Structural and Multidisciplinary Optimization (2018) 57:95–113
https://doi.org/10.1007/s00158-017-1873-3

RESEARCH PAPER

Kernel density estimation with bounded data

Young-Jin Kang[1] · Yoojeong Noh[1] · O-Kaung Lim[1]

Received: 20 March 2017 / Revised: 21 November 2017 / Accepted: 22 November 2017 / Published online: 8 December 2017

Abstract
The uncertainties of input variables are quantified as probabilistic distribution functions using parametric or nonparametric statistical modeling methods for reliability analysis or reliability-based design optimization. However, parametric statistical modeling methods such as the goodness-of-fit test and the model selection method are inaccurate when the number of data is very small or the input variables do not have parametric distributions. To deal with this problem, kernel density estimation with bounded data (KDE-bd) and KDE with estimated bounded data (KDE-ebd), which randomly generates bounded data within given input variable intervals for given data and applies them to generate density functions, are proposed in this study. Since the KDE-bd and KDE-ebd use input variable intervals, they attain better convergence to the population distribution than the original KDE does, especially for a small number of given data. The KDE-bd can even deal with a problem that has one data with input variable bounds. To verify the proposed method, statistical simulation tests were carried out for various numbers of data using multiple distribution types and then the KDE-bd and KDE-ebd were compared with the KDE. The results showed the KDE-bd and KDE-ebd to be more accurate than the original KDE, especially when the number of data is less than 10. It is also more robust than the original KDE regardless of the quality of given data, and is therefore more useful even if there is insufficient data for input variables.

Keywords Kernel density estimation · Nonparametric statistical modeling · Interval approach · Nonparametric distribution · Bounded data · Intersection area

1 Introduction

Uncertain quantification of random input variables is an important issue in the reliability analysis and reliability-based design optimization of physical systems. Accurate uncertainty quantification yields accurate reliability analysis results, and thus creates accurate design optimization results (Noh et al. 2010). Moreover, statistical model validation and calibration has recently been developed to improve the accuracy of computer aided engineering (CAE) analysis. In these processes, accurate statistical modeling of random input and output variables is required to verify the CAE analysis results and improve their accuracy (Youn et al. 2011).

✉ Yoojeong Noh
yoonoh@pusan.ac.kr

[1] School of Mechanical Engineering, Pusan National University, Pusan 609-735, South Korea

To quantify the uncertainties in random input variables, various statistical modeling methods have been proposed. The uncertainties can be quantified as probabilistic distributions or input variable intervals. A parametric method such as the goodness-of-fit (GOF) test and model selection method is the most commonly used statistical modeling method. In addition, the sequential statistical modeling (SSM) method that combines the GOF tests and model selection methods has been proposed (Kang et al. 2016). The GOF tests, such as the Kolmogorov-Smirnov (K-S) test and Anderson-Darling (A-D) test, determine the appropriateness of a candidate distribution for a given data set by accepting or rejecting a null hypothesis that a candidate distribution is a true model to represent the given data (Ayyub and McCuen 2012; Anderson and Darling 1952). A model selection method, such as the Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), and Bayesian method, determines relative appropriateness of candidate distributions by selecting the best fitted distribution for the given data set among several candidate distributions (Akaike 1974; Schwarz 1978; Burnham and Anderson 2004). The SSM method first

assesses the absolute fit of candidate models by a GOF test, and then selects the best fit distribution by using a model selection method among the selected candidate distributions accepted by the GOF test. If the number of given data is sufficient, i.e., larger than 30, and the data follows specific parametric distributions and the parametric method is accurate. However, if it is not, a nonparametric modeling method or interval approach needs to be used (Kang et al. 2017).

The nonparametric modeling method estimates a probabilistic distribution by only using given data without statistical parameters. Kernel density estimation (KDE) is the most commonly used nonparametric statistical modeling method. The KDE generates kernel functions based on the data with an optimal bandwidth, which is a smoothing parameter of the kernel functions, and then a KDE function is obtained by combining the kernel functions of all data. If input variables have nonparametric distributions and sufficient data is provided, the KDE is recommended to represent the given data set. However, in real applications, only a few of the data are applicable, and thus, it is difficult to use the existing parametric or nonparametric statistical modeling methods. Even though some input variables such as material properties are known to follow specific distributions, the distribution types and parameters can be different according to the dimensions or manufacturing process. Jang et al. (Jang et al. 2015) used the KDE to carry out the reliability-based design optimization of electric power steering motor and Cho et al. (Cho et al. 2016) used the nonparametric approach for uncertainty quantification in multidisciplinary design optimization. However, these methods have a limitation in expressing distribution functions when data is very limited and cannot deal with given data and information of intervals together.

Due to the lack of data, which is often common in engineering fields, an interval approach that quantifies uncertainties using input variables is often used. The interval approach such as the Dempster-Shafer theory, possibility theory, probability bounds approach (probability box theory), or a uniform distribution is most often used (Verma et al. 2010). These methods estimate intervals of input variables or given data, and then use them to create a reliability analysis or design. However, the empirical distributions obtained from the Dempster-Shafer theory cannot be used in numerical reliability analysis methods such as the first-order reliability method (FORM) or second-order reliability method (SORM) (Yao et al. 2013; Shah et al. 2015). Surrogate models for the empirical distributions have been applied in reliability analysis to overcome the problem of discretization, but there are still approximation errors (Agarwal et al. 2004; Zhang et al. 2014). Since the probability box method uses distributions with lower and upper bounds of estimated parameters, it requires more numbers for the reliability analysis than using one distribution with the estimated parameters only (Tucker and Ferson 2003; Karanki et al. 2009; Betrie et al. 2014, 2016). The interval approaches such as Dempster-Shafer theory and Probability bounds have been used for uncertainty quantifications, safety assessments, and design optimizations (Yao et al. 2013; Shah et al. 2015; Agarwal et al. 2004; Zhang et al. 2014; Tucker and Ferson 2003; Karanki et al. 2009; Betrie et al. 2014, 2016). However, these methods cannot overcome the imprecise probability for the output response, which is always expressed as its lower and upper bounds. To overcome the problem of the interval approach, the uniform distribution is often used, but it only includes the intervals of the input variables and the data distribution cannot be used to create a distribution function. Thus, a statistical modeling method that can use information from both given data and intervals of input variables is necessary in real engineering fields.

To overcome the limitations of the probabilistic approach and interval approach, the KDE-bd and KDE-ebd are proposed. This method combines the nonparametric statistical modeling method (KDE) and the interval approach using bounded data. In the proposed methods, a kernel density function is generated by combining kernel functions for each data set where the data includes given experimental data and bounded data. If the intervals of input variables are known, the bounded data are sampled from the given intervals, and the KDE-bd is used to generate the KDE functions by using the given and bounded data together. However, if intervals are unknown, the KDE-ebd is used to generate the KDE functions by using both the given data and the bounded data sampled from the estimated intervals. When the number of given data is very small, the density functions using the KDE-bd and KDE-ebd have moderate slopes and heavy tails because the shapes of the KDE functions are affected more by the bounded data than the given data. As the amount of data increases, the bounded data will not be used and the given data will mostly determine the shapes of the KDE functions, and thus, they become similar to the population distribution.

To verify the proposed method, various types of distributions are assumed to be true models in order to create sample data with various sizes that are randomly generated for a statistical simulation test. The simulation results using the proposed method are compared with those that used the original KDE. In real applications, 80 experimental data points of the compressive strength of aluminum-lithium are used to verify the validity of the proposed methods. A simple reliability analysis problem is used to show how the input models obtained from the KDE-bd and KDE-ebd affect the reliability analysis results by comparing them with those from the original KDE and uniform distributions.

In Section 2, the interval approach and KDE are discussed in more detail, and the maximum likelihood estimation (MLE) method is explained as a way of estimating bounds for generating bounded data in KDE-bd and KDE-ebd. Section 3 describes the KDE-bd and KDE-ebd process, and Section 4 presents statistical simulation results of the KDE, KDE-bd and

KDE-ebd in two cases, case I: given data with bounds and case II: given data without bounds. Finally, the KDE-bd and KDE-ebd are compared with the KDE and verified for accuracy of statistical modeling and reliability analysis through numerical examples in Section 5.

2 Statistical modeling methods

2.1 Various statistical modeling methods

The statistical modeling methods can be categorized as probabilistic, interval, parametric, or nonparametric approaches. Table 1 shows the categorized statistical modeling methods including the KDE-bd and KDE-ebd. The parametric approach uses parametric statistical models to model the data. The GOF test determines whether a specific statistical model fits the data or not, while the model selection method finds the best fitted model for the given data. The SSM uses the GOF test to select candidate models that satisfy absolute appropriateness, and then uses the model selection method to identify the best fitted distribution among the selected candidate distributions.

The p-box method generates bounds for the estimated cumulative distribution function (CDF) by using the lower and upper bounds of estimated parameters from the given data where the distribution type is known or can be identified from the probabilistic and parametric approach. The Dempster-Shafer theory assigns basic probability, which is given by the users' experience, to each piece of data and finds the bounds of the probability using the plausibility and belief functions (Verma et al. 2010). The basic probability assignment for each piece of data yields empirical CDF values, so that the lower and upper bounds of the CDFs are also empirical values.

Similarly, the interval representation deals with the uncertainty in the input variables using the interval numbers, which are the absolute bounds of the uncertain variables that one wants to explore in the analysis (Verma et al. 2010). The parametric or nonparametric approach can be classified as either a probabilistic or interval approach, which means that it cannot generate a statistical model to use both the given data and bound information together. On the other hand, the KDE-bd and KDE-ebd can use both the data and input variable bounds to generate a statistical model.

2.2 Kernel density estimation

The KDE is a nonparametric statistical modeling method that does not use parametric probability density functions (PDF) but only uses given data to create a statistical model. In other words, the KDE does not require statistical moments or specific probability density functions to estimate a probabilistic distribution. A kernel density function is obtained by combining kernel functions generated by each value. Since the KDE uses only data, it is useful when a parametric PDF cannot represent a distribution of given data. The KDE function is defined as (Silverman 1986; Wand and Jones 1994)

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

A simpler formula for (1) can be calculated by applying the rescaling notation $K_h(t) = (1/h)K(t/h)$. This equation is defined as (Silverman 1986; Wand and Jones 1994)

$$\hat{f} = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) \tag{2}$$

where X_i is given data, $\hat{f}$ is an estimated kernel density function, and $K(\cdot)$ is a kernel function satisfying $\int_{-\infty}^{+\infty} K(x)dx = 1$. h is a positive value and named as a bandwidth or smoothing parameter of the kernel function. If h becomes small, the kernel function becomes sharp. If h becomes large, the kernel function becomes smooth. The types of kernel functions and the bandwidth h are important factors that determine the accuracy of estimated kernel density functions. An inappropriate bandwidth yields under-smoothing or over-smoothing, and thus, an optimum bandwidth must be determined (Wand and Jones 1994). In this study, a second-order Gaussian kernel

Table 1 Various statistical modeling methods

	Probabilistic approach	Interval approach
Parametric approach	Goodness-of-fit (GOF) test, Model selection method, Sequential statistical modeling (SSM)	Probability bounds approach (p-box)
Nonparametric approach	Goodness-of-fit (GOF) test,	Dempster-Shafer theory,
	Kernel density estimation (KDE)	Interval representation
	*Kernel density estimation with bounded data (KDE-bd)	
	*Kernel density estimation with estimated bounded data (KDE-ebd)	

*Proposed methods in this study

function that has symmetric and non-negative kernels is used to obtain estimated kernel density functions because its mathematical formula is simple, and it is the most commonly used in the kernel density estimation (Chen 2015; Hansen 2009; Sheather 2004; Guidoum 2015). Moreover, since the bandwidth mostly affects the accuracy of the estimated kernel density functions more than any type of kernel function (Silverman 1986; Wand and Jones 1994; Chen 2015; Guidoum 2015), the Gaussian kernel makes it simple to calculate the optimum bandwidth and is preferred to other types of kernel functions (Silverman 1986; Wand and Jones 1994; Chen 2015). The original and rescaled Gaussian kernel function is expressed as

$$K(t) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}t^2\right\} \tag{3}$$

$$K_h(t) = \frac{1}{\sqrt{2\pi h}} \exp\left\{-\frac{1}{2}\left(\frac{t}{h}\right)^2\right\} \tag{4}$$

The optimal bandwidth (h^*) of the Gaussian kernel is calculated using Silverman's rule of thumb as follows (Silverman 1986).

$$h^* = \left(\frac{4}{3}\right)^{1/5} \hat{\sigma} n^{-1/5} \tag{5}$$

where n is the number of data samples, and $\hat{\sigma}$ is the estimated standard deviation of the data. Silverman's rule of thumb is sensitive to outliers, so that sample standard deviation, which is defined as the square root of variance of data values from its mean, is not appropriate for use. In this study, a corrected standard deviation, which is known as more robust (Analytical Methods Committee 1989), was used as follows.

$$\hat{\sigma} = \frac{Median(|X_i - Median(X_i)|)}{0.6745} \tag{6}$$

The Silverman's rule of thumb is describe in detail in Appendix 1.

2.3 Maximum likelihood estimation

The MLE is the most commonly used method to estimate statistical parameters of a distribution function using the concept of likelihood. The likelihood of a set of data is the probability of acquiring samples given the selected probability distribution functions. This method calculates the parameters of a chosen probability distribution model by maximizing the value of the likelihood function. The likelihood function is calculated by multiplying each probability density function by each sample. For convenience, the log-likelihood function is the most commonly used and is expressed as

$$\ln L\left(\theta; X_{1,\ldots,}X_n\right) = \sum_{i=1}^{n} \ln f(X_i|\theta) \tag{7}$$

where $f(\cdot)$ is a probability density function of the chosen distribution models and θ is the parameter vector.

Estimated parameters using the MLE method are expressed as

$$\left\{\hat{\theta}_{mle}\right\} \subseteq \left\{\underset{\theta\in\Theta}{\operatorname{argmax}}\, \ell(\theta; X_i)\right\} \tag{8}$$

where $\hat{\theta}_{mle}$ is the estimated parameter and ℓ is the log-likelihood function.

The MLE method can also be used to identify a probability distribution function, as well as to estimate parameters. Statistical modeling using the MLE method first calculates the maximum likelihood function values for all candidate models, and then a distribution model that has the largest value of likelihood is chosen as the most appropriate model.

3 Kernel density estimation with and without bounded data

A kernel density function can represent any type of distribution, but sometimes may have a very irregular shape especially for extremely small data, e.g., less than 10, which often occurs in real engineering fields. In addition, the distribution types and input variable parameters are usually unknown, and only the lower and upper bounds are known, which could be very important information, especially when there are few data. However, the original KDE cannot use the bound information because it only uses data to generate an estimated kernel density function.

In addition to the original KDE, there are various types of KDE that use bounded support such as the reflection method (Schuster 1985), boundary kernel method (Gasser and Müller 1979), pseudo data method (Cowling and Hall 1996), transformation method (Marron and Ruppert 1994), and generalized reflection method (Karunamuni and Alberts 2005a, 2005b; Karunamuni and Zhang 2008). These methods use bounded support to define the domain of a true density function and then generates a kernel density function only within the defined domain (Karunamuni and Zhang 2008). These methods have been developed to prevent long tailed density functions being distributed in the incorrect domains, and they are classified based on the manner in which their kernel density functions are generated. The KDE with the bounded support is intended

to allow the kernel density functions to be defined within the domain of input variables, but not for improving accuracy or ensuring conservativeness of modeling data distribution. If the bounded support region is similar to the data distribution region, the kernel density function may have very high densities near the bounded support, which can make the density function significantly differ from the true density function. The KDE with wide bounded support behaves like the original KDE and the one with narrow bounded support cannot be used to accurately model the data distribution. Thus, a new KDE needs to be developed to accurately and conservatively model the true density function for insufficient data by using both the data and bound information of input variables.

In this study, the KDE-bd method that combines the nonparametric-probabilistic approach (KDE) and the interval representation using input variable intervals is proposed. This method can be categorized as the KDE-bd method and the KDE-ebd method based on whether the input variable intervals are given or estimated from the given data. That is, if the intervals are known, the KDE-bd is used; otherwise, KDE-ebd is used. The proposed KDE-bd/ebd uses the boundary information of the data to make the kernel density function to be well fit to the data distribution as well as yielding conservative density functions, and thus, the estimated density function can be more accurate than KDE with bounded support in the entire domain and more robust than the original KDE. Detailed explanations of the KDE-bd and KDE-ebd are given in Section 3.1.

3.1 KDE-bd and KDE-ebd process

The KDE-bd randomly generates bounded data from a uniform distribution with given lower and upper input variable bounds and the KDE-ebd does the same but with the estimated lower and upper input variable bounds using given data. In the KDE-bd and KDE-ebd, the bounded data are added to the original data, and then the total data are used to generate estimated kernel density functions. The kernel density function is obtained by summing kernel functions generated on each sample of total data, where the optimum bandwidth (h^*) is calculated from the total data, unlike the bandwidth of the original KDE which does so only from the original data. The estimated kernel density function of KDE-bd and KDE-ebd is defined as

$$\hat{f}_h = \frac{1}{(n+m)\cdot h}\sum_{k=1}^{n+m} K\left(\frac{x-(XBD_k)}{h}\right) \tag{9}$$

where XBD_k is the total data for $k = 1,\ldots,n+m$, $XBD_k = \{X_i, BD_j\}$, X_i is the i^{th} given data for $i = 1,..,n$, and BD_j is the j^{th} bounded data for $j = 1,\ldots, m$.

Figure 1 shows the KDE-bd and KDE-ebd process. In Step (1), if the lower and upper bounds of the input variables, bd_L and bd_U, are given, they are selected as the two parameters, l and u, of a uniform distribution; otherwise, estimated parameters and their lower and upper bounds are calculated using MLE corresponding to the significance level, and thus, $\hat{a}$ and $\hat{b}$ become a minimum and maximum value of given data, respectively.

The point estimators, $\hat{a}$ and $\hat{b}$, are expressed as

$$\left\{\hat{a},\hat{b}\right\} \subseteq \left\{ \underset{\hat{a},\hat{b}\in\Theta}{\operatorname{argmax}} \left(\frac{1}{x_u - x_l}\right)^n \prod_{i=1}^{n} I_{\{x_l \le x \le x_u\}}(x_i) \right\} \tag{10}$$

where $I_{\{x_l \le x \le x_u\}}$ is an indicator function which has 1 if $x_l \le x \le x_u$ and has 0 otherwise, and the interval

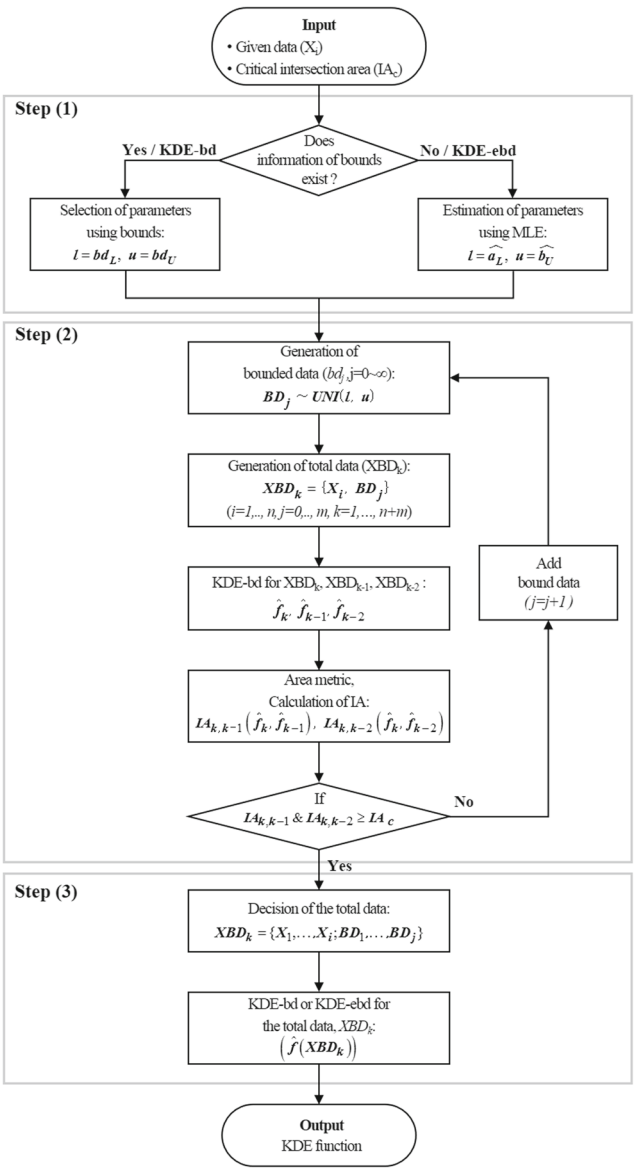

Fig. 1 KDE-bd and KDE-ebd process

estimators of $\hat{a}$ and $\hat{b}$ with a significance level (*ACI* and *BCI*) are expressed as

$$ACI = \{\hat{a}_L, \hat{a}_U\} = \left[\hat{b} - \frac{\hat{b}-\hat{a}}{\alpha^{1/n}}, \hat{a}\right] \tag{11}$$

$$BCI = \left\{\hat{b}_L, \hat{b}_U\right\} = \left[\hat{b}, \hat{a} + \frac{\hat{b}-\hat{a}}{\alpha^{1/n}}\right] \tag{12}$$

where the estimated parameters, l and u, become $\hat{a}_L$ and $\hat{b}_U$ respectively; n is the number of given data; α is the significance level.

If the significance level α is low, the range of the boundary region is wide and the estimated density function has a long tail, resulting in conservative results in reliability analysis. If the significance level is high, the range of the boundary region is narrow and the estimated density function has a short tail, resulting in less conservative results compared to those using a low significance level. When the number of data is very small, the parameters of the uniform distribution, $\left[\hat{a}, \hat{b}\right]$, can provide a very narrow range of bounded data if the data are densely distributed. Thus, the bounded data can be generated near the mode or mean, resulting in a narrower tail of the density function than the original KDE, which can produce unconservative results in reliability analysis. However, if the confidence intervals of the estimated parameters of the uniform distribution, $\left[\hat{a}_L, \hat{b}_U\right]$, are used to determine the boundary region for generating bounded data, the bounded data is widely distributed and the tail of the density function is thick and long. Accordingly, more conservative reliability results can be obtained. If the users need a density function more fit to the original data rather than the conservative density function, they can use a higher significance level to narrow the boundary region for sampling bounded data. However, a higher significance level does not necessarily guarantee accurate estimation of the density functions. A higher significance level generally yields more accurate density estimation than the lower significance level, but it may decrease the accuracy of estimating the density function because the estimated bounds could be too narrow especially for insufficient data.

In Step (2), bounded data are randomly generated from a uniform distribution with l and u. If the number of the original data is one or two, the initial number of bounded data is two or one, respectively. This is because the total number of data should be greater than or equal to three, due to the calculation of the intersection area for k-2, k-1, and k^{th} data; otherwise, the initial number of the bounded data is zero. After adding each bounded data to the original data, the kernel density functions f_k, f_{k-1}, and f_{k-2} are generated, where i is the number of given data, j is the added number of bounded data, and k is the number of total data. Whenever the bounded data are added, the kernel density function is generated and the intersection areas, $IA_{k,k-1}$ and $IA_{k,k-2}$, between the updated kernel function (f_k), k-1th, and k-2th kernel density function (f_{k-1} and f_{k-2}) are calculated. If all intersection areas, $IA_{k,k-1}$ and $IA_{k,k-2}$, are larger than the critical intersection area IA_c, then the additional bounded data do not affect the density estimation; thus, additional bounded data will not be generated to estimate the density function and Step (2) is terminated. Figure 2 shows how the final number of bounded data is determined. As in Fig. 2, if f_k (dotted line), f_{k-1} (solid line), and f_{k-2} (dash-dot line) are similar, the bounded data are not updated, and thus, Step (2) moves to Step (3). In Step (3), an estimated kernel density function is finally obtained using the final bounded and given data. An appropriate amount of bounded data can increase the intersection areas, but too much or too little bounded data can decrease the intersection areas. Therefore, it is necessary to determine the number of bounded data necessary to create an accurate density function model.

In this paper, f_k is compared with f_{k-1} and f_{k-2} by calculating $IA_{k,k-1}$ and $IA_{k,k-2}$, but it is also possible to compare f_k with f_{k-1} and f_{k-1} with f_{k-2} by calculating $IA_{k,k-1}$ and $IA_{k-1,k-2}$. However, since the updated density functions, f_k and f_{k-1}, are only compared with the previously updated density functions, f_{k-1} and f_{k-2}, respectively, f_k could satisfy the convergence criterion even if it does not tend to converge to the finally updated function obtained from original data and sufficient bounded data. If f_k and f_{k-1} are only compared, f_k could quickly satisfy the convergence criterion even though f_k is somewhat different from the finally updated function. Thus, the number of bounded data may not be enough to represent the conservativeness of the density function using KDE-bd/ebd. If f_k is compared with many previously updated density functions, f_{k-1}, f_{k-2}, f_{k-3} and so on, then it is difficult to satisfy the convergence criterion and it may require a large number of bounded data. Thus,

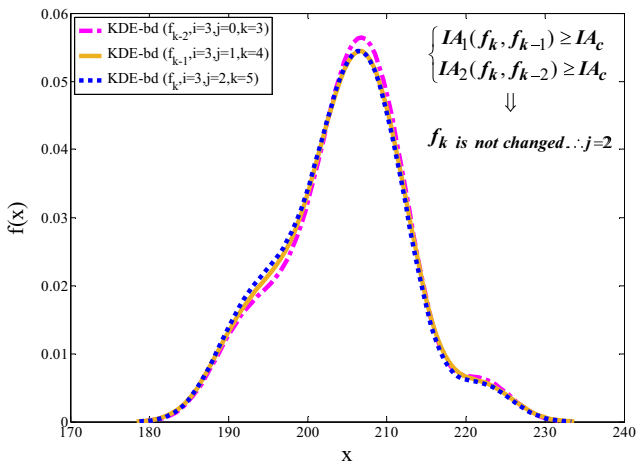

Fig. 2 Updated kernel density functions

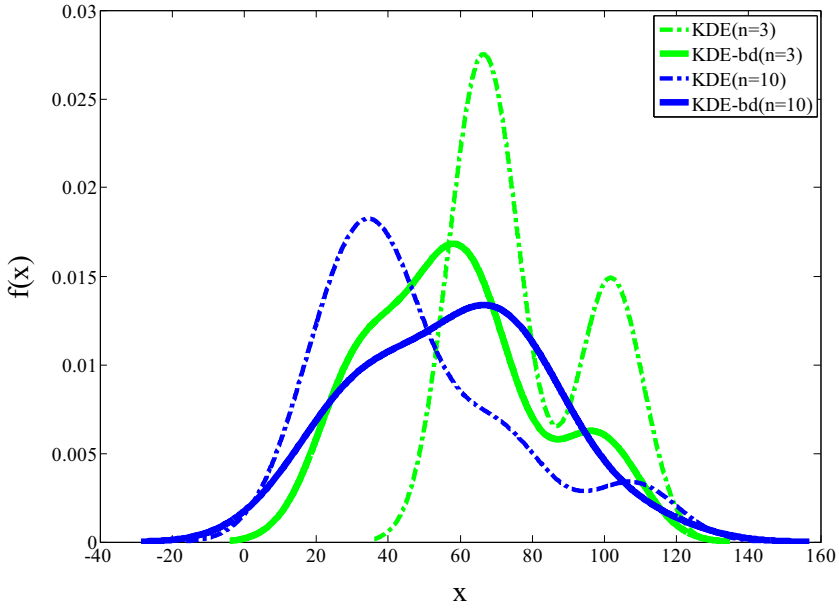

Fig. 3 KDE functions as KDE and KDE-bd

in this paper, f_k, f_{k-1}, f_{k-2} were only compared, and $IA_{k,k-1}$ and $IA_{k,k-2}$ were used to determine the reasonable number of bounded data.

Figure 3 shows the effect of the bounded data on the estimation of the kernel density functions. The kernel density functions using KDE (dash-dot line) are very irregular for small data samples such as $n = 3$ and $n = 10$, while density functions using KDE-bd (solid line) are smooth, due to the effect of the bounded data. When the number of original data is very small, since the estimated kernel density functions are sensitive to the additional data, a large number of the bounded data are required in the KDE-bd and KDE-ebd calculation process. However, as the number of original data increases, less bounded data is required, and then, the kernel density functions estimated from the total data finally converge to specific density functions and the bounded data is no longer required.

Figure 4 shows how to calculate the intersection areas, $IA_{k,\ k-1}$ and $IA_{k,\ k-2}$. The intersection areas are calculated using an area metric method to evaluate the coincidence rate between the two kernel density functions being compared. In this method, the intersection areas are obtained by calculating the overlapped area of the two kernel density functions, which can range from

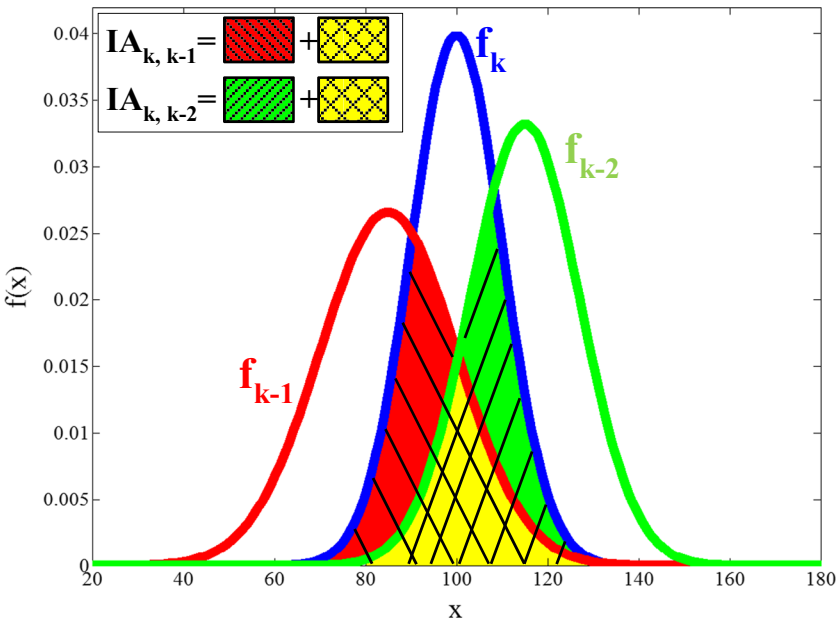

Fig. 4 Calculation of intersection areas

0 to 1. If two kernel density functions coincide, the intersection area is 1. If the two functions do not overlap, the intersection area is 0. First, an overlapped function ($f_{k,k-1}$) is defined from the two kernel density functions, f_k and f_{k-1}. Second, the closed interval [a, b] is uniformly divided into p subintervals where [a, b] encompasses the domain of the overlapped area, and then, the overlapped subinterval ($a = x_o <, \ldots, < x_p = b$) along the x-axis of the two kernel functions is obtained.

Finally, $IA_{k,k-1}$ and $IA_{k,k-2}$ are calculated using the Riemann integral of the overlapped function (Kang et al. 2016). The intersection area $IA_{k,k-1}$ is defined as (Jung et al. 2017)

$$IA_{k,k-1} = \sum_{l=1}^{p} f_{k,k-1}(x_l) \cdot (x_l - x_{l-1}) \tag{13}$$

$$f_{k,k-1}(x_l) = \min\{f_k(x_l), f_{k-1}(x_l)\} \tag{14}$$

4 Statistical simulation test

In the simulation tests, Birnbaum-Saunders (BS), generalized extreme value (GEV), log-normal (LOGN), logistic (LOG), normal (NORM), Rayleigh (RAY), and Weibull (WBL), which have different numbers of parameters and distinct shapes, are assumed as true models. Figure 5 shows the PDFs of the true models and the numbers next to the distribution name indicate the distribution parameters. In Fig. 5, the WBL distribution has a very narrow range over the random variable X and a large peak density, which makes it difficult to distinguish from other distributions, and thus, a larger size of other PDFs, except WBL distribution, is shown in the upper right hand corner of Fig. 5.

For the simulation tests, the KDE-bd requires random variable intervals that mostly cover the domain of the distribution. In general, the lower and upper bounds of the intervals are determined through extensive experience or accumulated knowledge of engineers or companies through fieldwork. In this study, the intervals are defined as confidence intervals according to the Guide to the Expression of Uncertainty in Measurement (GUM) published by ISO. The GUM introduces the expression of uncertainty as a measurement, and it recommends that the uncertainty of a measurement be expressed simply as $Y = y \pm U$, where Y is the estimated value attributable to the measurand, y is the mean value of the test, and U is the expanded uncertainty related to y (Gabauer 2000). This expression means confidence intervals at some probability level ranging from $y - U$ to $y + U$. A confidence interval of 95% is generally used (Gabauer 2000; Cox and Harris 2003).

In this study, the lower and upper values for the bounds are calculated from the 2.5 and 97.5-percentiles of the cumulative distribution function (CDF) of true models. Table 2 shows the lower and upper bounds corresponding to the 2.5 and 97.5-percentiles based on distribution types. If the bounds of X are known, the bounded data for the KDE-bd are expressed as

$$Bounded\ data = \left[X_{L|F(x)=0.025} \le BD_j \le X_{U|F(x)=0.975}\right] \tag{15}$$

If they are unknown, the estimated bounded data for the KDE-ebd are expressed as

$$Estimated\ Bounded\ data = \left[\max(X_i) - \frac{\max(X_i) - \min(X_i)}{\alpha^{1/n}} \le BD_j \le \min(X_i) + \frac{\max(X_i) - \min(X_i)}{\alpha^{1/n}}\right] \tag{16}$$

where X_i is the given data, and BD_j is the bounded data that are randomly generated from the given or estimated bounds for a uniform distribution based on given data sample used in the MLE.

To verify the performance of KDE-bd and KDE-ebd, each method is compared to the original KDE by randomly generating samples $n = 1$, 3, 5, 7, 10, 20, 30, and 50 with 1000 repetitions and by separately considering two cases, Case I: the given data with given bounds and Case II: the given data without given bounds. The generated samples are used to generate the kernel density functions using KDE, KDE-bd, and KDE-ebd, and then the intersection areas between the estimated kernel density functions and true PDFs are calculated to estimate accuracies of three methods.

4.1 Case I: the given data with information of bounds

In this section, the accuracy of the original KDE and KDE-bd are compared using intersection areas where both experimental data and input variable intervals are given. Tables 3, 4, and 5 show the average intersection areas between the estimated kernel density functions and the various true PDFs for the various numbers of sampled data when the true models are NORM, RAY, and GEV distributions. If the critical intersection area, IA_c, is too high, a large number of bounded data are required and the estimated kernel density functions using the bounded data become too smooth, and thus, their intersection areas could be too low. On the other hand, if IA_c is too low, a small number of bounded data are required and the estimated

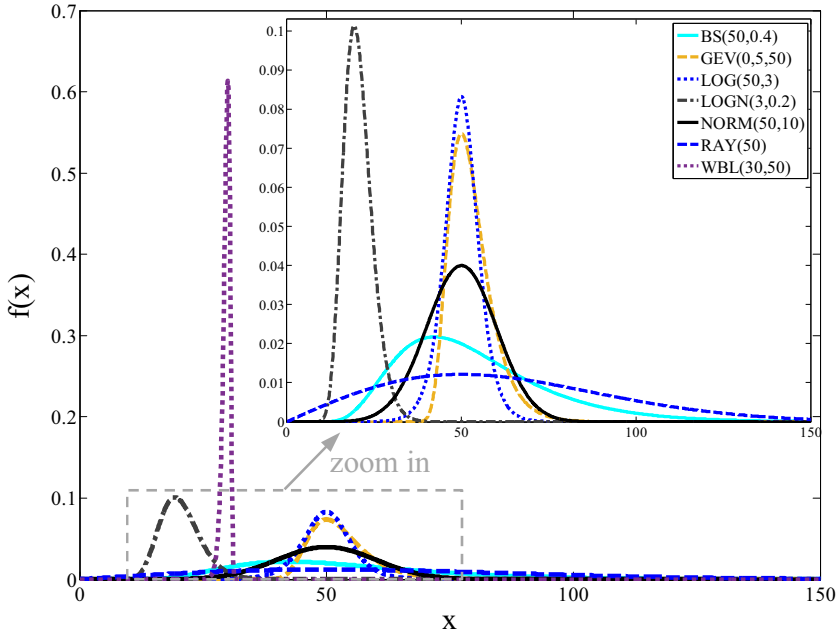

Fig. 5 PDFs of true distributions

kernel density functions become nonlinear, and thus, their intersection areas could be too low. In this study, IA_c was chosen as 0.95, which is known as a reasonable convergence criterion for the area metric (Jung et al. 2017). In Tables 3, 4 and 5, *BDn* indicates the average number of bounded data generated from the given confidence intervals for a critical intersection area of 0.95. All of the numbers in KDE, and *IA*s using KDE-bd, indicate the intersection areas between the true PDF and estimated kernel density function. The bold font indicates that the intersection areas calculated from the KDE-bd are larger than those from the original KDE. The underlined values indicate the cases that the intersections calculated from the KDE-bd and the original KDE are same.

First, the true model is assumed to be a NORM distribution. As shown in Table 3, as n increases, the intersection areas using the original KDE and KDE-bd increase and become close to one. The intersection areas using the KDE-bd are always larger than those using the KDE when $n \leq 30$. The intersection areas using the KDE increase considerably as n increases when $n \leq 20$, since the estimated kernel density functions are much affected by the additional data. The intersection areas increase slightly when $n > 20$ because the estimated density functions converge into a true density function and are less affected by the additional number of data. The intersection areas using KDE-bd also increase as n increases, but the rate of increase is slower than when the KDE was used, and the *BDn* decreases slightly until $n \leq 10$. Then, the *BDn* rapidly decreases from $n = 7$ to 20 because the estimated functions become quite robust for additional bounded data, and then, the *BDn* finally converges to zero at $n = 50$, where the intersection areas using KDE-bd become the same as those using the original KDE.

Figure 6 shows the change of the estimated density functions as more bounded data are added to the given data when

Table 2 Bounds of true models

True models	Lower ($X_{L \mid F(x)=0.025}$)	Upper ($X_{U \mid F(x)=0.975}$)
BS (50,0.4)	23.2575	107.2489
GEV (0,5,50)	43.5077	68.4101
LOG (50,3)	39.0240	61.0138
LOGN (3,0.2)	13.5776	29.7307
NORM (50,10)	30.4609	69.5776
RAY (50)	11.2470	135.7597
WBL (30,50)	27.8742	30.7940

Table 3 Comparison of intersection areas in KDE and KDE-bd: $X \sim$ NORM (50, 10)

n	KDE	KDE-bd	
		IA	*BDn*
1	–	**0.7794**	12.418
3	0.5845	**0.7946**	10.612
5	0.7142	**0.8083**	8.742
7	0.7699	**0.8153**	7.055
10	0.8104	**0.8305**	4.741
20	0.8602	**0.8627**	0.714
30	0.8844	**0.8846**	0.087
50	0.9047	0.9047	0

The bold font indicates that the KDE-bd or KDE-ebd is more accurate than the original KDE. The underlined entry indicates that the KDE-bd or KDE-ebd has the same accuracy to the KDE. The italicized entry indicates that the KDE-bd or KDE-ebd is less accurate than the KDE. The meaning of each entry is explained in the paper

Table 4 Comparison of intersection areas in KDE and KDE-bd: $X \sim$ RAY(50)

n	KDE	KDE-bd	
		IA	*BDn*
1	–	**0.7506**	12.375
3	0.5840	**0.7650**	10.656
5	0.7031	**0.7782**	8.709
7	0.7552	**0.7895**	7.030
10	0.8034	**0.8200**	4.848
20	0.8518	**0.8541**	0.806
30	0.8750	**0.8752**	0.093
50	0.8961	0.8961	0.002

The bold font indicates that the KDE-bd or KDE-ebd is more accurate than the original KDE. The underlined entry indicates that the KDE-bd or KDE-ebd has the same accuracy to the KDE. The italicized entry indicates that the KDE-bd or KDE-ebd is less accurate than the KDE. The meaning of each entry is explained in the paper

the true model is the NORM distribution with $n = 3$. In Fig. 6, the right legend depicts the density function of population (f_{POP}) and the updated kernel density functions (f_{k-5}~f_k) using the data from (k-5) to (k), and the left legend shows the combined given and added bounded data from (k-5) to (k). As the bounded data are added, the updated kernel density functions become smoother and tends to converge to the true density function, and $IA_{k,k-1}$ and $IA_{k,k-2}$ become larger than IA_c. Finally, the bounded data increases up to k^{th} data ($k = 16$) and the final density function is estimated using (k) bounded data by combining the original given data, $n = 3$.

When the number of data is only one, the KDE-bd is very similar to the true population using the bounds, while the KDE yields a very inaccurate density function using only one data.

Table 5 Comparison of intersection areas in KDE and KDE-bd: $X \sim$ GEV(0, 5, 50)

n	KDE	KDE-bd	
		IA	*BDn*
1	–	**0.6920**	12.501
3	0.5659	**0.7131**	10.567
5	0.6974	**0.7324**	8.891
7	0.7550	**0.7629**	7.303
10	0.7921	**0.7956**	5.232
20	0.8452	**0.8457**	0.873
30	0.8685	**0.8687**	0.121
50	0.8886	0.8886	0

The bold font indicates that the KDE-bd or KDE-ebd is more accurate than the original KDE. The underlined entry indicates that the KDE-bd or KDE-ebd has the same accuracy to the KDE. The italicized entry indicates that the KDE-bd or KDE-ebd is less accurate than the KDE. The meaning of each entry is explained in the paper

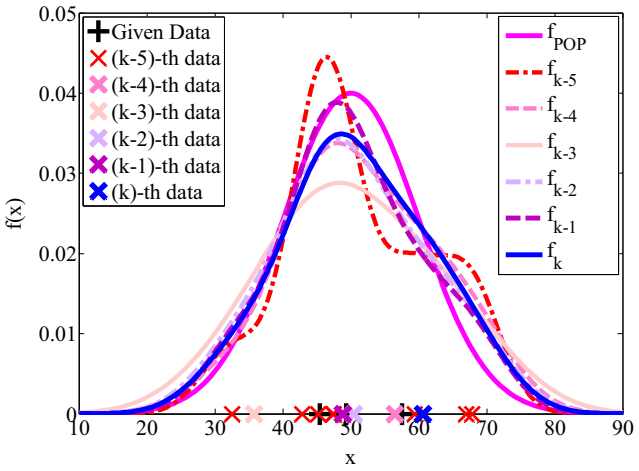

Fig. 6 Estimated density functions according to more bounded data with $n = 3$

Because its accuracy is too low for $n = 1$, the results using KDE are not shown in this paper. Figure 7 shows the estimated kernel density functions using KDE and KDE-bd when $n = 1$. The estimated density functions using KDE have very narrow density function shapes, and thus, the intersection area between the estimated density function (KDE in Fig. 7) and the true PDF (POP in Fig. 7) is very small. However, the estimated density function using the KDE-bd is similar to the true PDF shape when compared to the one using KDE, due to the additional bounded data.

Second, the RAY distribution is assumed to be the true model. The intersection areas using KDE and KDE-bd increase as n increases, while the *BDn* in KDE-bd decreases as in the NORM distribution, as shown in Table 4. However, the overall estimation accuracies using both methods for the RAY distribution are lower than those for the NORM distribution. This is because both methods use the Gaussian kernels and Silverman's rule of thumb, which uses a normal assumption for the true density function, and thus, the estimated functions fit well into symmetric distributions. In other words, since the

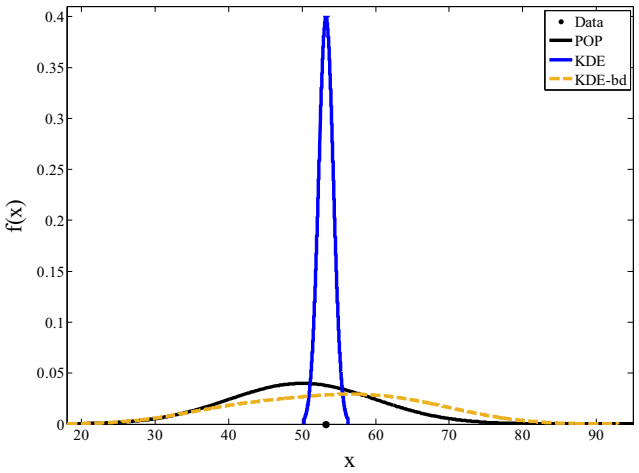

Fig. 7 Kernel density functions with $n = 1$

RAY distributions are skewed, the estimated density functions are less accurate. For skewed distributions, different kernels and bandwidth selection methods need to be used to calculate appropriate optimal bandwidth for accurately modeling the true density function. In addition to the Silverman's rule of thumb, the optimal bandwidth can be calculated by minimizing the objective function, which measures the error between the true density function and the KDE density function, using mean integrated squared error (MISE) and asymptotic MISE (AMISE) or modified functions of these measures. Most bandwidth selection methods such as biased cross-validation (BCV), unbiased cross-validation (UCV), complete cross-validation (CCV), and direct plug-in (DPI) uses the derivatives of the kernel functions to calculate the optimal bandwidth, and they are classified on the manner in which the true density functions are defined (Wand and Jones 1994; Scott and Terrell 1987; Hardle et al. 1990; Jones and Kappenman 1992; Sheather and Jones 1991).

Depending on the shape of the true density function, it is necessary to use the most appropriate kernel functions and bandwidth selection. However, the true density function is usually unknown; thereby making it difficult to select appropriate kernel types and bandwidth. Among various kernel types, the reason for using the Gaussian is because the Gaussian kernel can be used to easily calculate the bandwidth and the most widely used in KDE. In addition, it can be widely used to model various shapes of distributions without knowing the true model. Even though the Gaussian kernel assumes that the true model is the normal distribution, kernel functions still can approximate the data distribution through summation of kernel function values evaluated at each data and the effect of the kernel function decreases with increasing the number of data. Even though the modeling accuracy of KDE slightly decreases for the skewed distribution, the KDE-bd is still more accurate than the original KDE in comparison with intersection areas using the KDE.

Third, the GEV distribution is assumed to be the true model. As n increases, a similar tendency is observed as in the NORM and RAY distributions. The estimation accuracies using both methods are the lowest out of the three true distributions because the GEV distribution has very high skewness and kurtosis. However, the obtained intersection areas are not much different from those in the NORM or RAY distributions, and thus, KDE-bd is still more accurate and recommended over the KDE.

4.2 Case II: the given data without information of bounds

In this section, the accuracy of KDE and KDE-ebd are compared through the intersection areas when the intervals are unknown and only the experimental data are given. Tables 6, 7 and 8 show the average intersection areas when NORM, RAY, and GEV distributions are based on various numbers of data with 1000 repetitions, where IA_c is 0.95, same to Case I. The estimated intervals are used as uniform distribution parameters with $\alpha = 0.1$ to compromise the accuracy and conservativeness of the estimated density functions, and they are used to define the lower and upper bounds of the bounded data and their average values are presented in Tables 6, 7 and 8. The intersection areas of KDE-ebd are calculated for $n \geq 3$ since the estimated confidence intervals can be obtained for $n \geq 2$.

Table 6 Comparison of intersection areas in KDE and KDE-ebd: $X \sim$ NORM (50, 10)

n	KDE	KDE-ebd		
		IA	*BDn*	Estimated intervals [*l*, *u*]
1	–	–	–	–
3	0.5845	**0.6360**	10.376	[23.53, 76.53]
5	0.7142	**0.7181**	8.616	[25.57, 74.32]
7	0.7699	*0.7513*	7.176	[25.56, 74.04]
10	0.8104	*0.8007*	4.990	[26.49, 73.25]
20	0.8602	*0.8597*	0.756	[26.54, 73.38]
30	0.8844	*0.8842*	0.089	[26.15, 73.55]
50	0.9047	0.9047	0	[25.36, 74.58]

The bold font indicates that the KDE-bd or KDE-ebd is more accurate than the original KDE. The underlined entry indicates that the KDE-bd or KDE-ebd has the same accuracy to the KDE. The italicized entry indicates that the KDE-bd or KDE-ebd is less accurate than the KDE. The meaning of each entry is explained in the paper

First, when the NORM distribution is chosen as the true model, the intersection areas using KDE and KDE-ebd are shown in Table 6 and have the same intersection areas as the KDE, which are the same as in Case I. The intersection areas increase as n increases, while the *BDn* in KDE-ebd decreases, similar to Case I. The estimated intervals using KDE-ebd

Table 7 Comparison of intersection areas in KDE and KDE-ebd: $X \sim$ RAY(50)

n	KDE	KDE-ebd		
		IA	*BDn*	Estimated intervals [*l*, *u*]
1	–	–	–	–
3	0.5840	**0.6148**	10.446	[−25.58, 152.35]
5	0.7031	*0.6908*	8.878	[−17.03, 151.15]
7	0.7552	*0.7377*	7.193	[−10.055, 146.02]
10	0.8034	*0.7910*	5.033	[−6.27, 144.01]
20	0.8518	0.8519	0.838	[−0.07, 147.00]
30	0.8750	*0.8747*	0.108	[1.23, 150.27]
50	0.8961	0.8961	0.004	[2.23, 155.43]

The bold font indicates that the KDE-bd or KDE-ebd is more accurate than the original KDE. The underlined entry indicates that the KDE-bd or KDE-ebd has the same accuracy to the KDE. The italicized entry indicates that the KDE-bd or KDE-ebd is less accurate than the KDE. The meaning of each entry is explained in the paper

Table 8 Comparison of intersection areas in KDE and KDE-ebd: Population ~ GEV (0, 5, 50)

n	KDE	KDE-ebd		
		IA	*BDn*	Estimated intervals [*l*, *u*]
1	–	–	–	–
3	0.5659	**0.6150**	10.358	[35.96, 70.29]
5	0.6974	*0.6794*	8.766	[37.85, 69.49]
7	0.7550	*0.7256*	7.408	[39.41, 68.96]
10	0.7921	*0.7651*	5.317	[39.90, 69.81]
20	0.8452	*0.8425*	0.999	[40.93, 70.97]
30	0.8685	*0.8683*	0.127	[41.12, 72.15]
50	0.8886	0.8886	0	[41.29, 73.57]

The bold font indicates that the KDE-bd or KDE-ebd is more accurate than the original KDE. The underlined entry indicates that the KDE-bd or KDE-ebd has the same accuracy to the KDE. The italicized entry indicates that the KDE-bd or KDE-ebd is less accurate than the KDE. The meaning of each entry is explained in the paper

become narrow based on the increase in the number of data, unlike the intervals using KDE-bd, which do not vary with the number of data. The intersection areas using KDE-ebd are larger than those using KDE for $n \leq 5$, but they are slightly lower than those using KDE for $n > 5$, until $n = 30$ and they finally become the same for $n = 50$. The estimated intervals for the input variable X cover a wide range of the domain of X for a small amount of data, and thus, the KDE-ebd are more similar to the true model than the KDE. However, as the number of data increases, KDE fits the data distribution better than the KDE-ebd, meaning that the estimated model using KDE are more similar to the true model than the KDE-ebd.

Second, when the true model is the RAY distribution, the intersection areas using KDE and KDE-ebd are presented in Table 7. The intersection areas using KDE-ebd are larger than those using KDE when $n = 3$, but they become slightly smaller as n increases, until $n = 30$. Both results finally become the same when $n = 50$ similar to the results in the NORM distribution. Since the RAY distribution has a large variation shown in Fig. 5, the estimated intervals are wide. Similar to Case I, the RAY distribution is slightly skewed to the left and the intersection areas for RAY distributions are smaller than those of the NORM distribution.

Third, when the true model is the GEV distribution, the intersection areas using both methods are denoted in Table 8. A tendency similar to the NORM and RAY distribution is observed for the GEV distribution. Similar to Case 1, the GEV distribution has the lowest intersection areas among the three true models, owing to its nonlinear shape, but had the largest intersection with the RAY distribution. The KDE-ebd has better accuracy than the KDE when $n = 3$, but it has slightly lower accuracy until $n = 20$; however, both methods have similar accuracies when $n \geq 30$.

Consequently, to summarize the three methods, KDE-bd is the most accurate among the three methods in the NORM, RAY, and GEV distributions, based on the KDE-ebd and KDE calculated with all differing numbers of data, as long as the given intervals are appropriate enough to mostly cover the domain of the input variables. Even though the KDE-ebd has lower accuracy than KDE when $5 \leq n \leq 20$, its accuracy is similar with the KDE and it has a heavy tail in the kernel density function owing to a wide range of intervals, which makes the reliability analysis results using KDE-ebd more conservative than the KDE. In Section 5, it will be explained how density functions are estimated using KDE, KDE-bd, and KDE-ebd and how the estimated density functions yield reliability analysis results.

The KDE-bd and KDE-ebd were tested only for unimodal distributions, but they are also applicable for multimodal distributions like the original KDE. However, in order to model an accurate multimodal distribution, sufficient data is required and the KDE-bd/ebd becomes the same as the original KDE. If the data is very limited, it is difficult to distinguish whether the true model follows a unimodal or multimodal distribution from the data; thus, both original KDE and KDE-bd/ebd are difficult to accurately model multimodal distributions. However, when the true model is known as the multimodal distribution and given data are very insufficient, the KDE-bd/ebd might more smooth out the nonlinearity of the shape of the multimodal distribution compared to the original KDE. Nevertheless, it is still better to use the KDE-bd/ebd than the original KDE because the KDE-bd/ebd provides more conservative density estimation and yields more conservative statistical analysis results than the original KDE. If it is necessary to accurately model the data following the multimodal distribution, the KDE-ebd can be used by sampling the bounded data from non-uniform distributions that more fit to the original data than the uniform distribution. Likewise, if boundary information near the multiple modes is given, the KDE-bd also can provide an accurate density function even for insufficient data. The KDE-bd/ebd for modeling multimodal distributions will be tested in the future research.

4.3 Comparison of the three methods

To compare the three methods, statistical simulations were carried out for BS, LOG, LOGN, and WBL distributions as well as NORM, RAY, and GEV distributions. Figure 8 shows the average intersection areas between estimated kernel density functions and the seven true models after 1000 repetitions for $IA_c = 0.95$, using the KDE, KDE-ebd and KDE-bd. In Fig. 8, the upper left box in each row indicates the true models. The intersection areas are shown when $n = 1$, 3, 5, 7, 10, 20, 30, and 50, where a high amount of intersection indicates that the estimated

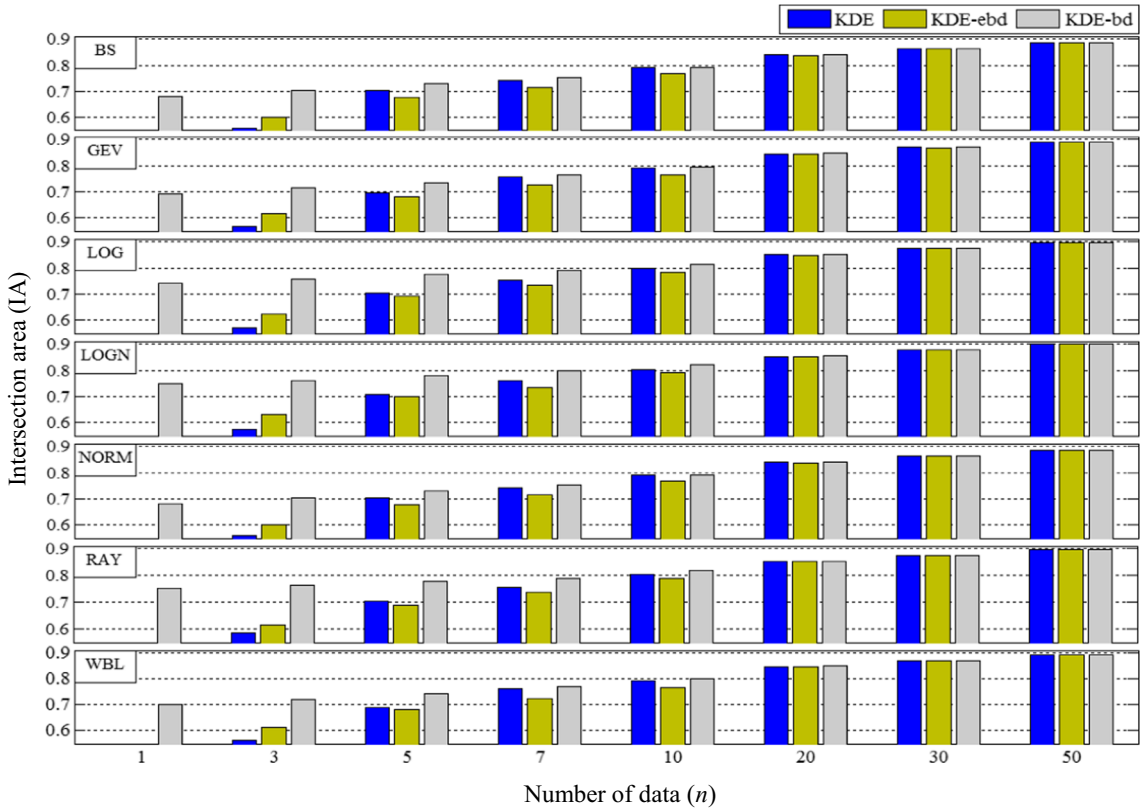

Fig. 8 Average intersection areas obtained using KDE, KDE-ebd, and KDE-bd

kernel density functions coincide well with the true PDFs. The average intersection areas using KDE-bd had the highest agreement with those using the true model, followed by KDE-ebd or KDE depending on the number of data. The KDE-bd is always the best method because the KDE-bd uses both given bounds and data for the input variables, whereas others only use data. Specifically, the KDE-bd can be only used when $n = 1$ and the KDE-bd and KDE-ebd are more accurate than the KDE for very small data, $n \leq 10$ and $n \leq 3$, respectively.

Comparing the accuracy of KDE-bd and KDE-ebd for different distribution types, the performance of the KDE-bd and KDE-ebd in the NORM, LOG, and LOGN distributions are superior to the others, since these models have low skewness and kurtosis, especially the NORM and LOG that have symmetricity (zero skewness). The KDE-bd and KDE-ebd have the lowest performance for the GEV distribution among all true models, since the GEV distribution has the highest skewness and kurtosis. Consequently, the KDE-bd and KDE-ebd are more accurate methods than the original KDE, especially for extremely small data. The KDE-bd can even deal with the special case of one data with given bounds of input variables.

Figure 9 depicts the range of intersection areas for the seven true distributions. The minimum, mean, and maximum intersection areas using three methods with 1000 repetitions are presented together. The mean values are expressed as triangle, cross, and circle markers for KDE, KDE-ebd and KDE-bd, respectively, and the minimum and maximum values are expressed as a horizontal bar. For all true distributions, the average intersection areas using the KDE-bd are much larger than those using KDE, and those using the KDE-ebd are slightly larger than KDE when $n \leq 3$ and smaller when $n \geq 5$. The ranges of the intersection areas using the KDE-bd are much narrower than those using the KDE. Those using KDE-ebd are slightly narrower than those of KDE when $n \leq 20$. The minimum intersection areas using the KDE-bd and KDE-ebd are larger than those using the KDE. Thus, the KDE-bd and KDE-ebd are robust methods regardless of the quality of data when compared to the original KDE. The KDE-bd is especially the most accurate and robust method if appropriate bounds are given. As mentioned in Section 4.2, for the GEV distribution, the KDE-ebd is slightly less accurate than the KDE, but the KDE-ebd is more robust regardless of data quality and also has heavier tails than the KDE, and thus, the KDE-ebd is superior to the KDE in terms of accuracy and robustness.

For skewed distributions such as GEV and RAY distributions, the average intersection areas using KDE-ebd are slightly lower than those using KDE. However, the minimum intersection area calculated using KDE-ebd is higher than that calculated using KDE, and the min and max range of the intersection areas using KDE-ebd is narrower than that using KDE. In other words, KDE-ebd is less sensitive to the number

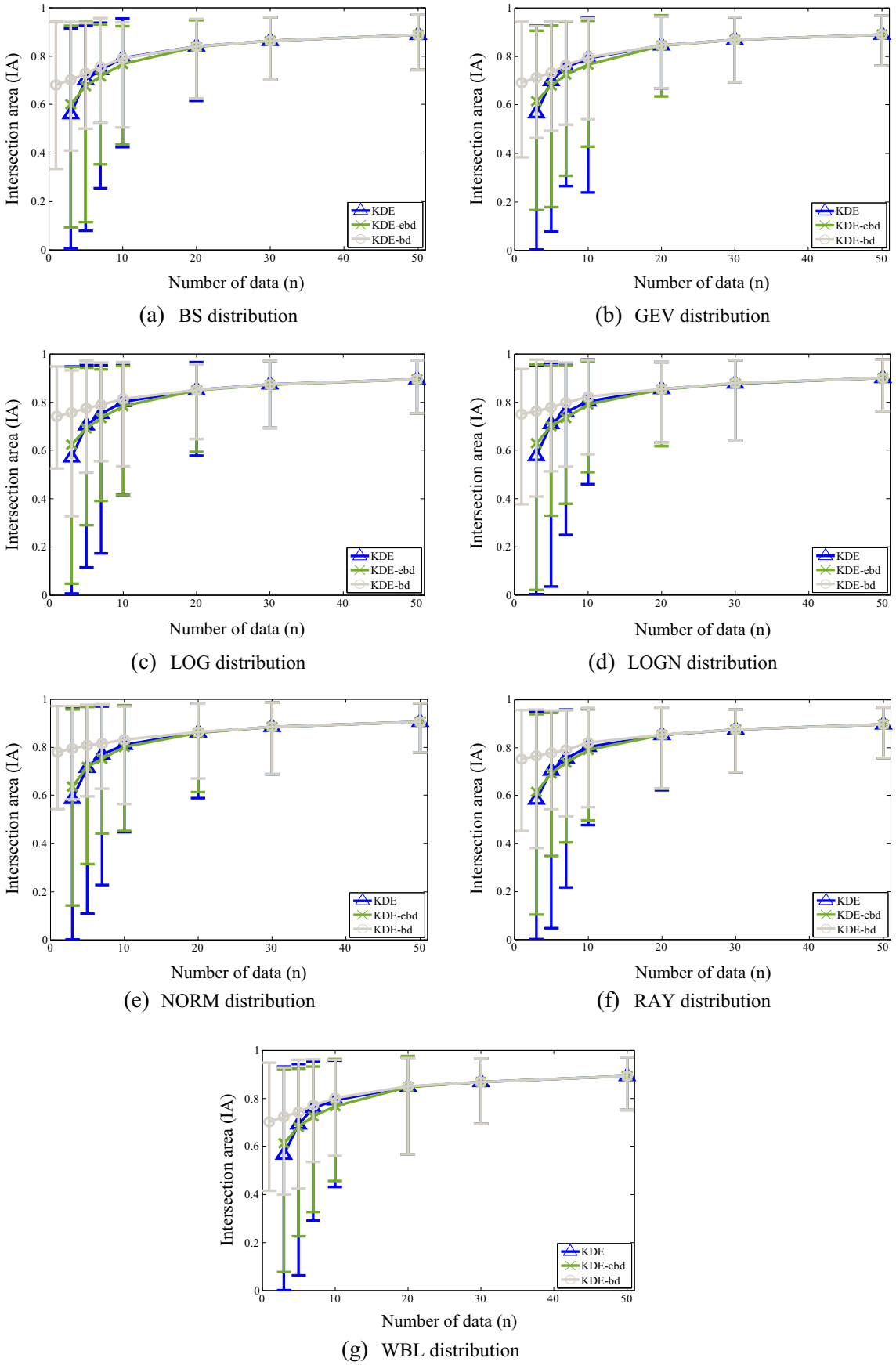

Fig. 9 Range of intersection areas according to sample size

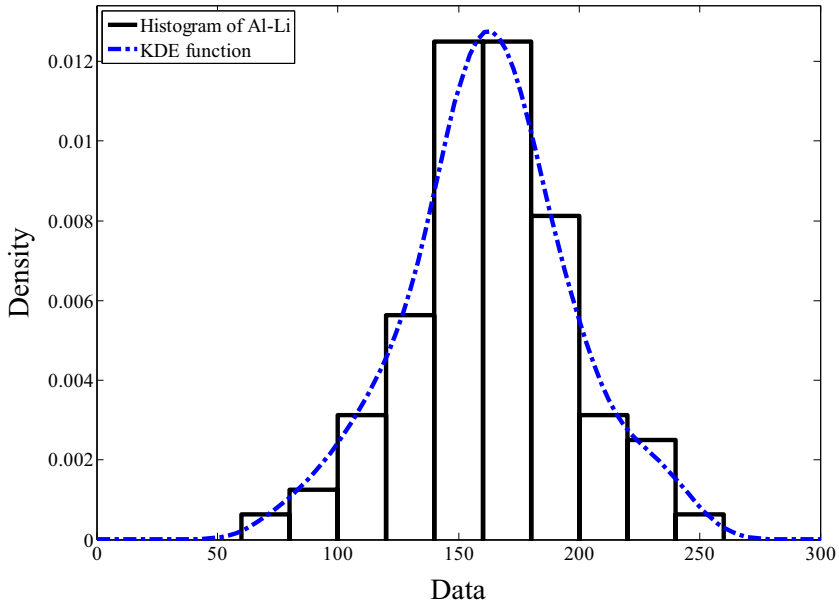

Fig. 10 Histogram and KDE function of 80 compressive strength

and quality of data compared to KDE. In general, the skewness of the data is not accurately estimated because the estimated density function using KDE-ebd has long tails due to the bounded data, which are randomly sampled from a uniform distribution. Instead, the long and thick tails of the density function could yield conservative results in reliability analysis and the accuracy of modeling the skewed distribution can be improved as the number of data increases. Because any statistical modeling methods have limitations in modeling the skewness of the distribution for insufficient data, more robust and conservative KDE-be/ebd is more suitable for modeling skewed distributions compared to the KDE.

5 Numerical examples

5.1 Compressive strength example

In this section, a numerical example is presented to demonstrate how the KDE, KDE-bd, and KDE-ebd methods are applied to model the density functions of 80 experimental data of the compressive strength of aluminum-lithium (Al-Li) alloy specimens (Montgomery and Runger 2003). In general, statistical models for input variables are unknown for real applications, and 80 data are not common. However, this example can be shown to compare and verify the statistical models using three methods by assuming that there are only a few data values, which are randomly generated from 80 data.

Figure 10 shows the histogram and estimated kernel density functions using KDE for 80 compressive strengths of Al-Li alloy which have a mean of 162.6625 and standard deviation of 33.7732, and is assumed as a population distribution in this example. For KDE-bd, the bounds of the compressive strength are defined by the assumed population, and it is the lower and upper bounds with 2.5-and 97.5-percentiles of CDFs of populations, 88.75 and 236.4, respectively.

To verify the performance of KDE-bd and KDE-ebd, both methods are compared with the original KDE by randomly sampling $n = 1, 3, 5, 7, 10, 20, 30, 50, 80$ from 80 total data with 1000 repetitions except $n = 80$, which was only repeated once. Based on the assumed true model, the intersection areas are calculated by comparing the estimated density functions with those obtained from $n = 80$. Table 9 shows average intersection areas between estimated kernel density functions and the true kernel density function for $IA_c = 0.95$. As n increases,

Table 9 Intersection areas according to sample size: compressive strength of Al-Li alloy

n	KDE	KDE-ebd		KDE-bd	
		IA	*BDn*	*IA*	*BDn*
1	–	–	–	**0.7510**	12.436
3	0.5934	**0.6459**	10.495	**0.7678**	10.627
5	0.7059	*0.7045*	8.881	**0.7854**	8.840
7	0.7707	*0.7536*	7.267	**0.8070**	7.501
10	0.8116	*0.7958*	5.523	**0.8288**	5.301
20	0.8740	*0.8730*	1.002	**0.8757**	0.918
30	0.9043	*0.9042*	0.152	0.9043	0.152
50	0.9402	0.9402	0	0.9402	0
80	1	1	1	1	0

The bold font indicates that the KDE-bd or KDE-ebd is more accurate than the original KDE. The underlined entry indicates that the KDE-bd or KDE-ebd has the same accuracy to the KDE. The italicized entry indicates that the KDE-bd or KDE-ebd is less accurate than the KDE. The meaning of each entry is explained in the paper

the intersection areas increase while the number of bounded data (*BDn*) decreases. Intersection areas using KDE-ebd are larger than those using KDE when $n \leq 3$, but smaller until $n = 20$, where both results become the same when $n \geq 30$. Intersection areas using KDE-bd are larger than KDE when $n \leq 20$ and then become the same when $n \geq 30$. Therefore, the KDE-bd method is the most accurate regardless of n, and KDE-ebd is better than the original KDE when $n \leq 3$. All methods yield similar results when $n \geq 30$. Since the density function obtained from 80 data entries is assumed as the true model, the intersection area becomes 1.0 when $n = 80$.

Figure 11 depicts the range of intersection areas. The range of the intersection areas using KDE-bd is the narrowest, followed by KDE-ebd and KDE. Although the average intersection area using KDE-ebd was slightly smaller than those from KDE, the ranges of the intersection areas using KDE-ebd are smaller than those using KDE. Accordingly, the density functions using KDE-ebd show a higher agreement with the population distribution at a higher rate than those using KDE. Although this example seems to be similar to the simulation tests, it is meaningful to apply the KDE-bd and KDE-ebd to real experimental data because the real experimental data may include measurement error, bias, or outliers. Since the obtained results in this example are similar to those in Section 4, the KDE-bd and KDE-ebd are still applicable to the real experimental data.

Figure 12 shows the original data, the population distribution function (POP), and the estimated kernel density functions using the three methods when $n = 3$. In Fig. 12, the KDE has a very irregular density function shape compared to KDE-ebd and KDE-bd, and the KDE-ebd has the widest density function shape due to the lower and upper bounds of the estimated confidence intervals. Thus, the KDE-bd is the most accurate and robust if there are given data with the bounds of the input variable. However, if the bounds of the input variable

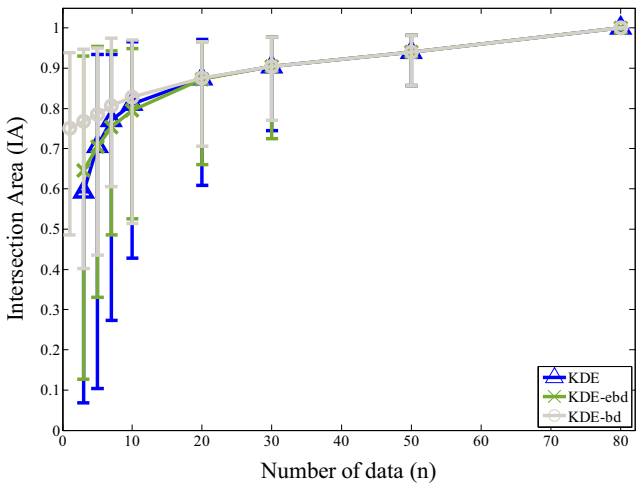

Fig. 11 Range of intersection areas of Al-Li alloy according to sample size

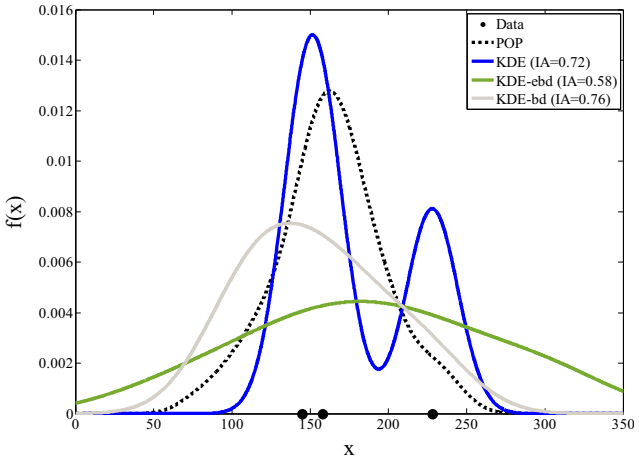

Fig. 12 Estimated KDE functions when $n = 3$

are unknown, the KDE-ebd needs to be used as an alternative method to KDE-bd. Even though the density function using KDE-ebd is different from the population distribution function, it is better than KDE because it has heavy tails and can show more conservative results in the reliability analysis than the KDE shown in Fig. 12.

5.2 A Cantilever example

In this section, a cantilever example was used to show how density functions using KDE, KDE-bd, and KDE-ebd affect reliability analysis results, shown in Fig. 13 (Eldred et al. 2007). The Young's modulus, geometric dimensions, and external loads of the beam are given as SI unit in Table 10.

The performance function for this problem $g_D(\mathbf{x})$ is defined as

$$g_D(\mathbf{x}) = \frac{4L^3}{Ewt}\sqrt{\left(\frac{Y}{t^2}\right)^2 + \left(\frac{X}{w^2}\right)^2} - D_0 \tag{17}$$

If the true model is used to carry out the reliability analysis, the exact probability of failure is 0.0693. If data are not given, it is common to use a uniform distribution. The two parameters for the uniform distribution are determined using

Table 10 Properties of geometric dimensions

Variables	Symbol	Value	Dist. type	COV
Length [m]	L	2.54	–	–
Width [m]	w	0.0635	–	–
Thickness [m]	t	0.0889	–	–
Displacement tolerance [m]	D_0	0.0546	–	–
Young's modulus [GPa]	E	13.885	NORM	0.05
Horizontal load [N]	X	2224.11	NORM	0.2
Vertical load [N]	Y	4448.22	NORM	0.1

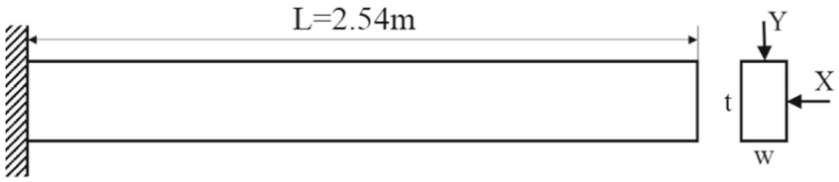

Fig. 13 A cantilever beam

confidence intervals with a confidence level of 95%, and a corresponding probability of failure calculated to be 0.1031. Figure 14 shows the ranges of the probabilities of failure using KDE, KDE-bd, KDE-ebd and a uniform distribution, where the parameters are determined using the lower and upper bounds of the estimated confidence intervals, and the horizontal dotted line that indicates the exact probability of failure. The boxplot depicts graphically grouped data for the quartiles of the data (Tukey 1977; Frigge et al. 1989). In the boxplot, the lower and upper bounds of the boxes indicate 1st quartile (Q_1) and 3rd quartile (Q_3), respectively. The centerlines of the boxes indicate the 2nd quartile (median), and the vertical dashed lines indicate the range of the data where the most data (over 98–99%) fall into the region bounded by $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$ where $IQR = Q_3 - Q_1$ and the point symbols indicate the outliers of the data.

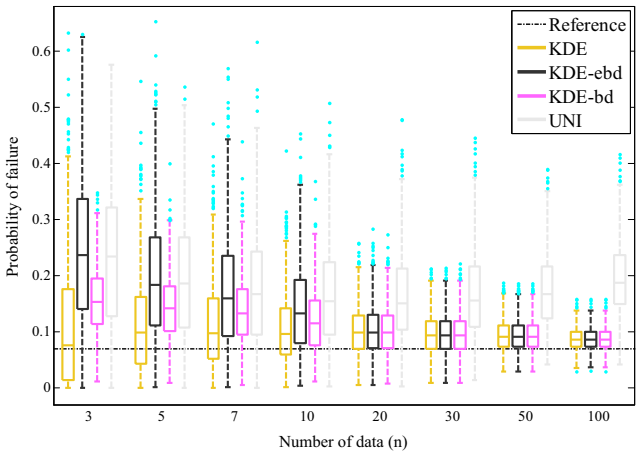

Fig. 14 Range of probability of failure according to the number of data

In Fig. 14, the average probabilities of failure using KDE converged most quickly to the exact probability failure, followed by the KDE-bd and KDE-ebd. Since the KDE-bd uses both the data and the bounds of the input variable, the ranges of the probabilities of failure are narrower than those using KDE. On the other hand, the probabilities of failure using the uniform distribution did not converge to the exact probability of failure even for a large number of data. This is because the lower and upper bounds of the estimated confidence intervals for the uniform distribution do not become narrow as the number of data increase and the data are widely spread. The KDE-ebd shows similar results to the uniform distribution for $n \leq 7$, but the ranges of the probabilities of failure converged to the exact probability of failure, while those using the uniform distribution did not. This is because the uniform distribution only uses the data to estimate its parameters whereas the KDE-ebd uses both the given data and bounded data to generate a density function. As the number of data increases, the effect of the given data on the estimation of the density function increases more than that of the bounded data and the density function becomes more similar with the true density function. In addition, the KDE-bd/ebd well describes the true density function using continuous and smooth kernel functions whereas the uniform distribution has a constant probability only on the interval and the CDF of the uniform distribution has discontinuous slopes at the lower and upper bounds of the interval. Accordingly, if the true model has a smooth and continuous probability curve, the uniform distribution will be rather different from the true model even if the number of data

Table 11 Percentiles of underestimating and overestimating the exact probability of failure

n		3	5	7	10	20	30	50	100
Methods									
KDE	Under	0.478	0.364	0.351	0.318	0.255	0.249	0.193	0.180
	Over	0.522	0.636	0.649	0.682	0.745	0.751	0.807	0.820
KDE-bd	Under	0.068	0.088	0.131	0.178	0.238	0.247	0.194	0.178
	Over	0.932	0.912	0.869	0.822	0.762	0.753	0.806	0.822
KDE-ebd	Under	0.117	0.124	0.171	0.194	0.238	0.246	0.195	0.179
	Over	0.883	0.876	0.829	0.806	0.762	0.754	0.805	0.821
UNI	Under	0.138	0.144	0.15	0.154	0.105	0.079	0.031	0.009
	Over	0.862	0.856	0.85	0.846	0.895	0.921	0.969	0.991

is enough. Since the KDE-bd/ebd can yield conservative density estimation for insufficient data and accurate destiny estimation for sufficient data, it will provide more desirable statistical analysis results than the uniform distribution.

Because of the heavy tail part of the density function estimated using the KDE-ebd, the KDE-ebd yields wide ranges of probabilities of failure. As the number of data increases, their ranges become narrower. It can be said that the KDE is better than the KDE-ebd in terms of the median values of the probabilities of failure, but the percentiles of underestimating the exact probability of failure using KDE-ebd are smaller than those using KDE, which means that KDE-bd and KDE-ebd have more conservative reliability analysis results than KDE.

Table 11 shows the percentiles of underestimating and overestimating the exact probability of failure for various numbers of data. The KDE underestimates the exact probability of failure with 47.8% accuracy, whereas the KDE-bd, KDE-ebd, and the uniform distribution do so with 6.8%, 11.7%, and 13.8%, respectively, when $n = 3$. As n increases, the percentiles of underestimating the exact probability of failure using KDE decrease and those using other methods increase. Finally, all methods yield similar percentile values of underestimating or overestimating the exact probability of failure. However, notice that the uniform distribution is still not converged and is too conservative in estimating the probabilities of failure.

6 Conclusions

In this study, the KDE-bd and KDE-ebd methods were proposed for estimating statistical input distributions. The KDE-bd and KDE-ebd combine the nonparametric statistical modeling method (KDE) and the interval approach using bounded data. To verify the accuracy of the KDE-bd and KDE-ebd, intersection areas using KDE, KDE-bd, and KDE-ebd were compared by performing statistical simulation tests for various distribution types and number of data. As a result, it was demonstrated that the KDE-bd and KDE-ebd methods were more accurate than the original KDE method, especially if there is a lack of data, in which case the estimated density functions converge to the population distribution as the number of data increases. Through numerical examples, it was shown that the KDE-bd and KDE-ebd estimate conservative density functions with heavy tails and thus yield more conservative reliability analysis results than the original KDE.

As a result, the following guideline of using KDE methods based on the number of data and information of boundary conditions can be proposed to estimate probabilistic density functions. If only the input variable intervals are given, only a uniform distribution can be used. If $n = 1$ and input variable intervals are given, the KDE-bd is recommended more than the uniform distribution. If n is greater than 2 and the intervals are known, the KDE-bd is still strongly recommended, but the KDE-ebd can also be used when the given intervals are undetermined. If n is greater than 2 and the bounds are unknown, the KDE-ebd is the most recommended.

The proposed methods can provide more accurate and conservative statistical models for reliability analysis, but such nonparametric methods can only be used limitedly in sampling-based reliability methods such as Monte-Carlo simulation (MCS) and importance sampling (IS). Accordingly, in future research, an integrated statistical modeling method combining parametric and nonparametric modeling will be further investigated in order to be applied to numerical reliability methods such as the first-order reliability method (FORM) and the second-order reliability method (SORM).

Acknowledgments This work was supported by the National Research Foundation of Korea (NRF) grant, funded by the Korean Government (NRF-2015R1A1A3A04001351) and by the Technology Innovation Program (10048305, Launching Plug-in Digital Analysis Framework for Modular System Design) and the Human Resources Development program (No. 20164030201230) of the Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant funded by the Ministry of Trade, Industry and Energy. This support is greatly appreciated.

Appendix 1: Silverman's rule of thumb

The Silverman's rule of thumb is a method which minimizes an objective function, mean integrated squared error (MISE), and it is probably the most popular one among the bandwidth selection methods (Schindler 2011). It assumes that true density is normally distributed therefore Silverman's rule will compute a bandwidth close to optimal if a random variable X is reasonably close to the normal distribution (Silverman 1986; Hansen 2009). It defines according to various kernel functions as follows (Hansen 2009).

$$h = \hat{\sigma} C_{\nu}(k) n^{-1/(2\nu+1)} \tag{18}$$

where $C_{\nu}(k)$ is the constant from Table 12, and ν is the order of the kernel.

Table 12 Constants of Silverman's rule (Hansen 2009)

Kernel	$\nu = 2$	$\nu = 4$	$\nu = 6$
Epanechnikov	2.34	3.03	3.53
Biweight	2.78	3.39	3.84
Triweight	3.15	3.72	4.13
Gaussian	1.06	1.08	1.08

References

Agarwal H, Renaud JE, Preston EL, Padmanabhan D (2004) Uncertainty quantification using evidence theory in multidisciplinary design optimization. Reliab Eng Syst Saf 85(1):281–294

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723

Analytical Methods Committee (1989) Robust statistics-how not to reject outliers. Part 1. Basic concepts. Analyst 114(12):1693–1697

Anderson TW, Darling DA (1952) Asymptotic theory of certain goodness of fit criteria based on stochastic processes. Ann Math Stat 23(2): 193–212

Ayyub BM, McCuen RH (2012) Probability, statistics, and reliability for engineers and scientists. CRC Press, Florida

Betrie GD, Sadiq R, Morin KA, Tesfamariam S (2014) Uncertainty quantification and integration of machine learning techniques for predicting acid rock drainage chemistry: a probability bounds approach. Sci Total Environ 490:182–190

Betrie GD, Sadiq R, Nichol C, Morin KA, Tesfamariam S (2016) Environmental risk assessment of acid rock drainage under uncertainty: the probability bounds and PHREEQC approach. J Hazard Mater 301:187–196

Burnham KP, Anderson DR (2004) Multimodel inference: understanding AIC and BIC in model selection. Sociol Methods Res 33(2):261–304

Chen S (2015) Optimal bandwidth selection for kernel density functionals estimation. J Probab Stat 2015:21

Cho SG, Jang J, Kim S, Park S, Lee TH, Lee M, Choi JS, Kim HW, Hong S (2016) Nonparametric approach for uncertainty-based multidisciplinary design optimization considering limited data. Struct Multidiscip Optim 54(6):1671–1688

Cowling A, Hall P (1996) On pseudodata methods for removing boundary effect in kernel density estimation. J R Stat Soc Ser B Methodol 58(3):551–563

Cox M, Harris P (2003) Up a GUM tree? Try the full monte! National Physical Laboratory, Teddington

Eldred MS, Agarwal H, Perez VM, Wojtkiewicz SF Jr, Renaud JE (2007) Investigation of reliability method formulations in DAKOTA/UQ. Struct Infrastruct Eng 3(3):199–213

Frigge M, Hoaglin DC, Lglewicz B (1989) Some implementations of the boxplot. Am Stat 43(1):50–54

Gabauer W (2000) Manual of codes of practice for the determination of uncertainties in mechanical tests on metallic materials, the determination of uncertainties in tensile testing. UNCERT COP7 report, Project SMT4-CT97-2165

Gasser T, Müller HG (1979) Kernel estimation of regression functions. Smoothing Techniques for Curve Estimation 757:23–68

Guidoum AC (2015) Kernel estimator and bandwidth selection for density and its derivatives. Department of Probabilities & Statistics, Faculty of Mathematics, University of Science and Technology Houari Boumediene, Algeria, https://cran.r-project.org/web/packages/kedd/vignettes/kedd.pdf

Hansen BE (2009) Lecture notes on nonparametrics. University of Wisconsin-Madison, WI, USA, http://www.ssc.wisc.edu/~bhansen/718/NonParametrics1.pdf

Hardle W, Marron JS, Wand MP (1990) Bandwidth choice for density derivatives. J R Stat Soc Ser B Methodol 52(1):223–232

Jang J, Cho SG, Lee SJ, Kim KS, Hong JP, Lee TH (2015) Reliability-based robust design optimization with kernel density estimation for electric power steering motor considering manufacturing uncertainties. IEEE Trans Magn 51(3):1–4

Jones MC, Kappenman RF (1992) On a class of kernel density estimate bandwidth selectors. Scand J Stat 19(4):337–349

Jung JH, Kang YJ, Lim OK, Noh Y (2017) A new method to determine the number of experimental data using statistical modeling methods. J Mech Sci Technol 31(6):2901–2910

Kang YJ, Lim OK, Noh Y (2016) Sequential statistical modeling for distribution type identification. Struct Multidiscip Optim 54(6): 1587–1607

Kang YJ, Hong JM, Lim OK, Noh Y (2017) Reliability analysis using parametric and nonparametric input modeling methods. J Comput Struct Eng Inst Korea 30(1):87–94

Karanki DR, Kushwaha HS, Verma AK, Ajit S (2009) Uncertainty analysis based on probability bounds (P-box) approach in probabilistic safety assessment. Risk Anal 29(5):662–675

Karunamuni RJ, Alberts T (2005a) On boundary correction in kernel density estimation. Stat Methodol 2(3):191–212

Karunamuni RJ, Alberts T (2005b) A generalized reflection method of boundary correction in kernel density estimation. Can J Stat 33(4): 497–509

Karunamuni RJ, Zhang S (2008) Some improvements on a boundary corrected kernel density estimator. Stat Probab Lett 78(5):499–507

Marron JS, Ruppert D (1994) Transformations to reduce boundary bias in kernel density estimation. J R Stat Soc Ser B Methodol 56(4):653–671

Montgomery DC, Runger GC (2003) Applied statistics and probability for engineers (3rd edition). Wiley, New York

Noh Y, Choi KK, Lee I (2010) Identification of marginal and joint CDFs using Bayesian method for RBDO. Struct Multidiscip Optim 40(1): 35–51

Schindler A (2011) Bandwidth selection in nonparametric kernel estimation. PhD Thesis. Göttingen, Georg-August Universität, Diss

Schuster EF (1985) Incorporating support constraints into nonparametric estimators of densities. Commun StatTheory Methods 14(5):1123–1136

Schwarz (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464

Scott DW, Terrell GR (1987) Biased and unbiased cross-validation in density estimation. J Am Stat Assoc 82(400):1131–1146

Shah H, Hosder S, Winter T (2015) Quantification of margins and mixed uncertainties using evidence theory and stochastic expansions. Reliab Eng Syst Saf 138:59–72

Sheather SJ (2004) Density estimation. Stat Sci 19(4):588–597

Sheather SJ, Jones MC (1991) A reliable data-based bandwidth selection method for kernel density estimation. J R Stat Soc Ser B Methodol 53(3):683–690

Silverman BW (1986) Density estimation for statistics and data analysis, vol 26. CRC press, London

Tucker WT, Ferson S (2003) Probability bounds analysis in environmental risk assessment. Applied Biomathematics, Setauket, New York, http://www.ramas.com/pbawhite.pdf

Tukey JW (1977) Exploratory data analysis. Pearson, New York

Verma AK, Srividya A, Karanki DR (2010) Reliability and safety engineering. Springer, London

Wand MP, Jones MC (1994) Kernel smoothing. CRC press, London

Yao W, Chen X, Quyang Q, Van Tooren M (2013) A reliability-based multidisciplinary design optimization procedure based on combined probability and evidence theory. Struct Multidiscip Optim 48(2): 339–354

Youn BD, Jung BC, Xi Z, Kim SB, Lee WR (2011) A hierarchical framework for statistical model calibration in engineering product development. Comput Methods Appl Mech Eng 200(13):1421–1431

Zhang Z, Jiang C, Han X, Hu D, Yu S (2014) A response surface approach for structural reliability analysis using evidence theory. Adv Eng Softw 69:37–45