

A Comparison of Methods for Representing and Aggregating Uncertainties Involving Sparsely Sampled Random Variables - More Results

Vicente Romero
Sandia National Laboratories

Joshua Mullins
Vanderbilt University

Laura Swiler and Angel Urbina
Sandia National Laboratories

ABSTRACT

This paper discusses the treatment of uncertainties corresponding to relatively few samples of random-variable quantities. The importance of this topic extends beyond experimental data uncertainty to situations involving uncertainty in model calibration, validation, and prediction. With very sparse samples it is not practical to have a goal of accurately estimating the underlying variability distribution (probability density function, PDF). Rather, a pragmatic goal is that the uncertainty representation should be conservative so as to bound a desired percentage of the actual PDF, say 95% included probability, with reasonable reliability. A second, opposing objective is that the representation not be overly conservative; that it minimally over-estimate the random-variable range corresponding to the desired percentage of the actual PDF. The presence of the two opposing objectives makes the sparse-data uncertainty representation problem an interesting and difficult one. The performance of a variety of uncertainty representation techniques is tested and characterized in this paper according to these two opposing objectives. Some test problems and results are presented from a study currently underway.

CITATION: Romero, V., Mullins, J., Swiler, L. and Urbina, A., "A Comparison of Methods for Representing and Aggregating Uncertainties Involving Sparsely Sampled Random Variables - More Results," *SAE Int. J. Mater. Manf.* 6(3): 2013, doi:10.4271/2013-01-0946.

I. INTRODUCTION¹

This paper discusses and tests various statistical concepts and techniques for expressing uncertainty due to random variability (aleatory uncertainty) when limited samples exist of a random-variable quantity. Limited sampling introduces an epistemic contribution of uncertainty to the problem of random-variable characterization. The importance of this topic extends beyond experimental uncertainty characterization to situations where the derived experimental information is used for model validation or calibration purposes.

If the random quantity is fully sampled such that its characteristic probability density function (PDF) is fully known, then well-known approaches can be used to express

and work with the probabilistic uncertainty. These include Monte Carlo and Quasi Monte Carlo approaches (often involving constructed response surfaces for non-linear response functions, e.g. references [1],[2],[3],[4]); linearized-response function approaches popular in experimental uncertainty quantification (e.g. [5],[6],[7],[8],[9],[10]); and more recent approaches like polynomial chaos and stochastic expansion methods for uncertainty propagation and aggregation (e.g. [11]).

However, when relatively few samples are available a substantial epistemic contribution of uncertainty exists in addition to the aleatory uncertainty due to the quantity's random variation. This epistemic uncertainty undermines accurate estimation of the underlying variability distribution or PDF. Hence, it is not practical to pursue a goal of accurate

¹This paper is a work of the United States Government and is not subject to copyright protection in the U.S.

PDF estimation from sparse data. Rather, a pragmatic goal is that the uncertainty representation should be **conservative** so as to bound a desired percentage of the actual PDF, say 95% included probability, with reasonable reliability. A second, opposing objective is that the representation not be overly conservative; that it minimally overestimate the random-variable range corresponding to the desired percentage of the actual PDF. The presence of the two opposing objectives makes the sparse-data uncertainty representation problem an interesting and difficult one.

A classical approach to working with sparse random data is to use statistical tolerance intervals ([12], [13]) to provide a reliably conservative interval representation for the combined epistemic and aleatory uncertainty associated with the limited data. The Tolerance Interval approach is tested in this paper, along with the very common practice of simply fitting the sample data with a normal distribution. A recently developed kernel-density estimation technique specifically designed to cope with limited data ([14]) is also tested, along with a “non-parametric” cubic-spline PDF method based on a likelihood function fit to the data ([15]). Finally, taking a cue from [16] and [17] an approach is tried that uses a Johnson-family four-parameter representation of PDFs ([18]). An approach introduced in [19] does not appear to address the type of sparse-data induced uncertainty that we are concerned with here.

The sparse-data approaches are described in section III of this paper. This follows a description in section II of the uncertainty representation problem and the test matrix for assessment of the approaches. Section IV presents results and performance comparisons. Previous results from [20] are summarized and new results are

II. TEST PROBLEMS AND PROCEDURES FOR SPARSE-DATA UNCERTAINTY REPRESENTATION METHODS

The set of problems studied is represented in Figure 1. Diverse PDFs of normal, uniform, and right-triangular shapes are randomly sampled and the samples are used by the methods under investigation to identify a range that ostensibly covers the 0.025 to 0.975 percentile ranges of the PDFs. The PDFs are sized and located relative to the r_{e0} reference lines shown in the figure such that the range between the said percentiles is 2 for all PDFs and the range is centered on the reference line. Accordingly, the 0.025 and 0.975 percentiles of the normal, uniform, and triangular PDFs reside at values of -1 and 1 .

The methods are tested for their performance in estimating, from random samples of the PDFs, ranges that contain the true 0.025 to 0.975 percentile range $(-1,1)$. The estimated ranges are compared to the true ranges and method performance is assessed and characterized according to the metrics and procedures presented in section IV of this paper.

Given our resource constraints the study involves sample sizes of $n = 2, 8$, and 32 samples from each PDF. Since the increase in sample size is a constant factor (of 4) each time, a rate of decrease of method estimation errors with increasing sample size can be calculated, but this is not done in this document.

The study also assesses method performance for fitting multiple sources of sparsely sampled probabilistic uncertainty (from the sets of PDFs in each row in Fig. 1) that aggregate to a total resultant uncertainty for a system. It was surmised that the aggregation of multiple sources of uncertainty may smooth out and reduce, or alternatively could amplify, the performance differences between the methods, so the test plan investigates this aspect. Since often only a handful of empirical samples of each source distribution exist, will the sparse-sample methods avoid a severe underestimation of the combined variability in the resultant exact PDFs loosely depicted at right in the figure? More specifically in this study, do the methods avoid underestimation of the true 0.025 to 0.975 percentile range of the combined variability at the system level?

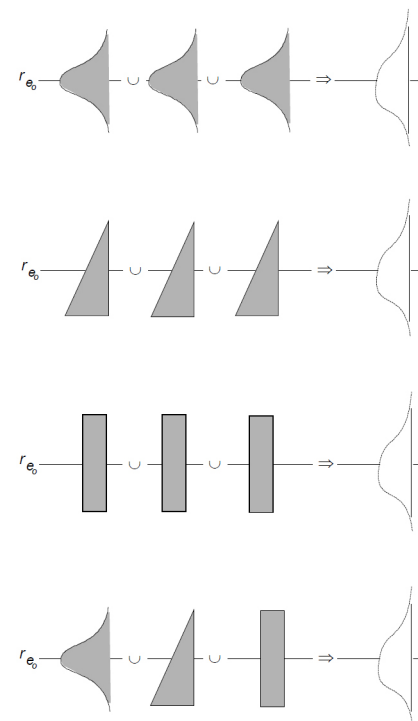


Figure 1. Test matrix for study of PDF representation performance of cited methods, under limited sample data from normal, uniform, and right-triangular test PDFs and convolutions of these as depicted in each row.

Broad experience finds that most system response quantities of project-level engineering interest are dominated by a few random variables even if many stochastic variables exist in the system. Therefore the study was designed to consider three dominant random variables having equal impact on the variability of system response (or nearly equal

impact in the case of the bottom row). A system response that is linear in the ranges of the uncertain variables is adopted to avoid any confounding of sparse-sample methodology results with errors from any inexactly captured non-linearities in uncertainty propagation procedures that would otherwise have to be employed. The three random variable sources are also specified to be independent of each other. The problem then defaults to a linear convolution problem. The four different convolution problems shown in [Figure 1](#) are studied, each with a different combination of three source uncertainties of normal, uniform, and/or right-triangular PDFs. “Exact” resultant PDF shapes are shown in [Figure 2](#). These are calculated from 10 million Monte Carlo samples as follows. A random sample from each of the three source PDFs per row in [Figure 1](#) is taken and the values of the three samples are added together. This sum constitutes one sample of the resultant PDF. Similarly 9,999,999 other samples of the resultant PDF are obtained. A Normal PDF approximation to each resultant PDF is also shown in [Figure 2](#). The Normal approximations are such that their .025 and .975 percentiles pass through the exact resultant PDFs' .025 and .975 percentiles listed in [Table 1](#). It is enlightening to see how close to Normal the resultant PDFs are from three source PDFs of normal, uniform, and/or right-triangular shapes.

Table 1. Locations of 0.025 to 0.975 percentiles of Exact Resultant PDFs from Monte Carlo Convolution of PDFs in Rows of [Figure 1](#).

PDFs convolved:	Location of 0.025 percentile	Location of 0.975 percentile
Row 1: three Normal PDFs	-1.732	1.732
Row 2: three Right-Triangle PDFs	-2.399	1.384
Row 3: three Uniform PDFs	-2.039	2.039
Row 4: one Normal, one Right Triangle, and one Uniform	-2.052	1.724

We note that the locations of the .025 and .975 percentiles of the resultant PDF from convolving the three Normal PDFs in row 1 of [Figure 1](#) can be determined analytically (exactly). The resultant is a Normal PDF with zero mean and a variance σ_{res}^2 that is the sum of the variances of the three Normal PDFs convolved. Thus the resultant standard deviation is $\sigma_{res} = [(1/1.96)^2 + (1/1.96)^2 + (1/1.96)^2]^{1/2} = 0.8837$. The 0.025 to 0.975 percentiles of the resultant Normal PDF lie at $0 \pm 1.96\sigma_{res} = (-1.732, 1.732)$. This agrees with the results in [Table 1](#) from 10 million Monte Carlo samples, providing a check on our implementation of the Monte Carlo convolution procedure.

The following procedure was undertaken for the convolution problems in the rows of [Figure 1](#) when very limited (sparse) data is available.

1. For a given row in [Figure 1](#) randomly sample each of the three PDFs with a specific number of samples prescribed in the test plan ($n = 2, 8, \text{ or } 32$). [Figure 3](#) shows an example for $n = 8$ samples and the problem depicted in the top row of [Figure 1](#).
2. Use one of the sparse-data treatment methods to estimate from the three sets of n samples (obtained in step 1) three ranges that ostensibly bound the 0.025 to 0.975 percentile range of the exact PDFs from which the samples are drawn.
3. Use the three estimated PDF percentile ranges to estimate the aggregate 0.025 to 0.975 percentile range of the resultant PDF from convolving the three PDFs in the given row in [Figure 1](#). The aggregation procedure for each method is explained in section III.
4. Using metrics and procedures from Section IV, assess and characterize the estimated aggregate percentile range against the true aggregate 0.025 to 0.975 percentile range of the exact resultant PDF.
5. Perform steps 1 - 4 1000 times (1000 “trials”) to characterize the random variability of method performance under different sets of random samples of the three involved PDFs.
6. Upon completion of step 5, 3000 estimated bounds will exist for the 0.025 to 0.975 percentile range of the source triangular, uniform, and Normal PDFs. Assess the 3000 estimated bounds against the true range $(-1,1)$ using the metrics and procedures presented in Section IV.
7. Perform steps 1 - 6 for each of the three sparse-data treatment methods.
8. Perform steps 1 - 7 for $n = 2, 8, \text{ and } 32$ samples per source PDF.
9. Perform steps 1 - 8 for each of the four rows in [Figure 1](#).

III. UNCERTAINTY ESTIMATION METHOD DESCRIPTIONS

III.A. Fitting a Normal PDF to the Sample Data

A very common approach in engineering practice is to simply fit the sample data with a normal PDF. This practice will be analyzed in Section IV for effectiveness with respect to the performance objectives previously discussed. The 0.025 and 0.975 percentiles of the fitted sample data triangular, uniform, and Normal PDFs are assessed against the true 0.025 and 0.975 percentile ranges $(-1,1)$ of the source PDFs.

To provide an estimated range for the 0.025 to 0.975 percentiles of the resultant PDF from convolving the three source PDFs depicted in a row of [Figure 1](#), the following procedure is used. For the three Normal PDF fits produced per “trial” (step 5 in Section II), their means relative to the

reference line r_{e0} in Fig. 1 are vectorially added to obtain the mean of the resultant PDF that comes from convolving the three approximating Normal PDFs. This makes use of a standard rule for convolution of PDFs. Likewise the variance of the resultant convolved PDF is obtained by summing the variances of the contributing Normal PDFs. Thus the mean and variance of the resultant PDF are easily obtained. It is also a fact that convolving Normal PDFs produces a resultant PDF that is Normally distributed. Because it is Normal, the obtained mean and variance of the aggregate PDF fully define its density function. The 0.025 to 0.975 percentile range of the aggregate PDE constructed in each trial is then compared against the exact 0.025 to 0.975 percentile range of the true aggregate PDF (from convolving the exact source PDFs in the applicable row of Fig. 1).

III.B. Tolerance Interval Method

Another simple approach is to use statistical tolerance intervals ([12], [13]) to provide a reliably conservative interval representation for the combined epistemic and aleatory uncertainty associated with sparse data. The length of tolerance intervals accounts for both the epistemic and aleatory elements of uncertainty accompanying limited data samples. Tolerance intervals (TIs) are characterized by two user-prescribed settings: one for “coverage” of a prescribed subset of the variability, and one for statistical “confidence” in covering or bounding at least that subset of the variability. For instance a .95-coverage/.90-confidence TI prescribes lower and upper values of a range of response that is said to have at least 90% odds that it covers or spans the .025 and .975 percentiles of the “true” PDF from which the random samples were drawn (for a large array of PDF types).

A 0.95/0.9 tolerance interval is constructed by multiplying the calculated standard deviation σ of the data samples by the appropriate factor f in Table 2 to create an interval of total length $2f\sigma$, where the interval is centered about the calculated mean of the data samples. The produced tolerance intervals are compared in Section IV to the true 0.025 and 0.975 percentile range $(-1,1)$ of the sampled PDF. Although constructed for Normal PDFs, 0.95/0.90 TI also span, with greater than 90% odds, the .025 to .975 percentile ranges of many other PDF types when sparsely sampled. This is shown in [21] for uniform and right-triangular PDFs and for PDFs resulting from convolutions in rows 2 - 4 in Figure 1.

Table 2. 0.95coverage/0.9assurance Tolerance Interval Factors versus Number of Samples of PDF (from [13]).

	$n = 2$ samples	$n = 8$ samples	$n = 32$ samples
factor, f	18.8	3.264	2.395

Figure 4 and Table 3 present a fuller look at how tolerance interval size varies with the number of data samples. The 0.95/0.9 tolerance intervals are very large at $n=2$ samples but quickly decrease in size when more samples are available. A knee in the rate of uncertainty decrease (per added sample) occurs somewhere between 4 to 6 samples,

with the rate of decrease being fairly small after $n=8$ samples. The tolerance interval has an asymptotic standard-deviation multiplier value of 1.96 for an infinite number of samples. This corresponds to the exact central .95 percentile range of the sampled normal PDF.

Table 3. 0.95coverage /0.90assurance Tolerance Interval Factors (standard deviation multipliers) vs. # of Samples of PDF

# samples	$f_{0.95/0.9}$
2	18.80
3	6.92
4	4.94
5	4.15
6	3.72
8	3.26
12	2.86
20	2.56
30	2.41
40	2.33
∞	1.96

The following procedure is used to provide estimated bounds on the exact 0.025 to 0.975 percentiles of the exact resultant PDFs in Figure 1. For the three tolerance intervals produced per trial (step 5 in Section II), their means are added to obtain the value of a resultant mean. This gives the center of a net tolerance interval that is compared in Section IV to the exact 0.025 to 0.975 percentile range of the true resultant PDFs. Using the classical linear-independent propagation rule of Kline & McClintock [5], the length of the net tolerance interval is the root sum of squares of the lengths of the three tolerance intervals calculated for that particular trial.

III.C. PDF Estimation by the Pradlwarter-Schuëller Kernel Density Method

Kernel density estimation (KDE, seminal papers by Rosenblatt [23], and Parzen [24]) is a technique used to estimate the density of a random variable X given n independent samples $X_1, \dots, X_n$ of it. Let $K(\cdot)$ be a kernel function which is a non-negative, real-valued, symmetric function that integrates to one $\int K(x)dx = 1$. Then the KDE is

$$f^{\{KDE\}}(x) = \hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right). \quad (1)$$

Here, h is known as the *bandwidth parameter* which controls the influence of each sample in providing a density estimate at a near-by point. Small h corresponds to a small region of influence; a large h to a large one. Common kernel functions include the Gaussian, uniform, triangular, and

Epanechnikov kernels. In this paper, we use a Gaussian kernel.

Selection of the bandwidth parameter is the subject of a vast literature. Popular methods include cross validation (see [25], [26]) and asymptotic analysis (see [26], [27]). A common approach to measuring the error in the estimation process is the mean integrated squared error (MISE):

$$MISE = E \int (\hat{f}_h(x) - f(x))^2 dx. \quad (2)$$

However, it is not feasible to minimize the MISE with respect to the bandwidth h unless the true density function $f(x)$ is known—which is not usually the case. Thus, in practice, people use cross-validation measures. In this approach, one maximizes a criterion such as the maximum-likelihood cross validation (MLCV) measure. At each observation X_i , the log-likelihood of the density at this point X_i is estimated based on all remaining observations except X_i . These log-likelihoods are then averaged over all observations:

$$MLCV = \frac{1}{n} \sum_{i=1}^n \log(\hat{f}_{-i}(X_i)). \quad (3)$$

Bandwidth estimation needs to be automated in some fashion. However, the use of automated methods such as optimization methods which maximize a log likelihood function (for example) may be subject to local optima depending on the starting bandwidth. Figure 5 shows two cases where a kernel density estimate was constructed over the same data set of four points, denoted with red X marks. The kernel density estimate shown with the green line is a KDE constructed with a small bandwidth (bandwidth value of 0.2), resulting in multimodal behavior of the output where the modes are around the actual data points. The density estimate shown with the blue line is constructed using a large bandwidth in the kernel (0.99), resulting in a smoothing of the density estimate.

In this work we did not want to use the cross-validation methods because of the small data sets (e.g. “leave one out” when you only have two data points leaves only one data point with which to construct the density estimate). We wanted to have an approach that was robust to small numbers of data points. Also, we wanted to ensure that density estimates based on the sparsest data sets were the most conservative. That is, we wanted to ensure that the tails of distributions derived from fewer data points would encompass or envelop the tails of the distributions derived from more data points. As Pradlwarter & Schuëller ([14]) state “In general, this condition is not fulfilled.” To handle low numbers of data points robustly, we used a bandwidth estimation approach that they proposed and implemented in [14]. This approach was designed to minimize the probability that future measurements will lead to data points outside the domain of existing data points. Specifically, Pradlwarter & Schuëller define the bounds a and b :

$$\begin{aligned} a &= x_{\min} - \frac{1}{2n-2}(x_{\max} - x_{\min}) \\ b &= x_{\max} + \frac{1}{2n-2}(x_{\max} - x_{\min}) \end{aligned} \quad (4)$$

where $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of the existing n data points. Given these bounds and the assumption that the existing data points are equivalent to n independent and identically distributed realizations of sample points from an unknown distribution $f(x)$, the probability that a point will fall outside the bounds a and b is:

$$p = 1 - \int_a^b f(x) dx. \quad (5)$$

The probability that n independent data points will fall

inside the domain is $\left(\int_a^b f(x) dx \right)^n$. To achieve a confidence level $(1-\alpha)$ that the probability of a new point falling outside the bounds will not exceed p , Pradlwarter and Schuëller suggest interpreting $\alpha = (1-p)^n$ as the confidence level or level of significance, and thus $P(\alpha, n) = 1 - \alpha^{1/n}$. Given this framework, the goal is to find a bandwidth h that will satisfy the following condition:

$$\int_{-\infty}^a f(x; h) dx + \int_b^{+\infty} f(x; h) dx = P(\alpha, n). \quad (6)$$

They use a Gaussian kernel, so that the KDE is specified as:

$$f^{\{KDE\}}(x) = \hat{f}(x; h) = \frac{1}{nh\sqrt{2\pi}} \sum_{i=1}^n \exp\left(-\frac{(x - X_i)^2}{2h^2}\right). \quad (7)$$

We found this formulation to be easy to optimize. We specified $\alpha=0.1$, and then calculated $P(\alpha, n)$ for a given number of data points n . Then, we found h which minimized the following expression:

$$abs \left[\left(\int_{-\infty}^a f(x; h) dx + \int_b^{+\infty} f(x; h) dx \right) - P(\alpha, n) \right]. \quad (8)$$

This is an easy optimization calculation, and appears robust to a limited number of data points. The KDE densities constructed in this fashion have the property that the density estimates are likely narrower for increasing number of points: the density estimates based on a few data points likely encompass the density estimates based on more points. Also,

the density estimates are wide and smooth when constructed with only a few points.

We constructed the KDE .025 to .975 percentile range by calculating the KDE values at input values between -5 and 5 , at increments of 0.01 . We took the KDE PDF and calculated the CDF (cumulative density function) from it. Then we constructed 10,000 random samples and interpolated where they would fall on the CDF curve (that is, we did an inverse mapping from the CDF back to the X values for the 10K samples). These 10K samples of a given KDE PDF formed the basis for locating its 0.025 to 0.975 percentiles. The percentile range was then compared against the exact 0.025 to 0.975 percentile range $(-1,1)$ of the true PDF to yield the error characterization described in Section IV.

To provide an estimated bounding range for the exact 0.025 to 0.975 percentiles of the resultant PDF from convolving the three source PDFs depicted in a row of [Figure 1](#), the 10K samples from each of the three KDE PDFs per “trial” (step 5 in Section II) were added in the following manner. The 1st sample from each of the three KDE PDFs was taken and the values of the three samples were added together. This sum constituted one sample of the resultant PDF corresponding to a convolution of the three KDE PDFs. Similarly, 9999 other such samples of the resultant PDF were obtained. The obtained 10K samples portray the resultant PDF for a given trial, 1000 of which trials were performed in the study. For each trial the 0.025 to 0.975 percentile range of the aggregate KDE PDF was compared against the exact 0.025 to 0.975 percentile range of the true resultant PDF in [Figure 2](#) to yield the error characterization described in Section IV.

III.D. PDF Estimation by the Sankararaman-Mahadevan Non-Parametric Method

The following methodology was developed by Sankararaman and Mahadevan [15] for the use of non-parametric distributions to fit point data. The following summary of the technique is paraphrased from [15]. For a more complete explanation of this approach see that reference.

Discretize the domain of X into a finite number of points, say θ_i , $i = 1: Q$. The domain is chosen based on the available data; the lowest value and the highest value are chosen as the lower bound and the upper bound of the domain, respectively. Assume that the PDF values at each of these Q points are denoted by $f_x(x = \theta_i) = p_i$ for $i = 1: Q$. Using an interpolation technique, the entire PDF $f_x(x)$ can be approximated for all $\theta \in X$, i.e. over the entire domain of X . Then the probability of observing the given point data is given by the likelihood, $L(\mathbf{p})$. This likelihood is a function of the following: (a) The discretization points selected, i.e. θ_i , $i = 1: Q$. (b) The corresponding PDF values p_i ; and (c) The type of interpolation technique used. For this work, the

discretization is fixed, i.e. uniformly spaced p_i values ($i = 1: Q$) over the domain of X are chosen in advance and the likelihood is maximized over the various values of p_i . The value of Q (number of discretization points) is chosen based on computational power—the larger the Q , the finer (in terms of flexibility) is the resulting interpolation of the PDF. In this paper, for the purpose of illustration, Q has been chosen to be equal to 11. Also, θ_1 is equal to the minimum of the available data ($X_{\min}$) and θ_Q is equal to the maximum of the available data ($X_{\max}$), and the intermediate θ 's are uniformly interspersed between $X_{\min}$ and $X_{\max}$.

The optimization problem is formulated as:

$$\begin{aligned} &\text{Given } \theta_i \in X \forall i, i = 1: Q \\ &\text{Max } L(\mathbf{p}) \end{aligned} \quad (9)$$

where $\mathbf{p} = \{p_1, p_2, p_3 \dots p_{Q-1}, p_Q\}$ & $f_x(x = \theta_i) = p_i$ subject to:

1. $p_i \geq 0 \forall i$
2. $f_x(x) \geq 0 \forall x$
3. $\int f_x(x) dx = 1$

Note: p_i at θ_i ($i = 1: Q$) is used to interpolate and calculate $f_x(x)$.

This optimization problem maximizes the likelihood function subject to three constraints. The first constraint states that the vector $\mathbf{p}$ (that contains probability values) needs to be positive. The second and third constraints state that $f_x(x)$ must be positive and the area under this curve should be equal to unity, so as to satisfy the properties of a PDF. The PDF may be constructed using various interpolation methods. Several interpolation techniques, such as linear interpolation, spline-based interpolation and Gaussian process interpolation could be used. In this paper a cubic spline is used for interpolation. The term “spline” refers to a wide variety of functions used for purposes of data interpolation and/or smoothing. Spline functions for interpolation are calculated as the minimizer of some suitable measure (e.g., the integral of the squared curvature) subject to some constraints (the interpolation constraints). For a more detailed explanation of splines, refer to Ahlberg et al. [28]. For the example presented here, eleven discretization points are chosen and the PDF is constructed. To find the .025 and .975 percentiles of the constructed PDFs and their convolved resultant PDFs, a similar procedure to the one described in the last paragraph of section III.C is performed.

III.E. Johnson PDF Method

The Johnson family of distributions was first introduced in [18] and has been further explored by many researchers. The Johnson family is comprised of four types of distributions: lognormal, normal, bounded Johnson, and unbounded Johnson. In practice, the special cases of normal

lognormal and normal distributions are rarely encountered when fitting sample data. Accordingly these two distributions were ignored in this study and each data set was fit to either a bounded or unbounded Johnson distribution. The bounded and unbounded Johnson distributions can assume many different PDF shapes, depending on the values of the four shape-defining parameters in the Johnson family of PDFs. It has been shown in [16] and elsewhere that sparse data often leads to a bounded distribution, and that result is also observed in the present study.

Four methods of estimating the Johnson parameter values are described in [29]: they are: (1) moment matching; (2) percentile matching; (3) least squares estimation; and (4) minimum L_p norm estimation. For this study the moment-matching approach was employed. The mean and the second, third, and fourth central moments of the data are calculated and equated to the respective quantities from a Johnson PDF parameterized by its four parameters. The system of four nonlinear simultaneous equations is solved to determine the four parameter values that cause the Johnson PDF's mean and moments to match those of the data set. In general the solution cannot be found analytically. A non-linear least squares (NLLS) optimization solver was used in this study. An NLLS solver is particularly well suited to this type of problem because it directly tracks the value of the entire residual vector at each iteration and calculates gradient information for each equation independently. Other solvers require a scalar objective function which would generally be some norm of the residual vector. Since it utilizes more of the residual information, the NLLS algorithm can be vastly more efficient than some of its counterparts. When it is well-behaved it converges rapidly to a solution. However, in some cases numerical noise arising from numerical integration to calculate moments of the Johnson PDFs can cause convergence issues and lead to very poor parameter estimates. In such cases it is useful to employ a gradient-free method to select a good starting point for the NLLS solver. A divided rectangles (DIRECT [30]) global optimization algorithm was utilized in this study for such cases. A steep increase in computational expense comes with such solvers, so DIRECT should not be used indiscriminately in all cases unless computational expense is of little or no concern. The solver convergence issues were rare enough that an efficient and robust algorithm was implemented which started with NLLS and switched to DIRECT on a case-by-case basis if/when convergence difficulty arose.

The study involved data sets of 2, 8, and 32 data points as previously mentioned. Numerical issues with the Johnson approach generally decrease as the number of data points increases. However, numerical issues do not completely vanish even with the relatively large sets of 32 samples. If the data set contains as few as 2 data points there is an additional complication stemming from uniqueness considerations. It can easily be shown that the third moment (skewness) of the data for any set of two samples is zero, and the fourth

moment (kurtosis) of any such set is one. These are attributable to the fact that the two samples will be on opposite sides of the mean, and equidistant from it. Then a bounded Johnson distribution will always be selected from the family. The γ parameter, which governs skewness, was automatically set to zero for all cases of two data points, and the mean was forced to fall at the midpoint of the two distribution bounds (governed by ξ and $\xi+\lambda$ respectively). Under these assumptions it is possible to achieve a converged solution for the four distribution parameters. No such assumptions are necessary when 8 or 32 data points are present. To find the .025 and .975 percentiles of the constructed Johnson PDFs and their convolved resultant PDFs, a similar procedure to the one described in the last paragraph of section III.C is performed.

IV. UNCERTAINTY ESTIMATION RESULTS, ERROR MEASURES, AND PERFORMANCE COMPARISONS

IV.A. Measures of Uncertainty Estimation Error

Figure 6 defines discrepancy or error quantities to be calculated by comparing the exact 0.025 to 0.975 percentile range of the true PDF against the percentile range calculated from the estimation methods being tested. For the i th trial the calculated discrepancies at the upper and lower ends of the exact percentile range are given by:

$$\Delta_{U-i} = r_{U-i} - r_{U-exact}$$

$$\Delta_{L-i} = r_{L-exact} - r_{L-i}$$

(10)

These equations return a positive value of discrepancy Δ_{L-i} when the estimated range extends beyond or “bounds” the true percentile at the lower end. The discrepancy Δ_{U-i} at the upper end is similarly positive when the estimated range extends beyond the true percentile at the upper end. In the opposite case where the true range extends beyond the estimated range, the discrepancy Δ is a negative value.

A primary goal for the methods here is that the estimated percentile range be conservative so as to envelope or bound the actual percentile range of the sampled PDF. Therefore, positive discrepancy values Δ from over-prediction of the actual percentile range are looked upon as more desirable errors than are negative values of Δ . Hence, we are looking for methodologies that will produce positive values of Δ_{U-i} and Δ_{L-i} with high reliability (i.e., in a high percentage of the trials). Nonetheless, a second opposing objective is that the positive values of Δ be small; that the method minimally over-estimate the true percentile range of the sampled PDF. This presence of the two opposing objectives makes the sparse-data uncertainty representation problem interesting and difficult. A clear “winner” would be the estimation method that encompasses the exact result the most times in

the trials and simultaneously has the least overshoot error (smallest average and maximum positive Δ values). Unfortunately, it will be seen that no method enjoys the best of both worlds, but some score significantly better than others over the set of attributes or measures of method performance characterized in the following.

IV.B. Example of Raw Results (for Normal PDFs in Row 1 of Figure 1)

The Normal PDFs in Figures 1 and 3 have a mean $\mu =$ zero and a standard deviation $\sigma = 1/1.96$ such that the $\mu \pm 1.96\sigma$ extents of the Normal PDFs, which mark their 0.025 to 0.975 percentiles, have values of -1 and 1 . Figures 7,8,9 plot the exact 0.025 to 0.975 percentiles at -1 and 1 , along with the first 20 of each method's estimated bounds for the said percentiles. For a fair comparison, each method is applied to the same sets of random samples. For example, the first three sets of samples for the $n=8$ trials are shown in Figure 3. These three sets of samples are used by each method to yield the first three intervals (in order left to right) in each method's plot in Figure 8.

The first 20 results from each method in Figures 7,8,9 give a glimpse of the relative performance of the methods. The Tolerance Interval (TI) method yields the largest proportion of conservative intervals at each # of trials, $n=2, 8, 32$. This will be quantified later. The Pradlwarter-Schueller (Pr-Sch) method yields the next highest proportions of conservative intervals for $n=2, 8, 32$. The Normal Fitting (NF) method, the non-parametric (NP) method, and the Johnson (JN) method all yield significantly smaller numbers of conservative intervals, even at the relatively large number of 32 samples in Fig. 9.

For $n=2$ samples the TI method yields the largest intervals on average, with many excessively large ones. The Pr-Sch method yields more moderately sized conservative intervals, although significantly more non-conservative intervals than the TI method (a detailed quantitative analysis is presented in the next section). For $n=2$ the NF, NP, and JN methods all yield a high proportion of non-conservative intervals that significantly underestimate the true .025 and .975 percentiles. Figures 8 and 9 show that the differences in interval size that the methods yield diminish markedly as the number of samples increases first to $n=8$ and then to $n=32$ samples. For $n=32$ samples the conservative intervals are actually slightly larger on average for Pr-Sch than for TI. For $n=32$ samples the NF, NP, and JN methods yield intervals that are not problematic in terms of significantly underestimating the true .025 and .975 percentiles.

Figures 10,11,12,13 show error histograms for TI, Pr-Sch, NF, and NP estimates of the 0.025 to 0.975 percentile range of the Normal source PDFs in the top row of Fig. 1 and of the convolution resultant PDF shown at top-left in Figure 2. (The histograms are from [20] where the JN method was not studied.) Each plot in Figures 10,11,12,13 has estimation errors separated into three histograms as defined next. The

integrated area under each histogram is proportional to the fraction of trial results binned into it.

- The green histograms contain the $+/+$ or “pos/pos” trial sums $S_i (= \Delta_{U-i} + \Delta_{L-i})$ where both Δ_{U-i} and Δ_{L-i} are positive in a trial. These results represent a trial in which the estimated interval bounds the true interval.
- The red histograms contain the $-/-$ or “neg/neg” trial sums $S_i (= \Delta_{U-i} + \Delta_{L-i})$ where both Δ_{U-i} and Δ_{L-i} are negative in a trial. These results represent a trial in which the estimated interval falls short of the true interval range at both upper and lower ends of the range.
- The blue histograms contain the “mixed” trial sums $S_i (= \Delta_{U-i} + \Delta_{L-i})$ where one of Δ_{U-i} and Δ_{L-i} is positive and the other is negative. These results represent a trial in which the true interval is bounded by the estimated interval at only one end or the other.

The following observations apply for the representations of the three source (Normal) PDFs and for the convolved resultant Normal PDF. The trends of method performance at $n=2$ samples (Figures 10 and 11) also generally apply at $n=8$ and $n=32$ samples (Figures 12 and 13 respectively) as will be quantified by more detailed processing in the next subsection. (The detailed processing in the next subsection will include Johnson method results even though JN histograms are not presented here.) The performance distinctions between the various methods decrease as more data samples are taken.

By observing the relative areas of the histograms it is apparent that the Pr-Sch and TI methods contain much greater proportions of positive results than do the Normal-Fitting and Non-Parametric methods. The detailed results processing in the next subsection will show that the TI method also contains significantly more positive results than does the Pr-Sch method. In a correlated result, the TI method contains significantly less negative and mixed results than all other methods.

The extents of the green $+/+$ histograms along the positive abscissa axis indicate the relative conservatism of the various methods' $+/+$ results. The greater the extent into positive territory, the more conservative on average the $+/+$ interval estimates are (this will be quantified in the next subsection). Likewise, the more negative the extents of the red $-/-$ histograms, the worse the average shortfall of the unconservative intervals.

IV.C. Detailed Processing and Comparison of Results

IV.C.1. Normal PDF results for Row 1 in Figure 1, All Methods

Relative frequencies of $+/+$, $-/-$, and mixed errors

Figures 14 and 15 plot the proportion of results in $+/+$, $-/-$, and mixed categories as a function of the number of data samples used. The solid lines show the net of taking the

number of $+/+$ results (desirable occurrences) and subtracting the number of $-/-$ results (undesirable occurrences) to get a net number of desirable results (as a percent of the total number of trials). The higher a solid line lies on the plots, the better the method's performance with respect to this measure. The Tolerance Interval method always ranks best by a considerable margin according to this measure, followed by the Pr-Sch method, then generally the Normal Fitting method, then the Non-Parametric method, and the Johnson method last. The NF, NP, and JN methods all have more undesirable $-/-$ results than desirable $+/+$ results, giving them net negative scores for this performance measure, even at the relatively large number of 32 samples.

The dashed lines show the percentage of trials in which mixed results occurred. Mixed results are considered to be undesirable because one end of the estimated interval does not bound the true percentile range. However, mixed results are not as undesirable as $-/-$ results. Because mixed results are undesirable, the lower the dashed line lies on the plots (i.e., the closer to zero), the better the method's performance with respect to this performance attribute. The TI method always performs best by a substantial margin according to this measure, followed by the Pr-Sch method, then relatively far behind are the NF, NP, and JN methods.

Note that two or more methods' solid-line results could occupy the same ordinate value at a given abscissa value of say $n=2$ samples, yet the methods' dashed lines can have different ordinate values there. For example, let one method yield (60% $+/+$, 30% $-/-$, 10% mixed) results and another method yield (55% $+/+$, 25% $-/-$, 20% mixed) results. Then both methods have the same solid-line plotted value of 30%, reflecting an equal score of "goodness" with regard to the desirable attribute of the number of very desirable $+/+$ occurrences minus the number of very undesirable $-/-$ occurrences. This tie score can be broken by considering the number of undesirable results comprising the mixed population. The method with 10% mixed results would be considered better than the method with 20% mixed results, hence the method that yielded (60% $+/+$, 30% $-/-$, 10% mixed) results is considered the overall better performer.

Accordingly, the virtual tie between the solid-line results of the Non-Parametric and Normal Fit methods at $n=32$ in Figure 14 is broken by considering their dashed-curve values at that abscissa value. The Normal Fit method has about 10% less mixed results than does the Non-Parametric method, making the Normal Fit method's performance better overall at $n=32$.

If one gives the same undesirability weight to the mixed results as is given to the $-/-$ results, then the graph and the ranking of method performance could be made substantially simpler by just quantifying the net value [(% of $+/+$ results) minus (% of $-/-$ results) minus (% of mixed results)]. The larger the value, the better the method has performed. However, it is not deemed reasonable that a mixed result is as bad or undesirable as a $-/-$ result. Therefore the two types of results should not be given the same undesirability weight.

Certainly a more appropriate weighting scheme could be devised and implemented, but in this paper we use a very simple performance ranking scheme described in Section V.

By comparing Figs. 14 and 15 the general tendency is that the aggregation (convolution) of multiple sources of variability somewhat washes out or mitigates the relative number of undesirable $-/-$ and mixed errors of the individual contributing sources.

Magnitude of $+/+$, $-/-$, and mixed overshoot and shortfall errors

Now the desirability of the methods is considered with respect to the magnitude of the overshoot and shortfall errors. For this performance category, mixed errors are quantified by adding the absolute values of Δ_{U-i} and Δ_{L-i} . This prevents large overshoot errors (large positive Δ) and large shortfall errors (large negative Δ), which are both undesirable, from summing to a small value $S_i = \Delta_{U-i} + \Delta_{L-i}$ and thereby giving the false appearance that the mixed errors are both small (which would be desirable).

Figures 16 and 17 plot the mean predicted 0.025 to 0.975 percentile ranges in the $+/+$, $-/-$, and mixed categories, revealing the corresponding mean overshoot and shortfall errors for the various methods. First consider the results for $n=2$ samples. The magnitudes of the mean $-/-$ and mixed errors are not dramatically different among all the methods. But for the $+/+$ overshoot errors, the Tolerance Interval method has large average overshoot in its large $+/+$ population of results. The mean overshoot error is much larger than for the Pr-Sch method, which also has a relatively large $+/+$ population of desirable results. The $+/+$ overshoot errors of the Normal-Fit, Non-Parametric, and Johnson methods are considerably less than for the other two methods. Unfortunately the proportions of desirable $+/+$ results are relatively small for the NF, NP, and JN methods.

The relative trends observed for $n=2$ samples are generally maintained at $n=8$ and $n=32$, within a context of decreasing performance distinctions between the various methods as the number of data samples increases. At $n=32$ the distinctions between method performance are small; the mean $-/-$ undershoot and mixed undershoot/overshoot errors are negligibly small for all methods, and the $+/+$ overshoot errors are negligibly small for the NF, NP, and JN methods and are only slightly bigger (but still small) for the TI and Pr-Sch methods.

Any mitigation or washing-out of individual representation error magnitudes in the aggregation (convolution) of multiple sources of variability can be assessed by comparing Figures 16 and 17. Normalized error magnitudes (as a % of the true percentile ranges) for the convolved results are typically slightly smaller than the individual PDF representation errors. This is usually the case for all methods, numbers of samples, and categories of $+/+$, $-/-$, and mixed errors. However, for $n=2$ samples the Tolerance Interval method shows a slight increase in