843 | 1.254 | 0.146 | 0.965 | 3.540 | 1.909 | 0.922 | 0.047 | 0.451 | 0.290 |
| Student's-*t* | 0.092 | 0.169 | 0.185 | 0.393 | 0.555 | 0.568 | 0.780 | 0.926 | 1.185 | 0.405 | 0.620 | 0.805 | 1.081 | 1.331 | 0.018 | 2.557 | 0.926 | 0.060 | 0.936 | 0.566 | 0.898 |
| Triangular | 0.050 | 0.874 | 0.625 | 1.224 | 0.775 | 1.002 | 0.743 | 1.360 | 0.898 | 0.940 | 0.682 | 1.294 | 0.833 | 0.743 | 0.416 | 2.991 | 1.360 | 0.374 | 0.505 | 0.101 | 0.317 |
| Truncated | 0.053 | 1.478 | 1.089 | 1.836 | 1.176 | 1.513 | 1.112 | 1.871 | 1.172 | 1.461 | 1.079 | 1.805 | 1.154 | 0.159 | 0.927 | 3.502 | 1.871 | 0.884 | 0.009 | 0.411 | 0.258 |
| Uniform | 0.073 | 0.557 | 0.147 | 0.912 | 0.161 | 0.586 | 0.165 | 0.944 | 0.188 | 0.537 | 0.145 | 0.883 | 0.143 | 1.060 | 0.000 | 2.575 | 0.944 | 0.042 | 0.917 | 0.517 | 0.639 |
| Von Mises | 0.073 | 1.020 | 1.043 | 1.366 | 1.334 | 1.387 | 1.414 | 1.745 | 1.703 | 1.249 | 1.316 | 1.644 | 1.635 | 0.483 | 0.801 | 3.376 | 1.745 | 0.758 | 0.117 | 0.279 | 0.045 |
| Weibull | 0.063 | 2.074 | 2.006 | 2.420 | 2.218 | 2.343 | 2.258 | 2.701 | 2.495 | 2.231 | 2.156 | 2.601 | 2.391 | 0.537 | 1.757 | 4.332 | 2.701 | 1.715 | 0.839 | 1.241 | 0.994 |

**Table 7**
The counts of wins, losses and ties of $RE_{I+II}$ compared with all discretization methods by using two-tailed $t$-test with the 99% confidence level.

| w/ t/l | EWD | | | | EFD | | | | KMCD | | | | OD | FFD | NDD | PD | WPD | MVSDD | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | M2 | M3 | M4 | M5 | M6 | M7 | M8 | M9 | M10 | M11 | M12 | M13 | M14 | M15 | M16 | M17 | M18 | M19 | M20 | M21 |
| 50 | 23/ | 22/ | 21/ | 21/ | 20/ | 20/ | 21/ | 24/ | 20/ | 20/ | 22/ | 23/ | 21/ | 20/ | 19/ | 21/ | 20/ | 19/ | 20/ | 19/ |
| | 0/1 | 0/2 | 0/3 | 0/3 | 1/3 | 0/4 | 0/3 | 0/0 | 0/4 | 0/4 | 0/2 | 0/1 | 0/3 | 0/4 | 0/5 | 0/3 | 0/4 | 0/5 | 0/4 | 1/4 |
| 100 | 23/ | 20/ | 23/ | 21/ | 20/ | 21/ | 21/ | 24/ | 22/ | 21/ | 20/ | 22/ | 21/ | 21/ | 22/ | 21/ | 21/ | 20/ | 21/ | 20/ |
| | 0/1 | 0/4 | 0/1 | 0/3 | 0/4 | 0/3 | 0/3 | 0/0 | 1/1 | 0/3 | 0/4 | 0/2 | 0/3 | 0/3 | 0/2 | 0/3 | 0/3 | 0/4 | 0/3 | 0/4 |
| 150 | 23/ | 22/ | 23/ | 22/ | 21/ | 22/ | 22/ | 24/ | 22/ | 22/ | 22/ | 23/ | 21/ | 18/ | 24/ | 22/ | 20/ | 21/ | 21/ | 20/ |
| | 0/1 | 0/2 | 0/1 | 0/2 | 0/3 | 0/2 | 0/2 | 0/0 | 0/2 | 0/2 | 0/2 | 0/1 | 0/3 | 0/6 | 0/0 | 0/2 | 0/4 | 0/3 | 0/3 | 0/4 |

Note that: M2: RE(I+II) vs. EWD(Stu) M3: RE(I+II) vs. EWD(Sco) M4: RE(I+II) vs. EWD(Squ) M5: RE(I+II) vs. EWD(Fre) M6: RE(I+II) vs. EFD(Stu) M7: RE(I+II) vs. EFD(Sco) M8: RE(I+II) vs. EFD(Squ) M9: RE(I+II) vs. EFD(Fre) M10: RE(I+II) vs. KMCD(Stu) M11: RE(I+II) vs. KMCD(Sco) M12: RE(I+II) vs. KMCD(Squ) M13: RE(I+II) vs. KMCD(Fre) M14: RE(I+II) vs. OD M15: RE(I+II) vs. FFD M16: RE(I+II) vs. NDD M17: RE(I+II) vs. PD M18: RE(I+II) vs. WPD M19: RE(I+II) vs. MVSDD($\alpha = 0$) M20: RE(I+II) vs. MVSDD($\alpha = 0.5$) M21: RE(I+II) vs. MVSDD($\alpha = 1$).

**Table 8**
Different $H_{\text{Space}} - H_{RE_{I+II}}$ and $H_{\text{Space}} + H_{RE_{I+II}} - 2H_{\text{True}}$ terms on five distributions.

| | Dataset size | $RE_{I+II}$ vs. $m$ SE | | $RE_{I+II}$ vs. $m_n$ SE | | $RE_{I+II}$ vs. NNDE | |
|---|---|---|---|---|---|---|---|
| | | $H_{m\text{SE}} - H_{\text{RE}_{I+II}}$ | $H_{m\text{SE}} + H_{\text{RE}_{I+II}} - 2H_{\text{True}}$ | $H_{m_n\text{SE}} - H_{\text{RE}_{I+II}}$ | $H_{m_n\text{SE}} + H_{\text{RE}_{I+II}} - 2H_{\text{True}}$ | $H_{\text{NNDE}} - H_{\text{RE}_{I+II}}$ | $H_{\text{NNDE}} + H_{\text{RE}_{I+II}} - 2H_{\text{True}}$ |
| Cauchy ($H_{\text{True}} = 2.531$) | 50 | 0.224 | −0.548 | −0.220 | −0.992 | 0.355 | −0.417 |
| | 100 | 0.186 | −0.462 | −0.155 | −0.803 | 0.257 | −0.391 |
| | 150 | 0.192 | −0.295 | −0.106 | −0.594 | 0.261 | −0.226 |
| Frechet ($H_{\text{True}} = 0.671$) | 50 | 0.008 | −0.892 | −0.212 | −1.111 | 0.060 | −0.840 |
| | 100 | 0.003 | −0.886 | −0.170 | −1.058 | 0.033 | −0.856 |
| | 150 | −0.024 | −0.860 | −0.158 | −0.995 | −0.020 | −0.856 |
| Hyperbolic Scant ($H_{\text{True}} = 1.166$) | 50 | −0.027 | 0.311 | −0.326 | 0.011 | 0.045 | 0.382 |
| | 100 | 0.003 | 0.319 | −0.226 | 0.090 | 0.041 | 0.357 |
| | 150 | 0.019 | 0.385 | −0.188 | 0.177 | 0.046 | 0.412 |
| Lognormal ($H_{\text{True}} = 1.419$) | 50 | 0.125 | −0.182 | −0.183 | −0.490 | 0.219 | −0.088 |
| | 100 | 0.093 | −0.204 | −0.138 | −0.435 | 0.141 | −0.156 |
| | 150 | 0.092 | −0.128 | −0.114 | −0.334 | 0.136 | −0.084 |
| Pareto ($H_{\text{True}} = 0.235$) | 50 | 0.165 | −0.100 | −0.046 | −0.311 | 0.222 | −0.043 |
| | 100 | 0.176 | −0.151 | 0.002 | −0.325 | 0.212 | −0.115 |
| | 150 | 0.142 | −0.136 | −0.010 | −0.288 | 0.164 | −0.114 |

is also described. We target on 10 density distributions from Table 1 as the positive distributions on which $RE_{I+II}$ has obtained better estimation performances. Based on each selected distribution, 100 datasets are generated under the different dataset sizes of $n = 50$, 100, and 150. For the specific dataset size, the notations $h_{min}, h_{max}, h_{ave}, h_I$, and $h_{I+II}$ denote the minimal value of all optimal bandwidths, the maximal value of all actual optimal bandwidths, the average value of all actual optimal bandwidths, the average value of all actual optimal bandwidths selected by $RE_I$, and the average value of all optimal bandwidths selected by $RE_{I+II}$. The experimental results are listed in Table 3.

From Table 3, we explain the implied meaning by taking Chi distribution as an example. The following relations are considered:

$$\text{Chi}: \ (\text{C1})\ h_I(\text{Chi}) \in [h_{\min}(\text{Chi}), h_{\max}(\text{Chi})]; (\text{C2}) h_{I+II}(\text{Chi}) \in [h_{\min}(\text{Chi}), h_{\max}(\text{Chi})]; (\text{C3}) |h_I(\text{Chi}) - h_{\text{ave}}(\text{Chi})| \succ |h_{I+II}(\text{Chi}) - h_{\text{ave}}(\text{Chi})|.$$

The expressions (C1) and (C2) imply that the optimal bandwidths selected by $RE_I$ and $RE_{I+II}$ are reasonable. That is to say, the optimal bandwidths selected by $RE_I$ and $RE_{I+II}$ fall into a feasible interval of optimal bandwidth. And, the inequality (C3) shows that the optimal bandwidth $h_{I+II}$ (Chi) selected by $RE_{I+II}$ is better than the optimal one $h_I$ (Chi) selected by $RE_I$, because $h_{I+II}$ (Chi) is more close to the actual optimal bandwidth than $h_I$ (Chi). The observation that $h_{I+II}$ (Chi) is more appropriate than $h_I$ (Chi). The tuning of bandwidth $h_I$ (Chi) occurs when the new strategy $RE_{I+II}$ is used to select the optimal parameter for RE. This kind of tuning to all the optimal bandwidths selected by $RE_I$ is effective on the distributions employed in this additional experiment. According to the experimental observations, the following conditions can be summarized for the positive distributions on which $RE_{I+II}$ can obtain the better estimation performances than $RE_I$:

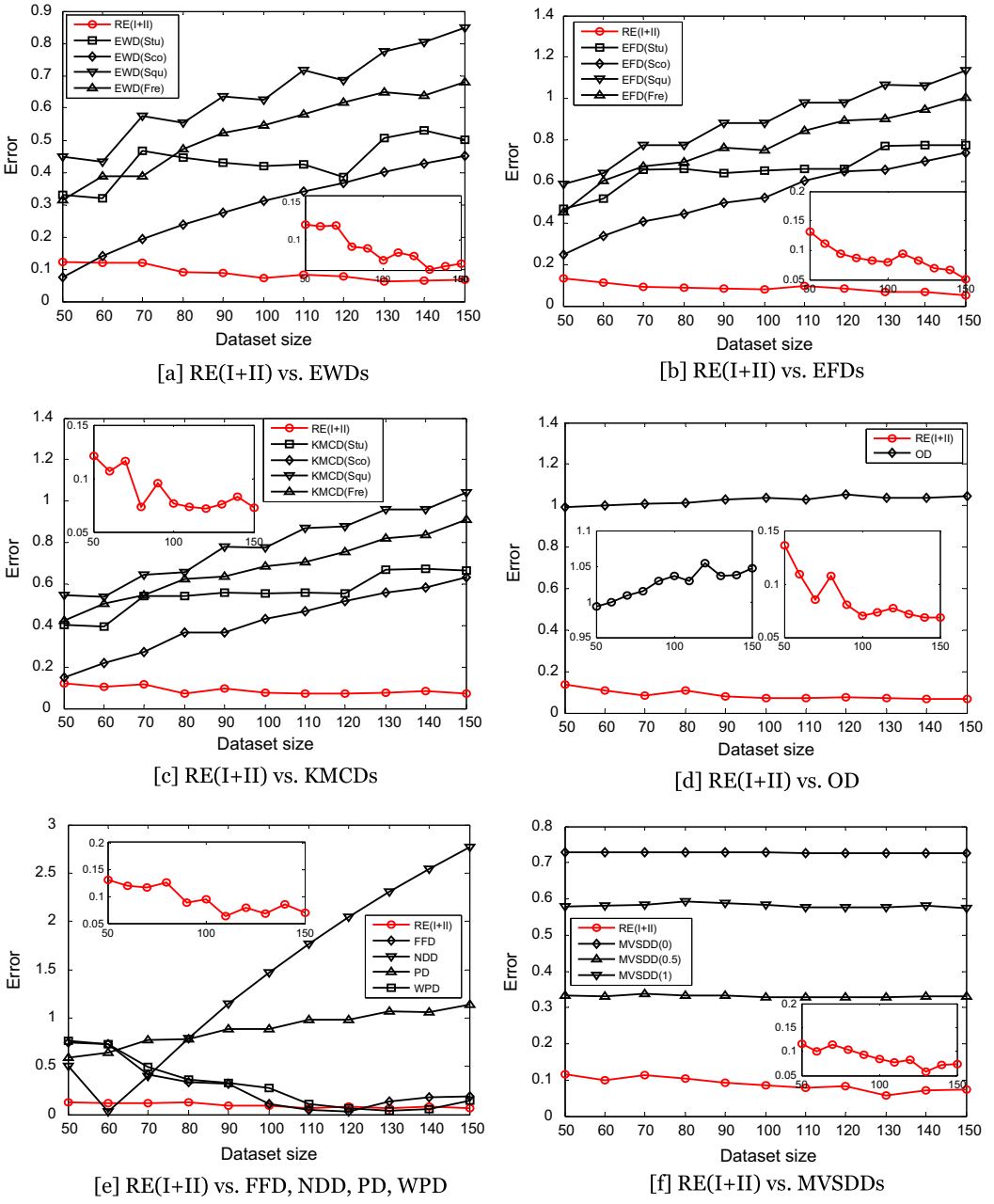



**Fig. 8.** The learning curves of $RE_{I+II}$ and different discretization methods on normal density.

$$\text{Positive distribution}: \ (\text{P1})\ h_{I}(\text{Pos}) \in [h_{\min}(\text{Pos}), h_{\max}(\text{Pos})];\ (\text{P2})\ h_{I+II}(\text{Pos}) \in [h_{\min}(\text{Pos}), h_{\max}(\text{Pos})];\ (\text{P3})\ |h_{I}(\text{Pos}) - h_{ave}(\text{Pos})| \succ |h_{I+II}(\text{Pos}) - h_{ave}\ (\text{Pos})|.$$

However, when the above conditions are not satisfied this tuning is not always helpful to the negative distributions on which $RE_{I+II}$ cannot obtain the better estimation performances in comparison with $RE_{I}$, for example, on Hyperbolic Secant, Truncated Normal, Uniform distributions. Firstly, for Hyperbolic Secant and Uniform distributions, the relations are:

$$\text{Hyperbolic Secant}: \ (\text{H1})\ h_{I}(\text{Hyp}) \notin [h_{\min}(\text{Hyp}), h_{\max}(\text{Hyp})];\ (\text{H2})\ h_{I+II}(\text{Hyp}) \notin [h_{\min}(\text{Hyp}), h_{\max}(\text{Hyp})];\ (\text{H3})\ |h_{I}(\text{Hyp}) - h_{ave}(\text{Hyp})| \prec |h_{I+II}(\text{Hyp}) - h_{ave}(\text{Hyp})|.$$

$$\text{Uniform}: \ (\text{U1})\ h_{\text{I}}(\text{Uni}) \notin [h_{\min}(\text{Uni}), h_{\max}(\text{Uni})];\ (\text{U2})\ h_{\text{I+II}}(\text{Uni}) \notin [h_{\min}(\text{Uni}), h_{\max}(\text{Uni})];\ (\text{U3})\ |h_{\text{I}}(\text{Uni}) - h_{\text{ave}}(\text{Uni})| \prec |h_{\text{I+II}}(\text{Uni}) - h_{\text{ave}}(\text{Uni})|.$$

We can see that on these two distributions, the optimal bandwidths selected by $\text{RE}_{\text{I}}$ and $\text{RE}_{\text{I+II}}$ all fall out the corresponding feasible intervals. And, for Truncated Normal distribution, there are:

$$\text{Truncated Normal}: \ (\text{T1})\ h_{\text{I}}(\text{Tre}) \in [h_{\min}(\text{Tre}), h_{\max}(\text{Tre})];\ (\text{T2})\ h_{\text{I+II}}(\text{Tre}) \notin [h_{\min}(\text{Tre}), h_{\max}(\text{Tre})];\ (\text{T3})\ |h_{\text{I}}(\text{Tre}) - h_{\text{ave}}(\text{Tre})| \prec |h_{\text{I+II}}(\text{Tre}) - h_{\text{ave}}(\text{Tre})|.$$

The tuned bandwidth of $\text{RE}_{\text{I+II}}$ is also not acceptable by the feasible interval. Secondly, for these three distributions, the process of tuning the bandwidth makes the optimal bandwidths far from the actual optimal bandwidths. Overall, it states that negative distributions are not satisfied with the summarized conditions for positive distributions so that $\text{RE}_{\text{I+II}}$ obtain the worse estimation performances on Hyperbolic Secant, Truncated Normal, Uniform distributions.

### 6.2. Compare $\text{RE}_{\text{I+II}}$ with different discretization methods

Because the discretization is used as the necessary pre-processing technology in many literatures [5–12] when the continuous entropy needs to be estimated, the different discretization methods are also introduced in our comparisons. In Section 2, 9 typical discretization methods have been summarized. For the setup of forementioned parameters in these methods, we give the following descriptions. For EWD, EFD, and KMCD, the number of discretized intervals $k$ and the width of interval $m$ need to be determined beforehand. And, these two parameters will produce effects on the estimated performances of EWD, EFD, and KMCD. There is not any best number of intervals, and different interval widths can reveal different features of the data. Some theoreticians [26–28] have attempted to determine an optimal number of intervals, but these methods generally make strong assumptions about the shape of the distribution. Depending on the actual data distribution and the goals of the analysis, different interval widths may be appropriate, so experimentation is usually needed to determine an appropriate width. However, there are various useful guidelines that can be used to determine the number of discretized intervals $k$ or the width of interval $m$:

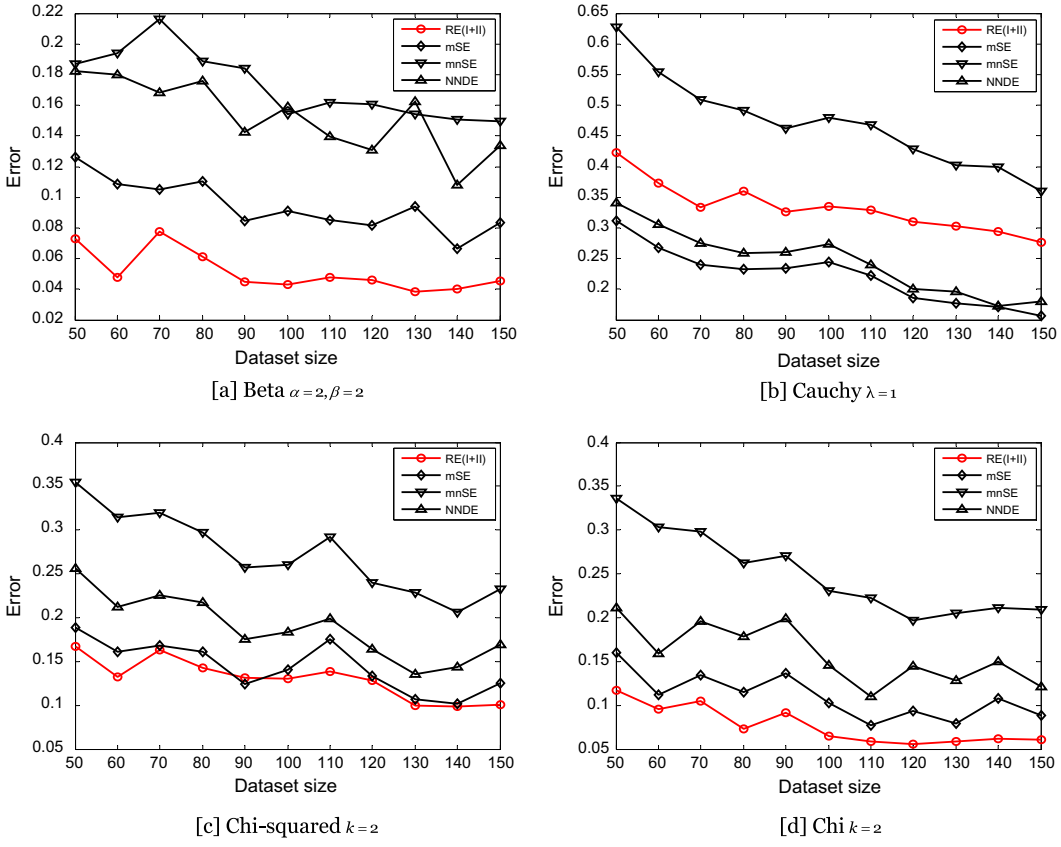



**Fig. 9.** The learning curves of $\text{RE}_{\text{I+II}}$, $m$ SE, $m_n$ SE, and NNDE on 24 different densities.

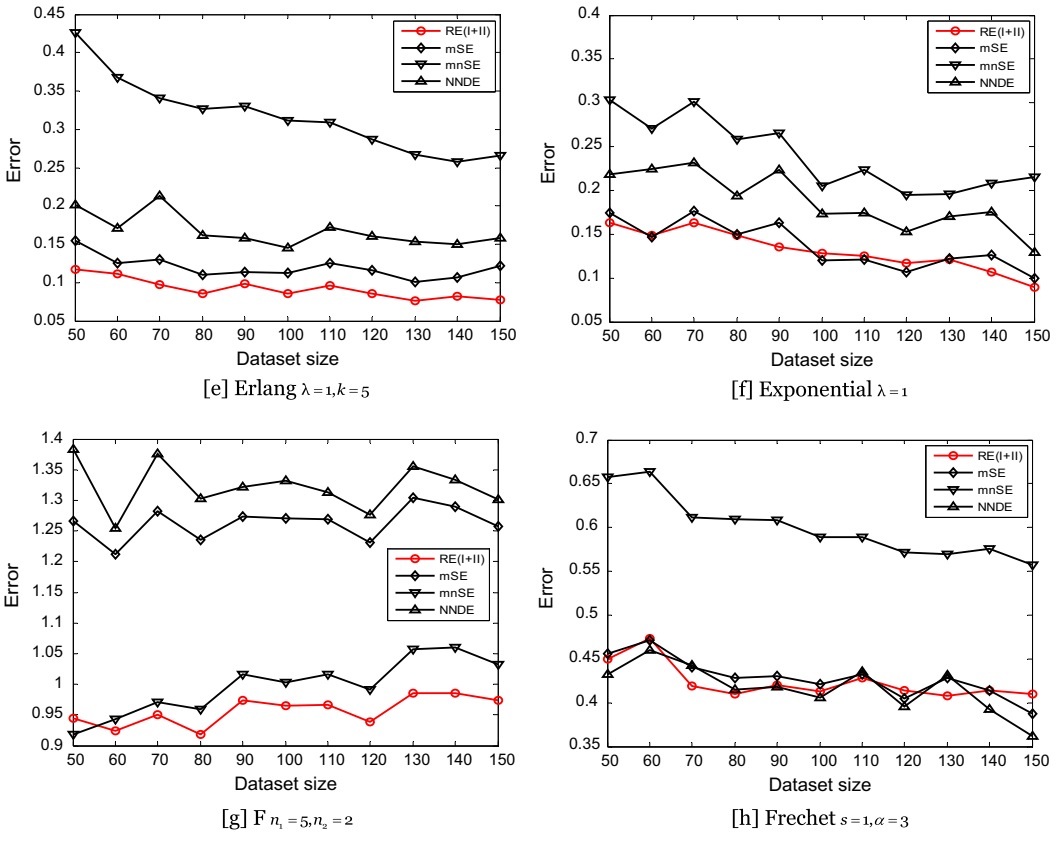
(1) Sturges' formula [53]: $k = \lceil \log_2 n + 1 \rceil$, where $n$ is the size of dataset $X$, and $\lceil u \rceil$ denotes the rounding of the element $u$ to the nearest integers towards infinity;
(2) Scott's choice [54]: $m = 3.5\sigma/n^{1/3}$, where $\sigma$ is the standard deviation of dataset $X$;
(3) Square-root choice: $k = \lceil \sqrt{n} \rceil$;
(4) Freedman-Diaconis choice [55]: $m = 2\mathrm{IQR}(X)/n^{1/3}$, where $\mathrm{IQR}(X)$ denotes the inter-quartile range [50] of dataset $X$.

In the following comparison, these four rules are respectively used in EWD, EFD, and KMCD to determine the number of discretized intervals or the width of interval. For example, EWD(Stu) denotes EWD with Sturges' interval number. In addition, because OD depends on the primary discretization, EWD with 10 discretized intervals is employed. In Peng's web homepage [51], three parameters $\alpha = 0$, or 0.5, or 1 are provided, so, we apply these three parameters in MVSDD in our experiment.

The experimental procedures are arranged as follows. We investigate the entropy estimation performances of different methods under the dataset sizes $n = 50$, 100, and 150. The estimation performance is measured by the error between the true entropy and the estimated entropy. Under the specific dataset size, 100 datasets are generated randomly for every density distribution listed in Table 1. By discretizing the continuous observations in advance, the estimated entropies can be calculated by the different discretization methods according to the Eq. (1.1). The experimental results are summarized in Tables 4–6.

For the specific dataset size and density distribution, we compare $\mathrm{RE}_{\mathrm{I+II}}$ with the different discretization methods via the two-tailed $t$-test with the 99% confidence level based on the comparative results on 100 datasets. According to the statistical theory, we speak of two results for a dataset as being significantly different only if the probability of significant difference is at least 99%. Then, based on the statistical results under the given dataset size, Table 7 records entries of $w/t/l$ which means that $\mathrm{RE}_{\mathrm{I+II}}$ wins in $w$ densities, ties in $t$ densities, and loses in $l$ densities.

From our experiments, we can see that the performances of $\mathrm{RE}_{\mathrm{I+II}}$ are overall the best among the related models. And, with the increase of dataset size, the advantages of $\mathrm{RE}_{\mathrm{I+II}}$ are advanced gradually. For example, compared with EWD(Squ), the changed series of win number of $\mathrm{RE}_{\mathrm{I+II}}$ is $21(n = 50) \to 23(n = 100) \to 23(n = 150)$; With EFD(Sco), $20 \to 21 \to 22$; With KMCD(Sco), $20 \to 21 \to 22$; With NDD, $19 \to 22 \to 24$; and with MVSDD($\alpha = 0$), $19 \to 20 \to 21$, and so on. The empirical results reflect that with the increase of dataset size, the estimated error of $\mathrm{RE}_{\mathrm{I+II}}$ will decrease gradually, while the




[e] Erlang $\lambda = 1, k = 5$ [f] Exponential $\lambda = 1$

[g] F $n_1 = 5, n_2 = 2$ [h] Frechet $s = 1, \alpha = 3$

**Fig. 9.** (*continued*)

discretization methods will lead to the increase of estimated error. In order to present this conclusion vividly, an additional simulation is also carried out. The normal distribution $N(0,1)$ is employed. Let $n$ denote the size of dataset and range from 50 to 150 in step of 10. For different values of $n$, 100 datasets are generated randomly and the entropy estimations are carried out base on these 100 times random repetitions. All the entropy estimations corresponding to different parameter $n$ are averaged. The experimental results are listed in Fig. 8:

The empirical observations from Fig. 8 demonstrate the conclusion mentioned above: the discretization methods will lead to the increase of estimated error gradually with the increase of dataset size. Based on the experimental results, two disadvantages of discretization methods are summarized as follows:

(1) For the discretization methods, it is very central to determine the number of intervals and the width of interval. Depending on the actual data distribution and the goals of the analysis, different interval widths may be appropriate. So, these existing determination rules are not always effective. It shows that there is not any best number of intervals, and different interval widths can reveal different features of the data. Compared with those discretization methods, $RE_{I+II}$ is a more flexible estimation method in which the only parameter $h$ is dependent on the given dataset and can be obtained by solving the Eq. (5.20);
(2) In Comparison with the related discretization methods, $RE_{I+II}$ obtains the best estimation performance. And, we find that the discretization methods are not appropriate for the large dataset. With the increase of dataset size, the estimated errors of different discretization methods also increase gradually. On the contrary, $RE_{I+II}$ can reduce the estimated error with the increase of dataset size.

Now, we give an analysis of the time complexity of the above-mentioned 9 discretization methods and $RE_{I+II}$. Because EWD, EFD, OD, FFD, NND, PD, WPD, and MVSDD are dominated by sorting, their complexities are of order $O(n\log_2 n)$ [25], where $n$ is the size of given dataset on which the entropy estimation is conducted based. When $k$ is fixed, the complexity of k-means clustering is $O(n^{(k+1)}\log_2 n)$ [58], where k is the number of clusters. KMCD uses the k-means clustering to determine intervals for the continuous observations. Thus, the complexity of KMCD is $O(n^{(k+1)}\log_2 n)$. In order to solve the optimal bandwidth for $RE_{I+II}$, we must find the minimization of $E_{I+II} = E_I^* + E_{II}^*$. From the Eqs. (36) and (38), we know that there are $n^2$

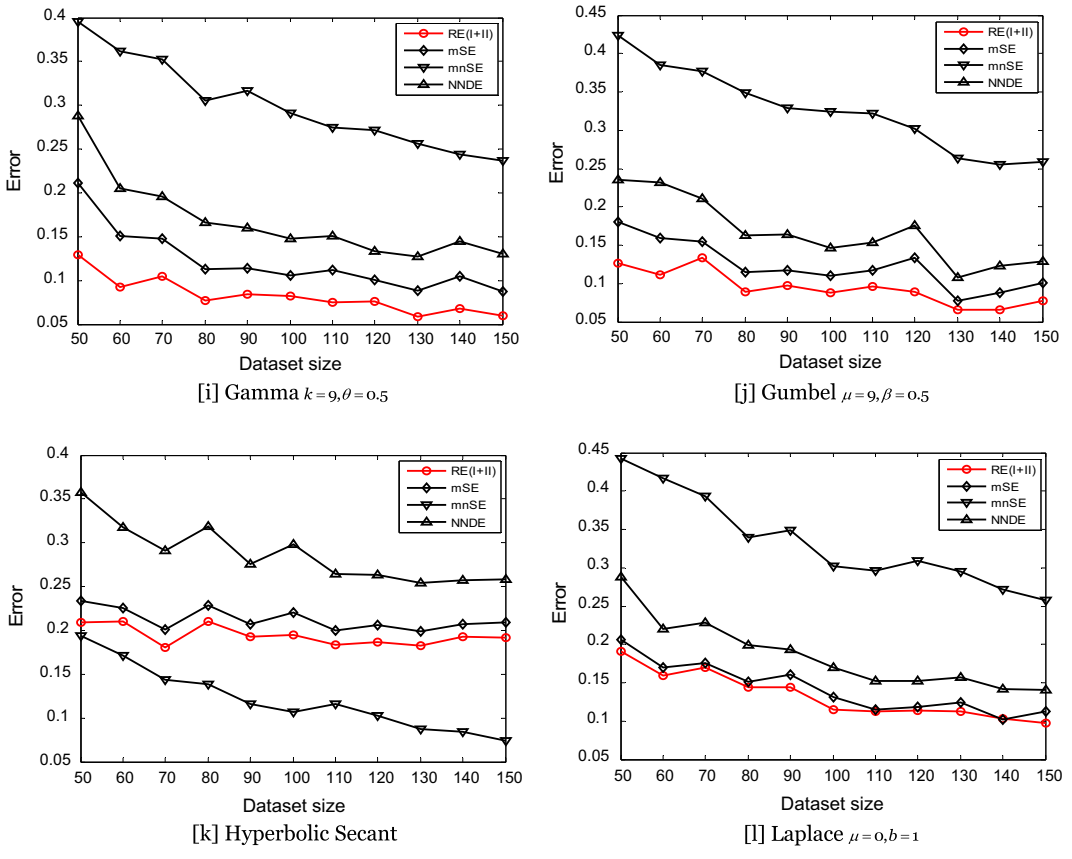



**Fig. 9.** (*continued*)

exponential components which need to calculate accordingly. Therefore, the computational complexity of $RE_{I+II}$ is of order $O(n^2)$. Thus $RE_{I+II}$ has complexity lower than KMCD and higher than EWD, EFD, OD, FFD, NND, PD, WPD, and MVSDD.

### 6.3. Compare $RE_{I+II}$ with mSE, $m_n$ SE, and NNDE

The other three mostly used entropy estimators based on space-$m$ SE [36], $m_n$ SE [37], and NNDE [38]-are also compared their performances with $RE_{I+II}$. The approximation rules of these methods are listed in the Eqs. (12)–(14) respectively. By arranging the experimental procedures as follows, the comparative results will be listed in Fig. 9.

Step 1: For every probability density listed in Table 1, $n$ random samples are generated $X^{(p)} = \left\{x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)}\right\}$, $(p = 1, 2, \ldots, 24)$;

Step 2: Compute the estimated entropies $H_{RE_{I+II}}(X^{(p)}, h_{I+II}), H_{mSE}(X^{(p)}), H_{m_nSE}(X^{(p)})$, and $H_{NNDE}(X^{(p)})$ by using four different estimators: $RE_{I+II}$, $m$ SE, $m_n$ SE, and NNDE;

Step 3: Calculate the estimated errors:

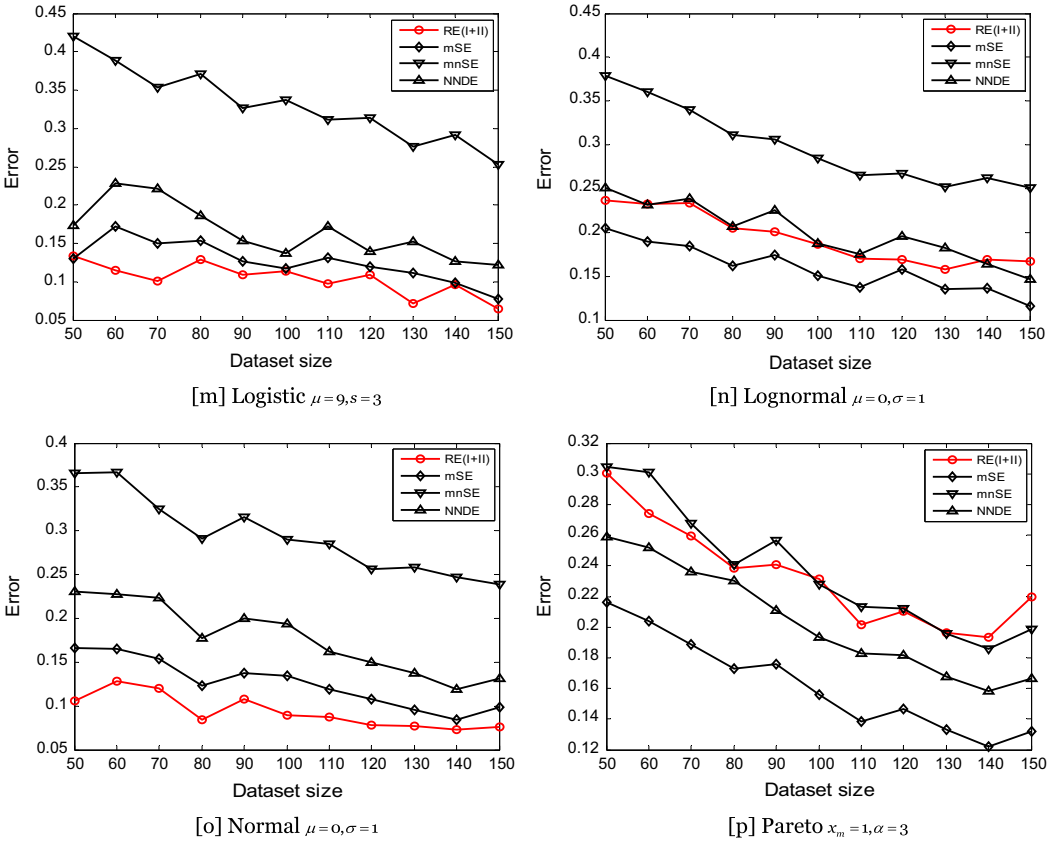
$$RE(I+II) = [H_{RE_{I+II}}(X^{(p)}, h_{I+II}) - H^{(p)}]^2,\ mSE = [H_{mSE}(X^{(p)}) - H^{(p)}]^2,\ m_nSE = [H_{m_nSE}(X^{(p)}) - H^{(p)}]^2,\ \text{and NNDE} = [H_{NNDE}(X^{(p)}) - H^{(p)}]^2,$$

where $H^{(p)}(p = 1, 2, \ldots, 24)$ is the true entropy of the $p$-th density listed in Table 1;

Step 4: Set the size of used dataset $n$ ranging from 50 to 150 in step of 10. Repeat the following Steps 1–3 for different values of $n$ and 100 times for each value. The average values of $RE(I+II), mSE, m_nSE$, and NNDE terms on the 100 repeated trials are recorded respectively.

From the comparative results, we can summarize the following experimental observations:

(1) Compared with $m$ SE, $RE_{I+II}$ has obtained the better estimation performances on 20 density distributions except Cauchy $(\lambda = 1)$, Frechet $(s = 1, \alpha = 3)$, Lognormal $(\mu = 0, \sigma = 1)$, and Pareto $(x_m = 1, \alpha = 3)$ distributions;




[m] Logistic $\mu=9, s=3$ [n] Lognormal $\mu=0, \sigma=1$

[o] Normal $\mu=0, \sigma=1$ [p] Pareto $x_m=1, \alpha=3$

**Fig. 9.** (*continued*)

(2) Compared with $m_n$ SE, $\mathrm{RE}_{\mathrm{I+II}}$ has obtained the better estimation performances on 23 density distributions except hyperbolic secant distribution;
(3) Compared with NNDE, $\mathrm{RE}_{\mathrm{I+II}}$ has obtained the better estimation performances on 20 density distributions except Cauchy ($\lambda = 1$), Frechet ($s = 1, \alpha = 3$), Lognormal ($\mu = 0, \sigma = 1$), and Pareto ($x_m = 1, \alpha = 3$) distributions.

Now, we will try to discuss the reason of incapability of $\mathrm{RE}_{\mathrm{I+II}}$. Let $H_{\mathrm{Space}}$ denote the estimated entropy with space based entropy estimator ($H_{mSE}, H_{m_nSE}$, or $H_{\mathrm{NNDE}}$). The incapability of $\mathrm{RE}_{\mathrm{I+II}}$ can be reflected from the larger estimation error. So, the following inequality can be derived:

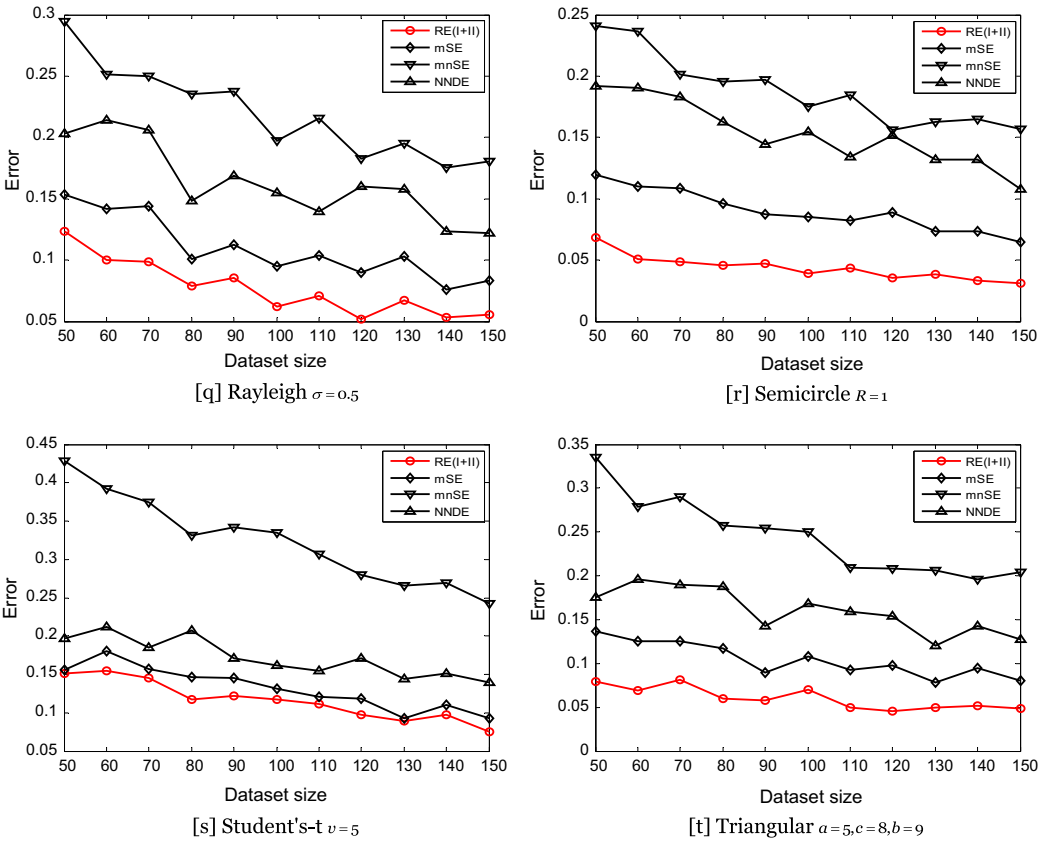
$$\begin{aligned}(H_{\mathrm{Space}} - H_{\mathrm{True}})^2 \prec (H_{RE_{\mathrm{I+II}}} - H_{\mathrm{True}})^2 &\Rightarrow H^2_{\mathrm{Space}} - 2H_{\mathrm{Space}}H_{\mathrm{True}} + H^2_{\mathrm{True}} \prec H^2_{RE_{\mathrm{I+II}}} - 2H_{RE_{\mathrm{I+II}}}H_{\mathrm{True}} + H^2_{\mathrm{True}} \\ &\Rightarrow H^2_{\mathrm{Space}} - 2H_{\mathrm{Space}}H_{\mathrm{True}} \prec H^2_{RE_{\mathrm{I+II}}} - 2H_{RE_{\mathrm{I+II}}}H_{\mathrm{True}} \Rightarrow (H_{\mathrm{Space}} - H_{RE_{\mathrm{I+II}}})(H_{\mathrm{Space}} + H_{RE_{\mathrm{I+II}}} - 2H_{\mathrm{True}}) \prec 0. \end{aligned} \quad (41)$$

The inequality (41) shows that when $H_{\mathrm{Space}} - H_{RE_{\mathrm{I+II}}}$ and $H_{\mathrm{Space}} + H_{RE_{\mathrm{I+II}}} - 2H_{\mathrm{True}}$ terms keep the opposite signs, the estimated quality of $\mathrm{RE}_{\mathrm{I+II}}$ will be inferior compared with space based entropy estimator. In order to validate this presentation, we carry out a numerical experiment on Cauchy ($\lambda = 1$), Frechet ($s = 1, \alpha = 3$), Hyperbolic Secant, Lognormal ($\mu = 0, \sigma = 1$), and Pareto distributions ($x_m = 1, \alpha = 3$). For every distribution, 100 random datasets are generated and the estimated entropy $H_{\mathrm{Space}}$ is the average of 100 times of repeated trials. Then, the $H_{\mathrm{Space}} - H_{RE_{\mathrm{I+II}}}$ and $H_{\mathrm{Space}} + H_{RE_{\mathrm{I+II}}} - 2H_{\mathrm{True}}$ terms are calculated based on the estimated entropy $H_{\mathrm{Space}}$. The experimental results are summarized in Table 8.

From the experimental results we can find that our proposed explanation is acceptable and can reflect the estimation mechanism of $\mathrm{RE}_{\mathrm{I+II}}$. For example, on Hyperbolic Secant distribution, $m_n$ SE has obtained the better estimation compared with $\mathrm{RE}_{\mathrm{I+II}}$, while $m$ SE and NNDE are second to $\mathrm{RE}_{\mathrm{I+II}}$. Then, we can find the following experimental observations:

$$\begin{aligned} &m\mathrm{SE}(\mathrm{Hyp}) : [H_{mSE}(\mathrm{Hyp}) - H_{RE_{\mathrm{I+II}}}(\mathrm{Hyp})][H_{mSE}(\mathrm{Hyp}) + H_{RE_{\mathrm{I+II}}}(\mathrm{Hyp}) - 2H_{\mathrm{True}}(\mathrm{Hyp})] \succ 0, \\ &m_n\mathrm{SE}(\mathrm{Hyp}) : [H_{m_nSE}(\mathrm{Hyp}) - H_{RE_{\mathrm{I+II}}}(\mathrm{Hyp})][H_{m_nSE}(\mathrm{Hyp}) + H_{RE_{\mathrm{I+II}}}(\mathrm{Hyp}) - 2H_{\mathrm{True}}(\mathrm{Hyp})] \prec 0, \\ &\text{and } \mathrm{NNDE}(\mathrm{Hyp}) : [H_{\mathrm{NNDE}}(\mathrm{Hyp}) - H_{RE_{\mathrm{I+II}}}(\mathrm{Hyp})][H_{\mathrm{NNDE}}(\mathrm{Hyp}) + H_{RE_{\mathrm{I+II}}}(\mathrm{Hyp}) - 2H_{\mathrm{True}}(\mathrm{Hyp})] \succ 0. \end{aligned}$$

Similarly, on Cauchy ($\lambda = 1$) distribution, $m_n$ SE has obtained the worse estimation compared with $\mathrm{RE}_{\mathrm{I+II}}$, while $m$ SE and NNDE are better than $\mathrm{RE}_{\mathrm{I+II}}$. Then, we can also find the following experimental observations:




[q] Rayleigh $\sigma = 0.5$ [r] Semicircle $R = 1$

[s] Student's-t $v = 5$ [t] Triangular $a = 5, c = 8, b = 9$

**Fig. 9.** (*continued*)

$$m\mathrm{SE}(\mathrm{Cauchy}):[H_{mSE}(\mathrm{Cauchy})-H_{RE_{\mathrm{I+II}}}(\mathrm{Cauchy})][H_{mSE}(\mathrm{Cauchy})+H_{RE_{\mathrm{I+II}}}(\mathrm{Cauchy})-2H_{\mathrm{True}}(\mathrm{Cauchy})]\prec 0,$$
$$m_n\mathrm{SE}(\mathrm{Cauchy}):[H_{m_nSE}(\mathrm{Cauchy})-H_{RE_{\mathrm{I+II}}}(\mathrm{Cauchy})][H_{m_nSE}(\mathrm{Cauchy})+H_{RE_{\mathrm{I+II}}}(\mathrm{Cauchy})-2H_{\mathrm{True}}(\mathrm{Cauchy})]\succ 0,$$
$$\text{and } \mathrm{NNDE}(\mathrm{Cauchy}):[H_{\mathrm{NNDE}}(\mathrm{Cauchy})-H_{RE_{\mathrm{I+II}}}(\mathrm{Cauchy})][H_{\mathrm{NNDE}}(\mathrm{Cauchy})+H_{RE_{\mathrm{I+II}}}(\mathrm{Cauchy})-2H_{\mathrm{True}}(\mathrm{Cauchy})]\prec 0.$$

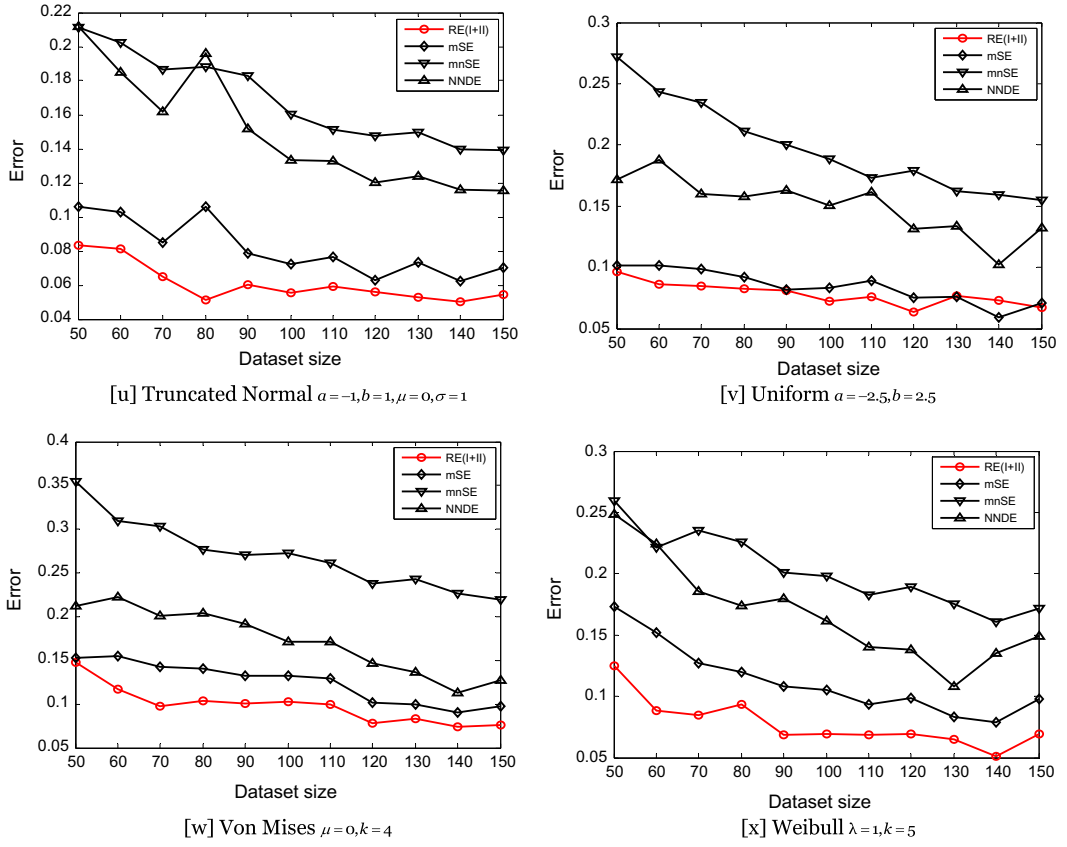
For the distributions on which $m$ SE and NNDE have obtained the better estimations, there is such a data characteristic which can be extracted: $H_{\mathrm{Space}}-H_{RE_{\mathrm{I+II}}}\succ 0$. The following discussions will give a detailed and insightful explanation of this characteristic. Let the current dataset be $X=\{x_1,x_2,\ldots,x_n\}$. And, sort the original dataset $X=\{x_1,x_2,\ldots,x_n\}$ in ascending order and the ordered dataset is: $\overline{X}=\{\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_n\}$, where $\bar{x}_1\leqslant\bar{x}_2\leqslant\cdots\leqslant\bar{x}_n$. For the elementary function $y(X)=\exp\left(-\frac{1}{2}x^2\right),x\in(-\infty,+\infty)$, we can find that when $|x|\geqslant 4.714$ (The value of 4.714 is not designated, that any value larger than 4.714 is also available), the value of $y(X)\leqslant 0.00001$ will be very close to 0. So, we think it is expected that when $|x|\geqslant 4.714, y(X)\to 0$. Based on the ordered dataset $\overline{X}=\{\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_n\}$, $\mathrm{RE}_{\mathrm{I+II}}$ can be described as:

$$\widetilde{H}_{RE_{\mathrm{I+II}}}=-\frac{1}{n}\sum_{i=1}^{n}\ln\left\{\frac{1}{nh_{\mathrm{I+II}}}\sum_{j=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\bar{x}_i-\bar{x}_j}{h_{\mathrm{I+II}}}\right)^2\right]\right\}=-\frac{1}{n}\sum_{i=1}^{n}\ln\left\{\frac{1}{\sqrt{2\pi}nh_{\mathrm{I+II}}}\sum_{j=1}^{n}\exp\left[-\frac{1}{2}\left(\frac{\bar{x}_i-\bar{x}_j}{h_{\mathrm{I+II}}}\right)^2\right]\right\}. \tag{42}$$

When $\min_{\substack{i=1,2,\ldots,n\\ j=1,2,\ldots,n\\ i\neq j}}\left|\frac{\bar{x}_i-\bar{x}_j}{h_{\mathrm{I+II}}}\right|\geqslant 4.714$, we can get the component $\underset{i=1,2,\ldots,n}{\frac{\sum_{j=1}^{n}\exp\left[-\frac{1}{2}\left(\frac{\bar{x}_i-\bar{x}_j}{h_{\mathrm{I+II}}}\right)^2\right]}{}}\overset{\text{when}\bar{x}_i=\bar{x}_j}{\longrightarrow}1$. Then, the approximation of the Eq. (42) is:

$$\widetilde{H}_{RE_{\mathrm{I+II}}}\approx-\frac{1}{n}\sum_{i=1}^{n}\ln\left[\frac{1}{\sqrt{2\pi}nh_{\mathrm{I+II}}}\right]=\ln\left[\sqrt{2\pi}nh_{\mathrm{I+II}}\right]. \tag{43}$$

Because $\min_{\substack{i=1,2,\ldots,n\\ j=1,2,\ldots,n\\ i\neq j}}\left|\frac{\bar{x}_i-\bar{x}_j}{h_{\mathrm{I+II}}}\right|\geqslant 4.714$, then $\min_{\substack{i=1,2,\ldots,n\\ j=1,2,\ldots,n\\ i\neq j}}|\bar{x}_i-\bar{x}_j|\geqslant 4.714h_{\mathrm{I+II}}$. The approximations of $m$ SE and NNDE can also be obtained by bringing $\min_{\substack{i=1,2,\ldots,n\\ j=1,2,\ldots,n\\ i\neq j}}|\bar{x}_i-\bar{x}_j|\geqslant 4.714h_{\mathrm{I+II}}$ into the Eqs. (12) and (14), let $m=1$ in the Eq. (12):




[u] Truncated Normal $a=-1,b=1,\mu=0,\sigma=1$ [v] Uniform $a=-2.5,b=2.5$

[w] Von Mises $\mu=0,k=4$ [x] Weibull $\lambda=1,k=5$

**Fig. 9.** (*continued*)

$$\widetilde{H}_{mSE} = \frac{1}{n}\sum_{i=1}^{n-1} \ln[n(\bar{x}_{i+1} - \bar{x}_i)] - \Psi(1) \approx \ln[4.714 n h_{\mathrm{I+II}}] - \Psi(1), \tag{44}$$

and

$$\widetilde{H}_{\mathrm{NNDE}} = \frac{1}{n}\sum_{i=1}^{n} \ln(n d_i) + \ln 2 + \gamma_E \approx \ln[4.714 n h_{\mathrm{I+II}}] + \ln 2 + \gamma_E. \tag{45}$$

Then, by comparing the Eq. (43) with Eqs. (44) and (45), we can find $\widetilde{H}_{mSE} - \widetilde{H}_{RE_{\mathrm{I+II}}} \succ 0$ and $\widetilde{H}_{\mathrm{NNDE}} - \widetilde{H}_{RE_{\mathrm{I+II}}} \succ 0$ with the increase of component $nh_{\mathrm{I+II}}$. This shows that the imprecise bandwidth selection for $\mathrm{RE}_{\mathrm{I+II}}$ on some distributions, for example, Cauchy ($\lambda = 1$), Frechet ($s = 1, \alpha = 3$), Lognormal ($\mu = 0, \sigma = 1$), and Pareto ($x_m = 1, \alpha = 3$) distributions, will lead to the conclusion $H_{Space} - H_{RE_{\mathrm{I+II}}} \succ 0$. So, such distributions satisfying the condition $\min_{\substack{i=1,2,\ldots,n \\ j=1,2,\ldots,n \\ i\neq j}} \left|\frac{\bar{x}_i - \bar{x}_j}{h_{\mathrm{I+II}}}\right| \geqslant 4.714$ will hamper $\mathrm{RE}_{\mathrm{I+II}}$ to manifest its advantages.

## 7. Conclusions and future research

In this study, we have investigated a new strategy of selecting an optimal bandwidth for re-substitution entropy estimator (RE). Two types of generated errors, entropy estimation error (type-I error) and density estimation error (type-II error), are considered to merge together. The new estimator named $\mathrm{RE}_{\mathrm{I+II}}$ aims to make a trade-off between these two errors. On 24 typical probability distributions, the estimation performance of $\mathrm{RE}_{\mathrm{I+II}}$ is demonstrated. The 9 mostly used unsupervised discretization methods and 5 sophisticated entropy estimators are employed as the competitors. The experimental results show that $\mathrm{RE}_{\mathrm{I+II}}$ can indeed attain the better estimation performance among the involved methods. We summarize some highlights briefly as follows: (1) the optimal bandwidth used in $\mathrm{RE}_{\mathrm{I+II}}$ indeed performs better than the singled bandwidth by minimizing type-I error. And, compared with the sophisticated splitting data estimator (SDE) and cross-validation estimator (CVE), $\mathrm{RE}_{\mathrm{I+II}}$ can also obtain a satisfactory estimation. The simulations confirm the validity and effectiveness of the derivation error measure; (2) the discretized estimators are sensitive to the dataset size. It is not feasible to build the entropy estimation when given with a large-scale dataset. With the increase of dataset size, the estimation error will increase significantly. On the contrary, $\mathrm{RE}_{\mathrm{I+II}}$ can reduce the estimation error with the augment of dataset size; (3) the overall performance of $\mathrm{RE}_{\mathrm{I+II}}$ also goes beyond the estimation behaviors of $m$-spacing estimator ($m$ SE), $m_n$-spacing estimator ($m_n$ SE), and nearest neighbor distance estimator (NNDE). The application conditions of $m$ SE, $m_n$ SE, and NNDE are also discussed with the experimental comparisons. The empirical analysis demonstrates that $\mathrm{RE}_{\mathrm{I+II}}$ is more insensitive to data and a better generalizable way for the estimation of continuous entropy. Our scheduled further developments in the research are to: (1) extend the new entropy estimation strategy into multivariate domain; and (2) apply the new entropy estimator in some machine learning and data mining algorithms, such as, decision tree and Bayesian classifier.

## Acknowledgements

The authors would like to thank the editors and anonymous reviewers. Their valuable and constructive comments and suggestions helped them in significantly improving this paper. This research is partially supported by the National Natural Science Foundations of China (60903088 and 61170040) and the Natural Science Foundations of Hebei Province (F2010000323, F2011201063, and F2012201023). The authors are also grateful for the partial supports of GRF Grant 5237/08E and CRG Grant G-U756 of The Hong Kong Polytechnic University.

## References

[1] C.E. Shannon, A mathematical theory of communication, Bell System Technical Journal 27 (1948) 379–423. 623–656.
[2] C.E. Shannon, Prediction and entropy of printed English, Bell System Technical Journal 30 (1951) 50–64.
[3] J.H. Lin, Divergence measures based on the Shannon entropy, IEEE Transactions on Information Theory 37 (1) (1991) 145–151.
[4] D.J.C. MacKay, Information Theory, Inference, and Learning Algorithms, Cambridge University Press, 2004.
[5] J.R. Quinlan, Induction of decision trees, Machine Learning 1 (1) (1986) 81–106.
[6] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufman Publishers, 1993.
[7] X.Z. Wang, C.R. Dong, Improving generalization of fuzzy if–then rules by maximizing fuzzy entropy, IEEE Transactions on Fuzzy Systems 17 (3) (2009) 556–567.
[8] A. Ratnaparkhi, Learning to parse natural language with maximum entropy models, Machine Learning 34 (1-3) (1999) 151–175.
[9] K. Nigam, J. Lafferty, A. Mccallum, Using maximum entropy for text classification, in: IJCAI-99 Workshop on Machine Learning for Information Filtering, 1999, pp. 61–67.
[10] H.M. Lee, C.M. Chen, J.M. Chen, Y.L. Jou, An efficient fuzzy classifier with feature selection based on fuzzy entropy, IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics 31 (3) (2001) 426–432.
[11] P. Mitra, C.A. Murthy, S.K. Pal, Unsupervised feature selection using feature similarity, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (3) (2002) 301–312.
[12] U.M. Fayyad, K.B. Irani, On the handling of continuous-valued attributes in decision tree generation, Machine Learning 8 (1) (1992) 87–102.
[13] H. Singer, Maximum entropy inference for mixed continuous-discrete variables, International Journal of Intelligent Systems 25 (4) (2010) 345–364.
[14] P.K. Li, B.D. Liu, Entropy of credibility distributions for fuzzy variables, IEEE Transactions on Fuzzy Systems 16 (1) (2008) 123–129.
[15] X.Z. Wang, J.H. Zhai, S.X. Lu, Induction of multiple fuzzy decision trees based on rough set technique, Information Sciences 178 (2008) 3188–3202.
[16] J. Catlett, On changing continuous attributes into ordered discrete attributes, Lecture Notes in Computer Science 482 (1991) 164–178.

[17] R. Kerber, ChiMerge: discretization of numeric attributes, in: Proceedings of the Tenth National Conference on Artificial Intelligence (AAAI-1992), 1992, pp.123–128.
[18] J. Dougherty, R. Kohavi, M. Sahami, Supervised and unsupervised discretization of continuous features, in: Proceedings of the Twelfth International Conference on Machine Learning (ICML-1995), 1995, pp. 194–202.
[19] C.N. Hsu, H.J. Huang, T.T. Wong, Why discretization works for naive bayesian classifiers, in: Proceedings of the Seventeenth International Conference on Machine Learning (ICML-2000), 2000, pp. 399–406.
[20] C.N. Hsu, H.J. Huang, T.T. Wong, Implications of the Dirichlet assumption for discretization of continuous variables in naive Bayesian classifiers, Machine Learning 53 (3) (2003) 235–263.
[21] L. Torgo, J. Gama, Search-based class discretization, Lecture Notes in Computer Science 1224 (1997) 266–273.
[22] J.A. Hartigan, M.A. Wong, Algorithm AS 136: A K-means clustering algorithm, Journal of the Royal Statistical Society-Series C: Applied Statistics 28 (1) (1979) 100–108.
[23] E. Frank, I.H. Witten, Making better use of global discretization, in: Proceedings of the 16th International Conference on Machine Learning (ICML-1999), 1999, pp. 115–123.
[24] S.A. Macskassy, H. Hirsh, A. Banerjee, A.A. Dayanik, Using text classifiers for numerical classification, in: Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-2001), 2001, pp. 885–890.
[25] Y. Yang, G.I. Webb, Discretization for Naive–Bayes learning managing discretization bias and variance, Machine Learning 74 (1) (2009) 39–74.
[26] Y. Yang, G.I. Webb, Non-disjoint discretization for Naive–Bayes classifiers, in: Proceedings of the 19th International Conference on Machine Learning (ICML-2002), 2002, pp. 666–673.
[27] Y. Yang, G.I. Webb, Proportional k-interval discretization for Naive–Bayes classifiers, in: Proceedings of the 12th European Conference on Machine Learning (ECML-2001), 2001, pp. 564–575.
[28] Y. Yang, G.I. Webb, Weighted proportional k-interval discretization for Naive–Bayes classifiers, in: Proceedings of the Seventh Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-2003), 2003, pp. 501–512.
[29] Y. Yang, G.I. Webb, A comparative study of discretization methods for Naive–Bayes classifiers, in: Proceedings of the 2002 Pacific Rim Knowledge Acquisition Workshop in PRICAI 2002 (PKAW-2002), 2002, pp. 159–173.
[30] Y. Yang, Discretization for Naive–Bayes learning, The School of Computer Science and Software Engineering of Monash University, 2003.
[31] H.C. Peng, F.H. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (8) (2005) 1226–1238.
[32] I. Ahmad, P.E. Lin, A nonparametric estimation of the entropy for absolutely continuous distributions, IEEE Transactions on Information Theory 22 (3) (1976) 372–375.
[33] H. Joe, Estimation of entropy and other functionals of a multivariate density, Annals of the Institute of Statistical Mathematics 41 (1989) 683–697.
[34] L. Györfi, E.C. van der Meulen, Density-free convergence properties of various estimators of entropy, Computational Statistics and Data Analysis 5 (4) (1987) 425–436.
[35] P. Hall, S. Morton, On the estimation of entropy, Annals of the Institute of Statistical Mathematics 45 (1) (1993) 69–88.
[36] J. Beirlant, M.C.A. van Zuijlen, The empirical distribution function and strong laws for functions of order statistics of uniform spacings, Journal of Multivariate Analysis 16 (3) (1985) 300–317.
[37] O. Vasicek, A test for normality based on sample entropy, Journal of the Royal Statistical Society-Series B: Methodological 38 (1) (1976) 54–59.
[38] A.B. Tsybakov, E.C. van der Meulen, Root-n consistent estimators of entropy for densities with unbounded support, Scandinavian Journal of Statistics 23 (1) (1996) 75–83.
[39] E. Parzen, On estimation of a probability density function and mode, Annals of Mathematical Statistics 33 (3) (1962) 1065–1076.
[40] D.W. Scott, Multivariate Density Estimation: Theory, Practice, and Visualization, John Wiley and Sons, Inc., 1992.
[41] M.P. Wand, M.C. Jones, Kernel Smoothing, Chapman and Hall, 1995.
[42] R.C. Holte, Very simple classification rules perform well on most commonly used datasets, Machine Learning 11 (1993) 63–91.
[43] H. Liu, F. Hussain, C. Lim Tan, M. Dash, Discretization: an enabling technique, Data Mining and Knowledge Discovery 6 (4) (2002) 393–423.
[44] S.A. Macskassy, H. Hirsh, A. Banerjee, A.A. Dayanik, Converting numerical classification into text classification, Artificial Intelligence 143 (1) (2003) 51–77.
[45] E. Frank, M. Hall, A simple approach to ordinal classification, Lecture Notes in Computer Science 2167 (2001) 145–156.
[46] J.H. Friedman, On bias, variance, 0/1-loss, and the curse-of-dimensionality, Data Mining and Knowledge Discovery 1 (1997) 55–77.
[47] C. Ding, H.C. Peng, Minimum redundancy feature selection from microarray gene expression data, Journal of Bioinformatics and Computational Biology 3 (2) (2005) 185–205.
[48] M. Abramowitz, I.A. Stegun, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, Dover, New York, 1972. 9th printing.
[49] J. Sondow, An antisymmetric formula for Euler's constant, Mathematics Magazine 71 (3) (1998) 219–220.
[50] G. Upton, I. Cook, Understanding Statistics, Oxford University Press, 1996.
[51] H.C. Peng's, mRMR Feature Selection Site, <http://penglab.janelia.org/proj/mRMR/>, 2011.
[52] A.V. Lazo, P. Rathie, On the entropy of continuous probability distributions, IEEE Transactions on Information Theory 24 (1) (1978) 120–122.
[53] H.A. Sturges, The choice of a class interval, Journal of the American Statistical Association 21 (153) (1926) 65–66.
[54] D.W. Scott, On optimal and data-based histograms, Biometrika 66 (3) (1979) 605–610.
[55] D. Freedman, P. Diaconis, On the histogram as a density estimator: $L_2$ theory, Probability Theory and Related Fields 57 (4) (1981) 453–476.
[56] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes: the Art of Scientific Computing, third ed., Cambridge University Press, New York, 2007.
[57] A. Pérez, P. Larrañaga, I. Inza, Bayesian classifiers based on kernel density estimation: flexible classifiers, International Journal of Approximate Reasoning 50 (2) (2009) 341–362.
[58] M. Inaba, N. Katoh, H. Imai, Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering, in: Proceedings of 1994 ACM Symposium on Computational Geometry (SoCG-1994), 1994, pp. 332–339.
[59] P.P. Guan, H. Yan, A hierarchical multilevel thresholding method for edge information extraction using fuzzy entropy, International Journal of Machine Learning and Cybernetics (2011), http://dx.doi.org/10.1007/s13042-011-0063-7.
[60] S.T. Wang, Z.H. Deng, F.L. Chung, W.J. Hu, From Gaussian kernel density estimation to kernel methods, International Journal of Machine Learning and Cybernetics (2012), http://dx.doi.org/10.1007/s13042-012-0078-8.