A Nonparametric-based Approach for the Characterization and Propagation of Epistemic Uncertainty due to Small Datasets

Conference Paper · January 2019

DOI: 10.2514/6.2019-1490

CITATIONS
7

READS
180

4 authors, including:

Zhenyu Gao
University of Texas at Austin
25 PUBLICATIONS **151** CITATIONS

SEE PROFILE

Dongwook Lim
Georgia Institute of Technology
31 PUBLICATIONS **167** CITATIONS

SEE PROFILE

Dimitri N. Mavris
Georgia Institute of Technology
1,555 PUBLICATIONS **13,000** CITATIONS

SEE PROFILE

AIAA SciTech Forum
7-11 January 2019, San Diego, California
AIAA Scitech 2019 Forum

10.2514/6.2019-1490

A Nonparametric-based Approach for the Characterization and Propagation of Epistemic Uncertainty due to Small Datasets

Zhenyu Gao *, Dongwook Lim †, Katherine G. Schwartz ‡, and Dimitri N. Mavris §
School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, Georgia, 30332

Quantification of uncertainty is ideally performed by the use of highly precise and consistent information, which is rarely available in many applications due to lack of knowledge and/or resources. When only small datasets are available for characterizing the underlying probability distributions of uncertainty sources, related epistemic uncertainty needs to be characterized and propagated with weaker assumptions and greater flexibility. This paper proposes a nonparametric-based approach to further facilitate the characterization and propagation of epistemic uncertainty due to the lack of sufficient data. The first part of this paper presents the use of Kernel Density Estimation (KDE) and Bootstrap to estimate probability distributions of random variables based on small datasets. Two types of density estimates are provided for uncertainty propagation: an optimal density estimate representing the best estimate of the true distribution, and a conservative density estimate representing risk and uncertainty that is inherent in small datasets. In the second part, copulas and inversion method are applied to model the dependence structure among random variables, to mitigate the overestimation or underestimation of uncertainty caused by incorrect independence assumption. The proposed method is illustrated by various illustrations and a challenging problem in aviation environmental impact analysis.

I. Introduction

In the design and analysis of complex engineered systems, the term *uncertainty* refers to both the probability that certain assumptions made during the design and analysis process are incorrect, and the presence of entirely unknown facts that might have a bearing on the future state of a product or system.[1] Since uncertainty has a significant impact on the success or failure of engineered systems, uncertainty quantification and analysis plays a crucial role in many disciplines of aerospace engineering, such as aircraft design, systems integration of new technologies, and fleet-wide performance assessments in air transportation.[2–5] The high degree of complexity and uncertainty associated with aerospace engineering applications has driven researchers towards the use of probabilistic and statistical analysis methods in many recent projects.[6–10]

Uncertainty quantification (UQ) is the process of characterizing and propagating uncertainty in a computational or real world system to provide information about quantity of interests. The objective of UQ is to analyze how those uncertain factors can affect the accuracy of the computational results, for reducing uncertainties or making better decisions. Different sources of uncertainty are generally categorized as either *aleatory uncertainty* or *epistemic uncertainty*.[11] Aleatory uncertainty is the uncertainty due to intrinsic randomness of nature, and is beyond people's ability to reduce it by collecting additional information. Epistemic uncertainty is the uncertainty due to lack of complete knowledge, and is possible to be reduced by gathering more information. An UQ process typically consists of uncertainty identification, characterization, propagation, analysis and reduction.[12] Among two categories of uncertainty and different steps in the UQ process, this work mainly focuses on the challenge of characterizing and propagating epistemic uncertainty

*Graduate Research Assistant, School of Aerospace Engineering, Georgia Institute of Technology, AIAA Student Member
†Research Engineer II, School of Aerospace Engineering, Georgia Institute of Technology, AIAA Senior Member
‡Research Engineer II, School of Aerospace Engineering, Georgia Institute of Technology, AIAA Member
§S.P. Langley Distinguished Regents Professor, School of Aerospace Engineering, Georgia Institute of Technology, AIAA Fellow

resulting from lack of sufficient statistical data (small datasets).[13] After the uncertainty sources have been identified, uncertainty characterization refers to a process to determine how uncertainty in each uncertainty source should be mathematically represented.[14] The literature point to a variety of ways of representing epistemic uncertainty, either in probabilistic (frequentist, Bayesian, probability boxes, etc.) or non-probabilistic treatments (evidence theory, fuzzy sets, interval methods, etc.)[15] Uncertainty propagation is the process of mathematically mapping uncertainties in all levels of the model to the uncertainties in the simulation results.[12] Information for characterizing uncertainty sources is usually provided by experimental data, data from supporting models, literature, or subject-matter expert (SME) opinions.[16] However, a challenge arises in the probabilistic approach when only small datasets are available to characterize the uncertainty sources, because both the frequentist and Bayesian statistical theories are well-suited for problems with large sample size, which is rarely available for many real world applications.[15] Small datasets lead to a specific form of epistemic uncertainty, which is the inability to assign a proper probability distribution to represent an uncertainty source. This problem could happen for various reasons, such as:

1. **Cost of collecting data is high (money and/or time)**: for example, in industry, experiments for collecting one data point costs $2,000. With a limited budget of $100,000 for all experiments, only 50 data points can be collected to assign a distribution. Also, the time required for collecting the complete dataset may take too long, making it an impossible task within the time frame of a project.

2. **Loss of information/unable to collect complete information**: for example, in aviation environmental impact study, the database does not cover all aircraft models

With small datasets or incomplete information, a common flaw in aerospace engineering applications is the arbitrary or unjustified use of probability distributions.[4] Due to a lack of knowledge and/or resources, normal distribution and some "lack of knowlegde distributions", such as Triangular distribution and Student's t-distribution, are often used. Nevertheless, incorrect assumptions of distribution types may introduce unwarranted information into the uncertainty propagation process, potentially leading to inaccurate UQ results and poor decisions. The observations above highlight the need for less assumptions and greater flexibility in accommodating distributions for uncertainty sources. The problem of accommodating and efficiently propagating epistemic uncertainty resulting from small datasets has received attention from the engineering community and has been studied by several researchers in the fields of civil and mechanical engineering.[10,15,17,18] Various approaches have been proposed to quantify and propagate uncertainty resulting from sparse and imprecise data, and interval data.

Within the category of probabilistic approach, however, most researchers have focused on the use of parametric-based methods, which have some notable limitations. The more general nonparametric methods haven't been widely considered so far, because of several challenges, such as the identification of an optimal nonparametric probability model from small dataset without overfitting. This papers presents a nonparametric-based approach which provides another angle to characterize and propagate epistemic uncertainty due to small datasets with less assumptions and greater flexibility.

II. Current Progress and Limitations

The following sections describe the latest development of parametric-based methods on the propagation and quantification of uncertainty resulting from small datasets, and their limitations. This part starts with a discussion between parametric and nonparametric methods.

A. Parametric vs. Nonparametric Statistics

Parametric Statistics is a branch of statistics with the assumption that the sample data comes from a population that follows one of the *parametric distributions* - probability distributions that are based on a fixed set of parameters.[19] Most of the well-known statistical methods and distributions are parametric,[20] such as Normal (Gaussian) distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ and Exponential Distribution $X \sim Exp(\lambda)$. *Nonparametric Statistics* is broadly defined to include all statistical methods that are not based on any parametric distribution family. For *nonparametric distributions*, some representative ones include histogram, empirical distribution, and kernel distribution. When comparing these two categories, a key difference is that nonparametric distributions are not based on parametric assumptions, and the distribution is completely

determined by the finite knowledge. This feature makes nonparametric models a better choice under some circumstances. For example, when data generated from complicated experiments are not attributed to any well-known parametric distributions, making parametric assumptions when people are not sure about the underlying distribution of the data is not reasonable.

B. The Parametric Approach

The parametric-based approach in the literature is based on parametric distributions. Since it is difficult to identify a single parametric distribution to represent the full uncertainty based on lack of sufficient data, the parametric-based approach is a probabilistic treatment that disassembles the total uncertainty into two parts: *model-form uncertainty* and *parameter uncertainty*. These two types of uncertainty represent two dimensions of the challenge to assign a unique parametric distribution based on small datasets. The first challenge is to assign a specific model form (parametric distribution type) to an uncertainty source. In most cases, the distributions are selected from a great variety of distributions that have been used in many applications to model uncertain variables, such as Normal distribution, Beta distribution, Weibull distribution, and others.[4] However, one single distribution assigned by experts and researchers based on limited knowledge and judgments can still be arbitrary and unjustified. A more scientific and widely accepted way is to select a set of candidate parametric models through a comparative down-selection process.[15] In this multi-model approach, the uncertainty associated with model selection and the lack of confidence to select the single best model is called *model-form uncertainty*.[21] Within the set of candidate parametric models, estimation of the model parameters for each model is another dimension of uncertainty, called *parameter uncertainty*. The parameters for a given model can be estimated as either deterministic values with confidence bounds in the frequentist approach, or joint parameter distributions in the Bayesian approach.[15] From the perspective of uncertainty propagation, the probabilistic multi-model approach is computationally expensive,[22] making efficient uncertainty propagation methods another significant concern.

The latest parametric-based approach for the propagation of uncertainty due to small datasets consists of four steps: *multi-model inference*, *parameter inference*, *establishment of a finite model set*, and *uncertainty propagation*. Details of these four steps are introduced below and shown in Figure 1.

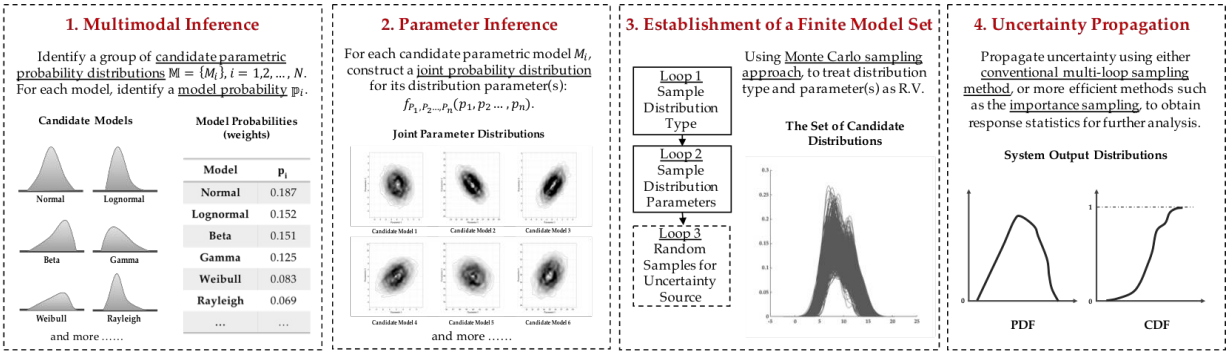

Figure 1. Main steps and elements of the parametric-based approach

With the available small dataset, it is difficult to identify a single best model for the true underlying distribution of the uncertainty source. The first step in the parametric approach is to identify a group of candidate probability models $\mathbb{M} = M_i, i = 1, 2, ..., n$. Examples of the candidate probability model include Normal distribution, Beta distribution, Gamma distribution, Weibull distribution, etc. When there is more than one candidate probability model, a *multi-model inference* (model selection) process that compares the validity of candidate models is needed. To do that, a suitable model selection criterion must be established, using criteria such as the Akaike Information Criterion (AIC, based on the Kullback-Leibler information theory), and Bayesian Information Criterion (BIC, based on Bayes factors/likelihood functions).[15,23] The content below summarizes the procedure of using a critical extension of AIC to estimate model probabilities based on small datasets (a similar procedure exists for BIC but is not explained here):[23]

$$AIC_c = -2\log(\mathcal{L}(\hat{\theta}|\mathbf{d}, M)) + 2K + \frac{2K(K+1)}{n-K-1} \tag{1}$$

$$\Delta_A^{(i)} = AIC_c^{(i)} - AIC_c^{min} \tag{2}$$

$$p_i = p(M_i|\mathbf{d}) = \frac{\mathcal{L}(M_i|\mathbf{d})}{\sum_{i=1}^{N}\mathcal{L}(M_i|\mathbf{d})} = \frac{\exp(-\frac{\Delta_A^{(i)}}{2})}{\sum_{i=1}^{N}\exp(-\frac{\Delta_A^{(i)}}{2})} \tag{3}$$

where $\mathcal{L}(\hat{\theta}|\mathbf{d}, M)$ is the likelihood function with the maximum likelihood estimate of parameters $\hat{\theta}$, $\mathbf{d}$ is the data given, M is the probability model, N is the sample size of the small dataset, and K is the number of parameters in $\hat{\theta}$. Model probabilities obtained through Equation 3 can be interpreted as the probability that model M_i is the best model for the data under the model selection criterion. The model probabilities p_i here is also called *Akaike weights*. Results from this step represent the *model-form uncertainty*. Other methods to quantify the model-form uncertainty associated with multiple completing distributions include the Bayesian model averaging (BMA) approach.[10,21]

For each candidate parametric probability model M_i, parameters of this candidate model are then estimated through statistical inference using available data.[10] When large datasets are available, it is possible to obtain deterministic estimates of the parameters using methods such as method of moments and method of maximum likelihood. Yet when only small datasets are available, uncertainty associated with parameter estimates exists. This part of uncertainty is called *parameter uncertainty* or *second-order uncertainty*, and is determined by a *parameter inference* process. Compared to the frequentist statistical approach which estimates the uncertainty in distribution parameters using statistical confidence intervals, the Bayesian approach constructs probability distributions for the distribution parameters,[10] and makes it possible to also treat parameter uncertainty in a probabilistic representation. With a candidate probability model M_i and its uncertain parameters θ_i, given data $\mathbf{d}$, the posterior distribution for the parameters is given by[15]

$$p^*(\theta_i|\mathbf{d}, M_i) = \frac{p(\mathbf{d}|\theta_i, M_i)p(\theta_i; M_i)}{p(\mathbf{d}; M_i)} \propto \mathcal{L}(\theta_i|\mathbf{d}, M_i)p(\theta_i; M_i) \tag{4}$$

$$p(\mathbf{d}; M_i) = \int \mathcal{L}(\theta_i|\mathbf{d}, M_i)p(\theta_i; M_i)d\theta_i \tag{5}$$

where $p(\theta_i; M_i)$ is the prior distribution that reflects the current belief or knowledge of the parameters, and $\mathcal{L}(\theta_i|\mathbf{d}, M_i) = p(\mathbf{d}|\theta_i, M_i)$. In the Bayesian approach, the knowledge of parameters θ_i for model M_i is updated using the available dataset. The Markov Chain Monte Carlo (MCMC) sampling method can be used to draw samples from $p^*(\theta_i|\mathbf{d}, M_i)$ for the distribution parameters.

After both the model-form and parameter uncertainties are quantified and expressed in probabilistic representations, the classical Monte Carlo sampling approach is used to create a finite set of candidate distributions. In this process, model form is treated as the first discrete random variable, and is selected randomly using model probabilities $p_i, i = 1, 2, ..., n$. In the next loop, parameters of the model are then randomly drawn from the corresponding joint parameter distribution $p^*(\theta_i|\mathbf{d}, M_i)$. The two loops of sampling identifies a unique parametric probability density function for the dataset $\mathbf{d}$. After this process is repeated for N times, a *finite model set* that contains N probability density functions is created. The finite model set creates probability bounds from the cloud of candidate distributions, and can be used for different purposes of uncertainty propagation. It is important to draw a sufficiently large amount of candidate distributions, such that the full range of candidate model forms and a large enough range of parameters can be covered.[15]

For *uncertainty propagation*, propagating different types of uncertainty through the multi-model approach can be computationally very expensive. In the conventional Monte Carlo-based approach, multiple loops of sampling are necessary.[10] Apart from the first two loops that identifies a unique probability density function, an additional third loop is needed to draw random samples from the constructed density. Samples generated from the third loop are used to represent the distribution of the uncertainty source, and propagated through a system model. A more efficient uncertainty propagation method first identifies a single *optimal sampling density* $q^*(\mathbf{x})$ that represents all the models in the finite set by minimizing the total expected mean square difference (MSD). Then the *importance sampling* technique is used to propagate the full set of models with optimal sampling density $q^*(\mathbf{x})$ and importance weights.[15]

C. Notable Limitations of the Parametric Approach

It is without doubt that by utilizing several classical statistical methods, the parametric-based approaches are solid contributions to the problem. Even with greater flexibility than a single probability model, however, the parametric-based approach still has three notable limitations:

1. **Limitation 1**: it still needs people to manually supply a group of candidate models based on limited knowledge. This approach starts with the identification of a group of candidate parametric models, which inevitably involves subjective judgments. Even with a relatively rigorous down-selection process and opinions from subject-matter experts, assumptions on model types based on limited data may still introduce unwarranted information in uncertainty quantification.

2. **Limitation 2**: the set of candidate models is constrained to well-established parametric probability distributions. Although the parametric models can be utilized to model a wide range of events that happen in nature, the complexity of uncertainty sources in real applications makes parametric models insufficient to span the epistemic uncertainty.

3. **Limitation 3**: the method is unable to model multimodal distributions (distributions with more than one peak). This is a higher-level need in uncertainty quantification, as most uncertain variables are presumed unimodal.

III. Proposed Nonparametric Approach

Compared to parametric statistical methods and distributions, nonparametric methods and distributions still haven't been widely used in the engineering community for related challenges. However, it must be noted that efforts have been made by several researchers in the past to build a nonparametric distribution to represent small datasets. Sankararaman and Mahadevan[17] built a method to fit nonparametric distribution to point data using likelihood function and interpolation techniques. Pradlwarter and Schuller[24] established a nonparametric PDF for insufficient data using kernel density and confidence interval. A common ground for these nonparametric methods is that they are designed and have advantages on extremely small and sparse datasets (sample size $N < 20$), which is not the case for many applications as the evidence is sometimes too little to initiate a study. The scope of this work is to propose an effective nonparametric approach for small datasets with more than 20 data, an order that is closer to a rigorous study and tackled by the current parametric methods. Although limited by the properties of parametric models, establishment of a finite model set (Step 3 in Figure 1) is a good indication of uncertainty when estimating the underlying distribution from small datasets. In the nonparametric approach, such a finite model set can be built through re-sampling (using bootstrap) and kernel distributions (using kernel density estimate). The following sections introduce the key nonparametric statistical elements and the proposed nonparametric approach.

A. Nonparametric Statistical Methods

1. *Kernel Density Estimation (KDE) and Bandwidth Selection*

A *kernel distribution* is a nonparametric representation of the probability density function (PDF) of a random variable.[25] It is used when a parametric distribution can not properly describe the data, or when people want to avoid making assumptions on the underlying distribution of the data. *kernel density estimation* (KDE) is the method to estimate the nonparametric PDF of a random variable. A kernel distribution is define by a *kernel function* (base shape of the PDF at each data value) and a *bandwidth value* (value that controls the smoothness of the kernel distribution).[26] The KDE estimates the smooth and continuous PDF of a random variable by adding the smoothing functions for each data value (shown in Figure 2), and is given by:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h}) \tag{6}$$

where x_i, $i = 1, 2, ..., n$, is an independently, identically distributed random sample from an unknown distribution, $K(\cdot)$ is the kernel function, and h is the bandwidth value. The kernel function satisfies $\int K(u)du = 1$, $\int uK(u)du = 0$, $\int u^2K(u)du > 0$, and has four basic forms: Normal, Triangular, Box, and Epanechnikov. Since the selection of kernel function K is not very sensitive to the shape estimator,

and therefore not crucial compared to the selection of bandwidth h,[27] the normal kernel function is used throughout this work.

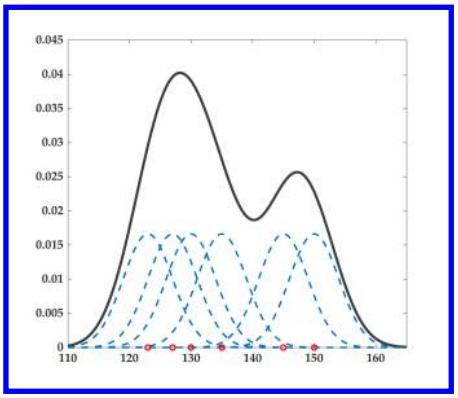

Figure 2. Kernel density estimation

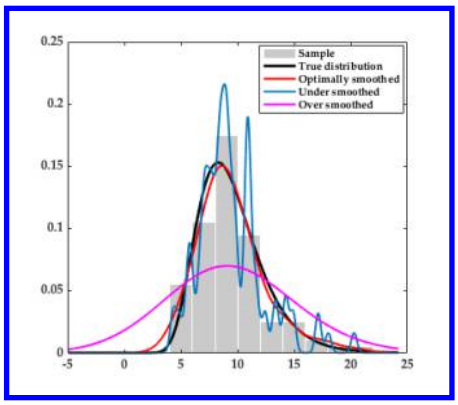

Figure 3. Effect of bandwidth selection

Selection of the bandwidth has been the most challenging part in KDE, because it is influential on the shape of the kernel estimator (shown in Figure 3). When the bandwidth value h is too small, we will obtain an under-smoothed density curve with high variability. On the contrary, when the bandwidth value is too large, the resulting density curve will be over-smoothed such that important structures of the true distribution can be obscured. In the small dataset case, it is especially challenging to identify an optimal bandwidth value without overfitting (producing under-smoothed curve). An optimal bandwidth value can be selected by minimizing the Mean Integrated Squared Error (MISE), given by[26]

$$\mathrm{MISE}(\hat{f}_h) = \mathrm{E}\Big\{ \int (f(x) - \hat{f}_h(x))^2 dx \Big\} = \int \mathrm{Bias}^2(\hat{f}_h(x))dx + \int \mathrm{Var}(\hat{f}_h(x))dx \tag{7}$$

by assuming that $f(x)$ is continuous, it can be shown that

$$\mathrm{Bias}\{\hat{f}_h(x)\} = \frac{h^2}{2} u_2(K) f''(x) + O(h^2) \tag{8}$$

$$\mathrm{Var}\{\hat{f}_h(x)\} = \frac{1}{nh} R(K) f(x) + O\Big(\frac{1}{nh}\Big) \tag{9}$$

where $R(K) = \int K^2(x)dx$ and $u_2(K) = \int x^2 K^2(x)dx > 0$. By adding the leading variance and squared bias terms appeared in Equations 8 and 9, we get the Asymptotic Mean Squared Error (AMSE) as[28]

$$\mathrm{AMSE}\{\hat{f}_h(x)\} = \frac{1}{nh} R(K) f(x) + \frac{h^4}{4} u_2(K)^2 [f''(x)]^2 \tag{10}$$

Assume the integrability on f, it gives us the Asymptotic Mean Integrated Squared Error (AMISE) as

$$\mathrm{AMISE}\{\hat{f}_h(x)\} = \frac{1}{nh} R(K) + \frac{h^4}{4} u_2(K)^2 R(f'') \tag{11}$$

where $R(f'') = \int [f''(x)]^2 dx$. And the bandwidth value that minimizes the AMISE is given by[28]

$$h_{\mathrm{AMISE}} = \Big[\frac{R(K)}{u_2(K)^2 R(f'')} \Big]^{1/5} n^{-1/5} \tag{12}$$

Equation 12 is a closed-form solution for the optimal bandwidth value, however, the $R(f'')$ is unknown if we do not know what the true distribution is. As a result, other methods for choosing a global optimal bandwidth value are needed. Here we briefly review the following main bandwidth selection methods: *rules of thumb*, *cross-validation methods*, and *plug-in methods*.

Among various bandwidth selection methods, the *rules of thumb* are computationally the simplest.[28] It replaces $R(f'')$, the only unknown part of h_{AMISE}, by the values from a parametric distribution family. Results of the rules of thumb lead to relationships between the optimal bandwidth value with some descriptive statistics of the sample, such as sample standard deviation S, sample interquartile range IQR, and sample

size n. With the normal distribution assumption for f and follow-on performance studies, the most widely-used rule of thumb was developed as[28]

$$h_{\mathrm{SNR}} = 1.06Sn^{-1/5} \tag{13}$$

where n is the sample size and S is the sample standard deviation. In order to not miss the bimodality,[29] a second rule of thumb was later built with some modifications on Equation 13, given by

$$h_{\mathrm{SROT}} = 0.9An^{-1/5} \tag{14}$$

where $A = \min\{S, IQR/1.34\}$. The h_{SROT} recommends a smaller factor (reduced from 1.06 to 0.9) and uses the smaller of two descriptive statistics S and $IQR/1.34$. The third rule of thumb was developed based on the maximal smoothing principle and chooses the largest degree of smoothing compatible,[30] given by

$$h_{\mathrm{OS}} = 1.144Sn^{-1/5} \tag{15}$$

Objective of the h_{OS} is to produce over-smoothed density curves,[30,31] even though it is only 1.08 time larger compared with h_{SNR}. These three rules of thumb, h_{SNR}, h_{SROT}, and h_{OS} own different characteristics and will be used later for different purposes.

Cross-validation is another category of classical bandwidth selection methods for KDE. The idea of cross-validation is to use part of data (training set) for training the model, and the remaining data (validation set) for estimating test error. Two common types of cross-validation techniques are *K-fold cross-validation* (K-fold CV) and *Leave-one-out cross-validation* (LOOCV).[32] In K-fold CV, the data is divided randomly into K equal-sized groups (or folds). Each time, part k is left out and treated as the validation set, and the model is fitted on the remaining $k-1$ folds. After this is done for each part $k = 1, 2, ..., K$, the test error results are combined. The LOOCV is a special case of K-fold CV, in which k is set to equal n. The advantage of using $k = 5$ or $k = 10$ (two most widely used K-fold CV) rather than $k = n$ is the computational cost at a large n, since the LOOCV requires fitting and validating the model for n times. Nevertheless, the K-fold CV also has one potential drawback: the test error result can be variable due to the randomness in data splitting. LOOCV is a better choice when the dataset size is relatively small, because it provides exhaustive (all but one) data to train the model each time. In this work, LOOCV is used when the dataset size n is less than or equal to 100. When $n > 100$, the 10-fold CV is used for better computational efficiency.

The use of cross-validation for choosing an optimal bandwidth value starts with the Integrated Squared Error (ISE), given by[28]

$$\mathrm{ISE}(\hat{f}_h) = \int (f(x) - \hat{f}_h(x))^2 dx = \int (\hat{f}_h(x))^2 dx - 2\int \hat{f}_h(x) f(x) dx + \int f^2(x) dx \tag{16}$$

In Equation 16, the last term on the right-hand side does not involve h. An expression was later proposed to estimate the two other terms, given by[33]

$$\frac{1}{n}\sum_{n=1}^{n} \int (\hat{f}_{-i}(x))^2 dx - \frac{2}{n}\sum_{n=1}^{n} \int \hat{f}_{-i}(X_i) \tag{17}$$

where $\hat{f}_{-i}(x)$ denotes the kernel estimator constructed with X_i excluded. This method of choosing a h that minimizes Equation 17 is referred to as *least squares cross-validation* (LSCV). Then, with[34]

$$\frac{1}{n}\sum_{n=1}^{n} \int (\hat{f}_{-i}(x))^2 dx = \int (\hat{f}_h(x))^2 dx + O_p\Big(\frac{1}{n^2 h}\Big) \tag{18}$$

a simpler to compute LSCV-based criterion is given by[28]

$$\mathrm{LSCV}(h) = \int (\hat{f}_h(x))^2 dx - \frac{2}{n}\sum_{n=1}^{n} \int \hat{f}_{-i}(X_i) \tag{19}$$

The version in Equation 19 is used by most of the researchers in statistics, and the resulting value of h that minimizes LSCV(h) is denoted by h_{LSCV}. Examples of some other cross-validation criteria include:[35]

$$\mathrm{LCV}(h) = -\sum_{n=1}^{n} \log \hat{f}_h(X_i) - \sum_{n=1}^{n} \log(1 - K(0)/(nh\hat{f}_h(X_i))) \tag{20}$$

$$\mathrm{AIC}(h) = -\sum_{n=1}^{n} \log \hat{f}_h(X_i) - \sum_{n=1}^{n} K(0)/(nh\hat{f}_h(X_i)) \tag{21}$$

Since LSCV has a slow rate of convergence (even though h_{LSCV} has the best possible convergence rate within cross-validation), a category of faster converging methods called *plug-in methods* was invented.[28] It replaces the unknown quantity $R(f'')$ in h_{AMISE} using an estimate $R(\hat{f}_g'')$, with a pilot estimate bandwidth g chosen by the user. Although the plug-in methods have overall good computational performance, its claimed superiority is challenged by some researchers because the methods could fail when the arbitrary specification of pilot bandwidth is wrong.[35]

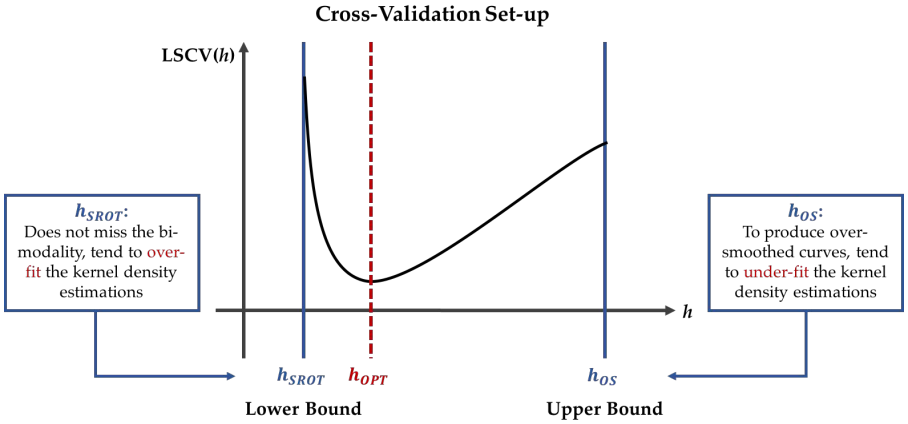

Figure 4. The final cross-validation set-up

Considering the fact that these bandwidth selection methods all have their advantages and disadvantages, in this work, we select the optimal bandwidth h_{opt} for KDE using a combination of rules of thumb and cross-validation methods. Previous research showed that if used properly, the classical cross-validation methods are far more informative than more recent works such as the plug-in methods.[35] However, the cross-validation methods have two potential drawbacks: the tendency to under smooth (overfit), and large computational cost. Studies performed by several researchers show that bandwidths obtained from least squares cross-validation are variable and tend to under smooth the density estimates.[28,35] To avoid over-smoothed or under-smoothed density estimates and reduce computational cost in least squares cross-validation, in this work we narrow down the range of bandwidth selection by setting lower and upper bounds for h. Recall that in the rules or thumb, h_{SROT} is the smallest among the three (Equations 13-15). It is an attempt not to miss bimodality and tends to produce under-smoothed density estimates. In contrast, h_{OS} is the largest among the three that produces over-smoothed density estimates. Identifying an optimal bandwidth value h_{opt} between h_{SROT} and h_{OS} is a way to both avoid dramatically over-smoothed or under-smoothed estimates, and reduce computational cost. Figure 4 shows how the optimal bandwidth for KDE is selected in this work. The value of h within the interval $[h_{\mathrm{SROT}}, h_{\mathrm{OS}}]$ that minimizes the least squares cross-validation function $\mathrm{LSCV}(h)$ is selected as the optimal bandwidth h_{opt}.

2. *Bootstrap*

When only small datasets are available and the density estimation method is chosen, a technique is needed to quantify the uncertainty of an estimator without generating additional data from the population. The *bootstrap* is a statistical technique that relies on *random sampling with replacement*.[26] It can be used to quantify the uncertainty associated with a certain estimator or statistical learning method.[32] Bootstrap was first mentioned by Bradley Efron in 1979, and the theory behind it is sophisticated, being based on Edgeworth Expansions.[36] In a new bootstrap sample, some of the N items in the original sample can appear more than once.[26] Figure 5 shows an example of using bootstrap to estimate the variability of a statistic without

generating additional samples from the population. From the original sample X with N observations, we randomly draw N observations with replacement to produce a bootstrap sample $X_{b,1}$. We then use $X_{b,1}$ to compute the value of the statistic, denoted by $\hat{\alpha}_{b,1}$. After this procedure is repeated B times, B different bootstrap samples $X_{b,1}$, $X_{b,2}$, . . . , $X_{b,B}$ and their corresponding statistic estimates $\hat{\alpha}_{b,1}$, $\hat{\alpha}_{b,2}$, . . . , $\hat{\alpha}_{b,B}$ can be obtained. The variability of statistic α is then estimated from $\{\hat{\alpha}_{b,i}, 1 \leq i \leq B\}$, using the empirical distribution, standard error, confidence interval, etc. In bootstrap, B should be large enough such that the final result is not affected by the randomness due to resampling.

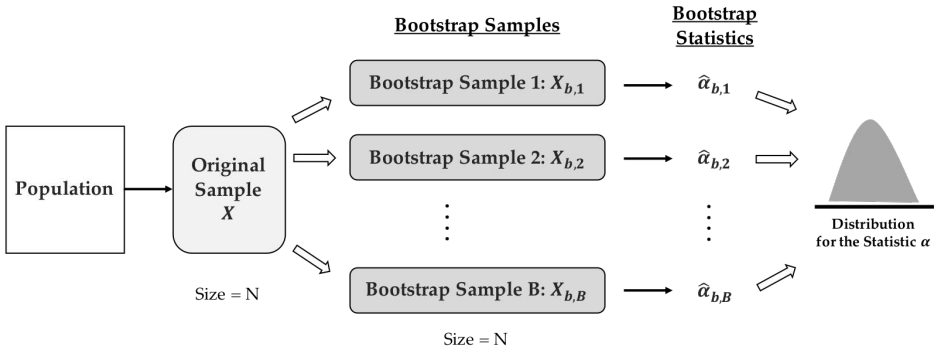

Figure 5. The bootstrap

Bootstrap is a powerful technique in quantifying uncertainties caused by incomplete information, but it should be used under one condition: the original sample should be "representative" of the population. Two standards for a representative sample are: a relatively enough sample size, and no sampling bias. Although by proposing the nonparametric approach, we are trying to minimize the use of assumptions for small datasets, it is still important that people who use bootstrap have contextual knowledge or belief on whether the sample is representative (at least an unbiased sample) of the population.

B. The Optimal and Conservative Density Estimates

Density estimation is the construction of an estimate of the underlying probability density function from the observed data,[37] and is the core of uncertainty characterization based on small datasets. In the nonparametric approach, we first aim to provide a nonparametric *optimal density estimate* that represents the best estimate of the true distribution through KDE with the previously defined optimal bandwidth value h_{opt}. When the optimal density estimates of the uncertainty sources (system inputs) are propagated through a system model, distributions of the responses (system outputs) serve as the optimal approximates. Then, in view of the additional risk and uncertainty in small datasets compared to an UQ process with enough data, this work proposes another estimate called the *conservative density estimate*. The objective of the conservative density estimate is to provide an estimate with the "maximum uncertainty" based on available data.

Small dataset is a form of incomplete information. Better precision in the characterization of uncertainty source is achieved when more information is available. When there is absolutely no information, *uniform distribution* is usually assumed based on the *principle of maximum entropy*.[38] The uniform distribution is therefore also called as the *maximum entropy distribution* on any interval $[a, b]$. Under the maximum entropy principle, we must use the probability distribution with the maximum entropy when making inference based on partial information.[39] The maximum entropy distribution can also be treated as a distribution that can best reflect the current state of knowledge with the maximum uncertainty. The conservative density estimate constructed in this work shares similar thinking with the maximum entropy principle.

The combination of KDE and bootstrap makes it possible to make inference about the shape of a distribution based on incomplete information. Given a small dataset $\mathbf{d}$, bootstrap is first used to generate a new bootstrap sample $\mathbf{d}_{b,1}$ with the same size $\mathbf{N}$. A nonparametric kernel density $\hat{f}_{b,1}(x)$ is then estimated by applying KDE on $\mathbf{d}_{b,1}$. After this procedure is repeated for B times, a family of bootstrap densities $\mathbb{f} = \{\hat{f}_{b,i}(x), i = 1, 2, ..., B\}$ is obtained. For example, when $B = 5,000$, a family of 5,000 bootstrap densities estimated from a small dataset of 50 data is shown in Figure 6. This bootstrap density family is highly informative in inferring the shape of the true underlying distribution, because a sufficiently large density family can provide wide enough probability bounds. Then, since bootstrap may yield inaccurate results

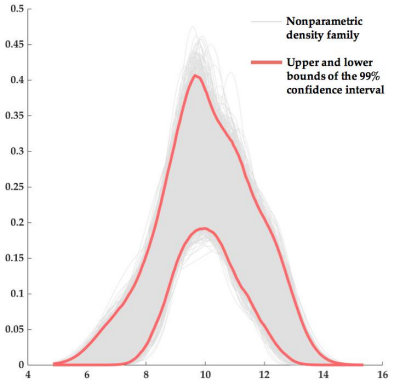

Figure 6. A family of 5,000 nonparametric kernel densities

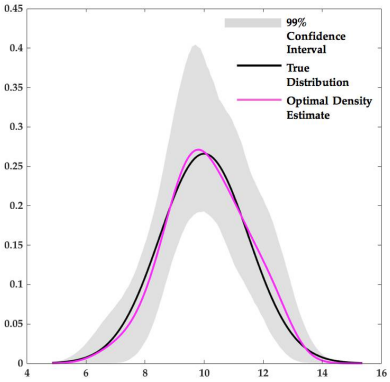

Figure 7. The true density, optimal estimate, and 99% CI

when estimating extreme situations, a 99% confidence interval (99% CI, also shown in Figure 6) is further constructed using percentile method. Note that the actual coverage might be less than the nominal coverage 99% as the bias are ignored. By showing how uncertain the estimation is at each point, the 99% CI is a measure of uncertainty in density estimation, and is usually wider when less data is available. When the 99% CI in Figure 6 is further extracted and overlapped with the true distribution and the optimal density estimate of this example, their relationships are shown in Figure 7.

The 99% CI constructed from a large amount of bootstrap densities provides useful information about the shape of the true distribution. This distinguishes the small dataset case from the total ignorance case, when there is absolutely no information and the uniform distribution is assumed according to the principle of maximum entropy. Similar to the maximum entropy distribution, now the conservative density estimate is targeted to represent a situation with the maximum uncertainty within the confidence interval, as shown in Figure 8. In the parametric case, people usually first assume a parametric distribution type, and then select the distribution with the largest entropy by maximizing the continuous entropy $h(p) = -\int p(x) \log p(x) dx$.[38] Yet in the nonparametric case, such a nonparametric distribution has to be identified through other methods. Under the property of KDE, the optimal density estimate (blue curve in Figure 9) can in fact be treated as the "average density" of the whole bootstrap density family (gray curves). Similarly, the conservative density estimate (red curve) can be constructed as the "average density" of a smaller portion of the bootstrap density family that contains the most uncertainties (pink curves).

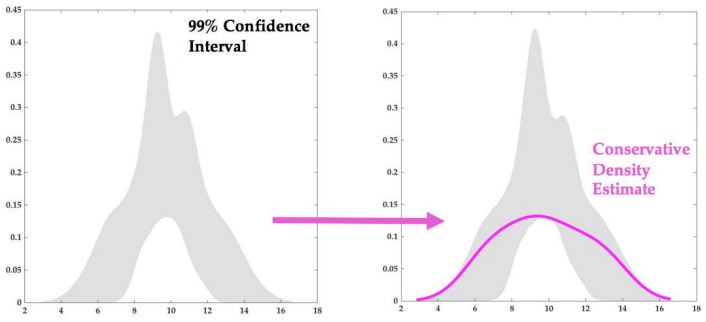

Figure 8. The 99% CI and the conservative density estimate

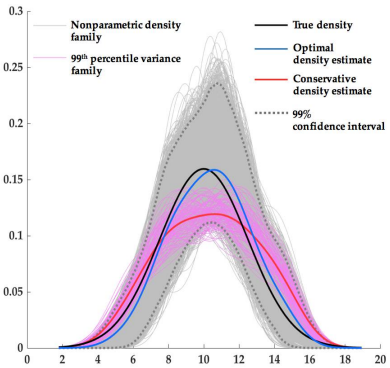

Figure 9. The two density estimates

Here, the simplest and most direct measure of uncertainty for a probability distribution is its variance. When we select out bootstrap densities whose variance are in the 99th percentile (in agreement with the CI), a much smaller density family is formed. The conservative density estimate is the "average density" of the 99th percentile variance family, and is obtained through a process in Figure 10. In this process, the B bootstrap samples are labeled by the variance of their corresponding kernel density estimates and ranked

accordingly. Then, bootstrap samples with the top 1% largest variance are drawn out to form a new dataset $\mathbf{d}_{new}$. For example, if there is a total of 10,000 bootstrap samples (each with size 50), 100 of them with the largest kernel density variances are selected to form a new dataset with 5000 data. The conservative density estimate is obtained through applying KDE on $\mathbf{d}_{new}$ with the optimal bandwidth.

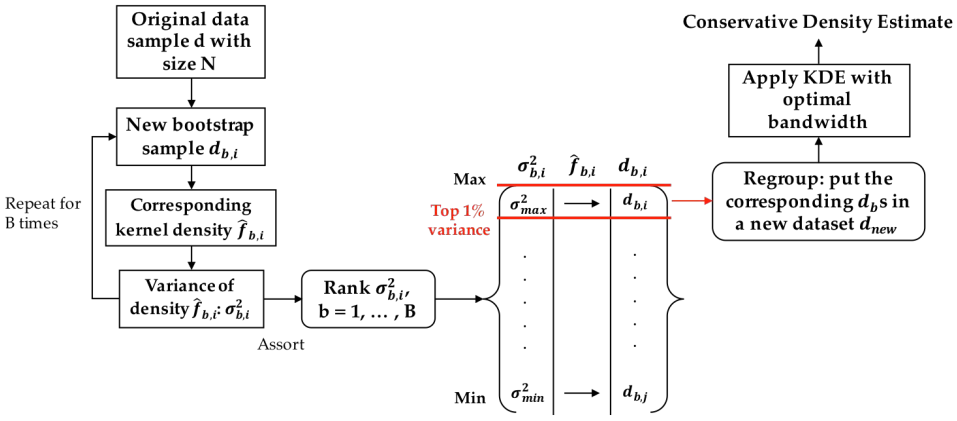

Figure 10. The process of getting the conservative density estimate

It is worth mentioning that properties of the conservative density estimate have similarities with the Student's t-distribution. The Student's t-distribution is used when a normally distributed population is assumed and when the sample size is small.[40] With larger degrees of freedom ($v = \mathbf{N} - 1$), the Student's t-distribution approaches the normal distribution, as shown at the left of Figure 11. Smaller values of v (or sample sizes) yield t-distributions with heavier tails.[41] The heavier tails and more spread out shapes of the Student's t-distribution are used to account for additional uncertainty in small samples. Similar characteristics can also be found in a series of the nonparametric conservative density estimates, shown at the right of Figure 11. As the sample size increases, the nonparametric conservative density estimate also approaches the true distribution. Both the Student's t-distribution and the conservative density estimate account for the uncertainty in small samples by underestimating the true distribution in the center and overestimating it on the tails. The difference is that for the conservative density estimate, there is no need to assume a parametric distribution for the sample. Also, the Student's t-distribution is already almost identical to the normal distribution when the sample size is greater than 20. The conservative density estimate can provide a model with distinguishable larger uncertainty than the true distribution at larger sample sizes.

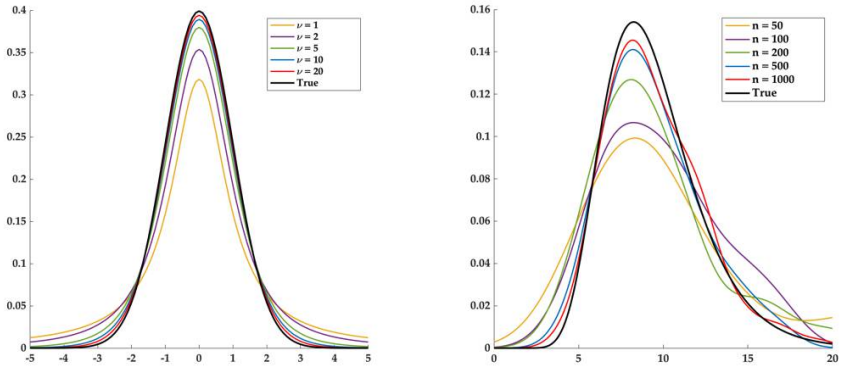

Figure 11. The Student's t-distribution (left) and the conservative density estimate (right)

C. Motivation for Dependence Modeling

The assumption of independence between random variables is common in aerospace-related probabilistic assessments. In most cases, the dependence between random variables are intentionally ignored due to the difficulty of characterizing such relationships as well as the computational expense.[4] However, the

complexity of aerospace applications often make the independence (or modest dependence) assumption between random variables unjustifiable. For example, in the aviation environmental impact analysis (which we will highlight later), there is a strong correlation between the two most significant inputs: Take-off Gross Weight (TOGW) and engine thrust. Correlations also exist between other variables such as the engine overall pressure ratio (OPR) and NO_x emission, and the engine bypass ratio (BPR) and NO_x emission, as shown in Figure 12. Neglecting dependent relationships between the correlated variables yields noteworthy changes in the probabilistic assessment results,[4] as uncertainty in the responses can be either underestimated or overestimated. The substantial effect that independence assumption may have on probabilistic assessment and uncertainty quantification results calls the need for effective and efficient dependence modeling.

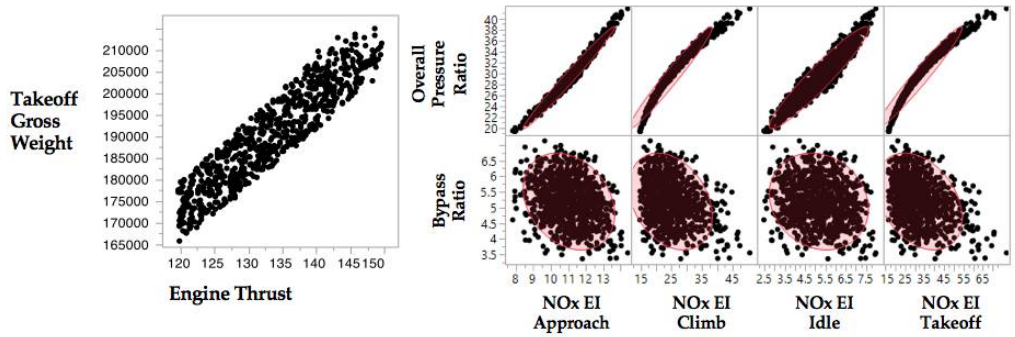

Figure 12. Correlations in aviation environmental impact analysis[42]

With small datasets, dependence modeling becomes more challenging for two reasons: 1. the inability to assign a common multivariate distribution based on limited information; 2. the need to account for additional uncertainty in small samples. To create joint probability distributions for correlated variables, *copulas* and *The Sklar's theorem* are now introduced to make the difference.

D. Copulas, Sklar's theorem and Inversion Method

The word *copula* means *tie* or *bond* in Latin. Copula is well known within the statistics literature, as it provides a natural way to measure and model dependence between random variables.[43] More specifically, copulas are functions that link one-dimensional marginal distribution functions together to form the joint multivariate distribution function.[44] When used in the form of cumulative distribution function (CDF), an n-dimensional copula is a function C with domain $[0,1]^n$ such that:[43] 1. C is grounded and n-increasing, and 2. C has margins C_k, $k = 1, 2, ..., n$ that satisfy $C_k(u) = u$, $\forall u \in [0,1]$.

The copula C can also be interpreted as a distribution function on $[0,1]^n$ with uniformly distributed margins. Copulas are important and useful because of the *Sklar's Theorem*: Let $X_1, X_2, ..., X_n$ be random variables with joint CDF

$$H(x_1, x_2, ..., x_n) = P(X_1 < x_1, X_2 < x_2, ..., X_n < x_n) \tag{22}$$

and marginal CDFs

$$F_i(x) = P(X_i \leq x), i = 1, 2, ..., n \tag{23}$$

then $(X_1, X_2, ..., X_n)^T$ has a copula C such that

$$H(x_1, x_2, ..., x_n) = C(F_1(x_1), F_2(x_2), ..., F_n(x_n)) \tag{24}$$

If $F_i(x), i = 1, 2, ..., n$ are all continuous, the copula C is unique. From the Sklar's Theorem we see that it allows us to separate the modeling of the univariate marginal distributions from the multivariate dependence structure, which is represented by a copula C. The copula density c can be estimated when C is differentiable, which happens when the marginal and joint CDFs are parametric or nonparametric smoothed,[45] given by

$$c(u_1, u_2, ..., u_n) = \frac{\partial^n}{\partial u_1 \partial u_2 ... \partial u_n} C(u_1, u_2, ..., u_n) \tag{25}$$

where $u_i = F_i(x_i), i = 1, 2, ..., n$. And the copula density satisfies

$$\frac{h(x_1, x_2, ..., x_n)}{f_1(x_1)f_2(x_2)\cdots f_n(x_n)} = c(u_1, u_2, ..., u_n) \tag{26}$$

There are three most frequently used classes of copulas: *Archimedean*, *Gaussian*, and *t-copula*. A copula is called an Archimedean copula if it has the form[43]

$$c(u_1, u_2, ..., u_n) = \phi_\theta^{[-1]}(\phi_\theta(u_1), \phi_\theta(u_2), ..., \phi_\theta(u_n)) \tag{27}$$

where ϕ is a generator function, and θ is the correlation parameter. Different generator functions define different families within the Archimedean copulas. The most commonly used families in Archimedean copula are Clayton, Frank, Gumbel, and Joe. On the other hand, a copula is called a Gaussian copula or a t-copula if F in the form

$$c(u_1, u_2, ..., u_n) = F(F_1^{-1}(u_1), F_2^{-1}(u_2), ..., F_n^{-1}(u_n)) \tag{28}$$

is a n-variate normal distribution $N_n(\mu, \Sigma)$ or a n-variate t-distribution $t_v(\mu, \Sigma)$, respectively. Among the three classes of copulas, Archimedean copula is the most popular one in the literature, because of the simplicity of mathematical formulation, and the ease to create.[4] Due to the nonparametric nature of this study, the Gaussian and t-copulas are excluded from being used to model dependence between variables. Through various experiments and tests, the *Frank family* within the Archimedean class is selected and used in dependence modeling.

Since copula works with uniformly distributed margins ($\mathbf{I}^n$ space), before a copula is fitted to data, a transformation process is taken to transform the data into the copula scale ($\mathbf{I}^2$). The *inversion method* can be used to transfer any univariate marginal distribution type into the uniform marginal distribution using that distribution's inverse CDF.[4] The inversion method can be applied to univariate samples separately without influencing the multivariate dependence structure estimated by a copula, and provides a way to specify correlations between random variables for any univariate distributions.

E. Summary of The Proposed Nonparametric Method

With the key nonparametric statistical methods and new proposed concepts introduced in the previous sections, the complete uncertainty characterization methods for small datasets are summarized in Figure 13-14. The main outcomes of this step include an optimal density estimate and a conservative density estimate. When the uncertain variable is independent, method in Figure 13 is applied on one-dimensional dataset for the estimation. The method in Figure 14 is used on two-dimensional dataset when the dependence between two variables is significant.

In consistent with the 1-D case, the 2-D uncertainty characterization for small datasets also yields two outcomes: a joint optimal density estimate, and a joint conservative density estimate, as shown in Figure 15. The joint optimal density estimate is created by the two marginal optimal density estimates and a copula through inversion method. Similarly, the joint conservative density estimate is created by the two marginal conservative density estimates and the same copula. The conservative joint density estimate extends the idea of the maximum uncertainty (or maximum entropy) estimate from 1-D to 2-D, and is used to reflect the current state of knowledge with maximum uncertainty based 2-D small datasets.

Due to two different density estimates from uncertainty characterization, two separate runs of uncertainty propagation are needed to provide corresponding estimates of the output distributions. Among various uncertainty propagation approaches, the Monte Carlo approach is used most widely. Figure 16 illustrates how the Monte Carlo approach is applied in this study to propagate the uncertainties represented by the two sets of input estimates. Surrogate models are utilized to reduce the computational effort when the original system modeling code is complex. When sampling inputs from the optimal density estimates and propagating uncertainty through the system model, the resulting output distributions also serve as the optimal estimates. In the second run, the inputs are sampled from the conservative density estimates, and the resulting output distributions serve as the conservative estimates. It can be imagined that if the inputs are defined by the conservative estimates with more uncertainty, the conservative output results, as a function of inputs, will too contain more uncertainty than the optimal results. As the dataset size grows, the difference between the optimal and the conservative results is expected to diminish.

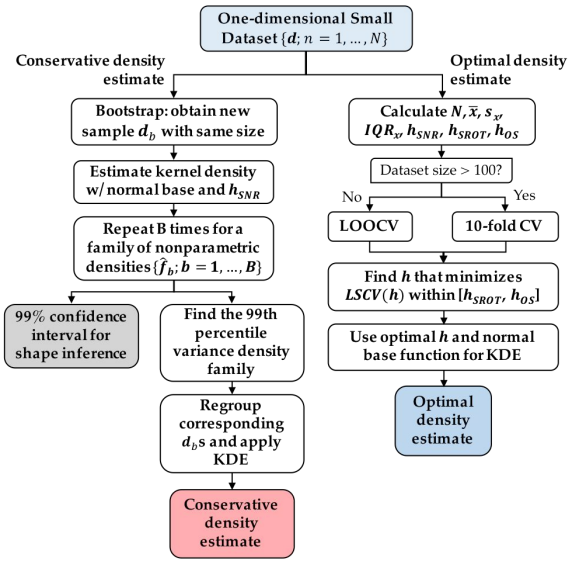

Figure 13. Flowchart of the 1-D nonparametric uncertainty characterization method for small datasets

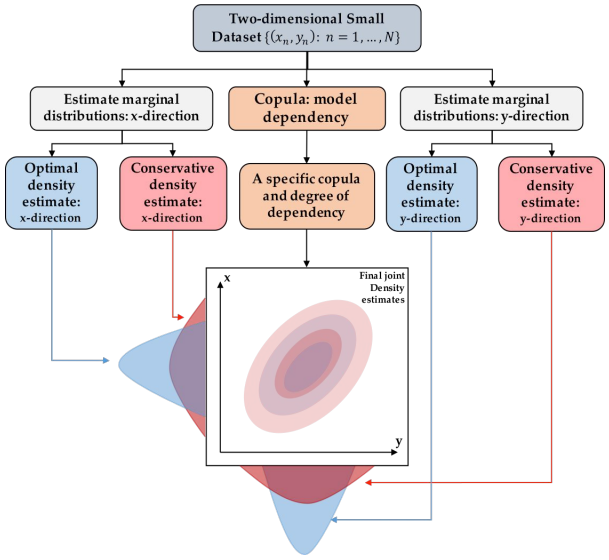

Figure 14. Flowchart of the 2-D nonparametric uncertainty characterization method for small datasets

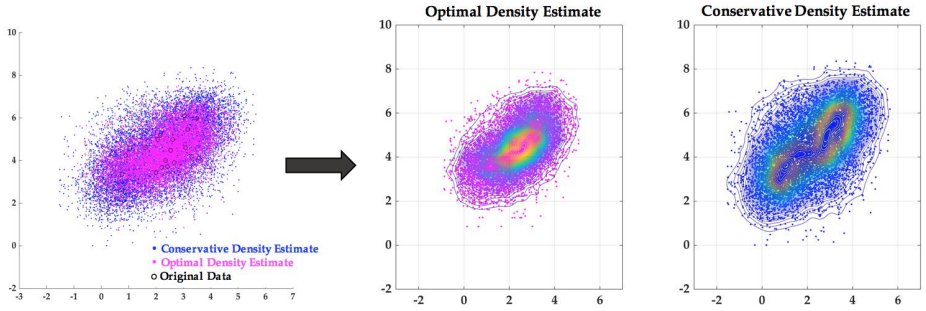

Figure 15. The joint optimal and conservative density estimates

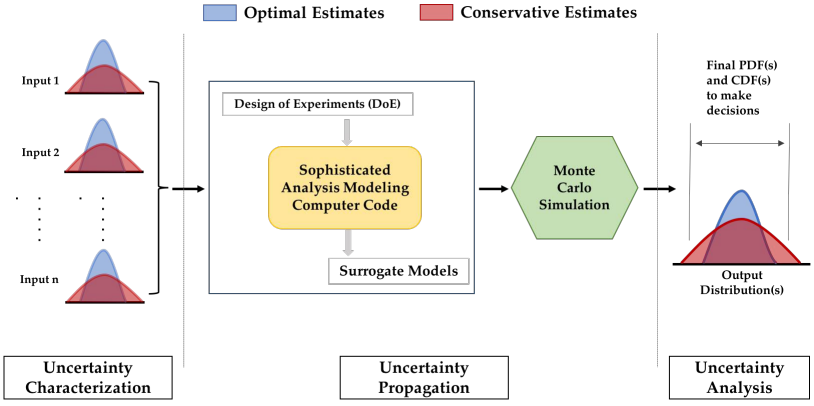

Figure 16. MCS-based uncertainty propagation for small datasets

IV. Uncertainty Characterization and propagation Illustrations

The proposed methods in Figure 13-14 need to be illustrated through different cases. The following sections present five illustrations (three for 1-D and two for 2-D) on uncertainty characterization and one illustration on uncertainty propagation based on small datasets. Observations that provide evidence on the effectiveness of the methods are made based on the results.

A. Uncertainty Characterization Illustrations

Density estimation is at the core of uncertainty characterization based on small datasets. In the uncertainty characterization illustrations, our target is to provide effective estimates of the true density using small datasets. The following questions should be considered while assessing the performance of the proposed methods:

1. How well does the optimal density estimate represents the true density based on small datasets? This can be measured by comparing statistical moments like the mean, variance, skewness, etc.

2. How valid is the conservative density estimate? What will happen to the two density estimates (optimal and conservative) when more data are added in?

3. How well does the nonparametric method estimate multimodal distribution?

4. How well does the joint estimates from 2-D method capture the dependency and estimate the joint distribution of two dependent variables?

The first illustration uses a Normal distribution as the true density, because it is the most commonly used distribution. Figure 17 shows the histograms and corresponding density estimates for six small datasets with different sizes. Each time, we create a small dataset by randomly generating N data from the true density. The proposed uncertainty characterization method is then applied on the small dataset to generate density estimates. Density plots in Figure 17 give visual comparisons between the true density (solid black curve), optimal density estimate (solid magenta curve), conservative density estimate (dotted magenta curve), and 99% CI (gray area). It is clearly seen that each conservative density estimate has a more spread out shape compared to the corresponding optimal density estimate. When more data are gradually added in (dataset size grows from 20 to 1000), the 99% CI narrows, and the conservative density estimate converges towards the optimal density estimate. This evolution illustrates the reduction of uncertainty as more data are added. Statistical moments of the above estimates are summarized in Table 2 presented in Appendix. It can be seen from Table 2 that the optimal density estimate has a good overall performance in estimating the mean, standard deviation, and skewness of the true density base on this group of small datasets. More importantly, the conservative density estimate always has a reasonably larger standard deviation compared to the corresponding optimal density estimate, while maintaining close estimations in mean and skewness.

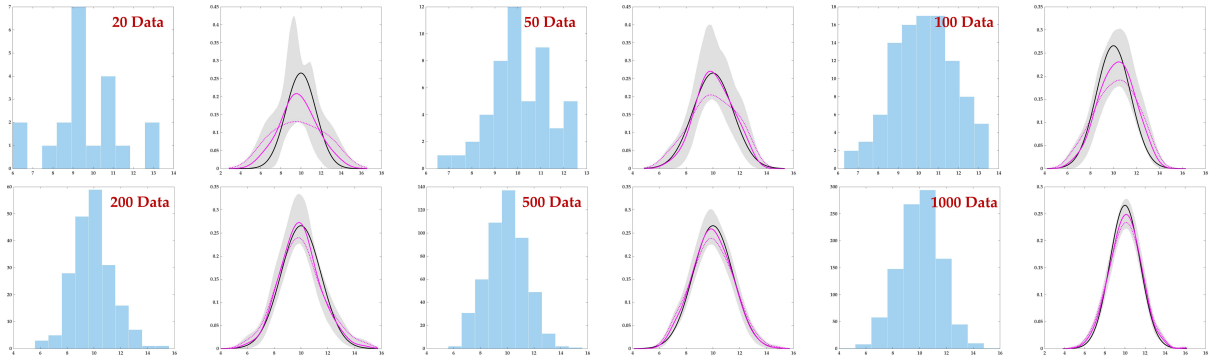

Figure 17. Illustration 1: Normal Distribution, density estimates from different dataset sizes

The second illustration uses a Lognormal distribution as the true density, as it has a non-zero skewness compared to the Normal distribution. Similar visual and statistical moments comparisons are provided by

Figure 18 and Table 3 in Appendix respectively. From the visual comparison, we observe similar characteristics and trends as the first illustration. Table 3 shows that both the optimal and the conservative density estimates have skewness values with the correct sign (+ or -) even at very small dataset sizes. This is an indication that the nonparametric method has the capability of estimating skewed distributions, which is common in numerous applications.

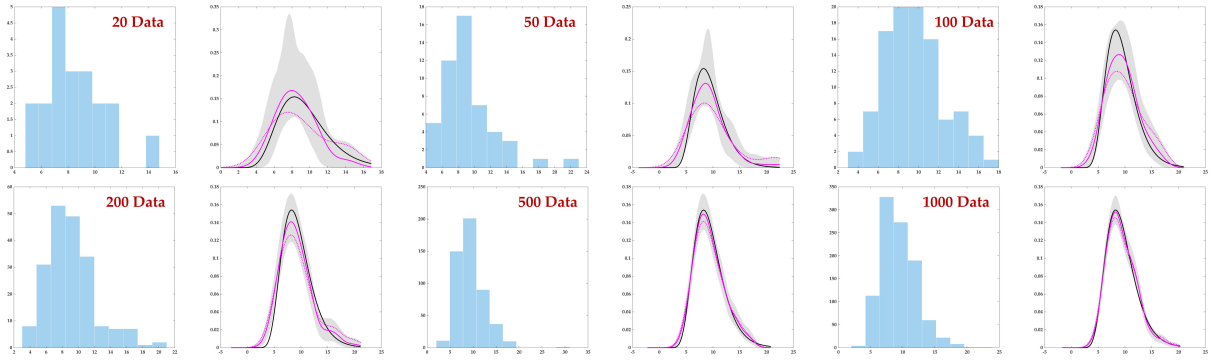

Figure 18. Illustration 2: Lognormal Distribution, density estimates from different dataset sizes

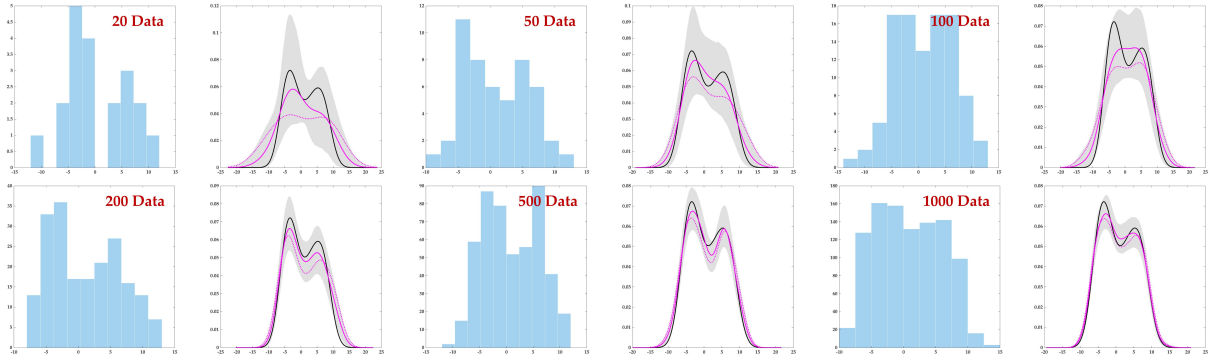

Figure 19. Illustration 3: Multimodal Distribution, density estimates from different dataset sizes

The third illustration uses a multimodal distribution as the true density. This illustration is specifically designed for the proposed nonparametric method, because it is beyond the scope of the parametric method in the literature. In this illustration, we only compare different density estimates visually in Figure 19. The result shows that when certain amount of data is collected (200 in this example), the method is capable of characterizing multimodal distributions.

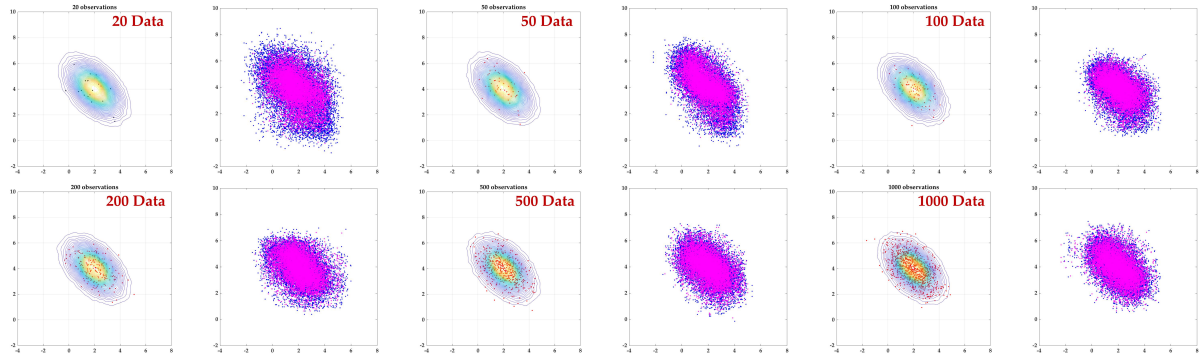

Figure 20. Illustration 4: Weak linear correlation (negative), density estimates from different dataset sizes

Two other illustrations (illustration 4 and 5) are presented for the 2-D uncertainty characterization method. Since the primary target for the 2-D method is to capture the correlation and estimate the joint

distribution of dependent variables, two extreme cases (very strong and very weak linear relationships) are used here. Figure 20 shows the result of applying the proposed method on 2-D small datasets to estimate a joint normal distribution that describes two weakly correlated variables. Two observations need to be noted here. First, across all the dataset sizes, the method appears to effectively capture the weak correlation between the two variables. Second, compared to the optimal density estimate (magenta), the conservative density estimate (blue) is apparently more spread out in both dimensions when the dataset size is small. The conservative density estimate again converges towards the optimal density estimate as the dataset size increases.

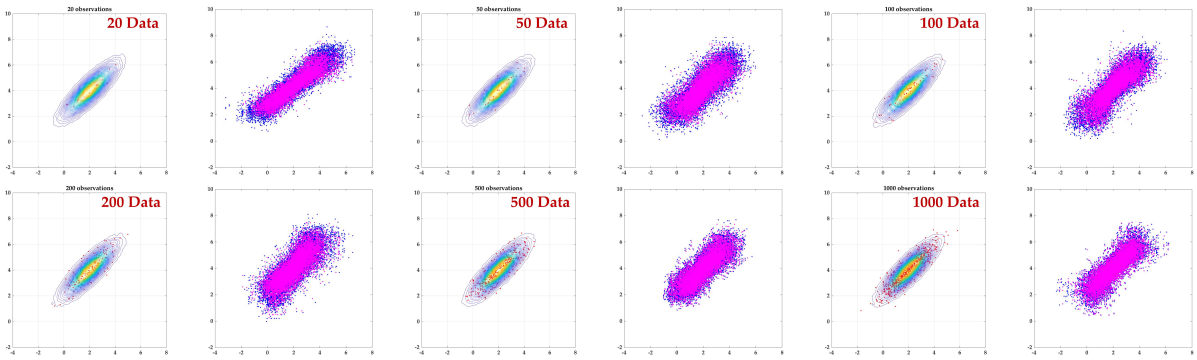

Figure 21. Illustration 5: Strong linear correlation, density estimates from different dataset sizes

A summary of statistical moments for the fourth illustration is given in Table 4 in Appendix. Result in Table 4 shows that in 2-D case, the overall performance of both the optimal density estimate and conservative density estimate meet the expectations as previously defined. Similar results can also be observed in the fifth illustration (strongly correlated variables) as shown in Figure 21 and Table 5 in Appendix.

B. Uncertainty Propagation Illustration

A complete uncertainty propagation through a complex system model is also conducted. The model is a response surface model that has nine inputs (five independent and four dependent) and one output. Assumed true distributions for the inputs are summarized in Table 6 in Appendix. At each dataset size, small datasets for the inputs are created by randomly sampling N data from their respective true distributions. Proposed uncertainty characterization methods in Figure 13-14 are then applied on the small datasets, and the output distributions are estimated through the uncertainty propagation process shown in Figure 16. Output distribution results at six different dataset sizes are shown in Figure 22. It can be observed that when only 20 data are available for characterizing each uncertain input, variability of the conservative estimate of the output distribution is apparently larger than the optimal estimate. This indicates a large degree of uncertainty caused by very small datasets. As the dataset size grows, while the optimal estimate maintains close estimation of the true output distribution, the conservative estimate gradually converges to the optimal estimate.

V. Application to Uncertainty Quantification in Aviation Environmental Impact Analysis

A. Application Formulation

In the past half-century, the development of global air transportation have forever changed the way people travel and deliver cargo. In the 2035 time frame, air transportation is expected to keep its growth momentum. Forecast released by The international Air Transportation Association (IATA) predicts a 3.7% annual Compound Average Growth Rate (CAGR) in the next 20 years, and the number of air travelers is predicted to double during this period.[46] While enjoying the convenience of global air transportation, its negative impact on the environment has become a major concern internationally. Emissions such as nitrogen oxides (NOx), sulfur oxides (SOx), and green-house gases (GHG) can exacerbate health-harming air pollution and climate change. Accurate modeling of aircraft fuel burn and emissions is a key factor in informing a number

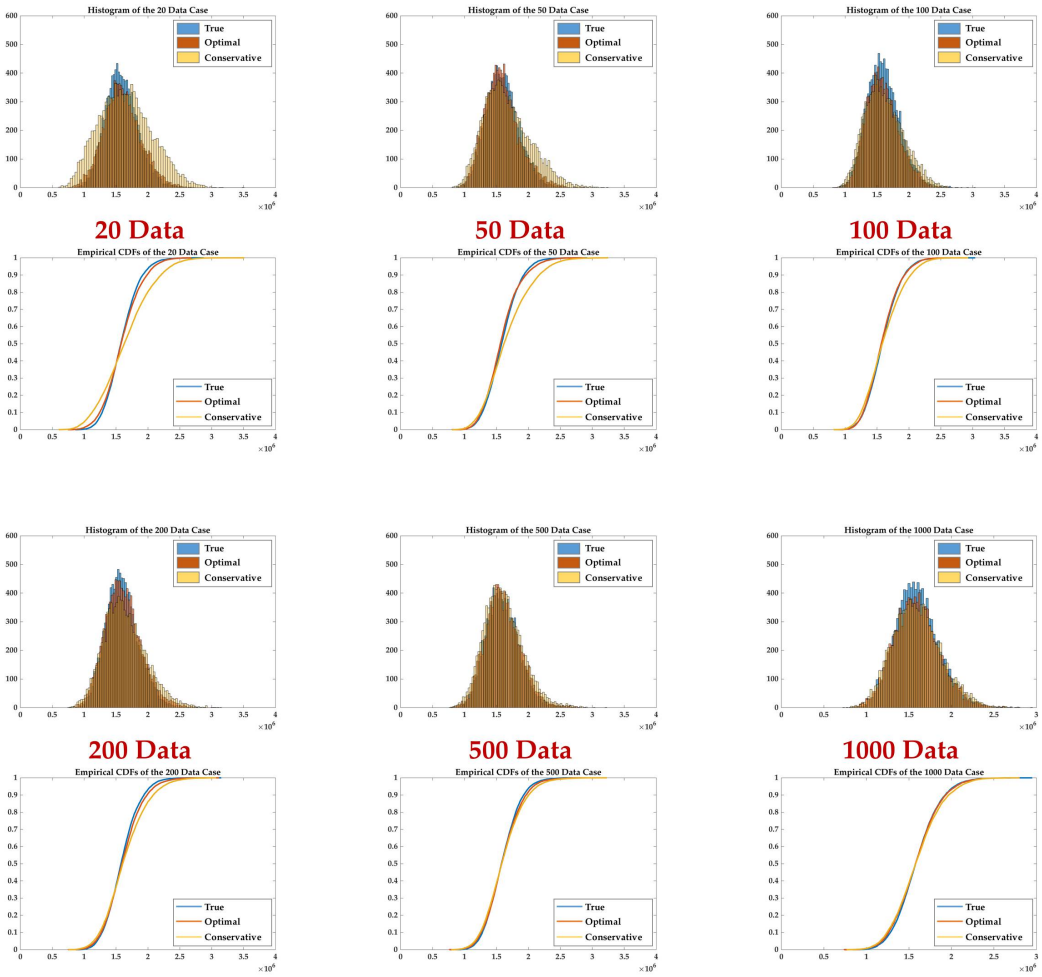

Figure 22. Uncertainty propagation illustration through a system model

of new operational procedures and policies to abate negative environmental impacts from air transportation. The Aviation Environmental Design Tool (AEDT) developed by the Federal Aviation Administration's Office of Environment and Energy (FAA/AEE) is a comprehensive software package which models aircraft operations in time and space to calculate the fuel burn, emission, and noise.[47] Within environmental impacts modeling, the modeling of departure operations around airports is of great interests to policy makers, airport authorities, and the communities around the airports. Sensitivity analysis in a previous research shows that accurate calculations of departure fuel burn, NOx emissions, and Sound Exposure Level (SEL) noise contour areas are significantly affected by two AEDT input parameters: takeoff weight and takeoff thrust.[48] Improved quantification and propagation of the uncertainties in takeoff weight and takeoff thrust is crucial in modeling the uncertainties in the stated environmental impacts around airports.

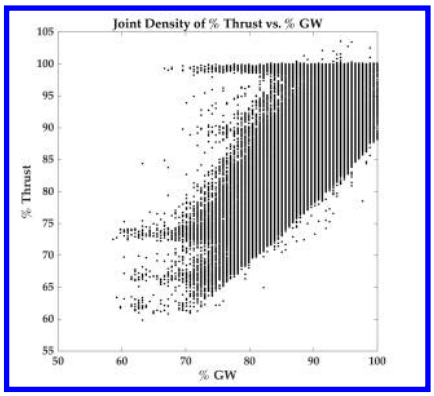

Figure 23. Joint distribution of % Thrust vs. % TOGW at all airports

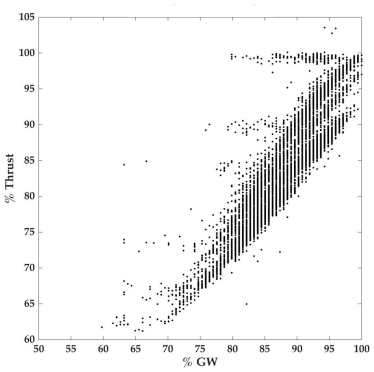

Figure 24. Joint distribution of % Thrust vs. % TOGW at a selected airport

Flight data collected from airlines provides a main source for analyzing the distributions of takeoff weight and thrust in real-world operations. Figure 23 depicts the joint distribution of percent takeoff thrust and takeoff gross weight (TOGW) for an airline's aircraft Boeing 737-800 throughout a year, with a total of 62,493 data. The % Thrust is the ratio between the takeoff thrust used and the maximum takeoff thrust available at the given atmospheric condition for each of the flights. The % TOGW is the ratio of the TOGW of the given flight to the maximum certified TOGW. Such a distribution can reflect the reality because most airlines practice reduced thrust takeoff, in which only part of the full thrust available is used to extend engine life and save maintenance cost.[47] The extent of thrust reduction varies from 1% to 40%, and is determined by several factors, including aircraft type, takeoff weight, airport, and weather. The complete population of data shown in Figure 23 takes a year to collect. In addition, if a more comprehensive nationwide study is wanted, takeoff weight and thrust data must be collected from different airlines and locations for all aircraft types. All these factors contribute to the difficulty of collecting complete and sufficient data for relevant studies, making small dataset a reality. Especially in some preliminary studies, people usually have limited time budget or resources for data collection, and have to do estimations based on a small subset of data. In this section, the proposed uncertainty characterization and propagation approach for small datasets is demonstrated via a hypothetical study that estimates the distributions of fuel burn and NOx emissions throughout a year for Boeing 737-800 at an airport using one day's data.

Looking at the joint distribution of % Thrust and % TOGW, it can be observed that the majority of the data displays strong positive correlation between the two inputs. Yet overall, this distribution is a difficult one to be estimated from small datasets. Unlike more common joint distributions shown in Figure 12, the distribution in Figure 23 has several boundaries due to the operation characteristics. For example, existence of the maximum and minimum thrust limits constraints the distribution within between 60%-100% in percent thrust. Some other performance limits (oblique boundaries in Figure 23) and randomnesses in real operations further complicate the distribution. Without knowing most of such contexts, it is extremely difficult to estimate the whole population from small datasets.

Figure 24 is a subset of the population shown in Figure 23, and contains 7,344 data from one selected airport. Compared to the whole population that reflects nationwide operations of this airline, data from this airport displays a stronger linear correlation between % thrust and % TOGW, and in general has relatively larger % TOGW levels.

B. Application Results

To create small datasets, four days are randomly selected from different seasons between March 2014 and March 2015, with 24 - 38 out of 7,344 data on each day (0.32% - 0.52%). Then, the proposed uncertainty characterization method is applied to the small datasets. The four groups of original data and their corresponding density estimates are shown in Figure 25. In Figure 25, both the optimal (magenta) and conservative (blue) estimates are plotted against distribution of the true population (black). Due to different number of flights and the differences in the frequency of certain air routes on each day, the four groups of estimations are slightly different from each other. Overall, the marginal estimates are not far from the true marginal distributions, and the correlation between the two inputs is well captured. In the next step, the joint density estimates from December 15, 2014 is used in an uncertainty propagation process through AEDT, to estimate the distributions of fuel burn and NOx emission throughout the whole year. Results of the uncertainty propagation are shown in Figure 26-27 and Table 1 below.

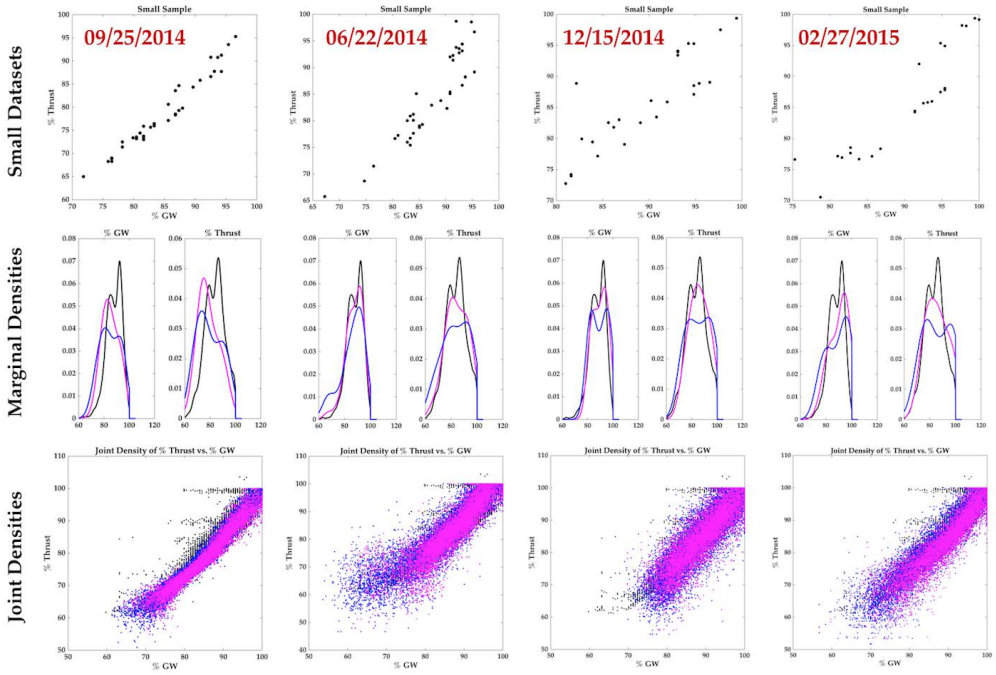

Figure 25. % Thrust vs. % TOGW: Small datasets, marginal densities and joint densities for four selected days

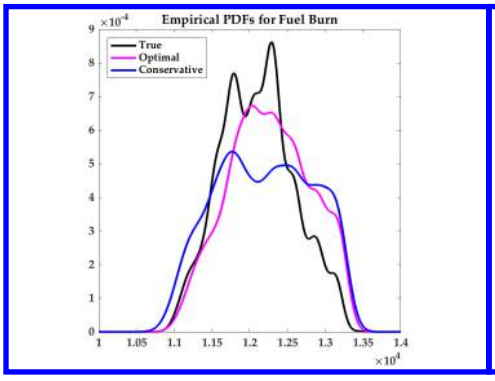

Figure 26. Fuel burn distribution results

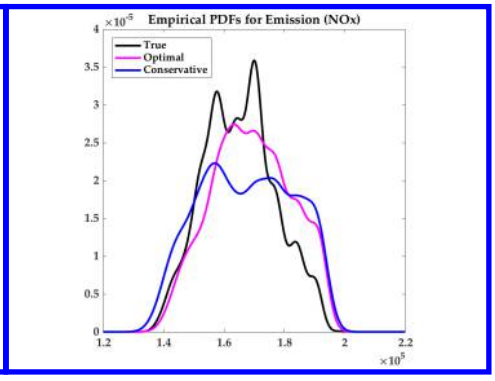

Figure 27. NOx distribution results

The uncertainty propagation results show that compared to the true distributions of fuel burn and NOx obtained by propagating the true distribution (the whole population) of % thrust and % TOGW, this group

Table 1. Comparison between the true and optimal results

	Statistics	True	Optimal	% Difference
Fuel Burn	Mean	1.2101×10^4	1.2259×10^4	1.31
(kg)	Standard Deviation	4.9099×10^2	5.4564×10^2	10.89
Emission	Mean	1.6524×10^5	1.6909×10^5	2.33
(g)	Standard Deviation	1.1933×10^4	1.3196×10^4	10.58

of the optimal results obtained by propagating the optimal density estimate perform well in estimating the mean values of the two true distributions. For the standard deviation, the optimal estimates are around 10% larger than the true standard deviation values, indicating that the uncertainties in the outputs are overestimated. Nevertheless, it is worth emphasizing that this study case has a quite difficult setting, as: 1. the distribution of % thrust and % TOGW in Figure 24 is difficult to be estimated without knowing relevant operations contexts; 2. data from one day is a too little portion of the whole population. The most important objective of this work is that, if such a small dataset is all that we have, an approach has to be found to still do the job. Methods proposed by this work are able to provide such a possibility.

C. Discussion: what if the sample is not representative?

In an earlier section where bootstrap is introduced, it is pointed out that bootstrap works well when the sample is representative of the population. In fact, a representative sample is vital to all methods for small datasets. Although this is somewhat difficult to judge in some situations, people who are not able to acquire a complete dataset should make use of the contextual knowledge when selecting data, or strictly control the experiment process that produces data, to make the sample a representative one. Going back to the complete population of % Thrust and % TOGW data at all airports shown in Figure 23, what subset of the population should be selected as a representative dataset when it is impossible to obtain the complete data? In aviation impact study, two common constraints that prevent us from obtaining the complete dataset here are time and space. For example, people may only collect such data from certain days instead of a whole year, or from certain airports instead of the whole nation. Then knowing the data is provided by only one airline, which subset is more representative: the one from a certain day, or one from a certain airport?

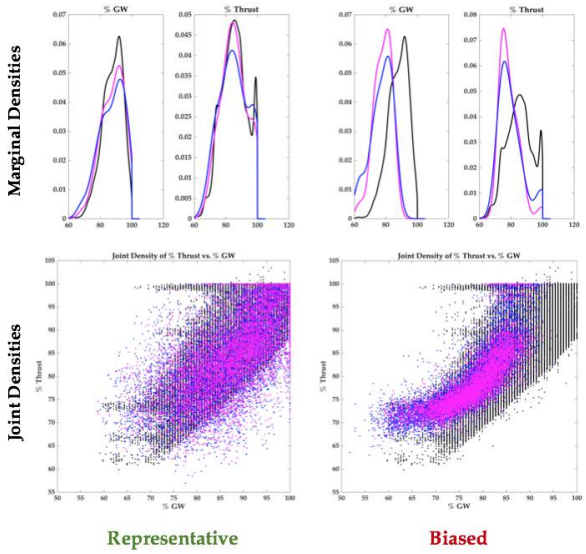

Figure 28. Results from representative and biased small datasets

In the first example, the distribution in Figure 23 is estimated from data from a randomly selected day, December 06, 2014, with 209 out of 62,493 data (0.33%). Results of the marginal and joint density estimates

are plotted against the true distributions and shown at the left of Figure 28. These results suggested that in this case, a subset of the population constrained by time is relatively representative, because our density estimates (magenta and blue) are close to the true distributions (black) in both marginal and joint results. More importantly, in the joint result, the density estimates have a relatively good coverage of the area of the true joint distribution.

In the second example, the same population distribution is estimated from data from another selected airport, with 59 out of 62,493 data (0.09%). Now, results at the right of Figure 28 show that data from this airport is not a representative sample. In the marginal results, our density estimates tend to have lower means than the true marginal distributions. In addition, in the joint result, the joint density estimates can only partially cover the the area of the true distribution. Since an airline's operations at different airports may have different features, the data provider may mainly operate short haul flights at this airport, which makes the sample not representative of its nationwide operations.

VI. Conclusions

This paper presents a nonparametric-based approach to characterize and propagate epistemic uncertainty resulting from small datasets. Under the nonparametric framework, the proposed method has three advantages. First of all, there is no need to manually supply a group of candidate parametric probability models based on limited knowledge and judgments. We avoid such assumptions that may take additional time and resources, and are in some cases, impossible to make. The second advantage is that the method does not introduce any unwarranted information that is inconsistent with the current knowledge, and the results are loyal to the available data. Lastly, the method has the capability of estimating unconventional distributions, such as multimodal distributions, which can not be estimated by the current parametric approach. Another crucial contribution of this work is the proposed conservative density estimate. All the prior works in the literature have been focusing on a single estimate in uncertainty quantification based on small datasets, which can not reflect the additional uncertainty that is inherent specifically in small datasets. Compared to the optimal estimate, the conservative estimate is another dimension that tries to reflect the current state of knowledge with the maximum uncertainty. By propagating the conservative estimates, corresponding estimates of the output distributions give people an idea of how large the uncertainty can be under the influence of small datasets.

Several uncertainty characterization illustrations and an application in aviation environmental impact analysis demonstrated the effectiveness of the proposed method through various typical cases, and showed that the method has strong potential for solving practical problems in many fields. The greater flexibility in the nonparametric methods makes it worth the effort to further make the framework more complete in many details. One potential avenue for future work is to apply boundary correction for nonparametric density estimation to represent uncertain variables with physical boundaries, which could happen in many real applications.

Appendix

Table 2. Illustration 1: Normal Distribution, comparison of the statistical moments

20 Data	True	Optimal	Conservative	200 Data	True	Optimal	Conservative
Mean	10	9.74	9.79	**Mean**	10	9.93	9.97
Standard Deviation	1.5	2.00	2.69	**Standard Deviation**	1.5	1.59	1.81
Skewness	0	0.05	0.03	**Skewness**	0	0.22	0.24
50 Data	**True**	**Optimal**	**Conservative**	**500 Data**	**True**	**Optimal**	**Conservative**
Mean	10	10.11	10.04	**Mean**	10	9.93	9.93
Standard Deviation	1.5	1.47	1.81	**Standard Deviation**	1.5	1.53	1.66
Skewness	0	-0.04	-0.15	**Skewness**	0	0.05	0.07
100 Data	**True**	**Optimal**	**Conservative**	**1000 Data**	**True**	**Optimal**	**Conservative**
Mean	10	10.26	10.21	**Mean**	10	10.03	10.01
Standard Deviation	1.5	1.65	1.93	**Standard Deviation**	1.5	1.61	1.71
Skewness	0	-0.12	-0.15	**Skewness**	0	-0.03	0.02

Table 3. Illustration 2: Lognormal Distribution, comparison of the statistical moments

20 Data	True	Optimal	Conservative	200 Data	True	Optimal	Conservative
Mean	9.44	8.50	9.10	**Mean**	9.44	9.11	9.42
Standard Deviation	2.89	2.44	3.48	**Standard Deviation**	2.89	3.25	3.79
Skewness	0.95	0.45	0.36	**Skewness**	0.95	0.84	0.84
50 Data	**True**	**Optimal**	**Conservative**	**500 Data**	**True**	**Optimal**	**Conservative**
Mean	9.44	9.30	10.18	**Mean**	9.44	9.38	9.54
Standard Deviation	2.89	3.67	5.20	**Standard Deviation**	2.89	3.04	3.59
Skewness	0.95	1.11	0.96	**Skewness**	0.95	1.07	1.89
100 Data	**True**	**Optimal**	**Conservative**	**1000 Data**	**True**	**Optimal**	**Conservative**
Mean	9.44	9.53	9.80	**Mean**	9.44	9.43	9.50
Standard Deviation	2.89	3.13	3.69	**Standard Deviation**	2.89	2.89	3.11
Skewness	0.95	0.39	0.34	**Skewness**	0.95	0.77	0.89

Table 4. Illustration 4: Weak linear correlation, comparison of the statistical moments

20 Data	True	Optimal	Conservative	200 Data	True	Optimal	Conservative
Mean: μ_x	2	1.79	1.82	**Mean:** μ_x	2	2.08	2.08
Mean: μ_y	4	3.89	3.75	**Mean:** μ_y	4	3.95	3.92
Covariance: σ_{xx}	1	1.18	2.12	**Covariance:** σ_{xx}	1	1.06	1.38
Covariance: σ_{xy}	-0.5	-0.53	-1.05	**Covariance:** σ_{xy}	-0.5	-0.41	-0.53
Covariance: σ_{yy}	1	1.33	2.72	**Covariance:** σ_{yy}	1	0.98	1.25
50 Data	**True**	**Optimal**	**Conservative**	**500 Data**	**True**	**Optimal**	**Conservative**
Mean: μ_x	2	1.82	1.83	**Mean:** μ_x	2	2.01	1.99
Mean: μ_y	4	4.19	4.09	**Mean:** μ_y	4	3.99	3.99
Covariance: σ_{xx}	1	1.03	1.45	**Covariance:** σ_{xx}	1	1.05	1.22
Covariance: σ_{xy}	-0.5	-0.65	-1.02	**Covariance:** σ_{xy}	-0.5	-0.47	-0.55
Covariance: σ_{yy}	1	1.24	2.06	**Covariance:** σ_{yy}	1	0.97	1.17
100 Data	**True**	**Optimal**	**Conservative**	**1000 Data**	**True**	**Optimal**	**Conservative**
Mean: μ_x	2	2.11	2.11	**Mean:** μ_x	2	1.97	1.97
Mean: μ_y	4	3.89	3.81	**Mean:** μ_y	4	3.99	3.98
Covariance: σ_{xx}	1	0.88	1.21	**Covariance:** σ_{xx}	1	1.02	1.18
Covariance: σ_{xy}	-0.5	-0.39	-0.56	**Covariance:** σ_{xy}	-0.5	-0.52	-0.59
Covariance: σ_{yy}	1	0.85	1.22	**Covariance:** σ_{yy}	1	1.07	1.23

Table 5. Illustration 5: Strong linear correlation, comparison of the statistical moments

20 Data	True	Optimal	Conservative	200 Data	True	Optimal	Conservative
Mean: μ_x	2	2.01	2.03	**Mean:** μ_x	2	1.99	1.99
Mean: μ_y	4	4.18	4.25	**Mean:** μ_y	4	4.02	4.06
Covariance: σ_{xx}	1	1.73	2.81	**Covariance:** σ_{xx}	1	1.11	1.41
Covariance: σ_{xy}	0.8	1.19	2.02	**Covariance:** σ_{xy}	0.8	0.92	1.17
Covariance: σ_{yy}	1	1.01	1.74	**Covariance:** σ_{yy}	1	1.22	1.57
50 Data	**True**	**Optimal**	**Conservative**	**500 Data**	**True**	**Optimal**	**Conservative**
Mean: μ_x	2	1.91	1.93	**Mean:** μ_x	2	2.03	2.04
Mean: μ_y	4	3.99	4.01	**Mean:** μ_y	4	3.99	4.01
Covariance: σ_{xx}	1	1.22	1.92	**Covariance:** σ_{xx}	1	1.08	1.23
Covariance: σ_{xy}	0.8	0.97	1.48	**Covariance:** σ_{xy}	0.8	0.87	0.99
Covariance: σ_{yy}	1	1.21	1.75	**Covariance:** σ_{yy}	1	1.06	1.21
100 Data	**True**	**Optimal**	**Conservative**	**1000 Data**	**True**	**Optimal**	**Conservative**
Mean: μ_x	2	2.04	2.07	**Mean:** μ_x	2	2.02	2.03
Mean: μ_y	4	4.04	4.01	**Mean:** μ_y	4	4.01	4.03
Covariance: σ_{xx}	1	1.19	1.65	**Covariance:** σ_{xx}	1	1.12	1.23
Covariance: σ_{xy}	0.8	0.99	1.41	**Covariance:** σ_{xy}	0.8	0.92	1.01
Covariance: σ_{yy}	1	1.21	1.72	**Covariance:** σ_{yy}	1	1.11	1.22

Table 6. True distributions of the 9 input variables in the complete uncertainty propagation test

Input Name	Dependency Status	True Distribution
x_1	Independent	$X \sim Lognormal(5.0, 0.30^2)$
x_2	Independent	$X \sim Lognormal(4.0, 0.20^2)$
x_3	Independent	$X \sim Normal(300, 40^2)$
x_4	Independent	$X \sim Gamma(80, 2.5)$
x_5	Independent	$X \sim Weibull(180, 6)$
x_6	Dependent	x of $X \sim MVN([100, 120], [1\ 0.85; 0.85\ 1])$
x_7	Dependent	y of $X \sim MVN([100, 120], [1\ 0.85; 0.85\ 1])$
x_8	Dependent	x of $X \sim MVN([60, 80], [1\ -0.5; -0.5\ 1])$
x_9	Dependent	y of $X \sim MVN([60, 80], [1\ -0.5; -0.5\ 1])$

References

[1]de Weck, O., Eckert, C., and Clarkson, J., "A classification of uncertainty for early product and system design," *International Conference on Engineering Design, ICED'07*, August 2007.

[2]Schwartz, K. G. and Mavris, D. N., "Planning Technology Development Experimentation through Quantitative Uncertainty Analysis," *AIAA SciTech Forum - 54th AIAA Aerospace Sciences Meeting, San Diego, CA*, No. AIAA 2016-0536, AIAA, January 2016.

[3]Cai, Y., Gao, Z., Chakraborty, I., Briceno, S., and Mavris, D. N., "System-Level Assessment of Active Flow Control for Commercial Aircraft High-Lift Devices," *Journal of Aircraft*, Vol. 55, No. 3, May 2018, pp. 1200–1216.

[4]Zaidi, T., Jimenez, H., and Mavris, D., "Quantifying Random Variable Dependence Structure Through Copulas Theory for Probabilistic Assessment," *14th AIAA Aviation Technology, Integration, and Operations Conference*, No. AIAA 2014-2171, AIAA, June 2014.

[5]Allaire, D., Noel, G., Willcox, K., and Cointin, R., "Uncertainty quantification of an Aviation Environmental Toolsuite," *Reliability Engineering and System Safety*, Vol. 126, June 2014, pp. 14–24.

[6]Li, C., Mahadevan, S., Ling, Y., Choze, S., and Wang, L., "Dynamic Bayesian Network for Aircraft Wing Health Monitoring Digital Twin," *AIAA Journal*, Vol. 55, No. 3, 2017, pp. 930–941.

[7]Huan, X., Safta, C., Sargsyan, K., Geraci, G., Eldred, M. S., Vane, Z. P., Lacaze, G., Oefelein, J. C., and Najm, H. N., "Global Sensitivity Analysis and Estimation of Model Error, Toward Uncertainty Quantification in Scramjet Computations," *AIAA Journal*, Vol. 56, No. 3, 2018, pp. 1170–1184.

[8]Farhat, C., Bos, A., Avery, P., and Soize, C., "Modeling and Quantification of Model-Form Uncertainties in Eigenvalue Computations Using a Stochastic Reduced Model," *AIAA Journal*, Vol. 56, No. 3, 2018, pp. 1198–1210.

[9]Chaudhuri, A., Lam, R., and Willcox, K., "Multifidelity Uncertainty Propagation via Adaptive Surrogates in Coupled Multidisciplinary Systems," *AIAA Journal*, Vol. 56, No. 1, 2018, pp. 235–249.

[10]Sankararaman, S. and Mahadevan, S., "Distribution type uncertainty due to sparse and imprecise data," *Mechanical Systems and Signal Processing*, Vol. 37, No. 1-2, May-June 2013, pp. 182–198.

[11]Kiureghian, A. D. and Ditlevsen, O., "Aleatory or epistemic? Does it matter?" *Structural Safety*, Vol. 31, No. 2, March 2009, pp. 105–112.

[12]L Oberkampf, W. and Roy, C., *Verification and Validation in Scientific Computing*, Campridge University Press, New York, NY, January 2010.

[13]Gao, Z., *A Nonparametric-based Approach on the Propagation of Imprecise Probabilities due to Small Datasets*, Master's thesis, Daniel F. Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, USA, May 2018.

[14]Schwartz, K. G. and Mavris, D. N., "Facilitating Technology Development Progression through Quantitative Uncertainty Assessments," *14th AIAA Aviation Technology, Integration, and Operations Conference*, No. AIAA 2014-2170, AIAA, June 2014.

[15]Zhang, J. and Shields, M. D., "On the quantification and efficient propagation of imprecise probabilities resulting from small datasets," *Mechanical Systems and Signal Processing*, Vol. 98, January 2018, pp. 465–483.

[16]Roy, C. J. and Oberkampf, W. L., "A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing," *Computer Methods in Applied Mechanics and Engineering*, Vol. 200, No. 25-28, June 2011, pp. 2131–2144.

[17]Sankararaman, S. and Mahadevan, S., "Likelihood-based representation of epistemic uncertainty due to sparse point data and/or interval data," *Reliability Engineering and System Safety*, Vol. 96, No. 7, July 2011, pp. 814–824.

[18]Kiureghian, A. D., "Analysis of structural reliability under parameter uncertainties," *Probabilistic Engineering Mechanics*, Vol. 23, No. 4, October 2008, pp. 351–358.

[19]Geisser, S. and Johnson, W., *Modes of Parametric Statistical Inference*, JOHN WILEY and SONS, INC., 2006.

[20]Cox, D. R., *Principles of Statistical Inference*, Cambridge University Press, 2006.

[21]Park, I. and Grandhi, R. V., "Quantification of model-form and parametric uncertainty using evidence theory," *Structural Safety*, Vol. 39, November 2012, pp. 44–51.

[22]Halpern, E. F., Weinstein, M. C., Hunink, M. G., and Gazelle, G. S., "Representing Both First- and Second-order Uncertainties by Monte Carlo Simulation for Groups of Patients," *Medical Decision Making*, Vol. 20, No. 3, 2000, pp. 314–322, PMID: 10929854.

[23]Burnham, K. P. and Anderson, D. R., "Multimodel Inference: Understanding AIC and BIC in Model Selection," *Sociological Methods & Research*, Vol. 33, No. 2, 2004, pp. 261–304.

[24]Pradlwarter, H. J. and Schuller, G. I., "The use of kernel densities and confidence intervals to cope with insufficient data in validation experiments," *Computer Methods in Applied Mechanics and Engineering*, Vol. 197, No. 29-32, May 2008, pp. 2550–2560.

[25]MathWorks, *Kernel Distribution*, Statistics and Machine Learning Toolbox, 2017.

[26]Kvam, P. and Vidakovic, B., *Nonparametric Statistics with Applications in Science and Engineering*, JOHN WILEY and SONS, INC., 2007.

[27]Wasserman, L. A., *All of Nonparametric Statistics*, Springer, 2006.

[28]Sheather, S. J., "Density Estimation," *Statistical Science*, Vol. 19, No. 4, 2004, pp. 588–597.

[29]Silverman, B., *Density Estimation for Statistics and Data Analysis*, Statistics and Applied Probability, London: Chapman and Hall, 1986.

[30]Terrell, G. R., "The Maximal Smoothing Principle in Density Estimation," *Journal of the American Statistical Association*, Vol. 85, No. 410, June 1990, pp. 470–477.

[31]Terrell, G. R. and Scott, D. W., "Oversmoothed nonparametric density estimates," *Journal of the American Statistical Association*, Vol. 80, No. 389, March 1985, pp. 209–214.

[32]James, G., Witten, D., Hastie, T., and Tibshirani, R., *An Introduction to Statistical Learning: With Applications in R*, Springer, 2015.

[33]Bowman, A. W., "An alternative method of cross-validation for the smoothing of density estimates," *Biometrika*, Vol. 71, No. 2, August 1984, pp. 353–360.

[34]Hall, P., "Large Sample Optimality of Least Squares Cross-Validation in Density Estimation," *The Annals of Statistics*, Vol. 11, No. 4, December 1983, pp. 1156–1174.

[35]Loader, C. R., "Bandwidth Selection: Classical or Plug-In?" *The Annals of Statistics*, Vol. 27, No. 2, April 1999, pp. 415–438.

[36]Efron, B., "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, Vol. 7, No. 1, 01 1979, pp. 1–26.

[37]Silverman, B. W., *Density estimation for statistics and data analysis*, Statistics and Applied Probability, London: Chapman and Hall, 1986.

[38]Marsh, C., "Introduction to continuous entropy," Tech. rep., Department of Computer Science, Princeton University, 2003.

[39]Jaynes, E. T., "Information theory and statistical mechanics," *Physical review*, Vol. 106, No. 4, 1957, pp. 620.

[40]Lehmann, E. L., "Student and small-sample theory," *Statist. Sci.*, Vol. 14, No. 4, 11 1999, pp. 418–426.

[41]Papastathopoulos, I. and Tawn, J. A., "A generalised Students t-distribution," *Statistics & Probability Letters*, Vol. 83, No. 1, 2013, pp. 70 – 77.

[42]FAA, "Aviation Environmental Design Tool Version 2b - Uncertainty Quantification Report," Tech. rep., Federal Aviation Administration, August 2017.

[43]Embrechts, P., Lindskog, F., and McNeil, A., "Modelling dependence with copulas," Tech. rep., Department of Mathematics, ETH Zurich, 2001.

[44]Nelsen, R. B., *An introduction to copulas*, Springer Science & Business Media, 2007.

[45]Charpentier, A., Fermanian, J.-D., and Scaillet, O., "The Estimation of Copulas: Theory and Practice," Tech. rep., Center for Research in Economics and Statistics (CREST), September 2006.

[46]IATA, "20 Year Passenger Forecast," Tech. rep., The International Air Transport Association (IATA), October 2016.

[47]Lim, D., LeVine, M. J., Ngo, V., Kirby, M., and Mavris, D. N., "Improved Aircraft Departure Modeling for Environmental Impact Assessment," *2018 Aviation Technology, Integration, and Operations Conference, AIAA AVIATION Forum*, No. AIAA 2018-3503, AIAA, June 2018.

[48]Lim, D., Li, Y., LeVine, M. J., Kirby, M., and Mavris, D. N., "Parametric Uncertainty Quantification of Aviation Environmental Design Tool," *2018 Multidisciplinary Analysis and Optimization Conference, AIAA AVIATION Forum*, No. AIAA 2018-3101, AIAA, June 2018.