Structural reliability analysis from sparse data

Dimitris G. Giovanis
Postdoc, Dept. of Civil Engineering , Johns Hopkins University, Baltimore, MD, USA

Michael D. Shields
Asst. Professor, Dept. of Civil Engineering , Johns Hopkins University, Baltimore, MD, USA

ABSTRACT: Over the past several decades, major advances have been made in probabilistic methods for assessing structural reliability with a critical feature of these methods being that probability models of random variables are known precisely. However, when data are scant it is rear to identify a unique probability distribution that fits the data, a fact that introduces uncertainty into the estimation of the probability of failure since the location of the limit surface in the probability space is also uncertain. The objective of the proposed work is to realistically assess the uncertainty in probability of failure estimates of the First Order Reliability Method (FORM) resulting from the limited amount of data.

1. Introduction

Over the past several decades, major advances have been made in probabilistic methods for assessing structural reliability. They largely began with the widely-used First Order Reliability Method (FORM) and Second Order Reliability Method (SORM) (Haldar and Mahadevan (2000)) and have moved forward in recent years to include more sophisticated simulation-based methods such as importance sampling (Engelund and Rackwitz (1993), Au and Beck (2003)) and subset simulation (Au and Beck (2001), Papaioannou and Straub (2015)). These methods have advanced such that probability of failure can be efficiently estimated for many systems very precisely with very low coefficient of variation. But, this precision does not necessary beget accuracy, although it is often implied.

A critical feature of most reliability methods is that distributions of random variables are known precisely, and in most cases are prescribed such that the reliability calculation can be performed using standard normal distributions. This has led to standard metrics of reliability such as the Hasofer-Lind reliability index defined as $\beta = \frac{\mu_Z}{\sigma_Z}$ where the random variable $Z = g(X)$ denotes the performance function of the structure such that $Z < 0$ denotes failure, $Z > 0$ represents a safe operating condition, and $Z = 0$ is referred to as the limit surface. When the limit surface cannot be linearized, simulation-based techniques such as importance sampling and subset simulation can offer an efficient means to estimate P_f. These methods employ different methods to concentrate simulations in the vicinity of the limit surface and estimate P_f from these simulations. Importance sampling employs an alternate sampling distribution whose density is concentrated near the limit surface and corresponding re-weighting to estimate P_f while subset simulation decomposes the failure probability (which is usually quite small) into a product of higher conditional probabilities that are easier and more efficient to estimate.

However, when data are scarce, it is often impossible to identify a unique probability distribution for the data. This calls into question the assumptions made by existing methods for P_f estimation and, moreover, introduces uncertainty, or imprecision, into the estimate. The effect of this is that the uncertainty in distribution causes uncertainty in the location of the limit surface in the probability space.

Consequently, using FORM or SORM the reliability index for a given structure becomes uncertain or, in a simulation-based approach, P_f estimates may have large uncertainty. In reality, this uncertainty may be large in relation to the P_f estimate itsself.

In a recent work (Zhang and Shields (2018)), Bayesian and information theoretic methods are developed in order to address the problem of uncertainty quantification and propagation when data for characterizing probability distributions are scarce. This is achieved by applying the information theoretic multimodel inference method to identify plausible candidate probability densities and associated probabilities that each method is the best model and the joint parameter densities for each plausible model are then estimated using Bayes' rule.

The objective of the proposed work is to apply the aforementioned methods for imprecise probability in the context of First Order Reliability Method. This way we can assess the uncertainty in probability of failure estimates that results from a lack of data necessary to precisely quantify probability distributions for model input random variables. The result of these analyses are imprecise probabilities of given response quantities in the form of probabilities of probabilities that allows us to probabilistically bound response quantities of interest and therefore assess confidence in our probabilistic estimates.

2. Structural reliability

Reliability is defined as the probability of a performance function $g(\mathbf{X})$ greater than zeros, i.e $P(g(\mathbf{X}) > 0)$ and subsequently, the probability of failure is defined as the probability $P(g(\mathbf{X}) < 0)$. Given that the joint probability density function of the random vector $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ is $f_{\mathbf{X}}(\mathbf{x})$, the probability of failure is evaluated as:

$$P_f = P(g(\mathbf{X}) < 0) = \int_{g(\mathbf{X})<0} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \qquad (1)$$

One of the most widely used reliability analysis method is the First Order Reliability Method (FORM) which is presented next.

2.1. *First Order Reliability Method*

The first-order reliability methods (FORM), as stated by its name, utilizes a first-order Taylor series expansions for the performance function in a standard normal probability space to derive probability of failure estimates. The basic concept of FORM is to enhance the solution of the integral of in Eq.(1) by symplifying the quantity $f_{\mathbf{X}}(\mathbf{x})$ and approximating the performance function $g(\mathbf{X})$. Consider a model in standard normal space with performance function $g(\mathbf{U})$. The FORM approximates the performance function by:

$$g(\mathbf{U}) \approx L(\mathbf{U}) = g(\mathbf{u}^\star) + \nabla g(\mathbf{u}^\star)(\mathbf{U} - \mathbf{u}^\star)^T \qquad (2)$$

where $\mathbf{u}^\star$ is the point around which the series is expanded and its typically called the design point (needs to be found) and it corresponds to the point on the line $g(\mathbf{U}) = 0$ with the highest probability. $\nabla g(\mathbf{u}^\star)$ is the gradient of $g(\mathbf{U})$ evaluated at $\mathbf{u}^\star$

$$\nabla g(\mathbf{u}^\star) = \left(\frac{\partial g(\mathbf{U})}{\partial U_1}, \frac{\partial g(\mathbf{U})}{\partial U_2}, \ldots, \frac{\partial g(\mathbf{U})}{\partial U_n} \right) \Bigg|_{\mathbf{u}^\star} \qquad (3)$$

The expansion in Eq.(2) allows for a straightforward estimation of probability failure as:

$$P_f = \Phi(-\beta) \qquad (4)$$

Given that Eq.(2) is formulated in the space of standard normal random variables $\mathbf{U}$, formulation of a FORM estimate requires a nonlinear transformation from the generally non-normal parameter space $\mathbf{X}$ to the standard normal space. This transformation is based on the condition that the cumulative distribution functions (CDFs) of the random variables remain the same before and after the transformation. Equation (1) then becomes

$$P_f = P(g(\mathbf{U}) < 0) = \int_{g(\mathbf{U})<0} \phi_{\mathbf{U}}(\mathbf{u}) d\mathbf{u} \qquad (5)$$

where $\phi_{\mathbf{U}}(\mathbf{u})$ is the joint probability density function (PDF) of $\mathbf{U}$. Assuming independence of the random variables the joint PDF is the product of the individual PDFs of standard normal distribution given by:

$$\phi_{\mathbf{U}}(\mathbf{u}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}u_i^2\Big) \tag{6}$$

and the probability integration becomes

$$P_f = \underset{g(u_1,\dots,u_n)<0}{\int\cdots\int} \phi_{\mathbf{U}}(\mathbf{u})\,du_1\,du_2\dots du_n \tag{7}$$

In order to find the location of the design point in the standard normal space $\mathbf{U}$, the joint PDF $\phi_{\mathbf{U}}(\mathbf{u})$ at the limit-state of $g(\mathbf{U}) = 0$ needs to be maximized. The mathematical model that describes this problem is defined as

$$\begin{cases} \max_{\mathbf{u}} & \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\big(-\frac{1}{2}u_i^2\big) \\ \text{subject to} & g(\mathbf{u}) = 0 \end{cases} \tag{8}$$

which is equivalent to

$$\begin{cases} \max_{\mathbf{u}} ||\mathbf{u}|| \\ \text{subject to} \quad g(\mathbf{u}) = 0 \end{cases} \tag{9}$$

where

$$||\mathbf{u}|| = \sqrt{u_1^2 + u_2^2 + \ldots + u_n^2} \tag{10}$$

The design point then corresponds to the shortest distance point from the limit state $g(\mathbf{U}) = 0$ to the origin O in U-space. This distance is call reliability index β. At the design point $\mathbf{u}^\star$, $g(\mathbf{u}^\star) = 0$ and thus, Eq.(2) becomes

$$L(\mathbf{U}) = \sum_{i=1}^{n} \frac{\partial g(\mathbf{U})}{\partial U_i}\bigg|_{\mathbf{u}^\star} (\mathbf{U}_i - \mathbf{u}^\star) = a_0 + \sum_{i=1}^{n} a_i U_i \tag{11}$$

where

$$a_0 = -\sum_{i=1}^{n} \frac{\partial g(\mathbf{U})}{\partial U_i}\bigg|_{\mathbf{u}^\star} \mathbf{u}^\star, \quad a_i = \frac{\partial g(\mathbf{U})}{\partial U_i}\bigg|_{\mathbf{u}^\star} \tag{12}$$

Equation (11) indicates that the performance function is a linear function of standard normal variables and therefore, is also normally distributed. The mean value and standard deviation are given by

$$\mu = a_0 = -\sum_{i=1}^{n} \frac{\partial g(\mathbf{U})}{\partial U_i}\bigg|_{\mathbf{u}^\star} \mathbf{u}^\star \tag{13}$$

and

$$\sigma = \sum_{i=1}^{n} \sqrt{a_i^2} = \sqrt{\sum_{i=1}^{n} \bigg(\frac{\partial g}{\partial U_i}\bigg|_{u_i^\star}\bigg)^2} \tag{14}$$

The probability of failure can be calculated then as

$$P_f \approx \Phi\big(\frac{-\mu}{\sigma}\big) = \Phi\Big(\sum_{i=1}^{n} a_i u_i^\star\Big) = \Phi\Big(\mathbf{a}\mathbf{u}^{\star T}\Big) \tag{15}$$

where

$$\mathbf{a} = (a_1, a_2, \ldots, a_n) = \frac{\nabla g(\mathbf{u}^\star)}{||\nabla g(\mathbf{u}^\star||} \tag{16}$$

and

$$a_i = \frac{\frac{\partial g}{\partial U_i}\Big|_{u_i^\star}}{\sqrt{\sum_{i=1}^{n} \bigg(\frac{\partial g}{\partial U_i}\Big|_{u_i^\star}\bigg)^2}} \tag{17}$$

By definition, the design point $\mathbf{u}^\star$ is the tangent point of the curve $g(\mathbf{U}) = 0$ and the circle with the radius of β and therefore, it is perpendicular to $g(\mathbf{U}) = 0$ with direction given by the unit vector $\mathbf{u}^\star/||\mathbf{u}^\star||$. On the other hand, the direction of the gradient is also perpendicular to the curve at the design point, and its direction can be represented by the unit vector $\mathbf{a}$ (Eq.(16). That means

$$\frac{\mathbf{u}^\star}{\beta} = \mathbf{a} \tag{18}$$

and therefore,

$$P_f \approx \Phi\big(\mathbf{a}\mathbf{u}^{\star T}\big) = \Phi\big(-\beta\mathbf{a}\mathbf{a}^T\big) = \Phi\big(-\beta\big) \tag{19}$$

However, when distributions in the parameter space are imprecise (i.e. not known exactly), it is not even clear what is meant by the design point. There cannot, in such problems, be a single point along $g(\mathbf{U}) = 0$ having maximum probability for the simple reason that $g(\mathbf{U}) = 0$ itself is not precisely defined. That is, while the performance function in the parameter space is uniquely defined (e.g. by a finite element model), imprecise probabilities do not afford a unique mapping of this performance function to establish a precise form in the standard normal space.

3. BAYESIAN MULTIMODEL SELECTION

Knowledge of the specific distribution model M of the parameters $\mathbf{X}$ is mandatory in order to perform reliability analysis. However, in order to assign a proper distribution (model) collected data must be available. But one must always ask whether the assignment of a probability model from a small data set is justified. To this purpose, principles of Bayesian inference and multimodel selection can be utilized in order to select probability models that adequately represent the set of data. Two settings of model selection are considered, one based on Bayes factors/likelihood functions and the other on information-theoretic selection criteria based on the Kullback-Leibler (K-L) information theory Kullback and Leibler (1951).

3.1. Bayesian Inference

Consider the probability model M with corresponding random parameters θ. Given available data $\mathbf{d}$ Bayes' rule can be utilized in order to update our knowledge of the parameters θ for M. In this setting, we define a prior PDF $p(\theta|M)$ for θ that represents our prior knowledge on the parameters and we seek to find the posterior PDF $p^\star(\theta|\mathbf{d};M)$ given $\mathbf{d}$ that represents our updated belief:

$$p^\star(\theta|\mathbf{d},M) = \frac{p(\mathbf{d}|\theta;M)p(\theta;M)}{p(\mathbf{d};M)} \qquad (20)$$

where $p(\mathbf{d};M)$ is the evidence, i.e a normalizing factor computed by marginalizing the likelihood $p(\mathbf{d}|\theta,M)$ over the θ

$$p(\mathbf{d};M) = \int \mathscr{L}(\theta|\mathbf{d};M)p(\theta;M)d\theta \qquad (21)$$

In the general case, a Markov Chain Monte Carlo (MCMC) method is required for the solution of the integral equation in order to draw samples from $p^\star(\theta|\mathbf{d},M)$.

3.2. Information model selection

The information-theoretic scheme utilizes a criterion in order to select the appropriate model that encompasses the information loss resulting from the approximation of the true with the candidate. In this term, the criterion can be one that minimizes the information loss; In Akaike (1974), the Akaike Information Criterion was introduced based on the fact that the expected relative K-L information could be approximated by the maximized log-likelihood function with a bias correction.

$$\text{AIC} = -2log(\mathscr{L}(\hat{\theta}|\mathbf{d},M)) + 2K \qquad (22)$$

where $\mathscr{L}(\cdot) = p(\mathbf{d}|\theta,M)$ is the likelihood function given the maximum likelihood estimate of the parameters θ, M is the candidate probability model with uncertain parameters $\hat{\theta}$, $\mathbf{d}$ are the available data and $2K$ is a bias correction factor. It is important in model selection to establish a relative scale for the AIC such that:

$$\Delta_A^{(i)} = \text{AIC}^{(i)} - \text{AIC}^{\min} \qquad (23)$$

where $\text{AIC}^{(i)}$ is the AIC for candidate model M_i and $\text{AIC}^{\min} = \min(\text{AIC}^{(i)})$. This normalizes the best model to a value $\Delta_A^i = 0$. By applying the transformation $\exp\left(\frac{-\Delta_A^{(i)}}{2}\right)$ one can obtain the likelihood of the model M_i given the data and subsequently the corresponding probability p_i, by normalizing the likelihood

$$p_i = p(M_i|\mathbf{d}) = \frac{\exp\left(\frac{-\Delta_A^{(i)}}{2}\right)}{\sum_{i=1}^N \exp\left(\frac{-\Delta_A^{(i)}}{2}\right)} \qquad (24)$$

where N is the number of candidate probability models. These probabilities can be interpreted as the probability that model M_i is the K-L best model for the data. The AIC is an asymptotic quantity that imply large datasets. In Hurvich and Tsai (1989) a critical extension of the AIC has been developed for small datasets, that introduces a second-order bias correction term yielding

$$\mathrm{AIC}_c = -2log(\mathscr{L}(\hat{\theta}|\mathbf{d},M)) + 2K + \frac{2K(K+1)}{n-K-1} \tag{25}$$

where n is the sample size of the data.

4. Application: plate buckling strength problem

In this section we investigate the issue of uncertainty that exists in the probability of failure estimates in FORM. This uncertainty results from the lack of sufficient dataset size, required for the precise quantification of the probability distribution model of the random parameters. To this end, the multimodel approach described above provides a natural framework for clarifying this issue.

Initially, Bayes' rule is utilized in order to establish a set of possible probability models for the parameter space and associated probabilities for each model. Each of these unique models in the candidate set has an associated mapping to the standard normal space, and hence defines a unique $g(\mathbf{U}_i), \quad i = 1,\ldots,N$ where N is the number of candidate models. Consequently, for each model M_i it is possible to establish a design point $\mathbf{u}_i^\star$, a reliability index β_i, and associated probabilities p_i. The result is therefore a set of FORM probability of failure estimates, each with known probability of occurrence that is derived from a small dataset.

In order to study the effects of model-form and parametric uncertainty on probability of failure we selected the model of a simply supported rectangular plate under uni-axial compression considering certain geometric and material uncertainties. The probability of failure is defined as

$$P_f = P(\psi < 0.5) \tag{26}$$

where ψ is the buckling strength defined as (Carlsen (1977))

$$\psi = \left(\frac{2.1}{\lambda} - \frac{0.9}{\lambda^2}\right)\left(1 - \frac{0.75\delta_0}{\lambda}\right)\left(1 - \frac{2\eta t}{b}\right) \tag{27}$$

where

$$\lambda = \frac{b}{t}\sqrt{\frac{\sigma_0}{E}} \tag{28}$$

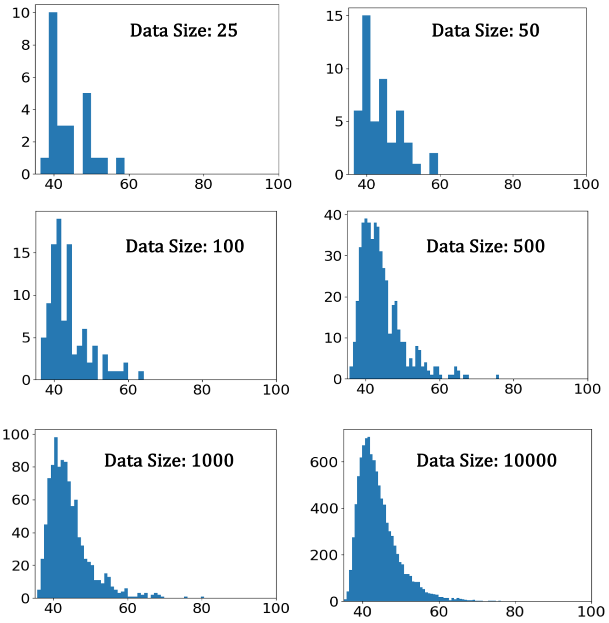

Figure 1: Randomly generated yield stress values that serve as the initial dataset for uncertainty quantification and propagation in plate buckling strength.

In Eq.(27) b is the width, t is the thickness, σ_0 is the yield stress, E is the Young's modulus, δ_0 is the initial deflection and η is the residual stress of the plate. Although all 6 variables are influencing the buckling strength, we will focus on a single material parameter, namely the yield stress σ_0, since in Zhang and Shields (2018) it was observed that the buckling strength is most sensitive to the yield stress and least sensitive to the plate width. Moreover, as suggested in Hess and Ayyub (2002), an approximation of the "true" mean yield stress is $\sigma_0 = \hat{\sigma}_0 + 34$ where $\hat{\sigma}_0$ follows a lognormal distribution with mean $\mu_{\hat{\sigma}_0} = 1.3 \times 34 - 34$ and standard deviation $\sigma_{\hat{\sigma}_0} = 0.1235 \times 1.3 \times 34 - 34$.

Data are generated from this distribution and are used in order to quantify the uncertainty in the probability of failure estimates. Fig.(1) depicts randomly generated yield yield stress values of different size (25, 50, 100, 500, $10^3, 10^4$), that serve as the initial dataset.

Since we have no knowledge of the "true" probability distribution model of the data we define a set of candidate models that consists of 8 distributions; lognormal, gamma, logistic, inverse Gauss,

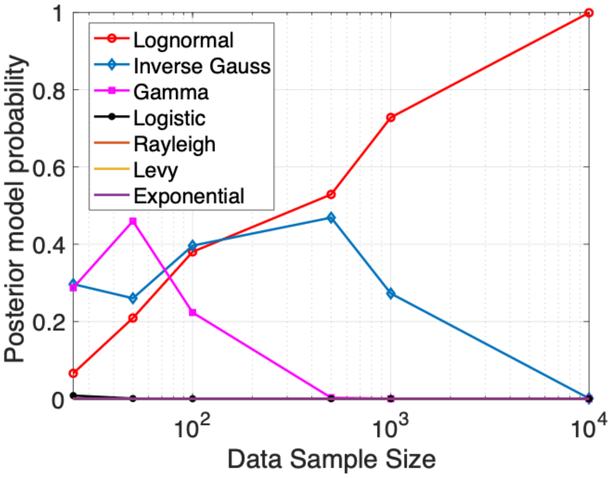

Figure 2: Posterior probability (AIC_c criterion) as a function of dataset size.

Rayleigh, Levy and exponential. The AIC model selection criteria is employed to rank the candidate distributions.

Fig.2 shows the AIC_c probabilities for each candidate model as a function of dataset size. It is obvious that the multimodel inference does not select the correct lognormal model conclusively until 1000 yield stress measurements are collected. In fact, prior to that point the inference placed a high level of probability on the incorrect Inverse Gaussian model. How, then does this uncertainty influence the estimation of the reliability of failure of the buckling strength?

Fig. 3 and highlights this influence in terms of the calculation of the failure probability. More specifically, Fig.3 shows the frequency of occurrence of the reliability index β for the lognormal candidate model and for different data size. For each case, Bayesian inference is employed to estimate the joint parameter probability densities for the model and then 1000 FORM problems are formulated and solved using the Hasofer-Lind approach for the approximation of the design point. From this figure we can see that the smallest the size of the data the highest the variance of the reliability index, which practically means that the calculated probability of failure has large variance. This is more pronounced in Fig.(4) where one can see the empirical CDFs for the probability of failure.

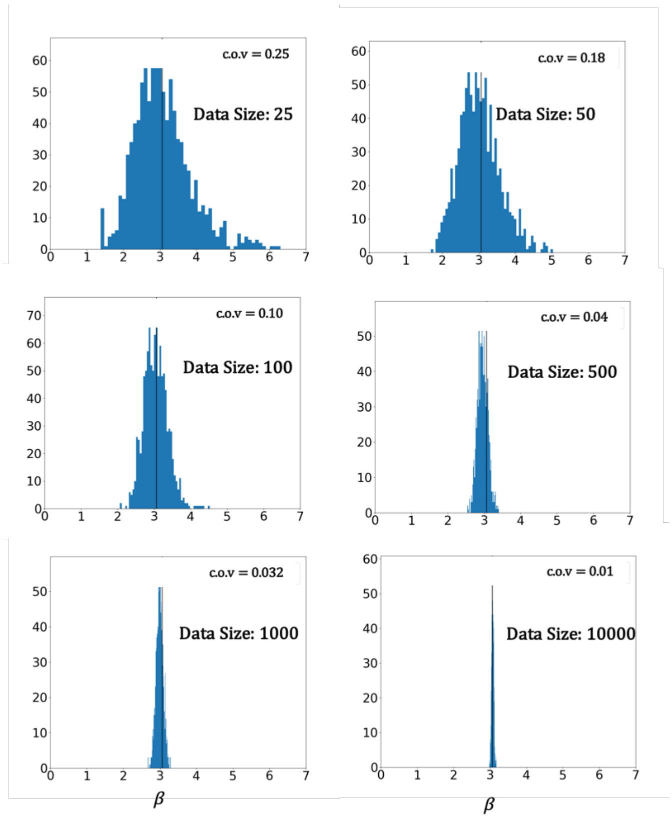

Figure 3: Posterior probability (AIC_c criterion) as a function of dataset size.

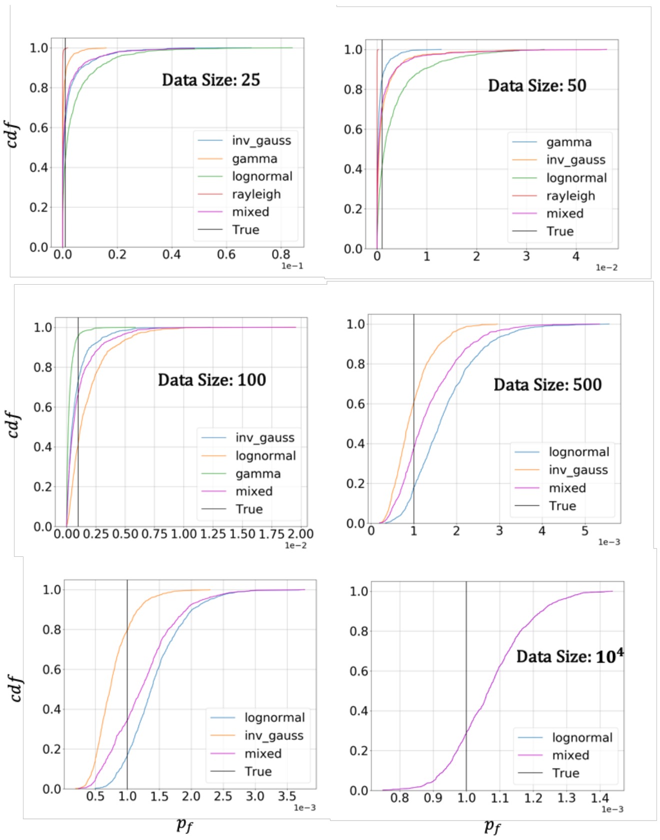

Figure 4: Empirical CDFs for the probability of failure occurs when $\psi < 0.5$.

5. CONCLUSIONS

This work addresses the issue of uncertainty in the reliability of a structural system as a result of insufficient amount of data required for the precise quantification of the probability model of the input parameters. To this purpose, methods from imprecise probabilities are utilized in order to select appropriate models that fit the data. An initial investigation in the context of FORM is performed and the results show (as expected) that if the number of available data is small then the estimation of the probability of failure has large variance.

ACKNOWLEDGEMENTS

Methodological developments presented herein have been supported by the Office of Naval Research with Dr. Paul Hess as program officer.

6. REFERENCES

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

Au, S. and Beck, J. (2001). “Estimation of small failure probabilities in high dimensions by subset simulation.” *Probabilistic Engineering Mechanics*, 16(4), 263–277.

Au, S. and Beck, J. (2003). “Important sampling in high dimensions.” *Structural Safety*, 25(2), 255–276.

Carlsen, C. A. (1977). “Simplified collapse analysis of stiffened plates.” *Norw. Maritime Res.*, 5(4).

Engelund, S. and Rackwitz, R. (1993). “A benchmark study on importance sampling techniques in structural reliability.” *Structural Safety*, 12(4), 255–276.

Haldar, A. and Mahadevan, S. (2000). *Probability, reliability, and statistical methods in engineering design*. John Wiley.

Hess, P.E., B. D. A. I. and Ayyub, B. (2002). “Uncertainties in material and geometric strength and load variables.” *Naval Eng. J.*, 2.

Hurvich, C. and Tsai, C.-L. (1989). “Regression and time series model selection in small samples.” *Biometrika*, 76(2), 297–307.

Kullback, S. and Leibler, R. A. (1951). “On information and sufficiency.” *Ann. Math. Stat.*, 22(1), 79–86.

Papaioannou, I., B. W. Z. K. and Straub, D. (2015). "MCMC algorithms for subset simulation." *Probabilistic Engineering Mechanics*, 41, 89–103.

Zhang, J. and Shields, M. D. (2018). "On the quantification and propagation of imprecise probabilities resulting from small datasets." *Mechanical Systems and Signal Processing*, 98, 465–483.