MODEL VALIDATION AND DESIGN UNDER UNCERTAINTY

By

Ramesh Rebba

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Civil Engineering

December, 2005

Nashville, Tennessee

Approved:

Professor Sankaran Mahadevan

Professor Prodyot K Basu

Professor Bruce Cooil

Professor Gautam Biswas

To my parents and teachers

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to my advisor, Prof. Sankaran Mahadevan, for his invaluable guidance, encouragement and lasting patience throughout my stay at Vanderbilt. He has been a tremendous source of inspiration in both academics and in personal life. I would also like to thank him for giving me an opportunity to present my work at various venues. I am very grateful for the technical support I received from Dr. Shuping Huang and Dr. Xiaomo Jiang who contributed very valuable discussions.

I would like to extend my thanks to Sandia National Laboratories, Albuquerque, NM and Department of Civil and Environmental Engineering at Vanderbilt for their financial support. Also, thanks to John McFarland for involving in the research and helping run simulations. Last but not the least; I would like to acknowledge the constant encouragement and friendship of Ned Mitchell, Natasha Smith, Yongming Liu and all IGERT students during my stay at Vanderbilt.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Overview

There has been an increased reliance on numerical models and simulation codes recently for predicting the behavior of complex engineering systems. With the advent of modern super computers, complex natural phenomena are sought to be modeled without actually performing full-scale experiments. The test-only based approach is very expensive and does not make use of available analytical models of system behavior, failure modes and sensitivities. Inexpensive modeling and simulation-based methods are able to use such information. However, with the approximations in the computational models and the limited amount of statistical data on the input variables, it is difficult to associate a high degree of confidence with prediction based only on computational methods. When physics is not well understood, selecting a wrong model could induce model form error while the discrete solution for a continuum domain could introduce numerical errors and convergence problems. The use of a mathematical or a computational model leads us to a common question: How good are these models? How valid are the models?

The performance of a model is judged by comparing the outcomes derived from the model with the observations made during the experiments. There is also uncertainty and error in the measurement of both input data and output response. These random effects and the approximations also affect the deviation of the model predictions from the nature. Verification and validation (V&V) under uncertainty thus involves quantifying

the error in the model prediction and effectively comparing the prediction with the experimental result when both prediction and test data are stochastic. The main goal of model validation is to assess the predictive capabilities of a computational code for specific applications. We also need to quantify the model errors using validation experiments. A key element in the model validation methodology is the definition of validation metrics or measures within a probabilistic framework. Also, the concept of model validation has to be extended to system-level problems where full-scale testing is impossible. Component-level validation results may be used to derive a system-level validation measure. This derivation again depends on the knowledge of inter-relationships between component modules. Another issue is the validation of statistical model or distribution as opposed to the validation of single response. The probability density function characterizing the uncertainty in a model prediction may be compared with a small set of experimental data that span the possible values of model response.

A computational model may also generate multiple response quantities (decision variables) at a single location or the same response quantity at multiple locations, and a validation experiment might yield corresponding measured responses in a single test. For instance, stress, strain, displacement and peak acceleration etc. are all derived from same finite element field. In each case, the multiple responses, being derived from same input, could be dependent on each other. In both the events, model validation involves comparison of multiple quantities of model prediction and test data (multivariate analysis). A single response quantity may be predicted at different points in the space, time or frequency domain. (e.g., spectral dynamic response, mode shapes of a structure etc). Validation in such cases involves comparison of curves or surfaces. Thus, validation

metrics need to be developed to compare multiple model outputs to the multiple data available. Also, each decision variable can be validated individually or a collective metric can be developed to validate the correlated quantities in order to judge the overall performance of the code. Markov Chain Monte Carlo methods appear to be a natural choice in dealing with multiple variables, and hence were investigated for this purpose. The metrics developed in this research will make use of classical and Bayesian hypothesis testing. Typically used point null hypothesis testing, where two quantities are tested for equality, can be practically not so useful for decision making and hence more practical interval-based hypothesis testing methods will be explored. Further one can calculate the probability that the model prediction falls within a certain range of data and vice versa. Thus a model reliability metric will also be proposed in this research.

One challenge in practical problems is to extend what we can learn about the model's predictive capability within the tested region to an inference about the predictive capability in the application or untested region. Confidence in the prediction near off-nominal region by a model, already validated in the nominal region, needs to be quantified. One approach is to construct a regression model for the test data in the validation domain, and to simulate test data in the untested region using this model. Inferences may be made in an incremental fashion from validation region to untested region, aided by bootstrapping and cross validation. However, this strategy may not work if there is a change in physics from the validation domain to the application domain. Therefore, proposed work in this direction will explore other extrapolation strategies under nonlinear behavior. If some linking or common variables can be established between the two domains, Bayesian methodology may offer some insight into the

extrapolation process. Bayesian networks will be explored for this purpose. Advanced methods such as adaptive rejection sampling, saddlepoint approximations and Laplace expansion methods offer alternatives to rigorous MCMC methods for Bayesian analysis.

Physical, information and model uncertainties, errors can introduce additional bias and variance in the model prediction. When continuum models are chosen to represent the reality and numerical methods such as finite element and finite difference methods are used to solve the continuum model, the approximations can result in numerical solution error. Similarly inadequate surrogate models can introduce random truncation errors in the model response. When the computational models are used for design in early stages, one must account for all the above mentioned uncertainties and errors. Reliability-based design optimization (RBDO) techniques ensure that the design is met with high confidence in light of various sources of uncertainties. As in any optimization problem, the constraints or the objective function can be a function of the model output. When probabilistic constraints are used, the approximate reliability analysis methods like FORM, SORM can also introduce additional errors. The study proposes to include model errors in the design explicitly.

The study investigates and develops methods for 1) Validation metrics appropriate for individual and multiple response quantities 2) Extrapolating or interpolating the validation inferences to untested regime, and 3) Quantifying model form, numerical solution and reliability analysis errors, uncertainties and 4) Incorporating model errors in the design.

1.2 Research objectives

Based on the discussions so far, the research objectives are summarized as below:

1. Develop classical and Bayesian statistics-based validation metrics for single and multivariate model outputs. Multivariate outputs may include single response at multiple locations or multiple outputs at single location. Extend Bayesian model validation to include outputs that are stochastic processes and fields;
2. Develop a methodology to assess the predictive capabilities of computational models in the application domain based on the data in validation domain. Bayesian networks will be explored to propagate uncertainties and inference across various domains.
3. Develop methods to quantify model form errors, numerical solution errors due to model resolution, and errors due to approximations in the reliability analysis. This is a part of verification process that must be carried out prior to validation. However the actual estimates of error can be used to assess the solution quality.
4. Incorporate various uncertainties and errors in the design. When limit-state based methods are used to estimate the probability of failure, model errors etc., can be explicitly used as additional variables in the reliability analysis.

1.3 Highlights of the research

The definitions of verification and validation in computational science now have been well established by several researchers at different private and government organizations.

However the progress on the actual implementation of those concepts has not been slow. Model validation itself is a hard statistical problem and a wide variety of techniques ranging from hypothesis tests to model reliability metric have been proposed in this research. Since most of the model responses are multivariate in nature, we address the simultaneous inferences of model output and test results. Whenever the existing statistical approaches are found to be inadequate to address the basic questions of validation directly, the study proposed some alternative. The proposed Bayesian validation methodology combines the prior information on model prediction with the observed data and updates our belief on confidence in the model.

Assessing the confidence in the model prediction for which we have no data is another challenge. Simulating field conditions in the laboratory is infeasible, and computer models are being used to design very complex future systems, for which historic data is not available. The confidence in such a design would depend on how the model behaved in the validation region and how "far" the validation domain is from the target application domain. We can then extrapolate the inferences made in validation region to application domain. The use of Bayesian networks in model validation is a unique concept and has promising use for extrapolation and in system-level model validation where full scale testing is often infeasible. The proposed method can include the change of physics or the sensitivity of the response in untested region to that from the validation domain and estimate the confidence in the extrapolated prediction. This is beneficial in using computer models for practical application.

While validation is necessary to assess the performance of the model, the ultimate goal is to use the computational model to design engineering systems. The proposed

study will thus explore the role of verification and validation in design. Even when actual model form errors are not available, if a parametric study of the design with respect to various errors in the model prediction can be conducted. Such sensitivity analysis can help one in resource allocation and identifying the areas of improvement.

1.4 Organization of dissertation

This dissertation is organized as follows: In Chapter 2, a review of concepts and definitions of validation are presented. A literature survey on validation metrics is conducted. Classical and Bayesian statistical methods are proposed and applied for univariate and multivariate model validation. Numerous examples are provided to demonstrate the proposed methodology.

Chapter 3 considers the various possible cases of extrapolation and extends the Bayesian validation methodology for model predictive assessment. Bayesian networks and Markov Chain Monte Carlo methods are explored for that purpose. A number of illustrative problems are presented to explain the proposed extrapolation method.

In Chapter 4, various sources of error in computational model prediction are identified first and methods to quantify uncertainties and errors are presented.

In Chapter 5, the model errors estimated using the methods described in Chapter 4 are used in RBDO. In the last chapter, some recommendations and research directions for the future work are presented.

CHAPTER II

VALIDATION METRICS

2.1 Overview

Various types of uncertainties and errors occur in computational model predictions that attempt to capture the behavior of real physical systems. The uncertainties arise due to model form inadequacies, lack of sufficient data, and inherent variabilities in the physical properties of the system. The corresponding experimental data needed to validate these computational models are also affected by experimental variability, measurement errors etc. Model validation under uncertainty thus reduces to comparing two uncertain quantities. Validation assessments can be made using qualitative methods, decision-theoretic methods, or statistical hypothesis testing methods. While a validation method should be able to provide an answer to the question whether the computational model accurately represents the reality, it should also verify whether the degree of confidence with which we accept a model is adequate for the intended model use. Several validation metrics are investigated in this study, focusing on their ability to address both accuracy and adequacy issues for engineering applications.

Depending on the nature or form of model output and experimental data, model validation may involve comparison of means or variances or even two or more probability distributions. The decision maker would like to know whether there is a significant difference between the prediction and observation. The validation metric would then provide a means for accepting or rejecting the model prediction. Both

classical and Bayesian methods will be explored for this purpose. In some problems, the model output may be a single response quantity that follows a statistical distribution that needs to be compared against single or repeated experimental observations of the same quantity. This may be termed univariate validation. In other problems, the model output and the corresponding validation data may be multivariate in nature. This study develops validation metrics (measures of comparison) for models with multivariate output also. Repeating univariate validation separately for several response quantities may give conflicting inferences for different quantities. Thus an overall performance measure for the computational model is possible only through aggregate validation.

In hypothesis testing, we usually formulate the null hypothesis as model prediction being exactly equal to an observation and the alternative hypothesis as model prediction being not equal to the observation. This is also referred to as "point null hypothesis" testing. A well known criticism of point null hypothesis tests is that the null hypothesis gets rejected even if the difference between the prediction and observation is small enough for all practical purposes. Thus a model rejected by such a test does not automatically render the model useless. Also we should expect the hypothesis test to increasingly 'punish' an invalid model and 'reward' a valid model with the availability of more data. If we allow for some acceptable difference between the model prediction and the observation, we can then test a null hypothesis that the data falls within certain bounds of the model prediction. The collection of more data should then increase or decrease the confidence consistently using the modified null hypothesis. The drawback in point null tests is that all models get rejected with increasing sample size and hence there is no incentive to do more tests or build better models. This study will investigate how

the interval formulation of hypothesis can address the issue of sample size. For practical purposes and ease of interpretation, a more direct approach that formulates model validation as a reliability analysis problem is also proposed. The proposed methodologies are illustrated throughout this chapter with several numerical examples.

2.2 Background

The fundamental concepts and terminology for validation and verification of computational codes have been established mainly by the ASCI (Accelerated Strategic Computing Initiative) program of the United States Department of Energy (DOE), American Institute of Aeronautics and Astronautics (AIAA, 1998), Defense Modeling and Simulation Office (DMSO) of the U.S. Department of Defense and American Society of Mechanical Engineers Standards Committee (ASME PTC#60) on verification and validation of computational solid mechanics etc.

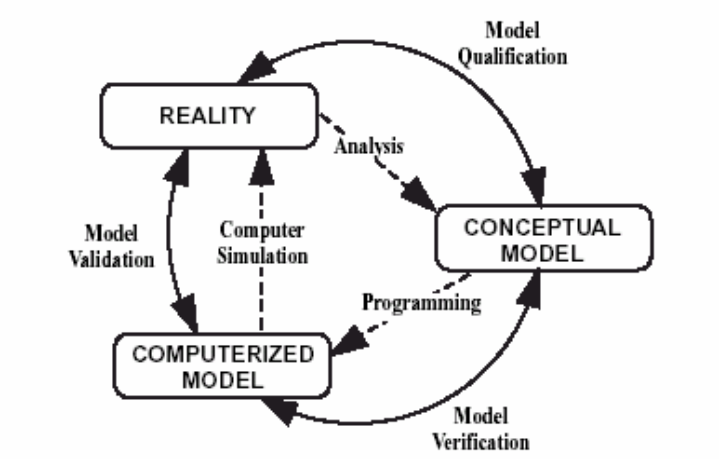


Fig. 2.1. Phases of Modeling and Simulation and the Role of V&V (AIAA, 1998)

Various other definitions of scientific validation also exist in the literature. Various requirements have been defined because of the large variety of applications for modeling. The first definition is given by the Institute of Electrical and Electrical Engineers (IEEE) for verification and validation (Boehm, 1984), with reference to the products of a software development cycle. DMSO (1996) gives definitions for V&V in the context of computational models. Verification is defined as the process of determining that a model implementation accurately represents the developer's conceptual description and specifications. Validation is defined as the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model. The study mainly emphasizes on model validation under uncertainty in computational mechanics with applications relevant to civil, mechanical and aerospace engineering systems only. The use of validation metrics comprises the most important part of a validation activity. A validation metric should quantitatively measure the degree of difference between model prediction and experimental data and should also include the uncertainties in them.

Several types of metrics have been proposed over the years for the validation of computational models. An attempt to collect and discuss various validation metrics was made by Oberkampf and Barone (2004) with a comprehensive list of studies by various researchers. In this chapter, details on specific metrics are provided, and practically useful validation metrics are proposed. Work on systematic model validation methods for engineering applications began in the field of fluid dynamics. Coleman and Stern (1997) combined various types of errors and uncertainties arising in CFD applications, and proposed a validation metric requiring the prediction error to be small. A comparison

error E is defined as the actual difference between prediction and data. Then the uncertainty associated with that error is computed through a combination of numerical errors (E_{SN}), modeling error (E_{SMA}), data or measurement error (E_D), and the uncertainties in previous data used to build the model (E_{SPD}). All these errors are assumed to be independent of each other and are combined linearly. The term uncertainty has been synonymously used with standard deviation in that paper. Thus the total uncertainty in the comparison error or the standard deviation of E is estimated as

$$\sigma_E = \sqrt{\sigma_{SPD}^2 + \sigma_{SMA}^2 + \sigma_{SN}^2 + \sigma_D^2} \tag{2.1}$$

The model prediction is said to be inadequate if $|E| < \sigma_E$. Simply stated, the metric verifies whether the actual prediction error is less than its standard deviation value. The confidence with which we accept or reject a model prediction is not reported with this metric.

Since the metric proposed in Eq. (2.1) does not give any measure of statistical significance of the result, hypothesis testing using classical statistics was found to be more appropriate for comparing data with prediction. For given prediction and data vectors $\boldsymbol{x}_{model}$ and $\boldsymbol{x}_{exp}$, a validation metric based on the Mahalanobis distance was proposed (Hills and Trucano, 2001):

$$r^2 = \left(\boldsymbol{x}_{model} - \boldsymbol{x}_{exp}\right)^T \left(\mathrm{cov}\left(\boldsymbol{x}_{model}\right) + \mathrm{cov}\left(\boldsymbol{x}_{exp}\right)\right)^{-1} \left(\boldsymbol{x}_{model} - \boldsymbol{x}_{exp}\right) \tag{2.2}$$

The model prediction is said to be close to the data when r^2 is less than some critical value $\chi_\alpha^2(n)$ where n is the number of data points or predictions. Thus r^2 follows a chi-square distribution with n degrees of freedom. This metric is valid under assumption that

data and prediction vectors are Gaussian and hence cannot be applied to all problems unless both the data and prediction are transformed into normal space. Also, the model is rejected at α significance level and thus p-value in this case can be computed as $P(r^2 > r_{obs}^2)$.

A quantitative comparison based on probability intervals has been suggested by Urbina *et al* (2003). In their approach, the probability distribution of the difference Δ between model prediction and the data is determined first. Then for any chosen proportion p such that $0 < p < 1$, the values of Δ corresponding to its CDFs $(1 - p)/2$ and $(1 + p)/2$ are estimated. Thus there will be a probability of p that Δ lies within that interval. If that $100 \times p\%$ probability interval contains zero, then the model prediction is said to be acceptable.

Zhang and Mahadevan (2004) applied Bayesian hypothesis testing and the Bayes factor metric for validation of limit state-based reliability prediction models. Suppose the model predicts a failure probability of p for a physical system based on the knowledge of various uncertainties. If we observed k failures out of n tests, then the validation metric or Bayes factor in this case is derived as $B = (n+1)\frac{n!}{(n-k)!k!} p^k (1-p)^{n-k}$. If B is greater than 1.0, we conclude that the data favors the model prediction. Recently the method has been extended to the validation of more generalized model outputs, both univariate and multivariate (Rebba *et al*, 2004). The validation metric in that case is ratio of posterior to prior densities of the model prediction. Other Bayesian approaches for model validation focused more on calibration of the model using the data and providing posterior probability intervals rather than a direct assessment of degree of match between

prediction and observation (Bayarri *et al*, 2003; Hasselman and Wathugala, 2002; Higdon *et al*, 2004). Similarly multivariate statistical methods are being extensively used in meteorological and climate modeling (Wilks, 1995) but their application in civil, mechanical and aerospace engineering has been limited.

2.3 Hypothesis testing for model validation

Depending on the availability of data, model output, the specific problem, validation may be considered in several ways: (i) the comparison of a set of discrete data with a single prediction (ii) comparing a continuous distribution (model output) with discrete data (test) or (iii) the comparison of multivariate model outputs with the corresponding observations. Null hypotheses that the difference between model and data is zero, can be appropriately constructed in each of these three cases and tested using some evidence (test data). Both classical and Bayesian hypothesis testing procedures can be used for this purpose. A careful attention is needed in formulating these hypotheses, satisfying the underlying assumptions and selecting a relevant significance test. This section discusses the issues in using hypothesis testing for model validation especially involving univariate comparisons, for the following two cases:

Case 1: Model prediction is a single number θ_0 while the data is $X = \{x_1, x_2, \ldots, x_n\}$ which are replicated experimental measurements taken with the same input as for the model.

Case 2: Model output follows a continuous distribution $f(\theta)$ while the data $X = \{x_1, x_2, \ldots, x_n\}$ is observed not for a particular value of input but for a wide range of input parameters during the experiment.

2.3.1 Classical point null testing

For case 1, the null and alternative hypotheses are formulated as H_0: $\bar{X} = \theta_0$ and H_a: $\bar{X} \neq \theta_0$. For case 2, we test H_0: $\bar{X} = \mu$ and $s_X^2 = \sigma^2$; H_a: $\bar{X} \neq \mu$ and $s_X^2 \neq \sigma^2$, where μ and σ are mean and standard deviation of θ respectively. The logic of classical hypothesis testing is as follows: First, a test statistic T is defined as a function of the difference between observation and prediction, and then the actual value of the statistic, t is estimated. Assuming that the null hypothesis is true, the probability of getting a test statistic value more than t is computed. Finally, this probability $P(T \geq t)$, also referred to as p-value, is compared to the significance level α (usually 0.01 or 0.05). If the p-value is less than or equal to the significance level, then the null hypothesis is rejected and the outcome is said to be statistically significant, i.e., not by chance. Practically, this can be interpreted as follows: if the p-value is too small, the t-statistic (a measure of the error) is too large to be acceptable under H_0; therefore we reject H_0.

For case 1, suppose the null hypothesis is true, then $\bar{X}$ may be assumed to follow a normal distribution, due to central limit theorem for a large n, with mean θ_0 and say known variance σ. Then the test statistic $T = \dfrac{\sqrt{n}\left|\bar{X} - \theta_0\right|}{\sigma}$ follows a Student t distribution with n - 1 degrees of freedom. The p-value corresponding to the observed statistic $|t|$ is calculated using the two tail regions shown in Fig. 2.2. Similar tests can be conducted for Case 2 as well by replacing θ_0 with μ and σ with s_X and defining an additional chi-square statistic $\chi^2 = \dfrac{(n-1)s_X^2}{\sigma^2}$ to compare the variances. Confidence intervals (CI) for the data mean can be constructed such that the area under the distribution curve of the test statistic

outside that interval is denoted as α. If the p-value is smaller than α, we can be sure that the actual observed test statistic falls outside the CI. Thus both confidence intervals and p-value can be used to accept or reject a model; the decision is the same based on either criterion. In Case 1 for instance, a tail-area less than 0.05 would mean that under the null hypothesis, there is very small chance of obtaining a true difference larger than the actually observed difference, and hence the true difference must not be zero. Since the true difference is not zero, we reject the null hypothesis. The interpretation is similar in Case 2.

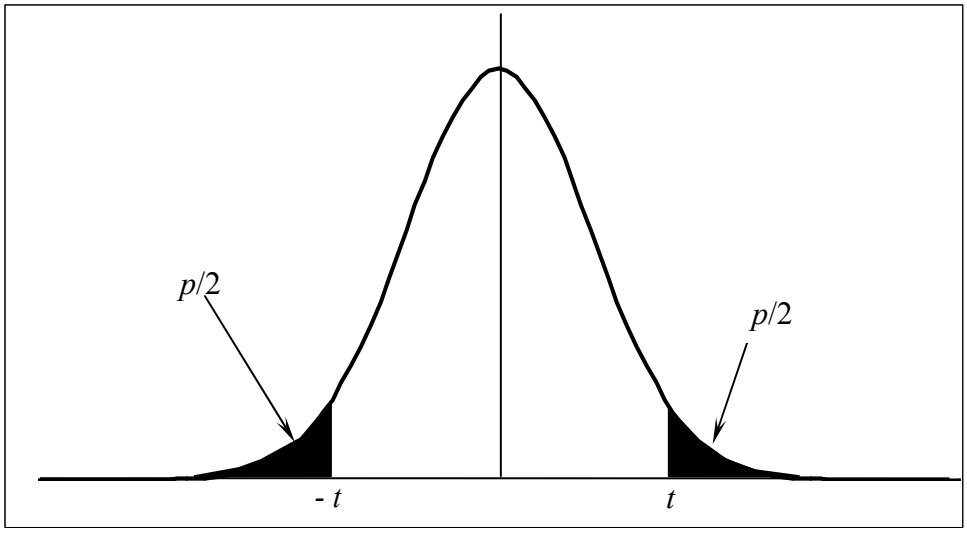


Fig. 2.2 p-value given by tail area

Although p-value may have legitimate meaning in other applications, its ability to explain the difference between prediction and observation under point null hypothesis testing has been debated (Berger and Delampady, 1987; Berger and Sellke, 1987). The use of p-value can be easily misused by decision makers. As often misinterpreted, the p-value is NOT the probability that the null hypothesis is true i.e., a small p-value does not

mean that there is a small probability that the null hypothesis is true. Thus a *p*-value can be used as a qualitative, indirect indicator for large or small standardized error but cannot be treated as a direct quantitative measure of strength of evidence for the null. Refer to the paper by Johnson (1999) for more details on definitions related to *p*-values.

2.3.2 Bayesian point null testing

In Bayesian hypothesis testing, we assign prior probabilities for the null and alternative hypotheses, let these be denoted as $P(H_0)$ and $P(H_1)$ such that $P(H_0) + P(H_1) = 1$. When an evidence or data D is obtained, the probabilities are updated as $P(H_0 \mid D)$ and $P(H_1 \mid D)$ using Bayes theorem. Then Bayes factor (Jeffreys, 1961) B is defined by the first term in the ratio in square brackets on the right hand side of Eq. (2.3).

$$\frac{P(H_0 \mid D)}{P(H_1 \mid D)} = \left[\frac{P(D \mid H_0)}{P(D \mid H_1)} \right] \frac{P(H_0)}{P(H_1)} \tag{2.3}$$

If B is greater than one, the data gives more support to H_0 than H_1. Also the confidence in H_0, based on the data, comes from the posterior null probability $P(H_0 \mid D)$, which can be rearranged from Eq. (2.3) as $\frac{P(H_0)B}{P(H_0)B + 1 - P(H_0)}$. Typically, in the absence of prior knowledge, we can assign equal probabilities to each hypothesis and thus $P(H_0) = P(H_1) = 0.5$. Then the posterior null probability can be further simplified to $B/(B+1)$. Thus a B value of 1.0 represents only 50% confidence in the null hypothesis being true. For the continuous, the Bayes factor can be derived in terms of probability density functions. Suppose $g(\theta)$ is the density under alternative hypothesis H_1: $\theta \neq \theta_0$; then the Bayes factor or weighted likelihood ratio of H_0 to H_1 is given by (Berger and Delampady, 1987)

$$B = \frac{f(x \mid \theta_0)}{\int f(x \mid \theta) g(\theta) dx} \tag{2.4}$$

In the absence of prior knowledge, $g(\theta)$ can be assumed to be $f(\theta)$. Thus Eq. (2.4) can be rewritten using Bayes theorem as $B = \frac{f(x \mid \theta_0)}{\int f(x \mid \theta) f(\theta) dx} = \left. \frac{f(\theta \mid x)}{f(\theta)} \right|_{\theta=\theta_0}$, i.e., the Bayes factor (validation metric) becomes simply the ratio of posterior to prior densities evaluated at the model prediction value.

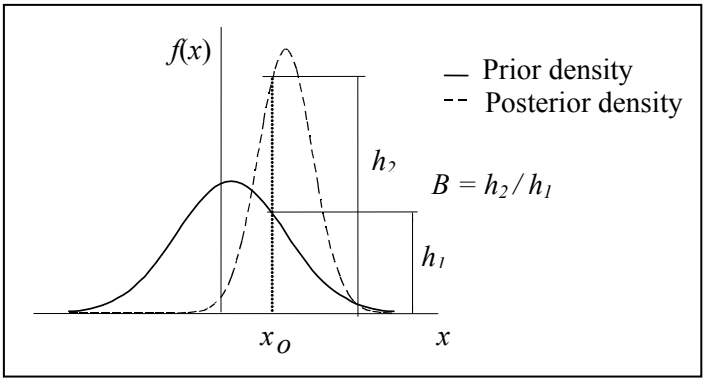


Fig. 2.3 Bayesian Validation Metric

Bayes factors are sensitive to the prior assumptions on the distributions. However, in the context of model validation, the priors are not assumed but the distributions predicted by the model output themselves are used as priors (Rebba *et al*, 2004). Non-informative priors like uniform distributions can always be used if the decision maker insists on giving more objective (frequentist) treatment to the validation problem.

Discussion: Numerical studies (Berger and Sellke, 1987) show that Bayesian and frequentist conclusions might agree or disagree depending on the problem. In the context of model validation and point null testing, what one should be concerned about is the

support for H_0 based on data and any prior information available. This is directly answered by $P(H_0 \mid D)$ whereas the *p*-value estimates the probability of obtaining an error statistic more than the actually observed statistic under null hypothesis H_0. Since the data is judged under the assumption of null being correct, we can say that *p*-value estimates $P(D \mid H_o)$. In the classical approach, we reject the null hypothesis because the error statistic is too large. In the Bayesian approach, we follow the hypothesis which is more probable.

In order to understand the above argument more mathematically, consider the following: the classical approach assumes $P(H_0) = 1$ and $P(H_a) = 0$ and estimates $P(D \mid H_0)$ to accept or reject H_0. Bayesians argue that prior belief in a null hypothesis is not entirely 100% but only 50% thus leaving a certain amount of disbelief in the null in the absence of any evidence or data. Hence it is more reasonable to assume $P(H_0) = P(H_a) = 0.5$. Upon the availability of data *D*, these prior beliefs are updated to compute the probabilities $P(H_0 \mid D)$ and $P(H_a \mid D)$. If $P(H_0 \mid D) > 0.5$, we have an increased confidence in the null hypothesis. Notice that we came to conclusion that "null hypothesis is more likely or more probable after the evidence." This is exactly the reason behind Bayesian hypothesis testing and appears to have more practical use. Computationally, *p*-values require minimal effort as opposed to Bayes factors that require prior information.

One argument on behalf of the classical approach is that no assumptions are made regarding the prior distributions of the model parameters even though we assume $P(H_0) = 1$. It should be remembered that in Bayesian model validation, we do not subjectively choose the priors but simply treat the probability distribution of the model output as the

prior. However, the likelihood function is sometimes assumed to follow normal distribution even when there is no evidence to support that assumption in case of small data sets. The criticism of classical methods is that suppose we have some prior information on the model, the metric ignores that information completely in the analysis. Thus no single method is better than the other for all the validation problems in practice. The ideal approach would be to use frequentist methods but resort to Bayesian methods when there is sufficient information on the prior distributions and likelihood functions.

2.4 Multivariate hypothesis testing for validation

For model response that changes with location (in space or time coordinates), the uncertainty can be characterized by a random field or process depending on the domain of interest. Numerically, the ensemble of realizations of the random field or process may be expressed in the form of a matrix with each realization represented as a row of values at discrete points in the domain. Validation in this situation is performed at finite number of locations or time instants since the experimental data are typically collected at a discrete points only in practice. The elements of each column of the model output matrix may follow a statistical distribution. Measurements made at discrete locations or times are often assumed to be taken independently. However, such observations are correlated in realistic situations and hence this study emphasizes the inclusion of correlations among measurements.

Multiple response quantities can be predicted at a single location. For example, various quantities like stress, mode shape, displacement etc may be computed from the derivatives or the integration of the finite element field combined with the structural

parameters. These different quantities are dependent, or in a first order sense correlated, since they are based on the same input. The uncertainties in the input parameters propagate to these derived quantities. Multivariate distributions can be used to represent such quantities. Similarly, the experimentally observed quantities resulting from the same input or experiment also have correlations. Since any experimental observation contains uncertainty, the measured quantities can be assumed to be correlated random variables. Thus, both computational model outputs and experimentally measured quantities form sets of correlated random variables.

2.4.1 Individual versus aggregate validation

Individual, or univariate, validation compares each model response prediction with a corresponding experimental observation. The validation metric value and hence the confidence measure for one variable may differ from that for another variable. This leads to a practical decision-making problem whether to reject or accept the model when different variables give conflicting inferences. When individual validation indicates that not all the responses match well with the data, it certainly exposes the deficiencies in the model and one must improve the model at that stage. At the same time, it is also important to incorporate the correlations among the model outputs introduced by the PDE or the computational model in the validation process. While the replications of the experimentally observed response may be independent of each other, the various measured response variables themselves could be correlated. We wish to capture these correlations among the data in the aggregate validation metric. Marginal or univariate comparisons do not incoporate the correlation information among multiple responses.

Also, a model that passes the univariate validation process may not pass an aggregate test. In a graphical sense, when the realizations of the model output and the data form two different clouds, the distance between their centroids may be small but the cloud orientations could be different and vice versa. Similarly, their orientations could be same but the scatter in each of the principal directions for each cloud could be different. Univariate or marginal comparisons may miss such observations. Thus the decision maker may use both types of validation metrics, univariate and aggregate, to detect different types of weaknesses in the model. In this regard, aggregate validation helps to assess the overall "quality" of the computational model by comparing all the model output variables simultaneously accounting for the model and data correlations as mentioned above.

In some practical cases involving the use of surrogate models (response surface-based), individual comparison may prove to be inadequate. For example, if we compare only mean values of two random processes for discrete time intervals individually, we would be neglecting the underlying correlation structure of the stochastic process entirely. Since any new model prediction (response at future time period) is based on an underlying correlation structure of the process, it is more sensible to include those dependency relations in the validation metric. It has been argued that multivariate methods limit the inflation of Type I experiment-wise error (Thompson, 1994) that is observed in multiple univariate analyses (such as *t*-tests, ANOVAs, etc). Each individual univariate analysis adds to the chance that one of these analyses will be due to error, hence, the inflation of Type I "experiment-wise" error. More precisely, experiment-wise error is the probability (P_{ew}) that one or more of a series of significance tests (say n)

result in Type I error. If in any single test, the Type I error probability is α, then $P_{ew} = 1-(1-\alpha)^n$.

The metrics discussed in this section are based on hypothesis testing where the null hypothesis is that the model is correct and the alterative hypothesis is that the model is not correct. Both classical and Bayesian methodologies can be used to derive such metrics. Individual validation is handled with univariate analysis while aggregate validation is handled with multivariate statistics.

2.4.2 Classical statistical methods

Let the multivariate output be represented using a matrix $\boldsymbol{X}$ of size $n \times p$ where n is the number of random realizations and p is the number of different response quantities, or the number of spatial or temporal points at which a single response is predicted or observed. Also $\boldsymbol{\mu_0}$ is the vector of mean values of each column of $\boldsymbol{X}$. Let the corresponding observation data matrix be represented as $\boldsymbol{Y}$ with same or different dimensions as $\boldsymbol{X}$. Let $\bar{\boldsymbol{Y}}$ be the mean vector and $\boldsymbol{S}$ be the covariance matrix of $\boldsymbol{Y}$. In this study, the discussion is limited to matrices of equal dimensions and only one-sample hypothesis testing is discussed. See Srivastava (2002) for various other cases. The first similarity measure discussed in this study is based on distance between the two matrices (observation and prediction). The Mahalanobis distance similarity measure is computed as (Srivastava, 2002)

$$d^2 = n(\bar{\boldsymbol{Y}} - \boldsymbol{\mu_0})^T \boldsymbol{S}^{-1} (\bar{\boldsymbol{Y}} - \boldsymbol{\mu_0}) \tag{2.5}$$

The observed data is said to belong to the population (model output) if d^2 is less than a critical value, $d_{crit}^2 = \frac{f \times p}{(f - p + 1)} \times F_{p, f-p+1, \alpha}$ where $f = n - 1$, and $F_{p, f-p+1, \alpha}$ comes from the F distribution. While Eq. (2.5) measures the difference between the centroids of the two data clouds, the covariance structure is a measure of linear dependence among the variables and defines the orientation or alignment of the data clouds. Thus a second metric, covariance similarity between the model output and observed data, based on the log-likelihood ratio test (LRT) is obtained as (Srivastava, 2002)

$$R = - (2m / f) \log (\lambda) \tag{2.6}$$

where $m = f - 2d$, $f = n$-1, and $d = \frac{2p^2 + 3p + 1}{12(p + 1)}$. For large n, R follows a chi-square distribution with $p(p+1)/2$ degrees of freedom, p being the number of model output variables under consideration. The likelihood ratio λ is estimated as

$$\lambda = e^{\frac{fp}{2}} |S|^{\frac{f}{2}} |\Sigma_0|^{-\frac{f}{2}} \left[etr\left(-\frac{1}{2} f \Sigma_0^{-1} S \right) \right] \tag{2.7}$$

where Σ_0 and S are the population and sample covariance respectively, *etr* represents exponential trace of a matrix. The hypothesis that the two matrices are similar is rejected at 100(1-α) % significance level, when R exceeds a threshold value chosen from the chi-square distribution with $p(p+1)/2$ degrees of freedom. Also, a test for comparison of the first few largest eigenvalues (Lawley, 1956) of the model output and data correlation matrices can be performed to check for any significant equality.

The mean and covariance similarity metrics can be combined into a single aggregate metric using the relation shown in Eq. (2.6) but with $f = n$ and

$d = \frac{2p^2 + 9p + 11}{12(p+3)}$ and R following a chi-square distribution with $p(p+3)/2$ degrees of freedom. Also the likelihood ratio λ has the extra term as shown below:

$$\lambda = e^{\frac{fp}{2}} |S|^{\frac{f}{2}} |\Sigma_0|^{-\frac{f}{2}} \left[etr\left(-\frac{1}{2} f \Sigma_0^{-1} \left(S + (\overline{Y} - \mu_0)(\overline{Y} - \mu_0)^T \right) \right) \right] \tag{2.8}$$

Similarly multivariate two-sample testing can also be performed for matrices of unequal sizes. The purpose of combining the two types of validation metrics (mean and covariance comparison) in Eq. (2.8) is just to arrive at a single metric that measures the overall quality of code rather than to avoid the conflicting inferences each metric may provide. The model shall be improved if any one of the two metrics fails to meet the accuracy requirements defined by the corresponding hypothesis test.

It should also be noted that in both univariate and multivariate cases, afore mentioned formulae for statistical tests are only valid under the assumption of normality i.e., x has to be normal and the matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ have to be jointly normal. This condition may not be easily satisfied in most practical engineering problems where the output and observations could be non-Gaussian and highly skewed. Further, when the two data sets have a large number of variables and the measured data is believed to have noise, the data should be filtered and excess variance may be removed. Principal component analysis (PCA) may achieve these two goals for reducing the dimension and noise in the data (Srivastava, 2002; Wentzell *et al*, 1997).

2.4.3 Bayesian method

Consider p outputs (x_1, x_2, x_3,…, x_p) obtained from a computational model; and each model output is treated as a random variable. The joint PDF of the multiple response

quantities is denoted by $f_X(x_1, x_2, x_3,\ldots, x_p)$. Similarly, experimentally observed response quantities may be treated as a set of correlated random variables $(y_1, y_2, y_3,\ldots, y_p)$ with each observation assumed to have a Gaussian zero-mean error for the sake of illustration, with constant variance σ^2. While the validation metric for a single response is simply the ratio of its posterior and prior densities evaluated at a particular model prediction value (Eq. (2.4)), this univariate case can be extended to a more general multivariate case where the overall metric is defined as the ratio of posterior joint probability density to the prior joint probability density. The likelihood function for the experimental observation was assumed to be proportional to the Gaussian density function in Eq. (2.6). When multiple observations (independent as well as dependent) are made, the overall likelihood is then proportional to the multi-dimensional Gaussian distribution. Then a collective comparison can be made using the Bayesian validation metric similar to Eq. (2.4) as

$$B = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{x})^T \boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{x})\right)}{\int\int\ldots\int \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{x})^T \boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{x})\right) f_X(x_1, x_2, \ldots, x_p)\, dx_1 dx_2 \ldots dx_p} \tag{2.9}$$

where $\boldsymbol{V}$ is the covariance matrix of the observed data. Again, B is evaluated at a particular model prediction set $(x_1, x_2, x_3,\ldots, x_p)_0$. The data is said to favor the model if B is greater than one. This metric can also be used for a single response quantity predicted at multiple locations of space and time by simply replacing any i^{th} response quantity x_i with $x(t)$.

One practical difficulty in the metric shown in Eq. (2.9) is the estimation of the joint probability density function for non-normal model outputs. For normal variables, an explicit expression for the joint PDF is available but the construction of joint PDFs for other non-normal cases is quite cumbersome. Also, the densities computed using non-

parametric or parametric methods (Scott, 1992; Tapia, 1978) tend to be either too small or too large in some cases, thus leading to numerical difficulties in the computation of the Bayesian validation metric. Several computational issues need to be resolved before implementing classical as well as Bayesian validation metrics to practical problems.

2.4.4 Transformation methods for non-normal data

The application of transformations to data to achieve normality has been suggested and discussed by several authors (e.g., Srivastava, 2002). The transformed variables can then be used in the proposed validation metrics. A few of the popular transformation methods are discussed in this section, briefly explaining the underlying assumptions in each of the methods. A literature survey reveals that each transformation technique is found to be suitable for a particular application and according to the researcher's preference.

Rosenblatt transformation

Let the non-normal random vector $\boldsymbol{x} = (x_1, x_2,\ldots, x_p)$ have a joint distribution function $F_{\boldsymbol{X}}(x_1, x_2,\ldots, x_p)$. A transformation (Rosenblatt, 1952) based on successive conditioning can be made as follows:

$$\Phi(u_1) = F_{x1}(x_1)$$

$$\Phi(u_2) = F_{x2}(x_2 \mid x_1)$$

$$\Phi(u_3) = F_{x3}(x_3 \mid x_1, x_2)$$

..

$$\Phi(u_p) = F_{xp}(x_p \mid x_1, x_2, \ldots, x_{p-1}) \qquad (2.10)$$

Using Φ^{-1} in each case, one can obtain an independent set of standard normal variables $\boldsymbol{u}$ = (u_1, u_2, u_3,..., u_p). This method requires the knowledge of exact full and conditional densities.

Nataf transformation

The Nataf transformation (Nataf, 1962) addresses practical problems where we usually know only the marginal densities and the correlation structure among various variables. Thus one can define standard normal variates $\boldsymbol{u}$ = (u_1, u_2, u_3,..., u_p) obtained by marginal transformations of (x_1, x_2,…, x_p) as

$$u_i = \Phi^{-1}(F_{xi}(x_i)) \tag{2.11}$$

Further assuming that all u's are jointly normal, one can construct the joint PDF of the model output variables using the relation

$$f_X(x_1, x_2, ..., x_p) = f_{x_1}(x_1) f_{x_2}(x_2) ... f_{x_p}(x_p) \frac{\varphi_p(u_1, u_{2,...,} u_p, \boldsymbol{C}')}{\varphi(u_1)\varphi(u_2)...\varphi(u_p)} \tag{2.12}$$

where $\varphi_p(u_1, u_2,\ldots, u_p, \boldsymbol{C'})$ represents a p-dimensional standard normal PDF and the elements of the equivalent covariance matrix $\boldsymbol{C'}$, are obtained by solving the equation

$$c_{ij} = \iint \left(\frac{F_{X_i}^{-1}(\Phi(z_1)) - \mu_{X_i}}{\sigma_{X_i}} \right) \left(\frac{F_{X_j}^{-1}(\Phi(z_2)) - \mu_{X_j}}{\sigma_{X_j}} \right) \varphi_2(z_1, z_2, c'_{ij}) dz_1 dz_2 \tag{2.13}$$

where c_{ij} and c_{ij}' are the elements of correlation matrices for the original and transformed variables respectively. Both posterior and prior joint PDFs may be derived using Eq. (2.12) to be used in the Bayesian aggregate validation metric given in Eq. (2.9). But one should ensure that u_1, u_2…etc in Eq. (2.12) are jointly normal before applying in Eq. (2.9).

Power and Modulus transformations

Box and Cox (1964) proposed a family of power transformations for the original data points $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$ to define the univariate transformed data as

$$u_i = \frac{x_i^{\lambda} - 1}{\lambda g_1^{\lambda-1}} \quad \text{for } \lambda \neq 0$$

$$= g_1 \log x_i \quad \text{for } \lambda = 0 \tag{2.14}$$

if the data is positive. Here g_1 is the geometric mean of the given data calculated as $g_1 = \left(\prod_{i=1}^{m} x_i \right)^{\frac{1}{m}}$ and λ is a parameter that needs to be estimated.

If some data points are negative, we may consider the transformation

$$u_i = \frac{(x_i + a)^{\lambda} - 1}{\lambda g_2^{\lambda-1}} \quad \text{for } \lambda \neq 0$$

$$= g_2 \log(x_i + a) \quad \text{for } \lambda = 0 \tag{2.15}$$

where $g_2 = \left(\prod_{i=1}^{m} (x_i + a) \right)^{\frac{1}{m}}$ and a is chosen such that $(x_i + a) > 0$ for all i. John and Draper (1980) proposed alternative transformations as

$$u_i = \frac{(|x_i - b| + 1)^{\lambda} - 1}{\lambda g_3^{\lambda-1}} sign(x_i - b) \quad \text{for } \lambda \neq 0$$

$$= g_3 \log(|x_i - b| + 1)\, sign(x_i - b) \quad \text{for } \lambda = 0 \tag{2.16}$$

where $g_3 = \left(\prod_{i=1}^{m} (|x_i - b| + 1) \right)^{\frac{1}{m}}$. The value of b is usually chosen as an arithmetic or geometric mean of the original data $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$. One easy way to find the

likelihood estimate of λ is to maximize the function $L(\lambda) = -(m - 1)\log(s^2_\lambda)$ where s^2_λ is the variance of the transformed data u_i.

The transformations shown in Eqs. (2.14), (2.15) and (2.16) can be applied to multivariate data to marginally transform the non-normal data into nearly Gaussian. However, marginally normal does not automatically mean jointly normal. Hence, we can define a vector of parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_p)$ that can be used to transform each of the random variables. Instead of obtaining the parameters one by one, we can obtain the entire vector in a single estimation (Andrews *et al*, 1971) by finding the maximum value of the function $L(\boldsymbol{\lambda}) = -(m - 1)\log|\boldsymbol{S}_\lambda|$ where $\boldsymbol{S}_\lambda$ is the covariance matrix of the transformed random variables $\boldsymbol{u}_\lambda$.. This transformation produces nearly jointly normal variates, but it is desirable to test the normality of transformed data. In both univariate and multivariate cases, the parameters can be estimated using any standard optimization routines such as steepest descent, Newton-Raphson method etc.

Suitability of transformation methods

While the Rosenblatt transformation is quite accurate, actual closed form conditional distributions are almost impossible to obtain in many cases. Also, in the context of the Bayesian metric, if we do not even know the exact joint PDF of the model output variables, constructing an explicit continuous joint CDF and conditional distributions is even more impossible for large p and hence this method is discarded.

The Nataf transformation has an advantage that one can obtain the normal data using marginal densities alone. But these marginally normal data need not be jointly normal in all cases and hence the method may be inaccurate in some situations. Thus Eq. (2.12) can be used only under the assumption of multivariate normality for the

transformed data (u_1, u_2, u_3,..., u_p). If an exact distribution (closed form) is not available for the model output, this method cannot be used easily. Hence this method should be used with caution and only after checking that the data is jointly normal using standard tests (Srivastava & Hui, 1987).

Thus, in this study, we can use the power transformations proposed by Box and Cox as they are mathematically tractable, simple to implement and do not have strict requirements or significant assumptions. It is also not required to know the exact closed form distributions for each of the model response variables in this method.

For the purpose of computing the Bayesian aggregate validation metric given in Eq. (2.9), the joint PDF can sometimes be either too small or too large, leading to numerical overflow or underflow problems. This computational hurdle can be overcome as follows. Using the concept of transformation of variables, any multivariate function $g(w_1, w_2,\ldots, w_n)$ can be rewritten in terms of a new set of variables $h(z_1, z_2,\ldots,z_n)$ using the relation

$$g(w_1, w_2,\ldots, w_n) = h(z_1, z_2,\ldots,z_n)\begin{vmatrix} \frac{\partial z_1}{\partial w_1} & \frac{\partial z_1}{\partial w_2}\ldots & \frac{\partial z_1}{\partial w_n} \\ \frac{\partial z_2}{\partial w_1} & \frac{\partial z_2}{\partial w_2}\ldots & \frac{\partial z_2}{\partial w_n} \\ .. & .. & .. \\ \frac{\partial z_n}{\partial w_1} & \frac{\partial z_n}{\partial w_2}\ldots & \frac{\partial z_n}{\partial w_n} \end{vmatrix} \tag{2.17}$$

In the context of individual transformations that take place according to the multivariate case described previously, the joint PDF (being a function of p random variables) can be expressed using the standard normal variates as

$$f_{\boldsymbol{X}}(x_1, x_2, \ldots, x_m) = f_{\boldsymbol{Y}}(u_1, u_2, u_3, \ldots, u_p) \prod_{i=i}^{p} \left| \frac{du_i}{dx_i} \right| \tag{2.18}$$

Since the product of various terms in Eq. (2.18) may lead to numerical instability, taking natural logarithm on both sides, the right hand side is reduced to a summation of several terms like $\left| \frac{du_i}{dx_i} \right|$ instead of a product. Also the exponential function evaluation can be omitted in the term $\log(f_{\boldsymbol{U}}(u_1, u_2, u_3, \ldots, u_p))$ since the vector of transformed random variables $\boldsymbol{u}$ form a joint normal distribution with a covariance matrix and zero means. Since the Bayes factor in Eq. (2.9) is the ratio of posterior to prior joint PDFs, taking the natural logarithm of the ratio also leads to significant mitigation of accuracy problems anticipated before. Under joint normality assumption for u_1, u_2, etc., even Eq. (2.12) can be used for the Bayesian aggregate validation metric. The posterior densities are usually derived using Markov Chain Monte Carlo simulation instead of exact analytical integration.

2.4.5 Numerical examples

Example 1: This example deals with the multivariate transformation of non-normal data. Suppose the model output is represented by 4 correlated random variables $\boldsymbol{x} = \{x_1, x_2, x_3, x_4\}$ with the same correlation coefficient of 0.75 between any two variables. Further, their marginal densities are found to be all exponential with mean values of 0.5, 0.25, 1 and 2 respectively. The corresponding experimental data points are generated intentionally from lognormal distributions and distance and covariance measures based on Eq. (2.5) and (6) are compared to check if the model output matrix matches with the experimental data matrix.

10,000 samples from the model output joint distribution are generated and transformed using the Box-Cox method (based on maximizing the logarithm of determinant of the covariance matrix) to obtain the population mean values for the normal variates as (-0.362, -0.325, -0.23, 0.36). Also the transformation parameters $\boldsymbol{\lambda}$ were jointly found to be (0.2645, 0.2636, 0.270, 0.2711). It was observed that the Box-Cox transformation produced nearly joint normal data, since several random linear combinations of the data produced normal densities. Using Eq. (2.14), the 50×4 experimental data points are transformed using this same $\boldsymbol{\lambda}$ vector, to obtain $\boldsymbol{U}$, with mean values of (-0.272, -0.263, -0.0182, 0.811) and sample covariance as

$$\boldsymbol{S} = \begin{bmatrix} 0.107 & 0.036 & 0.158 & 0.317 \\ 0.036 & 0.018 & 0.071 & 0.135 \\ 0.158 & 0.071 & 0.329 & 0.549 \\ 0.317 & 0.135 & 0.549 & 1.351 \end{bmatrix}$$

The sample means for the 50 experimental data points are (0.643, 0.327, 1.298, 2.636) and sample covariance matrix in the original space is

$$\boldsymbol{S} = \begin{bmatrix} 0.402 & 0.151 & 0.703 & 1.499 \\ 0.151 & 0.086 & 0.339 & 0.603 \\ 0.703 & 0.339 & 1.739 & 2.627 \\ 1.499 & 0.603 & 2.627 & 6.629 \end{bmatrix}$$

Distance similarity

The d^2-statistics are computed as per Eq. (2.5) and summarized in Table 2.1. Since the Nataf transformation cannot guarantee joint normal data, the transformation in this example is limited to Box-Cox method only. The critical value for d^2 is calculated as 10.967 for a significance level of 0.05, with degrees of freedom of $p = 4$ and $f = 49$. From Table 2.1, we can conclude that the experimental data and model output matrices do not

have identical mean; the transformed data suggests that there is more error than what we would have estimated from the original data.

Table 2.1. Aggregate comparison

Data type	Observed mean	Predicted mean	d^2	Result
Untransformed	(0.643, 0.327, 1.298, 2.636)	(0.5, 0.25, 1, 2)	3.79	Pass
Box-Cox	(-0.272, -0.263, -0.018, 0.811)	(-0.362, -0.325, -0.230, 0.360)	11.61	Fail

This is evident from the different d^2 values in Table 2. 1, and is the expected result, since the prediction and observation come from two different distributions. The discussion in this example so far relates to the aggregate validation metric from Eq. (2.5) and one can compare the marginal distribution statistics of data and model output as well. Thus each row of data matrix $\boldsymbol{Y}$ is compared to the marginal densities of $\boldsymbol{x}$ and results summarized in Tables 2.2 and 2.3. The critical value for t-statistic in this case is 2.009 for a significance level of 0.05 and 49 degrees of freedom.

Table 2.2. Individual comparison for untransformed data

Variable	Pred. mean	Obs. mean	Obs. std	t
x_1	0.5	0.643	0.64	1.574
x_2	0.25	0.327	0.296	1.87
x_3	1	1.298	1.332	1.608
x_4	2	2.636	2.601	1.739

Table 2.3. Individual comparison for transformed data (using Box-Cox method)

Variable	Pred. mean	Obs. mean	Obs. std	*t*
x_1	-0.362	-0.272	0.331	1.924
x_2	-0.325	-0.263	0.138	3.158
x_3	-0.23	-0.0182	0.58	2.583
x_4	0.36	0.811	1.174	2.715

The smaller the *t* value (last columns in Table 2.2 and 2.3), the more acceptable the model is. The results from Tables 2.2 and 2.3 indicate that the model passed the validation test more easily in the original space than in the normal space. Failure is the correct inference here.

Covariance Similarity

The covariance similarity measures in original and normally transformed space are computed following the same procedures as in Section 4.3.1 but with Eq. (2.6) instead of Eq. (2.5). The results are presented in Table 2.4. For this example, $p = 4$, and using Eq. (2.6), the model is rejected if R exceeds 18.307.

Table 2.4. Covariance similarity measure

Data Type	*R*	Result
Untransformed	33.56	Fail
Transformed	31.67	Fail

Combined metric

Next, a combined metric for distance and covariance similarities is calculated using Eq. (2.8) with the critical value for R as 18.307. The results are summarized in Table 2.5 below. In all cases, the validation inferences are correct.

Table 2.5. Combined metric for distance and covariance similarity

Data Type	***R***	**Result**
Untransformed	39.69 > 18.307	Fail
Transformed	40.05 > 18.307	Fail

Bayesian metric

Validation using Bayesian hypothesis testing is illustrated in this subsection. Aggregate multivariate as well as multiple univariate comparisons are considered. The computational model output x_i follows an exponential distribution with parameter θ_i. In the Bayesian context, the parameters $\theta_i's$ are assumed to be random variables with some joint density and the experimental data is used to update these random parameters. The priors are chosen from the Gamma distribution in this example, since it is a conjugate distribution to the exponential distribution (which is the density of each model response variable).

$$f(\theta_i) = \frac{b^a e^{-b\theta_i} \theta_i^{a-1}}{\Gamma(a)} \tag{2.19}$$

Also the correlation coefficient between any two random parameters θ_i and θ_j, is assumed to be 0.75 (same as for the variables x_i and x_j). Both assumptions regarding priors and correlation were numerically verified to be true for the current example, using Monte Carlo simulation. See Jeffreys (1961), Leonard & Hsu (1999) for detailed information on the selection of prior density. Table 2.6 shows the priors for the distribution parameters. The shape and scale factors (a, b) of the Gamma distribution can be derived from the assumed mean and standard deviation values.

Table 2.6. Priors for the parameters of exponential distribution

Var	μ	σ	***a***	***b***
θ_1	2	0.2828	50	25
θ_2	4	0.5657	50	12.5
θ_3	1	0.1414	50	50
θ_4	0.5	0.0707	50	100

Further, each data point z_j is assumed to have come from the exponential distribution (same as the density of model prediction). Thus the likelihood function in this case is

$$f(z_j \mid \theta_i) = \theta_i e^{-\theta_i z_j} \tag{2.20}$$

The ratio of posterior to prior joint probability densities of these parameters, evaluated at the value (2, 4, 1, 0.5) gives the aggregate validation metric, B. Alternatively, individual Bayes factors B_i can be computed as the ratio of posterior and prior marginal densities for

each of the variables x_i. In either case this factor should be greater than 1.0 in order to infer data support for the model prediction.

The Bayesian updating is performed using Gibbs sampling. We need to evaluate the joint density function in computing the aggregate validation metric. At the end of Gibbs sampling, we have only a large number of joint samples from which marginal densities can be constructed but a closed form equation for the joint PDF is very difficult to obtain especially when the output is non-normal. To overcome this difficulty, one can transform the joint output into multivariate normal space using the Box-Cox power transformation and use Eq. (2.18) to evaluate the joint density (both prior and posterior).

Table 2.7. Posteriors and priors for the parameter θ of exponential distribution

Var	**Prior**		**Posterior**		***B***
	a	***b***	***a***	***b***	
θ_1	50	25	139.54	84.98	0.092
θ_2	50	12.5	130.52	40.087	0.083
θ_3	50	50	137.30	169.55	0.062
θ_4	50	100	119.00	296.84	0.067

Table 2.7 shows the prior and posterior Gamma parameters for each response variable. The overall Bayes factor for the 4 variables was found to be $\mathbf{6.82\times10^{-5}}$ which indicates that model output is not acceptable overall and also the individual comparisons in Table 2.7 indicate almost no support for the hypothesis that the experimental data belongs to the same joint distribution as the model output. Among the several methods considered in

this example (univariate in the original space, multivariate in the transformed normal space, and the Bayes factor test), the methods in normally transformed space and aggregate Bayesian metric reached the correct validation inference.

Example 2: Validation of a three-parameter energy dissipation model for lap joints

The example provided in this section deals with the energy dissipation due to friction at the lap joints in a structure. Here we consider a three-parameter Smallwood model (Smallwood *et al*, 2001; Urbina *et al*, 2003) to study the accuracy of the mathematical model in predicting the loss of energy dues to friction in a lap joint. The purpose of the mathematical model is to predict the dissipation energy D released per cycle at the joint when subjected to harmonic force amplitude of F_0. The hysteresis curve (force vs. displacement graph) for the joint comprises of two symmetrical curves; upper and lower. The energy loss in the joint under one cycle of sinusoidal loading is found by integrating the area under the hysteresis curve and analytically derived as

$$D = k_n \left(\frac{n-1}{n+1} \right) \Delta z^{n+1} \tag{2.21}$$

where k_n is a nonlinear stiffness, n is a nonlinear exponent and Δz is the displacement amplitude obtained by solving the equation below:

$$2F_0 = k\Delta z - k_n \Delta z^n \tag{2.22}$$

where k is a linear stiffness term. The three parameters n, k_n (or $\log(k_n)$ in this case) and k are quantified from the experiments and the statistics are given in Tables 2.8 and 2.9. Each of these parameters is found to follow a normal distribution. Five levels of loading were applied in the experiment at 60, 120, 180, 240, and 320 lb that span the range of loadings the system may be exposed to.

Table 2.8. Statistics of Smallwood parameters

Variable	***n***	**$\log_{10}(k_n)$**	***k***
Mean	1.36	5.855	1172700
Std. Dev	0.068	0.1866	12865

Table 2.9. Correlation coefficients among Smallwood parameters

n	**$\log_{10}(k_n)$**	***k***
1	0.902	0.494
0.902	1	0.2295
0.494	0.2295	1

The same five levels are used in the model computation. 10,000 sets of correlated Smallwood parameters were generated and substituted into Eq. (2.21) to obtain 10,000 samples of energy dissipated per cycle for each force level and it was observed that $-\log_{10}D$ in each case followed a normal distribution. 12 sets of experimental data were obtained by dismantling the structure and reassembling it, thus simulating the stochastic properties of structure (Urbina *et al*, 2003). This test data for dissipation energy for a particular force level may be compared with the D values predicted by the Smallwood model. The comparisons were made marginally (at each of the five different loadings) as well as collectively, using both classical and Bayesian hypothesis testing. Since the model output is Gaussian, no transformation is needed. The results are summarized in Tables 2.10 to 2.13.

Table 2.10. Individual *t*-tests for the mean of predicted energy at each load level

F_0 (lb)	60	120	180	240	320
t	1.175	0.26	0.526	1.144	1.518
$t_{11,\ 0.05}$ (critical)	2.2	2.2	2.2	2.2	2.2
Result	Pass	Pass	Pass	Pass	Pass

Table 2.11. Individual *F*-tests for the variance of predicted energy at each load level

F_0 (lb)	60	120	180	240	320
F	11.09	8.82	8.53	9.78	13.13
$F_{11,\ 0.05}$ (critical)	19.675	19.675	19.675	19.675	19.675
Result	Pass	Pass	Pass	Pass	Pass

Table 2.12. Aggregate comparison in original space (5 variables)

Type of Comparison	statistic	critical value	Result
distance similarity	39.15	31.22	Fail
covariance similarity	-	18.307	N/A
distance + covariance	-	23.685	N/A

The covariance similarity metric in Table 2.12 could not be reported in some cases since the test statistic turned out to be a very large positive number indicating that the test would definitely fail. One possible explanation for this behavior is that the high

correlation among model output variables may have resulted in numerical instabilities during the covariance matrix inversion needed for computing the metric given by Eq. (2.6).

For the Bayesian model validation, the mean values of energy at different load levels (five in this case) are assumed to be random variables that are being updated using the available test data. The posterior and prior densities of those means can be used to calculate the marginal Bayes factors as shown in Eq. (2.4) or the collective metric given in Eq. (2.9). Table 2.13 shows the priors, and the posteriors obtained using a Markov Chain Monte Carlo simulation procedure.

Table 2.13. Bayes factors for energy dissipated at different force levels

Prior		**Posterior**		***B***	**Result**
μ_μ	σ_μ	μ'_μ	σ'_μ	$f(\mu'_\mu)/f(\mu_\mu)\|\mu_\mu$	
4.37	0.0275	4.35	0.0138	1.410	Pass
3.65	0.0275	3.63	0.0114	1.265	Pass
3.23	0.0275	3.21	0.0111	1.062	Pass
2.93	0.0275	2.91	0.0114	0.969	Fail
2.63	0.0275	2.611	0.0123	1.061	Pass

The aggregate Bayes factor as per Eq. (2.9) was found to be **0.013** indicating that the model prediction is not supported by the data in an overall sense, although individually it is slightly supported by the experimental data at several load levels.

For this particular application problem, the model passes when comparisons are made marginally but fails collective comparison. Suppose the correlation among model response variables at different load levels is close to zero, the overall Bayes factor is simply the product of individual Bayes factors (from Table 2.13) and hence the model passes the aggregate comparison test as well. This is an important result, showing that in multivariate model validation, the decisions at the end of the validation process are highly dependent on the correlation structure among the multiple response quantities of interest.

The chapter so far discussed various point-null hypothesis testing methods for comparing model predictions and test data. Both classical and Bayesian methods have been explored for this purpose. The following section address the issue of practical significance of a result as opposed to the statistical significance and also highlights the effect of sample size on the validation inference.

2.5 Interval-based hypothesis testing

2.5.1 Effect of sample size on inference

Apart from philosophical differences, both *p*-values and Bayes factor (or an indirect estimate of posterior null probability, $P(H_0 \mid D)$) are affected by the number of data samples. For example, a typical *p*-value for equality of means under normality assumptions can be computed as $2\left(1-\Phi\left(\frac{\sqrt{n}\,|\bar{x}-\theta_0|}{\sigma}\right)\right)$ for Case 1 described in Section 2.3.1. Although $\bar{x}$ converges quickly with increasing *n*, the *p*-value however can reject

the null hypothesis with a large value for n. Thus, even with very small difference between $\bar{x}$ and θ_0, the null hypothesis can still be rejected with increasing n.

In the Bayesian approach, the posterior null probability $C = P(H_0 \mid D)$ can be derived in this case with a particular choice of priors (Jeffreys, 1961) as $\left[1+(1+n)^{-\frac{1}{2}} \exp\left\{\frac{n^2|\bar{x}-\theta_0|^2}{2\sigma^2(1+n)}\right\}\right]^{-1}$. In the context of model validation however, the priors are not chosen by statisticians by experience but they come from the probability distributions of the computational model output. Since $\bar{x}$ converges very quickly with increasing n, for the sake of illustration, it is assumed that $|\bar{x}-\theta_0|/\sigma$ remains to be 0.1 and thus independent of n. Then one can plot p-value and Bayesian confidence measure C as a function of n as shown in Fig. 2.4.

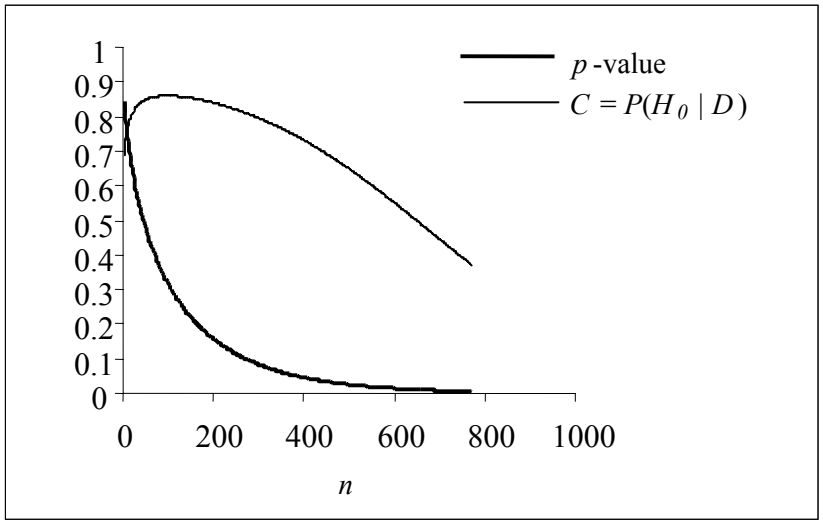


Fig. 2.4 Effect of sample size on p-value and C

As n increases both classical and Bayesian hypothesis tests increasingly reject the null hypothesis for this problem but one can notice that C reaches a maximum value for a particular n whereas p-value monotonously decreases. In order to understand the relation between the sample size and p-value or C, we need to explore the kinds of risks we take by using these metrics for decision making.

There are two kinds of errors related to any statistical significance test: Type I and Type II errors. If we reject a null hypothesis that is actually valid, we commit Type I error (false-positive) and when we accept an invalid hypothesis, we commit Type II error. Since we reject a null hypothesis whenever p-value is less than a significance level, say 0.05, it is said that we limit the probability of committing a Type I error to 0.05. In a classical setting, probabilities are never computed under the assumption that H_a is true and thus the p-value does not measure or consider the Type II error at all in making a statistical decision. As mentioned already in Section 2.3.2, classical hypothesis testing assumes $P(H_0) = 1$ or H_0 is completely true. The Bayesian approach on the other hand assumes both H_0 and H_a are equally likely in the absence of prior knowledge. Thus the computation of C involves likelihood estimation under both null and alternative hypotheses. The Bayesian metric thus includes both types of error in a statistical significance test.

Again it should be remembered that a rejection of null hypothesis would only mean that model prediction and experimental observation are not exactly equal. This does not automatically render the model useless since there is very low probability for two numerical quantities to be equal in practice. Since both p-value and C decrease with increasing n, even accurate models can get rejected under point null hypothesis testing.

However, that collecting additional data should increase the confidence in the decision to accept or reject a model depending on how good or bad the model was in the beginning. An interval based approach for hypothesis testing method is formulated below to address this problem and arrive at consistent decisions.

2.5.2 Formulation for interval-based testing

Interval-based model validation test hypotheses may be represented as H_0: $|\bar{X} - \theta_0| < \varepsilon$ versus H_a: $|\bar{X} - \theta_0| > \varepsilon$. From a classical testing perspective, *p*-value is calculated as the probability of a test statistic being greater than the observed value given the null hypothesis is true. Then it is not quite obvious what distribution *T* follows under this new H_0. When ε is zero, the test statistic *T* follows *t*-distribution or T^2 follows F-distribution with n - 1 degree of freedom for both numerator and denominator. As defined previously in this chapter, *p*-value may estimated as $P(T > t \mid H_0)$ and for the interval hypothesis, the null H_0 can be expressed as $((\bar{X} - \theta_0)^2 < \varepsilon^2)$. By defining $\delta = \frac{n\varepsilon^2}{s^2}$, one can make use of non-central *F* distribution for computing the *p*-value as

$$p\text{-value} = P\left(F_{1,n-1,\delta} > t^2\right) \tag{2.23}$$

As ε becomes smaller, the resulting *p*-value converges to the case of point null hypothesis.

The Bayesian formulation of hypotheses for this case would be H_0: $|\theta - \theta_0| < \varepsilon$. Then we update this hypothesis after observing the data. Here we are testing if the model prediction θ is in fact near θ_0. This is not exactly the same as testing H_0: $|\bar{X} - \theta_0| < \varepsilon$.

First we assume that the null and alternative have equal prior probabilities i.e., $P(|\theta - \theta_0| < \varepsilon) = 0.5$ and then calculate $P(|\theta - \theta_0| < \varepsilon \mid \bar{x})$. Again, note that we do not make any inference on $|\bar{X} - \theta_0|$ as opposed to the classical hypothesis testing.

Suppose we have some prior information on model prediction variable θ in the form of a probability distribution $f(\theta)$, then the Bayes factor can be defined as

$$B = \frac{\int_{\theta_0-\varepsilon}^{\theta_0+\varepsilon} f(\bar{x} \mid \theta) f(\theta) d\theta}{\int_{-\infty}^{\theta_0-\varepsilon} f(\bar{x} \mid \theta) f(\theta) d\theta + \int_{\theta_0+\varepsilon}^{+\infty} f(\bar{x} \mid \theta) f(\theta) d\theta} \tag{2.24}$$

Suppose for the sake of illustration we assume that $f(\bar{x} \mid \theta)$ is $N(\theta, \sigma^2/n)$ and our hypothesis is that θ is near θ_0 with density $N(\theta_0, \beta^2)$ where β is some constant.

Then the posterior null probability $C = P(H_0 \mid D) = \frac{B}{B+1}$ can be calculated as (Schervish, 1995)

$$C = \int_{\theta_0-\varepsilon}^{\theta_0+\varepsilon} f(\theta \mid \bar{x}) d\theta = \Phi\left(\frac{n\lambda[\bar{x} - \theta_0]}{\sigma^2} + \frac{\varepsilon}{\lambda}\right) - \Phi\left(\frac{n\lambda[\bar{x} - \theta_0]}{\sigma^2} - \frac{\varepsilon}{\lambda}\right) \tag{2.25}$$

where $\lambda = \frac{\beta\sigma}{\sqrt{n\beta^2 + \sigma^2}}$. The above expression given in Eq. (2.25) has been derived under a special case of Gaussian assumptions for the model output $f(\theta)$. One can numerically calculate a more general case of confidence measure using Eq. (2.24). This concept can also be extended to the multivariate case.

A "classical"-type solution corresponding to this formulation has been derived from Eq. (2.25) by simply assuming that θ has a very flat density (very large standard deviation) and hence indicating that information on θ_0 is purely objective (Berger and

Delampady, 1987; Schervish, 1995). Thus setting $\beta \to \infty$, the expression for C in Eq. (2.25) reduces to $\Phi\left(\frac{\sqrt{n}\,|\,\bar{x}-\theta_0\,|}{\sigma}+\frac{\varepsilon\sqrt{n}}{\sigma}\right)-\Phi\left(\frac{\sqrt{n}\,|\,\bar{x}-\theta_0\,|}{\sigma}-\frac{\varepsilon\sqrt{n}}{\sigma}\right)$. Thus if θ_0 is deterministic, one can use this expression to estimate the confidence in the null hypothesis. Although this solution tends to an objective, frequentist approach (flat prior), this is not the *p*-value in a classical hypothesis test.

2.5.3 Effect of sample size

In order to examine the effect of sample size on the Bayesian metric, consider the following example: Suppose β=1, ε = 0.2, σ = 1, then the Bayesian confidence measure C versus n for the case $|\bar{x}-\theta_0|$ = 0.1 and $|\bar{x}-\theta_0|$ = 0.25 are given in Fig. 2.5. In the first case, the difference $|\bar{x}-\theta_0|$ is less than ε, and in the second case, $|\bar{x}-\theta_0|$ is greater than ε. As the data set becomes larger, the first model should be increasingly acceptable with increasing n and the second model should be increasingly reject the model.

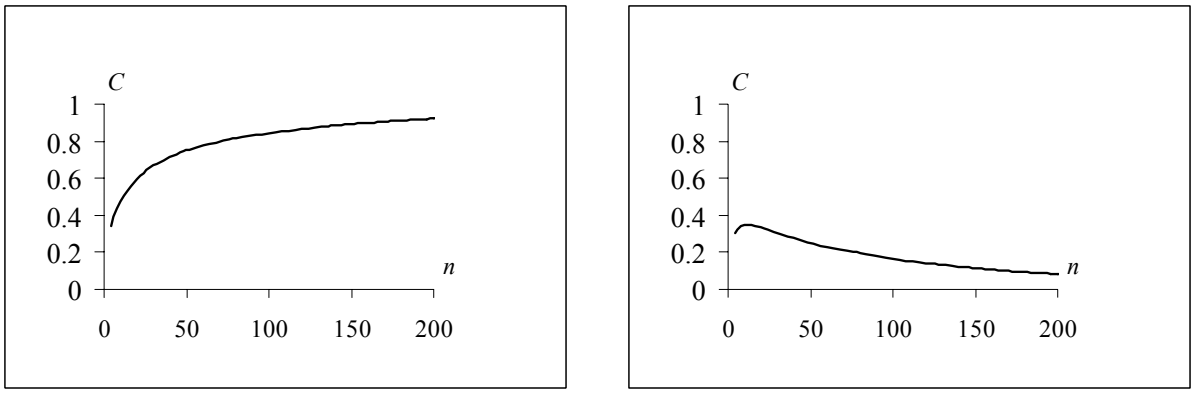


Fig. 2.5 Interval-based Bayesian formulation

It is clear from Fig. 2.5 that the confidence in the first case increases with large n while the confidence drops in the second case where the difference between the model prediction and the data is large. Thus the interval-based formulation gives a consistent result with increasing data size. Suppose we do not have any prior information on θ i.e., θ_0 is deterministic, then setting β to some very large value (10000), the plots are given to mimic a "classical" result in Fig. 2.5. Again additional data rejects a model that is originally not so accurate and accepts a model that is originally accurate enough. Thus the interval-based hypothesis testing formulation appears to be practically more useful and consistent with data size. The classical approach is difficult to implement with this formulation. But the Bayesian approach is easy to implement, and can even be extended to provide a frequentist result.

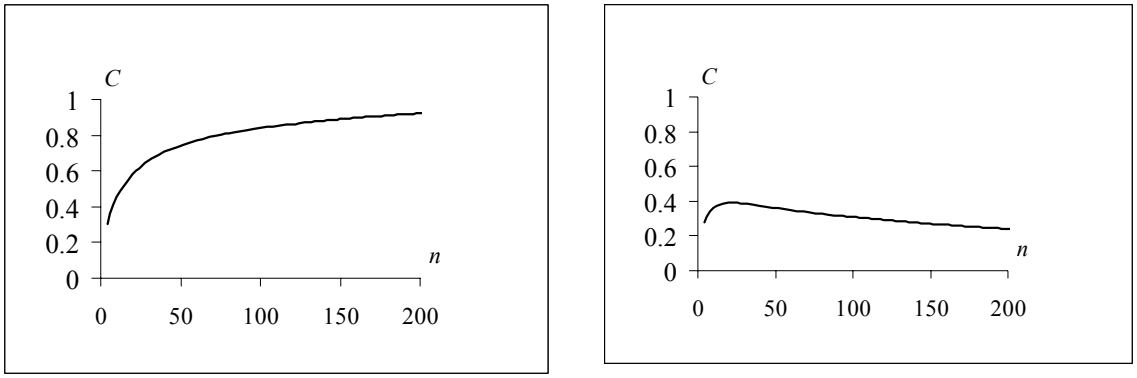


Fig. 2.6 Interval-based classical formulation

2.6 Alternatives to hypothesis testing

In this section, some alternatives to *p*-values and Bayes factors are explored for validation purposes. Each method has its drawbacks and advantages and the decision maker or

model developer has to choose the appropriate metric of comparison depending on the problem.

2.6.1 Decision Theoretic Approach

One approach is the use of decision-theoretic utility or loss functions instead of Bayes factors or *p*-values for testing the null and alternative hypotheses (Ferrandiz, 1985). Calling d_0 as a decision to accept the null H_0: $\theta = \theta_0$, and d_1 as a decision to accept the alternative H_1: $\theta \neq \theta_0$, one can define the utility function $u(d_i, \theta)$ of choosing d_i when θ is the parameter. Using the Bayesian approach, having observed the data x, the decision d_1 is the optimal decision if and only if $E[u(d_1, \theta) - u(d_0, \theta) \mid x] > 0$. The difference in the utility functions is usually chosen as a squared loss function or an absolute error metric (Schervish, 1995).

2.6.2 Equivalence Testing

Model validation is ultimately a test of how well model predictions match with experimental or historical observations. One would think that the burden of proof should rest with the model, to force it to show that it can make accurate predictions. It has been argued that traditional statistical tools (like classical null hypothesis testing) are inappropriate since their ability to detect differences between model output and observation is greatly influenced by the sample size. Thus if data is observed with high variance and has small sample size, the model easily passes the hypothesis test. Equivalence tests are along the lines of the practical formulation described in Section 2.4 and they stress on disproving the alternative hypothesis rather than rejecting a null.

Equivalence tests define an acceptable error unlike the point null hypothesis formulation (Wellek, 2002; Robinson and Froese, 2004). The null and alternative hypotheses for an equivalent test are as follows:

$$H_0: D > u_b \text{ or } D < u_l$$

$$H_1: u_l < D < u_b$$

where D is the difference between the data and prediction, and u_l and u_b are lower and upper limits that can be defined for accuracy requirements. If we reject the null on the basis of data, we conclude that H_1 is true and that model and data are statistically equivalent. First, confidence intervals (CI) are defined under normality assumptions for the difference D based on the observed difference d. Knowing the sample variance, sample size, and actual difference d, the CI can be estimated as $d \pm t_{1\text{-}\alpha} \frac{s}{\sqrt{n}}$ where $t_{1\text{-}\alpha}$ is some critical value determined from the t-distribution with n -1 degrees of freedom. If this CI falls entirely within the equivalence interval $[u_l, u_b]$, we say that model and data are statistically equivalent and hence accept the alternative hypothesis H_1. The concept is further illustrated in Fig. 2.7.

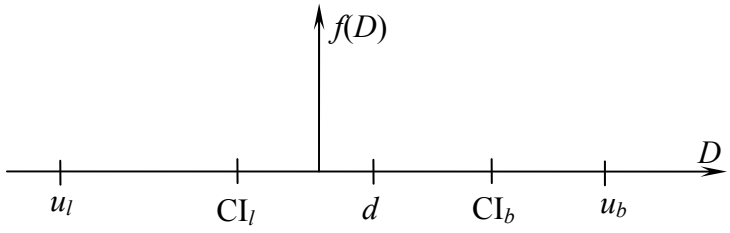


Fig. 2.7 Equivalence testing

In Fig. 2.7, CI_l and CI_b represent the lower and upper bounds of the confidence interval. Here the entire CI lies within the equivalence interval $[u_l, u_b]$ and hence we can conclude that the difference between model and data is smaller than the allowable difference.

Also with a large data set or increasing n and if d lies between u_l and u_b, the confidence interval of d converges and collapses to almost a single point near d and the model simply passes. That is, for a given observed difference d between u_l and u_b, the chances for the model to pass the validation test increase with increasing sample size. Although CIs and conducting equivalence testing seem to be more useful than mere reporting of the p-value, it should be remembered that the CI derivation requires $t_{1-\alpha}$ which is calculated at a significance level α. Practically, it is hard to estimate the distribution of the test statistic under this new null hypothesis and hence the p-values cannot be calculated.

There are also other subjective 'effect size' estimators of practical significance that have been defined as alternatives to p-values. These measures of association or correlation (Cohen, 1994; Kirk, 1996) have not become popular in the validation community since they are only qualitative indicators of difference between model and data and do not provide a quantitative measure of evidence for or against the null hypothesis. Some of the popular indicators include: Cohen's d estimate, defined as $d = \left(\bar{X}_1 - \bar{X}_2\right) / S_{pool}$ where X_1 and X_2 are two different sets of observations of quantities of interest whose difference we wish to test to be zero, and S_{pool} is the pooled variance of the difference. A d value of 0.2 indicates small 'effect size' where as d of 0.8 indicates large 'effect size'. Spearman's rank correlation (Spearman, 1904) is sometimes used as a correlational indicator between data and prediction. Similarly, variance effect-size

indicators are derived from the proportion of the variance in data or in prediction to the total variance such as $\eta^2 = \frac{S_{data}^2}{S_{model}^2 + S_{data}^2}$. Refer to Fern and Monroe (1996) for a detailed discussion on these indicators.

2.7 Model reliability metric

In Section 2.6.2, it was concluded that model prediction and observation can be considered equivalent if the confidence interval for their difference D falls within an "equivalence interval". Although this provides a means to pass or fail a model, it still does not give any estimate on the confidence with which we accept or reject the model prediction. The statistical significance level used in such tests should not be termed as the confidence measure. It would be more useful for a decision maker to have a quantitative measure of the "reliability" or probability of success of the model. Then one can be confident that the systems designed using highly reliable models are reliable as well. Also, expressing the validation results in terms of simple probabilities would be easier to interpret than the often misinterpreted, controversial terms like Bayes factors, posterior densities or *p*-values. Another justification for the use of simple metrics for comparison is that one can avoid the debate over major philosophical differences between frequentist and Bayesian approaches while performing model validation.

Along the lines of equivalence testing, we can define a simple metric $r = P(-\varepsilon < D < \varepsilon)$ to indicate the model reliability, i.e., the probability that the observed difference is within a small interval. The accuracy requirement here is ε, which helps to estimate the probability, and the adequacy (confidence) requirement is c such that we accept the

model prediction only when $P(-\varepsilon < D < \varepsilon) \geq c$. Depending on the nature of model output and data, the difference between them (D) will be an uncertain quantity. In this study, uncertainties are characterized using continuous probability distributions only. Also, when D is multivariate, multiple threshold values may be defined for ε. Similarly, single or multiple confidence requirements for c can be defined while performing marginal or collective comparisons. Several cases are considered below:

Case 1: The model prediction is a single number x_0 while the data is $X = \{x_1, x_2, \ldots, x_n\}$ which are replicated experimental measurements taken for the same input. The validation question in this case would be to ask if the condition $P(|\bar{X} - \theta_0| < \varepsilon) > c$ is satisfied. Suppose $\bar{X}$ follows a normal distribution $N\left(\bar{x}, \frac{s}{\sqrt{n}}\right)$ where $\bar{x}$ is the observed sample mean of X and s is the sample standard deviation. Then the model reliability is calculated as

$$P(H_0) = r = \Phi\left[\frac{\sqrt{n}\left(\varepsilon - |\bar{x} - \theta_0|\right)}{s}\right] - \Phi\left[\frac{\sqrt{n}\left(-\varepsilon - |\bar{x} - \theta_0|\right)}{s}\right] \tag{2.26}$$

It should be noted that $\bar{X}$ ideally follows a t-distribution but for the sake of simplicity shown as Gaussian in Eq. (2.26). Suppose $|\bar{x} - \theta_0|$ is 0.2 and the standard deviation of the data s is 2.0, from the example discussed in Section 2.5.3. Also ε is assumed to be 0.1. In a deterministic sense, the observed difference definitely does not lie in the interval [-0.1, 0.1] as an accuracy requirement. However since the observed difference is a random quantity, it is more rational to estimate the probability of the difference falling within a tolerance interval. A plot between model reliability r and sample size n is shown in Fig. 2.8. The trend looks similar to the Bayesian confidence

measure *C* given in Fig. 2.4, and the model prediction is judged to be of low reliability as *n* increases. This is the correct inference.

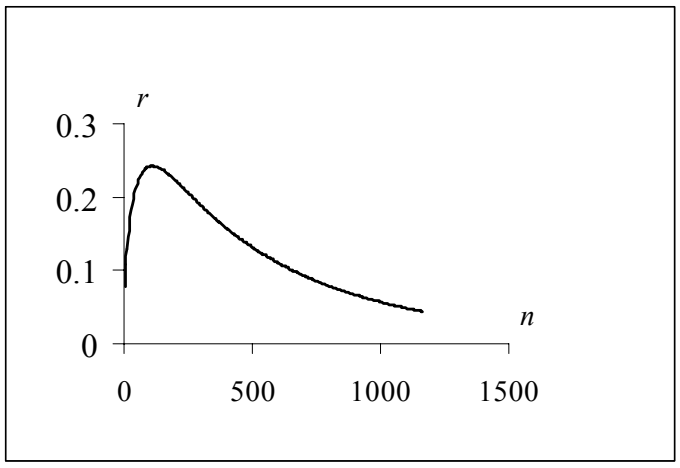


Fig. 2.8 Model reliability versus sample size

If the reliability requirement for this case had been defined at 95%, the model will not pass for any size of *n* as seen from Fig. 2.7. The maximum confidence we can report from the given information would be 24.2% when *n* is 108.

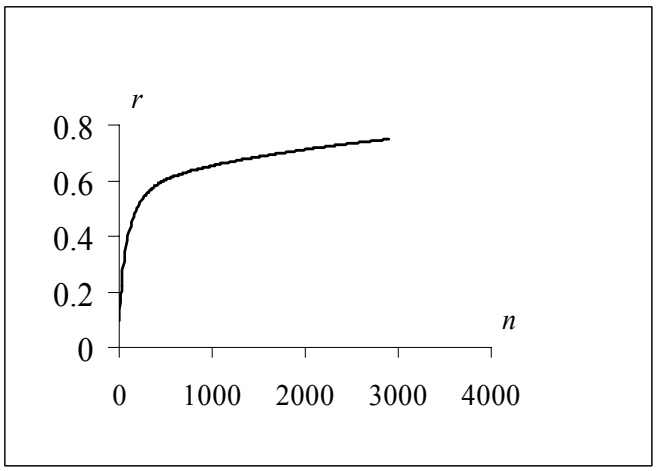


Fig. 2.9 Model reliability and sample size

Suppose $|\bar{x} - \theta_0|$ is 0.075 and it falls inside the interval [-0.1, 0.1], then the model reliability as a function of the sample size is plotted as shown in Fig. 2.9. Thus when the observed difference between the model prediction and the observation falls within the predefined interval, the probability of finding that difference in the interval increases with increasing sample size. Once again, increasing sample size confirms the correct inference. Point null hypothesis testing using classical and Bayesian approaches could not capture this feature and hence are of less practical use for model validation.

When $n < 30$, the normality assumption may not be valid and hence once can use the bootstrap estimate (Efron and Tibshirani, 1993) of sample mean for each resampling iteration and count the number of samples with the prediction difference smaller than the threshold ε, to compute the model reliability. However, it has been shown through some numerical studies (Davison and Hinkley, 1988; Young and Daniels, 1990) that bootstrap techniques can produce noticeably biased results and also require some computational effort for resampling a large number of samples. Analytical methods can also be used with saddlepoint approximations to compute the probabilities (Jensen, 1995).

In summary, the model reliability metric $P(-\varepsilon < D < \varepsilon)$ to compare the sample mean and a prediction can be calculated using any of these three methods:

a) closed form equation under normality assumption for large data sets

b) bootstrap estimates of sample statistics

c) analytical saddlepoint approximations.

Case 2: The model output follows a continuous distribution $f(x)$ while the data $Y = \{y_1, y_2, \ldots, y_n\}$ is observed not for a particular value of input but for a wide range of input parameters during the experiment. Sometimes, historical data may be used to validate a

model in which case also, the corresponding inputs to data and model prediction are unknown. The validation question in this case would be to ask if the test data or the sample belongs to the population $f(x)$ of model outputs. The classical solution is to compare the respective means and standard deviations of sample and population. Since two different probability density functions can have the same first two moments, this type of comparison is not rigorous enough. The comparison criterion should thus include the entire information on the probability density function. It has been recommended to compare the moment or cumulant generating functions of data and model output in such cases (Koutrouvelis, 1980; Cabana and Quiroz, 2005).

For any suitable value of λ, the cumulant generating function (CGF) for the model output x is given by $K(\lambda) = \log\left[\int_{-\infty}^{\infty} e^{\lambda x} f(x)dx\right]$ while the empirical CGF of the data is estimated using the relation $K_n(\lambda) = \log\left[n^{-1}\sum_{i=1}^{n} e^{\lambda y_i}\right]$. The model reliability in this case is defined as $r = P(K_n(\lambda) - k(\lambda)| < \varepsilon)$. By resampling the data Y using the bootstrap method, we can calculate $K_n(\lambda)$ several times and hence compute the model reliability r. It should be noted that ε here carries no physical meaning unlike in case 1 where ε was defined as a threshold for limiting predictive inaccuracy. In this case 2, ε can be defined as some percentage value of $K(\lambda)$. Since PDF and CGF are directly related, comparisons at different values of λ indirectly represent comparisons across a wide range of model predictions. Although asymptotic distributions can been derived for $K_n(\lambda)$, we will limit the analysis to bootstrap in this study for the sake of simplicity.

Multivariate Comparison: The univariate comparisons in cases 1 and 2 can easily be extended to multivariate problems. Instead of comparing one sample mean at a time to the corresponding model prediction, multiple sample means can be compared simultaneously to compute the overall model reliability.

For example, $P(|\bar{y}_1 - x_1| < \varepsilon_1 \cap |\bar{y}_2 - x_2| < \varepsilon_2 \cap ... \cap |\bar{y}_m - x_m| < \varepsilon_m)$ can be calculated as a system reliability problem as opposed to a component reliability formulation used in case 1. Similarly, CGF of the multivariate model output, i.e., $K(\lambda_1, \lambda_2, \lambda_3, .., \lambda_{m-1}, \lambda_m) = \log\left[\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} e^{\sum_{i=1}^{m} \lambda_i x_i} f(x_1, x_2, ..., x_m) dx_1 ... dx_m\right]$ can be compared with the corresponding empirical CGF $K_n(\lambda_1, \lambda_2, ..\lambda_m) = \log\left[n^{-1} \sum_{i=1}^{n} e^{\lambda_1 y_{i,1} + \lambda_2 y_{i,2} + ... + \lambda_m y_{i,m}}\right]$ using bootstrap or analytical approximations.

In summary, whether it is required to compare the data mean to the model prediction or to test if the data set belongs to the predicted distribution, bootstrap or saddlepoint approximations (Jensen, 1995) can be used in a reliability estimation formulation. Using this approach, both univariate and multivariate problems can be addressed. The model reliability assessment is quite different from the point null hypothesis testing approach. It clearly incorporates both accuracy and adequacy requirements to facilitate practical usefulness, and is not adversely affected by increasing sample size. The advantages and drawbacks in using some of the various metrics discussed so far in this section are summarized in the Table 2.14.

Table 2.14. Statistical methods for model validation

Method	Advantages	Disadvantages
Classical point null hypothesis testing (*p*-values)	Simple to estimate, well established mathematical methods, objective and frequentist, addresses adequacy and very stringent accuracy requirement	Confusing and often misused, ignores Type II error, cannot incorporate prior knowledge, no direct assessment of the model, cannot be used for extrapolation purposes
Bayesian point null hypothesis testing (Bayes factors)	Includes both Type I & II errors, incorporates prior knowledge, direct assessment of the model, addresses adequacy and very stringent accuracy requirement, can be used for extrapolation application domain	Relatively difficult to compute, likelihood is assumed sometimes, needs prior distributions (based on model, however)
Probability intervals (classical)	Simple to use, no prior assumptions on the distributions, addresses adequacy and no accuracy requirement	Not clear how to extrapolate to the application domain, possibility of misuse in interpreting the confidence level
Interval hypothesis testing (classical) Equivalence test	Better method than point null tests, easy to estimate and interpret the result, addresses adequacy and adequacy	No direct assessment of hypothesis, ignores Type II error, cannot incorporate prior knowledge, cannot be used for extrapolation purposes
Interval hypothesis testing (Bayesian)	Better than point null tests, includes both types of errors, incorporates prior knowledge (model prediction), direct inference on the null hypothesis, addresses accuracy and adequacy, can be used for extrapolation	Difficult to estimate, likelihood is assumed sometimes
Model reliability formulation	No priors required, direct assessment of model prediction quality, addresses adequacy and adequacy, relatively easy to compute	Not clear yet how to extrapolate to the application domain
Decision theoretic approach	Incorporates prior knowledge, easy to interpret the results, can include accuracy and adequacy	Subjective definition of utility functions, use for extrapolation purposes is questionable
Effect size indicators	Simple to compute, several available, addresses adequacy but very stringent accuracy	Qualitative measures, no direct assessment of confidence in model, cannot be used for extrapolation purposes

2.7.1 Numerical examples

Example 1: The objective of this example is to illustrate the implementation of the model reliability analysis formulated in Section 2.7. Consider a cantilever beam that has a

natural frequency Ω as a function of material and geometric properties given as (Cruse, 1997)

$$\Omega = 562\sqrt{\frac{Et^2}{12\rho L^4}} \tag{2.27}$$

where E is Young's modulus of the beam, ρ is the density of the material, t is the beam thickness and L is the length of the beam. The natural frequency can be treated as the model response for given random inputs: $E \sim N$ (30, 0.04), $t \sim N$ (1, 0.1), $\rho \sim N$ (0.01, 0.002) and $L \sim N$ (20, 2). The probability density function of Ω is calculated numerically using Monte Carlo simulation and found to approximately follow a lognormal distribution i.e., $\Omega \sim LN$ (3.119, 0.251). This however does not mean that the actual analytical distribution of Ω itself would be lognormal. Also it should be noted that the standard deviations of E and ρ are sufficiently small so that the simulation produced only samples that are positive and the problem of calculating the square-root of a negative quantity does not arise. Suppose a beam with properties $E = 28$, $\rho = 0.01$, $t = 1.1$ and $L = 22$ is tested to measure the natural frequency by some experimental procedure. Since measurement uncertainty cannot be captured from a single experimental observation, assume several repeated measurements are taken as $\boldsymbol{y}$ = (19.94, 20.03, 18.12, 20.65, 19.64). These data points are assumed to have come from a Gaussian distribution here, but in general can be from any distribution. The corresponding model output for the same input values is found to be $\Omega_0 = 19.51$ rad/s.

Now the null hypothesis $|\bar{\boldsymbol{y}} - \Omega_0| > \varepsilon$ can be verified using an equivalence test or the model reliability metric $r = P(-\varepsilon + \Omega_0 < \bar{\boldsymbol{y}} < \Omega_0 + \varepsilon)$. The accuracy threshold ε can be chosen as 5% of the model prediction i.e., 1.951. Thus the probability $P(18.534 < \bar{\boldsymbol{y}} <$

20.485) needs to be computed. Suppose each data point in $\boldsymbol{y}$ follows a normal distribution, $\bar{\boldsymbol{y}}$ would then follow normal distribution as well with statistics N(19.676, 0.4222). The summary of model reliability results using normality assumption, bootstrap method (million samples), and saddlepoint method are presented in Table 2.15. The difference between bootstrap and analytical results was found to be only 0.2% and this difference is expected to decrease for large sample sizes.

Table 2.15. Cantilever beam-results of model reliability analysis

Method	***r***
CLT	0.9680
Bootstrap	0.9914
Analytical	0.9941

Suppose several beams are tested with different materials and available dimensions that possibly cover the range of the input parameters. The observations form a small subset of all natural frequencies that are possible for the cantilever beam. If we have 15 observations $\boldsymbol{y}$ = {23.260, 33.091, 18.248, 16.422, 29.480, 16.338, 24.682, 21.045, 14.011, 33.832, 23.989, 17.969, 17.689, 22.509, 24.318} on natural frequencies available from some database or new experiments, validation in this case would mean testing whether these samples come from the probability distribution $f(\Omega)$. Since Ω from the model prediction follows a lognormal distribution, $\log_e(\Omega)$ follows normal distribution with a mean value of 3.119 and standard deviation of 0.251. Then we can test if each sample $\log(y_i)$ belongs to that particular normal distribution.

As suggested previously, we compare the empirical CGF of the data with that of the model prediction. For any Gaussian density function $N(\mu, \sigma)$, the CGF $K(\lambda)$ is defined as $(\mu\lambda + 0.5\sigma^2\lambda^2)$. The empirical CGF for the data is given by $K_{15}(\lambda) = \log\left[6.666\sum_{i=1}^{15} e^{\lambda y_i} \right]$. A plot of $K_{15}(\lambda)$ and $K(\lambda)$ is shown in Fig. 2.10.

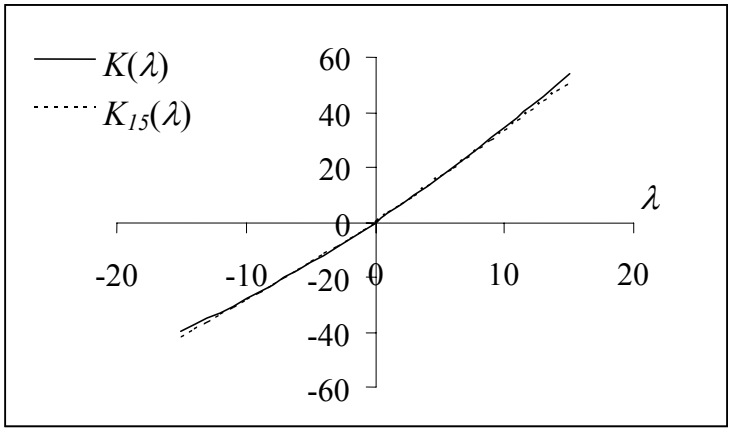


Fig. 2.10 Comparison of CGFs

Even graphically, the two CGFs match very well. By resampling ***y***, we can plot a family of $K_{15}(\lambda)$ curves and compare against $K(\lambda)$ numerically.

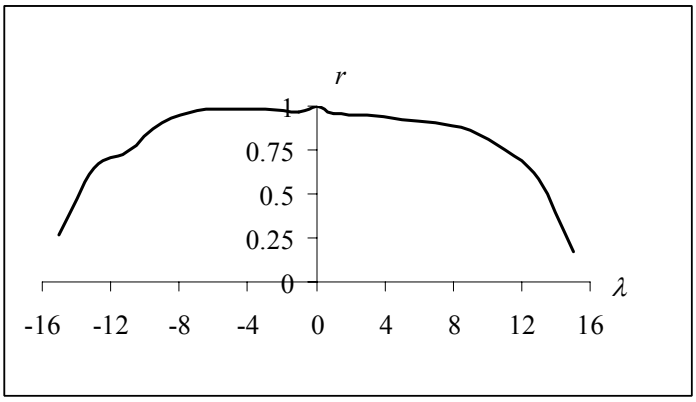


Fig. 2.11 Model reliability variation with λ

As discussed in Case 2 of Section 4, the model reliability $r = P(|K_n(\lambda) - K(\lambda)| < \varepsilon)$ is computed for different values of λ, and with ε chosen as 5% of $K(\lambda)$. The results are shown in Fig. 2.11.

Discussion: In the first part (Case 1) of the problem, there was nearly 99% chance that model prediction is close to the data mean as required by the accuracy limit ε. In the second part (Case 2) however, CGFs have been compared. Since all moments can be derived from the CGF, comparing the CGFs of data and prediction would be identical to comparing their respective moments simultaneously. It is advised that such bootstrap comparisons be made at smaller values of λ near zero. In this problem, there is 95-97% reliability in the vicinity of $\lambda = 0$. Also, the plot shown in Fig. 2.11 is not smooth since the data is discrete and hence the CGFs are not smooth.

As noticed in this example, the accuracy and adequacy thresholds that we define are still subjective. However, no distributions were assumed for the data or any where else during these calculations.

Example 2: The objective of this example is to illustrate how various types of validation metrics can lead to different conclusions for a multivariate comparison problem. Consider a three-parameter Smallwood mode1 shown in Section 2.4.5. The model and test data comparisons can be made individually (at each of the five different loadings) as well as collectively, using the proposed reliability metric. The mean values of energy predicted (×) at different load levels are plotted against the data (-) in Fig. 2.12.

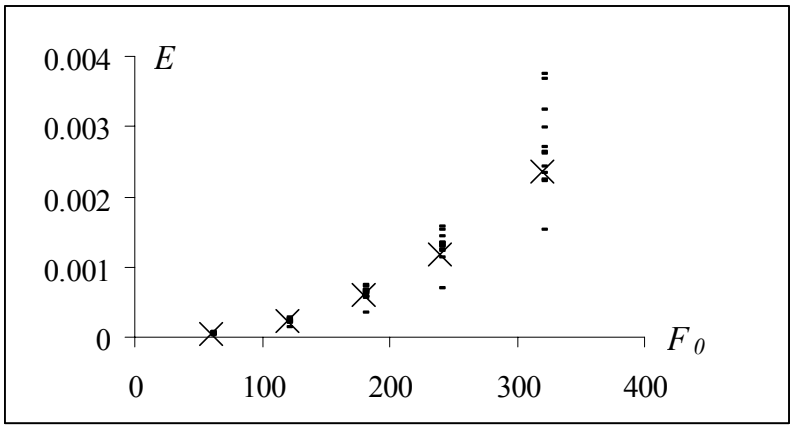


Fig. 2.12 Force amplitude vs. Energy

Individual probabilities $P(L_i < \bar{\boldsymbol{E}}_i < U_i)$ as well as overall probability $P\left(\bigcap_{i=1}^{5} L_i < \bar{\boldsymbol{E}}_i < U_i\right)$ can be computed where L_i and U_i represent lower and upper bounds defined for the experimentally observed mean energy $\bar{\boldsymbol{E}}_i$ at i^{th} load level. Also L_i and U_i are chosen as 0.95 times and 1.05 times the mean predicted energy as given in Table 2.16.

Table 2.16. Error bounds for model prediction

Load lb	**Mean prediction**	$\boldsymbol{L_i}$	$\boldsymbol{U_i}$
60	4.3421E-05	4.1250E-05	4.5592E-05
120	2.2652E-04	2.1519E-04	2.3784E-04
180	5.9662E-04	5.6679E-04	6.2645E-04
240	1.1877E-03	1.1283E-03	1.2471E-03
320	2.3659E-03	2.2476E-03	2.4842E-03

Table 2.17 shows various types of validation metrics determined under classical and Bayesian point null hypothesis testing, interval hypothesis testing and the new model

reliability formulation. The metrics have been calculated for model prediction and data observed at different load levels. If we ignore the correlations among the data and compute overall reliability of the model as the product of each reliability estimate, we get $r = \prod_{i=1}^{5} P\left(L_i < \bar{\boldsymbol{E}}_i < U_i\right)$=0.00386. However, a joint probability estimate (an analogy to system reliability problem) considering the bootstrap samples of the entire 12×5 data matrix yield the overall reliability estimate as $r = P\left(\bigcap_{i=1}^{5} L_i < \bar{\boldsymbol{E}}_i < U_i\right)$=0.02. Thus the individual reliabilities are much larger than the overall reliability estimate.

Table 2.17. Model validation metrics at each load level

Load lb	**60**	**120**	**180**	**240**	**320**
Point null *p*-value (classical)	0.2926	0.76	0.5097	0.1628	0.1013
Point null posterior probability (Bayes)	0.6723	0.7751	0.7442	0.5622	0.4521
p-value (Interval-based testing)	0.4037	0.999	0.916	0.12	0.037
C Bayesian interval testing	0.3922	0.8567	0.681	0.2916	0.1337
$r = P(L_i < \bar{\boldsymbol{E}}_i < U_i)$ (Model reliability method)	0.3384	0.7207	0.5714	0.2497	0.1111

In Table 2.17, the *p*-values computed using interval-based hypothesis testing are dependent on the choice of ε. The model reliability metric calculated the probability of observing the data within certain bounds of model prediction. When the model predictions at each load level are ranked based on the validation metric value, all the metrics give the same rank order indicating their general agreement, although their interpretations are different. However, for specific load values, different metrics can

result in different conclusions regarding the model validity. For example, the point null hypothesis test accepts the model prediction at all loads since the *p*-value is larger than 0.05, whereas the interval hypothesis test tends to reject the model for three of the loads.

2.8 Summary

This study investigated various statistical methods for model validation. The inadequacies of point null hypothesis testing are highlighted, and a more practical interval-based hypothesis formulation is argued for. Bayesian hypothesis testing is found to be a direct way to assess the strength of evidence to a model, as opposed to the use of *p*-values in classical hypothesis testing. A direct approach to estimate the model reliability as the probability of the data falling within a range of model prediction has been proposed. The model reliability metric and the interval-based Bayesian metric consistently reject an invalid model and accept a valid model as the sample size increases to large values.

This chapter also addressed the validation of computational models with multiple outputs using multiple observations from the experiments. Both univariate (individual) and multivariate (aggregate) comparisons can be implemented using hypothesis tests. In the case of classical hypothesis testing, when the normality assumption for the data is violated, the original samples are appropriately transformed to normal variates and test statistics are calculated. The aggregate Bayesian validation metric requires the ratio of posterior to prior joint probability density functions. While a closed form expression is available for the multi-normal density, the estimation of non-normal multivariate densities is often cumbersome involving series expansions or iterative techniques and also the PDF values tend to be too small or too large. In this case, the Box-Cox

transformation ensures that the joint PDF be expressed as a product of a multi-normal density and a correction factor. This simplifies the calculations and the construction of multivariate density without compromising accuracy.

CHAPTER III

EXTRAPOLATING VALIDATION INFERENCES TO APPLICATION DOMAIN

3.1 Overview

Models are often validated in a controlled environment conducting a limited number of small scale tests. Also, the response quantity of interest in the target application may be different from the validated response quantity. In some cases, validation data may be available in the nominal region and the field application may involve off-nominal (tail) behavior. When system-level tests are not feasible, component level data may be used to make partial inference on the validity of system-level prediction. In all of the above cases, inferences from the validation domain have to be extrapolated to the untested region. A Bayesian framework for drawing inferences for predictions in the untested domain is developed and implemented using Bayesian networks (BN) in this study. Also, the proposed Bayesian framework requires numerous evaluations of the computational model output or the joint densities, which could be very expensive. In this study, saddlepoint approximation and Laplace approximation-based techniques are used to carry out multivariate integrations needed for obtaining the marginal and conditional distributions. Also the uncertainty in the model output is quantified using saddlepoint approximations as well instead of more expensive response surface construction.

It should be noted that the work reported in this draft is fairly recent and hence no extensive literature is available at this stage. However the need for assessing extrapolation has been repeatedly stressed in several studies by Oberkampf and Trucano

(2002), U.S. Department of Defense (DMSO, 1996), Thacker and Huyse (2002), American Society of Mechanical Engineers Standards Committee (ASME PTC#60) on verification and validation of computational solid mechanics, etc. Extrapolation itself is not a fresh topic in data analysis and is of great importance in various applications. Methods have been developed in geographic information science (Pontius and Batchu, 2003; Pontius *et al*, 2003) to estimate the precision for an extrapolation into the future, based on the validation from a previous time. Combination of validation and calibration was used to linearly extrapolate land use changes for a future time period. Statistical extrapolation techniques have been widely used in climatic change simulation (Busch and Heimann, 2001). In environmental sciences, laboratory results were extrapolated to the field conditions and across various ecosystems (Livingston *et al*, 1985). The need for extrapolation in predictive exposure (risk) assessment has been identified by EPA (Beck *et al*, 1994). Linear extrapolation using time series forecasting is a well developed research topic in financial and management sectors (Box and Jenkins, 1974; Williams and Goodman, 1960).

All the extrapolation studies mentioned above assume a linear model behavior and restrict to spatial and temporal predictions. Typically, validation experiments are limited to a subset of physics and hence may not cover the range of physics required for model actual application. A mathematical link between the target application and validation experiments must be established (Hills and Leslie, 2003). With such knowledge, validation experiments can be weighted to better represent the target application (Hills and Trucano, 2001). First order sensitivity factors were taken by Hills

and Trucano (2001) as measure of dependency between the validation and extrapolated regions and the analysis was limited to Gaussian model outputs.

3.2 Extrapolation methodology

A Bayesian methodology is pursued in this section, for two cases of extrapolation. The first case deals with extrapolating validation inferences for one quantity to a different response quantity for which data is absent. The second case addresses the task of validation with change in the input conditions. Further this case can be divided into two categories: a) A model may be validated using nominal input values for the experimental set up while the decision variable could be the model prediction for tail inputs b) nature of input condition can be different in validation and target domains i.e., change in type of input loading, material etc. In both cases, a mathematical link between the target application and validation experiments is established using the Bayes network concept.

3.2.1 Case 1: Validated and decision variables are different

Often the quantity validated and the decision variable (quantity of interest in target application) are quite different. Experimental limitations may define the quantity to be measured for the purpose of validating a model. For example, one may validate the axial strain predicted by a model using strain measurements in the laboratory, but the variable that affects the design decision could be shear or torsional stress. Those decision variables can be directly or indirectly related to normal stress through some linking variables. Similarly a decision variable could be the probability of failure of the structure whereas validation may be limited to stress prediction. If the decision variable is not too

different from the validated variable, we can accept the model prediction in untested region with some confidence, if a mathematical link between the decision variable and validation domain can be established. When such an explicit relation cannot be established, sensitivity analysis could give a first order relation between the validation and decision variables. The confidence or updated belief in the extrapolation is then derived from the validation metric in the tested region.

Consider a computational model $y(x, \boldsymbol{\alpha})$ in the validated region. Inferences need to be made for a decision variable $h(x, \boldsymbol{\alpha}, \boldsymbol{\beta})$ with $\boldsymbol{\alpha}$ being a set of input random variables (x could represent space or time co-ordinates) and $\boldsymbol{\beta}$ an additional set of random variables in the application domain. Suppose the computational model y is validated using experimental observations z; then the density functions associated with y and hence those of $\boldsymbol{\alpha}$ can be updated using the Bayes theorem. Thus the joint probability distribution and hence the marginal densities of each of the input parameters in $\boldsymbol{\alpha}$ can be updated as

$$f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha} \mid z) = \frac{f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) f(z \mid y(x, \boldsymbol{\alpha}))}{\int f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) f(z \mid y(x, \boldsymbol{\alpha})) d\boldsymbol{\alpha}} \tag{3.1}$$

where $f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$ is the prior density, and $f(z \mid y(x, \boldsymbol{\alpha}))$ is the likelihood function. The updated parameters can then be used to estimate the updated distribution for h by generating input parameters from the posterior density $f_{\alpha}(\boldsymbol{\alpha} \mid z)$ and substituting them in $h(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid z)$. The new and old densities of h can then be compared similar to Eq. (2.4) to assess the predictive capability of the model in the application domain.

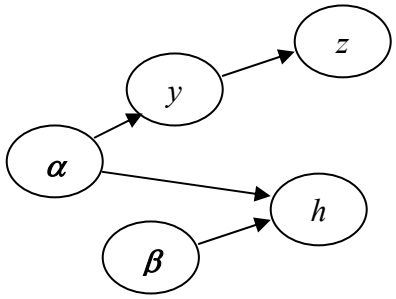


Fig. 3.1 Bayesian network representation of validation and extrapolation

The ratio, $B_h = f(h(\alpha, \beta \mid z))/ f(h(\alpha, \beta))$ is treated similar to the Bayes factor in Eq. (2.4) in assessing the confidence in the decision variable or the model in the application domain. The integration required in Eq. (3.1) can be calculated using Markov Chain Monte Carlo techniques. The quantities y, z, α, β, and h can be linked through a Bayes network as shown in Fig. 3.1.

Bayes networks have been used in artificial intelligence (Heckerman *et al*, 1994), engineering decision strategy (Jensen and Jensen, 2001), safety assessment of software-based systems (Dahll, 2000), and model-based adaptive control (Friis-Hansen *et al*, 2000). Bayes networks have also been applied to the risk assessment of water distribution systems, as an alternative to fault tree analysis (Castillo *et al*, 1999). Recently, the Bayes network concept was extended for structural system reliability reassessment by Mahadevan, Zhang, and Smith (2001) by including multiple failure sequences and correlated limit states. Both forward and backward propagation of uncertainty among the components and the system were accomplished.

Bayes networks are directed acyclic graphical representations (DAGs) with nodes to represent the random variables and arcs to show the conditional dependencies among

the nodes. Each node has a probability density function associated with it. The arc emanates from a parent node to a child node. Each child node thus carries a conditional probability density function, given the value of the parent node. The entire network can be represented using a joint probability density function. The network also facilitates the inclusion of new nodes that represent the observed data and thus the updated densities can be obtained for all the nodes.

The updating methodology is briefly discussed here as follows: Consider the Bayes network U with seven nodes a to g as shown in Fig. 3.2. Thus $U = \{a, b, ..., g\}$. Each node is assigned a probability density function as $f(a), f(b|\ a), f(c|\ a), f(d|\ c), f(e|\ b, d), f(f)$ and $f(g|\ e, f)$. In the context of this study, the variables or nodes a, b etc., may correspond to input random variables as well as quantities computed at each step of the computational process. The joint PDF of the entire network is the product of PDFs of various nodes in the network i.e,

$$f(U) = f(a) \times f(b|\ a) \times f(c|\ a) \times f(d|\ c) \times f(e|\ b, d) \times f(f) \times f(g|\ e, f) \qquad (3.2)$$

Note that for nodes b, c, d, e and g, only the conditional densities are defined and included in the joint PDF in Eq. (3.2).

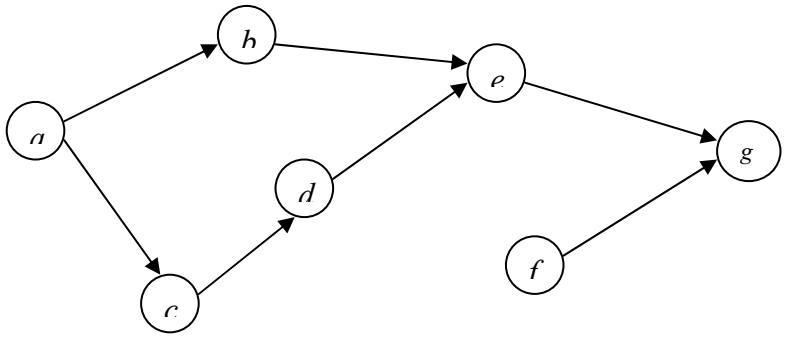


Fig. 3.2 Bayes network before data is collected

The marginal PDF of *b* (for example) can be obtained by the integration of the joint PDF over all the values of the remaining variables. This integration is conveniently done using Markov Chain Monte Carlo techniques (Gilks *et al*, 1996. The joint probability density function for the network can be updated using the Bayes theorem when data is available. Assume that some evidence or test data *m* for node *b* is available. A new node *m* is now added to the network (see Fig. 3.3); this new node is associated with a conditional density function $f(m|\, b)$. Then the joint PDF $f(U, m)$ for this new network is

$$f(U, m) = f(a) \times f(b|\, a) \times f(c|\, a) \times f(d|\, c) \times f(e|\, b, d) \times f(f) \times f(g|\, e, f) \times f(m|\, b) \qquad (3.3)$$

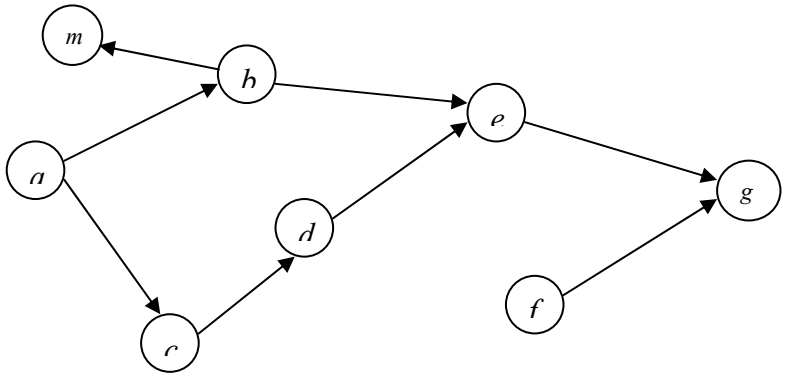


Fig. 3.3 Updated Bayes network with additional data node

With this new joint density, the posterior marginal densities of each of the nodes can be estimated by integrating the joint density over the range of values of all other nodes. Thus the node *b* represents the validated variable while node *g* represents the decision variable.

3.2.2 Case 2: Extrapolation for changes in input conditions

Sometimes the variable in the validation domain could be the model prediction y evaluated at the nominal value of input variable while the decision variable h could be the prediction made using the same model for the input from the tail region, or vice versa. For instance, in reliability analysis, failure may occur in the tail regions of the distributions of the input random variables, but experimental data may be available only at nominal values. Thus B_y could be $\frac{f(y \mid \boldsymbol{z})}{f(y)}$ evaluated at $\boldsymbol{\mu_\alpha}$ while B_h could be the ratio $\frac{f(y \mid \boldsymbol{z})}{f(y)}$ evaluated at $(\mu + 2\sigma)_\alpha$.

The Bayes network shown in Fig. 3.1 applies to this case as well. Now, h is basically the same variable as y; the distinction is that h is evaluated at the tail of the input probability density function and y is evaluated at nominal values of the input. Thus this is a special case of the general extrapolation in Case 1 where y and h could be physically different quantities.

Sometimes, the input variables in the validation and application domains could be completely different although the model response variable is the same quantity. For example, input conditions like type of loading (i.e., distributed vs. concentrated load), material properties (e.g., linear vs. nonlinear elasticity), geometry and boundary conditions (e.g., rigid vs. flexible joints) could be physically different. In all these cases, we need linking variables that connect the two domains.

Another case of extrapolation is system-level model assessment when only component-level data is available (Mahadevan & Rebba, 2005). A large system of codes can be decomposed into subsystems, components etc, and represented using a Bayesian

network. Once data is available on any of the component level nodes, then all nodes, including system level nodes can be updated. The posterior and prior distributions of the system level nodes can give an estimate of confidence in the code prediction of system-level quantities. Thus the Bayes network approach offers a rational and effective methodology to extrapolate inferences from the validation domain to the application domain, as long as the two domains have common, linking nodes.

Table 3.1. Various cases of extrapolation from validation to application

Case	**Validation Domain**		**Extrapolation Domain**	
	Validated variable	Input Conditions	Decision variable	Input Conditions
1	Component-level response (*energy dissipated in a single joint*)	loading type 1 (*sinusoidal*)	Component-level response (*energy dissipated in a single joint*)	loading type 2 (*shock/ impulse or arbitrary load*)
2	Component-level response (*energy dissipated in a single joint*)	loading type 1 (*sinusoidal*)	System-level response (*total energy dissipated in an assembly of joints*)	loading type 1 (*sinusoidal*)
3	Component-level response (*energy dissipated in a single joint*)	loading type 1 (*sinusoidal*)	System-level response (*total energy dissipated in an assembly of joints*)	loading type 2 (*shock/ impulse or arbitrary load*)
4	Response using model type 1 (*small deflection theory*)	load range 1 (*small loads*)	Same response quantity, using model type 1 (*small deflection theory*)	load range 2 (*large loads*)
5	Response using model type 2 (*large deflection theory*)	load range 1 (*small loads*)	Same response quantity, using model type 2 (*large deflection theory*)	load range 2 (*large loads*)
6	response using model type 1 (*small deflection theory*)	load range 1 (*small loads*)	Same response quantity, using model type 2 (*large deflection theory*)	load range 1 (*small loads*)
7	response using model type 1 (*small deflection theory*)	load range 1 (*small loads*)	Same response quantity, using model type 2 (*large deflection theory*)	load range 2 (*large loads*)
8	response quantity 1 (*temperature*)	input 1 (*parameters*)	Response quantity 2 (*flux, physics change*)	input 1 (*parameters*)
9	response quantity 1 (*stress*)	nominal conditions	Failure data	abnormal conditions

Each of these main cases can further be categorized into sub-cases depending on the nature of the problem. For instance, Table 3.1 various sub-cases were derived where the validation inferences made in the test domains need to be extrapolated to the untested domains; the list is not exhaustive. The terms in parentheses in italics indicate some examples that can possible be implemented to understand the concepts involved in the extrapolation.

Fig. 3.4 illustrates the cases 4 to 8 graphically where test data is available for validated model M_1 for an input i and inferences have to be made for the same model M_1 for input j or model M_2 with input i or j. Thus, M_2 may be treated as a decision variable and M_1 as validated variable. This describes the case of a univariate extrapolation.

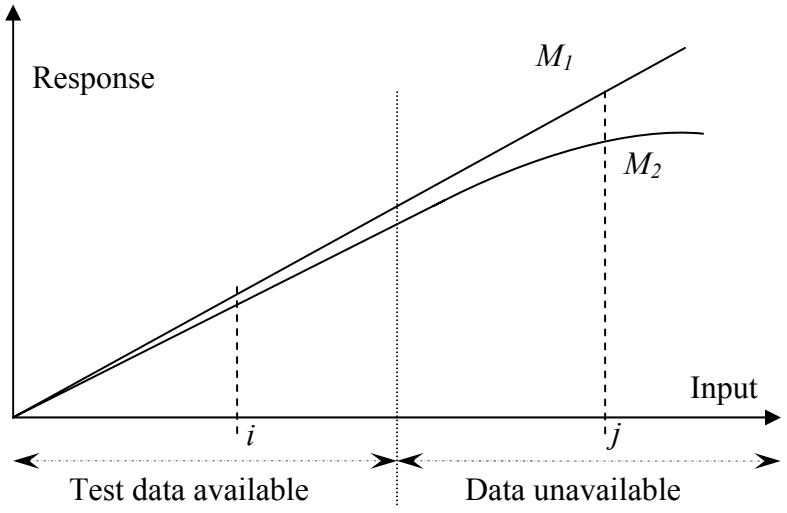


Fig. 3.4. Extrapolation from cases 4-8

As we find the ratio of posterior and prior densities at model prediction x_0 (from the validation domain), one can also determine the lower and upper bounds for the model prediction for which B will be greater than 1.0. Thus in Fig. 3.5, any prediction in the

range [x_L, x_U] (shaded portion) will have a probability greater than or equal to 50% being correct and all the model predictions in that range may be termed as 'close enough' to the data with more than 50% probability. Thus the interval acts as a 'domain' within which the model predictions are considered valid.

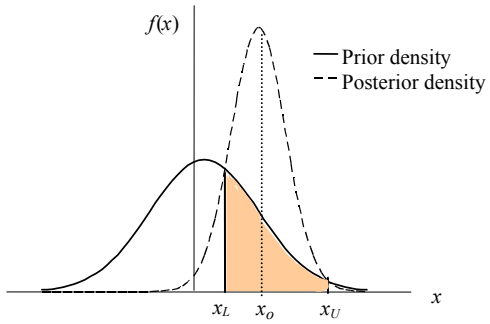


Fig. 3.5. Confidence interval for the prediction

3.3 Multivariate extrapolation

Sometimes, decisions are made based on several variables instead of a single variable. For example, the temperature profile across the width of a plate or response of a structure to a random load over an entire period of time may determine the design criteria instead of critical temperature or stress evaluated at a particular location of space and time. Thus when two or more variables interact in an application, both the validation and extrapolation must be carried out using multivariate analysis. Thus model response, under stochastic conditions, may be represented using a family of curves and having validated the 'mean curve' or 'mean surface', one may need to quantify the confidence in the other curves.

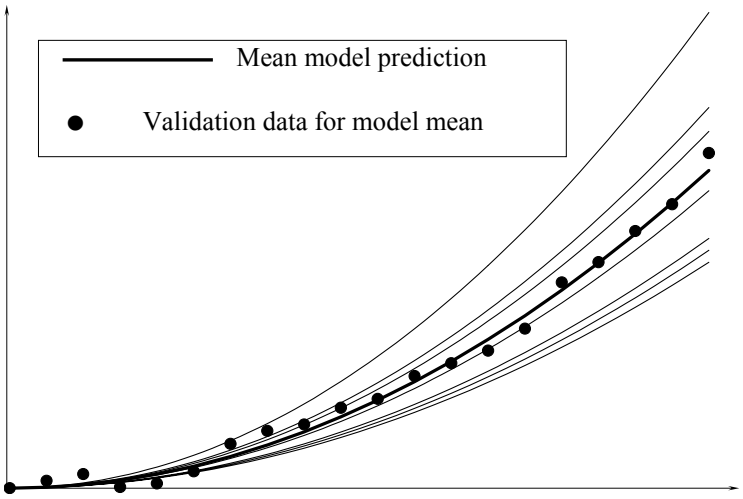


Fig. 3.6. Extrapolating a curve

Methodologies for multivariate extrapolation are given in this section. Note that the Bayes factor metric given in Eq. (2.4) can be extended to a multivariate case with m variables, as the ratio of posterior joint probability density to the prior joint probability density:

$$B(\boldsymbol{x}_o) = \left. \frac{f_X\left(x_1, x_2, ..., x_m \mid y_1, y_2, ..., y_m\right)}{f_X\left(x_1, x_2, ..., x_m\right)} \right|_{x_0} \tag{3.4}$$

Here B is evaluated at a particular model prediction set $\boldsymbol{x}_0 = (x_1, x_2, x_3, \ldots, x_m)_0$. The data is said to favor the model if B is greater than one. Similar to the procedure described in Section 3.2, both $B_{\boldsymbol{y}}$ and $B_{\boldsymbol{h}}$ can be calculated using Eq. (3.4) but the only difference now is that $\boldsymbol{y}$ and $\boldsymbol{h}$ represent a vector of variables. Now, the variable in the validation domain could be the model prediction vector $\boldsymbol{y}$ evaluated at its mean input while the decision variable $\boldsymbol{h}$ could be the prediction made using the same model $\boldsymbol{y}$ for the input from tail region. Thus $B_{\boldsymbol{y}}$ could be $\frac{f(\boldsymbol{y} \mid \boldsymbol{z})}{f(\boldsymbol{y})}$ evaluated at mean vector $\boldsymbol{\mu}_{\alpha v}$ while B_h could be the

ratio $\frac{f(\boldsymbol{y}\,|\,\boldsymbol{z})}{f(\boldsymbol{y})}$ evaluated at a tail vector $(\boldsymbol{\mu}+2\boldsymbol{\sigma})_{\alpha v}$. In other words, $\boldsymbol{h}$ and $\boldsymbol{y}$ come from the same model in this case.

In Section 3.2, an interval has been estimated in which model predictions have more than 50% chance of being correct. A similar 'region of acceptance' for the multivariate case can be established and this region defines the bounds within which extrapolation can be done with more than 50% confidence. As the model predictions are farther away from this region, their acceptance probability drops.

3.4 Dealing with large-scale models

Practical validation and extrapolation problems deal with very large scale models that bring up computational challenges. The Bayesian methodology requires a large number of function evaluations for the model output and joint density evaluations especially when sampling based methods are employed for Bayesian calculations and replacing multiple integrals. Gibbs sampling (Gilks *et al*, 1996) has been commonly used for deriving posterior marginal densities of random model input variables and output variables. This Markov Chain Monte Carlo (MCMC) technique involves a rejection sampling step to sample each random variable from a full-conditional distribution. Thus when the Bayesian network of variables is relatively large and each rejection step calls for a number of "black-box" type finite element code evaluations, the joint density of the variables in the BN may be difficult to evaluate. Typically, response surfaces are used as surrogates to the full-scale computational code. However the accuracy of such surrogate models, for highly non-linear problems and large numbers of input variables, is

questionable, and construction of the response surface might demand a significant number of code evaluations.

In this study, some of the advanced methods available in the literature are explored for improving the rejection sampling schemes and to some extent even avoiding sampling-based techniques by deriving closed-form analytical expressions for posterior marginal distributions. The applicability of adaptive rejection sampling methods, both derivative-based and derivative-free, (Gilks & Wild, 1992; Gilks, 1999) to the Bayesian extrapolation framework is investigated. The metric for confidence measure in the validation and application domains uses the posterior and prior densities of the model responses in their respective domains. This requires computing the posterior and prior marginal distributions of input random variables. Even adaptive Gaussian-quadrature techniques for numerical evaluation of multiple integrals can be prohibitively expensive due to the curse of dimensionality. Such methods however were found to be more accurate with low to moderate number of variables in the problem. Saddlepoint and Laplace-approximation methods (Tierney & Kadane, 1986) can be used for that purpose as an alternative to Gibbs sampling. The efficiency of these approximate techniques will be studied.

Further, with the proposed metric for confidence measure, one need not know the entire distribution function for the model output; only one density value needs to be evaluated. Also there is a need for eliminating the response surface construction as a way to represent the black-box model. Since the model output is a nonlinear function of random input variables, Saddlepoint and Laplace expansion techniques allow us to approximate the underlying nonlinear function using other simple closed-form

expressions. These methods typically use the gradients of model output with respect to the input variables and may require much less number of function (black-box code) evaluations.

Response surfaces are usually constructed for the model (or system) response with respect to the input variables. An alternative solution would be to directly sample a few model outputs and build a non-parametric model to compute the univariate probability density function. The basic idea behind this approach is that the total error in fitting a response surface for the model output (a hyper-surface) in the multidimensional space, will be more than the error that may result due to fitting a nonparametric model for the probability distribution function (a curve). However this needs to be verified for some problems of interest. In summary the key topics covered in this section are:

a. Approximate methods for posterior marginal distributions
b. Approximate methods for density of nonlinear functions
c. Adaptive rejection sampling techniques to improve MCMC simulation efficiency
d. Density estimation from limited samples through a non-parametric method

3.4.1 Posterior marginal density estimation

Consider a vector of random variables $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$ that can be partitioned into a variable θ_1 and an (m-1) dimensional vector $\boldsymbol{\theta}_2 = (\theta_2, \ldots, \theta_m)$. The joint probability density function of $\boldsymbol{\theta}$ is $\pi(\boldsymbol{\theta})$ and the observed data $\boldsymbol{x}$ is described using the log-likelihood function $L(\boldsymbol{\theta})$. We are interested in evaluating the marginal posterior density

$$\pi\left(\theta_1 \mid \boldsymbol{x}\right) = \frac{\int \pi\left(\theta_1, \boldsymbol{\theta}_2\right) e^{L(\boldsymbol{\theta})} d\boldsymbol{\theta}_2}{\int \pi\left(\theta_1, \boldsymbol{\theta}_2\right) e^{L(\boldsymbol{\theta})} d\boldsymbol{\theta}} \tag{3.5}$$

Since the integration in Eq. (3.5) is difficult to evaluate numerically, with the likelihood being a function of the computational model, Laplace's method (Tierney & Kadane, 1986) may be employed to calculate an approximate density function. Let $\hat{\boldsymbol{\theta}}$ maximize $\pi(\boldsymbol{\theta})\, e^{L(\boldsymbol{\theta})}$ and $\boldsymbol{\Omega}$ be the inverse of the Hessian of $[L(\boldsymbol{\theta}) + \pi(\boldsymbol{\theta})]$ evaluated at $\hat{\boldsymbol{\theta}}$. Now for a given θ_1, let $\hat{\boldsymbol{\theta}}_2(\theta_1)$ maximize $\pi(\theta_1, \boldsymbol{\theta}_2)\, e^{L(\theta_1, \boldsymbol{\theta}_2)}$, which is a function of $\boldsymbol{\theta}_2$ with constant θ_1 and $\boldsymbol{\Omega}(\theta_1)$ be the inverse of Hessian of $[L(\theta_1, \boldsymbol{\theta}_2) + \pi(\theta_1, \boldsymbol{\theta}_2)]$ evaluated at $(\theta_1, \hat{\boldsymbol{\theta}}_2)$. Then the approximate marginal density is given by

$$\pi(\theta_1 \mid \boldsymbol{x}) = \left(\frac{\left|\boldsymbol{\Omega}(\theta_1)\right|}{2\pi\left|\boldsymbol{\Omega}\right|} \right)^{1/2} \frac{\pi\left(\theta_1, \hat{\boldsymbol{\theta}}_2(\theta_1)\right) e^{L\left(\theta_1, \hat{\boldsymbol{\theta}}_2(\theta_1)\right)}}{\pi\left(\hat{\boldsymbol{\theta}}\right) e^{L(\hat{\boldsymbol{\theta}})}} \tag{3.6}$$

For the extrapolation problem described in the beginning Section 3.2.1, θ_1 could be α. Suppose we like to partition $\boldsymbol{\theta}$ into two vectors of dimensions k and m - k, the marginal posterior density of the first k variables is given by (Tierney *et al*, 1989)

$$\pi(\boldsymbol{\theta}_k \mid \boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{k}{2}}} \left(\frac{\left|\boldsymbol{\Omega}(\boldsymbol{\theta}_\kappa)\right|}{\left|\boldsymbol{\Omega}\right|} \right)^{1/2} \frac{\pi\left(\boldsymbol{\theta}_k, \hat{\boldsymbol{\theta}}_{m-k}(\boldsymbol{\theta}_k)\right) e^{L\left(\boldsymbol{\theta}_k, \hat{\boldsymbol{\theta}}_{m-k}(\boldsymbol{\theta}_k)\right)}}{\pi\left(\hat{\boldsymbol{\theta}}\right) e^{L(\hat{\boldsymbol{\theta}})}} \tag{3.7}$$

3.4.2 Approximate distributions of non-linear functions

Suppose a computational model y is a nonlinear function of $g(\boldsymbol{\theta})$ of k random input variables and let $\pi(\boldsymbol{\theta})$ be the joint probability distribution of $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}$ maximize the joint distribution $\pi(\boldsymbol{\theta})$, then the marginal density of this k dimensional function $y = g(\boldsymbol{\theta})$ is given by

$$f_y(y) = \frac{1}{(2\pi)^{\frac{k}{2}}} \left(\frac{|\boldsymbol{\Omega}(y)|}{|\boldsymbol{\Omega}| \left| \left(\nabla_{\hat{\boldsymbol{\theta}}(y)} g\right)^T \boldsymbol{\Omega}(y) \left(\nabla_{\hat{\boldsymbol{\theta}}(y)} g\right) \right|} \right)^{1/2} \frac{\pi\left(\hat{\boldsymbol{\theta}}(y)\right)}{\pi\left(\hat{\boldsymbol{\theta}}\right)} \qquad (3.8)$$

where $\boldsymbol{\Omega}$ is the inverse of Hessian of $\pi(\boldsymbol{\theta})$ evaluated at $\hat{\boldsymbol{\theta}}$. Further let $\hat{\boldsymbol{\theta}}(y)$ maximize $\pi(\boldsymbol{\theta})$ subject to the constraint $g(\boldsymbol{\theta}) = y$ and $\boldsymbol{\Omega}(y)$ be the inverse of Hessian of $\pi(\boldsymbol{\theta})$ evaluated at $\hat{\boldsymbol{\theta}}(y)$. Also $\nabla_{\hat{\theta}(y)} g$ is the gradient $\frac{\partial g}{\partial \theta_i}$ evaluated at $\hat{\theta}_i(y)$ for $i = 1$ to k. Note that $\hat{\boldsymbol{\theta}}(y)$ sometimes refers to the most probable point (MPP) corresponding to the limit-state $g(\boldsymbol{\theta}) - y = 0$ in the well-known first order reliability method (FORM) for estimating failure probability (Haldar & Mahadevan, 2000).

Now the Bayes factor computation requires both the prior and posterior densities of y. To compute the posterior density of y, the procedure is identical to the one describe earlier in this section except that the saddlepoints of $\pi(\boldsymbol{\theta})\, e^{L(\boldsymbol{\theta})}$ with and without using the constraint $g(\boldsymbol{\theta}) = y$ are used respectively for the numerator and denominator of Eq. (3.8). Here $L(\boldsymbol{\theta})$ represents the log-likelihood function for the data on y.

3.4.3 Improved sampling techniques for MCMC simulation

This section discusses various techniques used for generating samples from posterior marginal densities. Before adaptive rejection sampling (ARS) is described, the algorithm for rejection sampling is explained here first.

Rejection Sampling

Suppose we wish to draw a sample from a distribution $f(x)$, we choose a simplified sampling density function $g(x)$ and a constant M such that $f(x) \leq Mg(x) \ \forall \ x$. Then the following steps may be performed:

Step 1: Sample x^* from $g(x)$;

Step 2: Sample u from uniform $U(0, 1)$;

Step 3: if $u \geq f(x) / Mg(x)$, accept x^*;

else go to Step 1;

Since the probability of acceptance of a sample is equal to $1/M$ in this case, depending on the choice of M, many evaluations of $f(x)$ may be needed. For the BN shown in Fig. 3.1 and from Eq. (3.1), this could be the likelihood function $f(z \mid y(\alpha)) \times f_{\alpha}(\alpha)$ which is a function of black-box type model output and we would be sampling α from it. ARS reduces such evaluations by improving the sampling density $g(x)$ with the each iteration.

Derivative-based ARS

Suppose the target density $f(x)$ is unimodal and log-concave (which most common distributions are). ARS uses an updated candidate density $g(x)$ in the each iteration. The method requires selecting k points initially on the curve $h(x) = \log (f(x))$ and drawing tangents at those points. Fig. 3.4 shows an enveloping upper bound curve $u(x)$ and a lower bound curve $l(x)$. The piece-wise linear functions $u_i(x)$ constructed in the region $x \in [x_{i-1}, x_i]$ is given by

$$u_i(x) = h_k(x_i) + (x - x_i)\, h'_k(x) \tag{3.9}$$

The subscript k in the above Eq. (3.11) refers to the number of abscissa chosen in that particular iteration. Similarly, a lower bound for the curve $h(x)$ in the region $x \in [x_i, x_{i+1}]$ is given by the piece-wise linear function

$$l_i(x) = \frac{(x_{i+1} - x_i) h_k(x_i) + (x - x_i) h_k(x_{i+1})}{(x_{i+1} - x_i)}$$

Further, the points z_i's represent the intersections of tangents drawn at x_i and x_{i+1}:

$$z_i = \frac{h_k(x_{i+1}) - h_k(x_i) - x_{i+1} h_k'(x_{i+1}) + x_i h_k'(x_i)}{h_k'(x_i) - h_k'(x_{i+1})} \tag{3.10}$$

At the each iteration, the target density would be

$$g_k(x) = \frac{\exp(u_k(x))}{\int \exp(u_k(x^1)) dx^1} \tag{3.11}$$

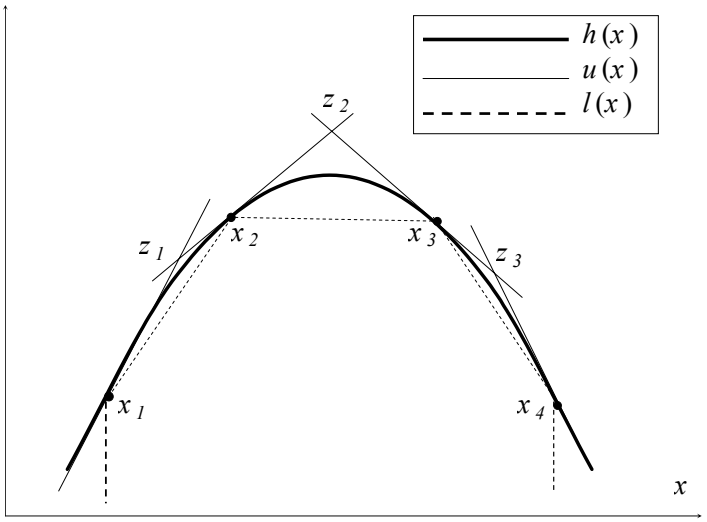


Fig. 3.7. Illustration of derivative-based ARS

Algorithm:

Step 1: Initialize $T_k = \{x_i: i = 1, 2, \ldots, k\}$ be the k starting points;

Step 2: Calculate the upper and lower bound piece-wise linear functions $u_k(x)$, $l_k(x)$ and $g_k(x)$;

Step 3: Sample x^* from $g_k(x)$ and u from uniform $U(0,1)$;

Step 4: Perform squeezing test:

If $u \leq \exp\{l_k(x^*) - u_k(x^*)\}$ accept x^*; else compute $h_k(x^*)$ and $h'_k(x^*)$;

Perform Rejection test:

If $u \leq \exp\{h_k(x^*) - u_k(x^*)\}$ accept x^*; else reject x^*.

Step 5: If both $h_k(x^*)$ and $h'_k(x^*)$ are computed in Step 4, include x^* in T_k to form T_{k+1}, $u_{k+1}(x)$, $l_{k+1}(x)$ etc;

Thus all the accepted samples x^* follow the target distribution function $f(x)$.

Derivative-free ARS

This method is similar to the derivative-based ARS but uses secants instead of tangents to form an upper-bound hull enveloping the log-concave target distribution. Fig. 3.8 shows the extended secants intersect at points z_i's and the title suggests, derivatives of $h(x)$ are not needed in this method. However, the savings in the computational effort by eliminating the derivative calculation may be partially compensated by slower convergence of the candidate density function $g(x)$. Thus both derivative-based and derivative-free methods have some tradeoffs in terms of number of function evaluations. Convergence studies for these different techniques show that the tangent method works slightly better (Gilks *et al*, 1996). While the above techniques are meant for univariate distributions only, sampling schemes for simultaneous multivariate distributions have also been developed in the literature but the details are omitted in this study.

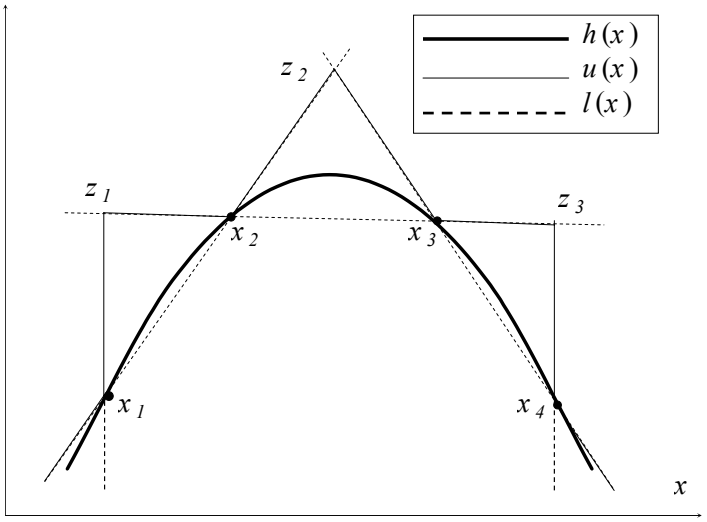


Fig. 3.8. Illustration of derivative-free ARS

3.4.4 Density estimation from limited samples

Sometimes, only a limited number of samples can be generated from Gibbs sampling or any uncertainty propagation technique, in order to save time and computational effort. Parametric models for distributions can be fit to those samples but such models suffer in accuracy due to small sample size. One or more models may have the same fit to the data in which case it is difficult to choose any particular parametric model. Several non-parametric methods like kernel density estimators and Box-Cox transformation techniques are available in literature (Devroye & Gyorfi, 1985) but some numerical studies (results not provided here) have shown that smoothing is a problem in kernel density estimators while the Box-Cox method provides a bad interpolating function. In this study, we adopt the orthogonal series expansion for arbitrary random variables. The series expansion provides a smoothing function in series form whose coefficients are determined by equating the moments on each side. To understand it

better, consider a random variable being expanded using Hermite Polynomials (Ghanem & Spanos, 1991) as

$$x = a_0 + a_1\xi + a_2(\xi^2-1) + a_3(\xi^3-3\xi) + a_4(\xi^4-6\xi^2+3) + \ldots. \quad (3.12)$$

where ξ follows standard normal distribution. If several samples of x are available, moments on both sides of Eq. (3.12) can be computed and equated to solve the coefficients a_0, a_1 etc. Once the coefficients have been estimated, several thousands of samples of ξ can be generated from standard normal density to obtain samples of x. Thus a smooth function for the distribution of x can be obtained from limited samples. The higher moments as a function of the coefficients in RHS of Eq. (3.12) can be obtained from symbolic integration using applications like MATLAB or MATHCAD. For example, a second order expansion will have a mean value a_0 and variance $(a_1^2 + 2a_2^2)$ and a skewness of $(8a_2^3 + 6a_2a_1^2)$. Equating those nonlinear expressions with the moments calculated from data, one can solve for a_0, a_1, a_2 respectively.

3.5 Numerical examples

3.5.1 Investigation of Structural Joints

The safety of critical aerospace components is dependent on their structural connections with the surrounding support structure. Several experimental studies are being investigated (Gregory *et al*, 2003) to understand the behavior of bolted-joints under dynamic loading. Analytical models are being developed to predict the component response to sinusoidal environmental loadings. Also, the energy dissipation in lap-joint type connections is of interest in improving the efficiency and safety of the aerospace

system. Experiments are conducted for single bolted connections under steady state sinusoidal loads to derive energy dissipation curves as a function of input force. These data are used to calibrate the analytical models for predicting the loss of energy at resonance due to friction in lap-joints. Several assemblies of the joint connections have showed the inherent variability (randomness) in the predicted dissipation energy (Urbina *et al*, 2003). In other words, parameters of empirical models for such phenomena are treated as random variables.

These empirical models have been validated using classical and Bayesian hypothesis testing methods (Rebba and Mahadevan, 2003; Urbina *et al*, 2003). The actual application in which the aerospace system will be operated is subject to random and shock loadings. It may not always be possible to test the bolted connection under such loads. Hence the validation inferences made for sinusoidal loadings in the laboratory need to be extrapolated for arbitrary load conditions. Due to safety concerns, the maximum acceleration transmitted to the component could be below a certain threshold. Thus two types of extrapolations --- 1) sinusoidal to arbitrary loading conditions 2) component to system-level validation --- are considered in this example.

High-fidelity computational models can be built to understand the response under single bolt connection to the structure. But most often, the critical components in the actual structure are supported by three or more connections and it is quite expensive to develop high-fidelity models to capture the physics of the system. Hence research is being done to formulate simple, low-fidelity models capable of capturing the dynamic response of the internal component supported by the surrounding structure (Segalman *et al*, 2003). Thus the maximum acceleration experienced by the critical component and the

total energy loss for the system under sinusoidal and arbitrary loadings computed using these models needs to be validated.

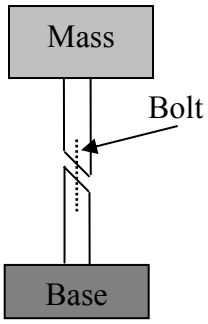


Fig. 3.9a. Single lap joint

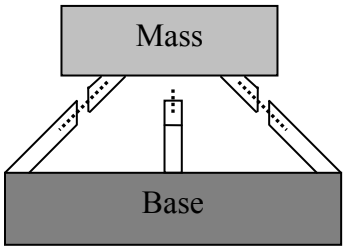


Fig. 3.9b. Three-legged system

Fig.3.9a shows the single inclined lap joint (component) validated in the laboratory and Fig. 3.9b shows the three-legged joint (system) to be used in the actual application. In both cases, the bolted joint is connected to a rigid base that will be excited using a known dynamic force and a mass representing the critical component is placed on top of the joint connection (See Pilch and Trucano (2001) for more details). This study addresses several issues in development of such models and includes various uncertainties. Also, the predictive capabilities of such low-fidelity system-level models need to be assessed using the already validated (Rebba and Mahadevan, 2003; Urbina *et al*, 2003), high fidelity models. The various validation and extrapolation activities to be conducted as part of this study are summarized in Table 3.2. These three cases show the increasing levels of tiers in the hierarchy of system-level model validation where simple models are validated for known loading conditions first and gradually extended to complex models and more uncertain conditions. The quantity of interest is maximum acceleration in all examples.

Table 3.2 System-level model validation activities

Case	Validation Domain	Application Domain	Response Quantity	Loading in Validation Domain	Loading in Application Domain
1	Single Leg	Single Leg	Acceleration	Sinusoidal	Arbitrary
2	Single Leg	3-Legged	Acceleration	Sinusoidal	Sinusoidal
3	Single Leg	3-Legged	Acceleration	Sinusoidal	Arbitrary

Three cases of extrapolation are considered in this example. In case 1, the response of a single spring is validated under sinusoidal loading and inferences need to be extrapolated to the acceleration under arbitrary loading for the same single leg structure. In Case 2, the application domain involves a three-legged system subject to sinusoidal loading. This can be treated as a system-level model assessment. Case 3 deals with system-level model assessment and change in input conditions at the same time. All the three cases in Table 3.2 are numerically illustrated using spring-mass systems. These numerical examples serve as the initial step to study and better understand the physics of joints behavior under dynamic loading, and to further verify the proposed system-level model validation methodology using actual system-level data that will be available subsequently.

The bolted joints are represented using springs with known stiffness k and damping coefficient c. The mass attached on the top of the joint is denoted by m. For the three-legged system, the individual bolts (or springs) are assumed to have identical properties and hence same statistics. The maximum force at any time for a given type of loading is limited to 100 lb. During the calculations however, the units are omitted for clarity. The sinusoidal excitation at the base has a frequency of Ω rad/s. The experimental

error in measuring the acceleration of the mass is assumed to be Gaussian with zero mean and variance of 9 in/sec^2. For the 3-leg system, each spring is assumed to be inclined making an angle θ_i for i = 1, 2, 3, to the horizontal. This angle is assumed to be random to model the uncertainties in the configuration of the connections and errors made in their assembly.

Table 3.3 Statistics of parameters in the spring-mass system

Parameter	**Type**	**Mean**	**Std. Dev**
k	Lognormal	1000	100
c	Lognormal	7	2
Ω	Normal	5	1
θ	Normal	45^o	5^o
m	Constant	5	

The statistics of various parameters are shown in Table 3.3. Fig. 3.10 shows the simplified models of the structural joints.

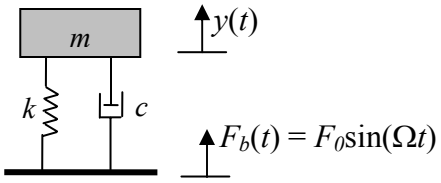


Fig. 3.10a Single leg

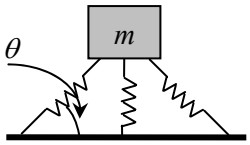


Fig. 3.10b Three-legged system

The computational model predicting the response (displacement, velocity etc) of the mass is denoted by the PDE

$$m\ddot{y} + ky + c\dot{y} = \frac{F_0 k}{m\Omega^2}\sin\Omega t + \frac{F_0 c}{m\Omega}\cos\Omega t \tag{3.13}$$

Further, the maximum acceleration transmitted to the mass for a sinusoidal loading is given by this component level model as

$$\ddot{y}_{\max} = \frac{F_0}{m}\frac{\sqrt{1+(2r\zeta)^2}}{\sqrt{(1-r^2)^2+(2r\zeta)^2}} \tag{3.14}$$

where the frequency ratio $r = \frac{\Omega}{w}$, the natural frequency $w = \sqrt{\frac{k}{m}}$ and damping factor $\zeta = \frac{c}{2mw}$. For the three legged system, the effective stiffness and damping coefficient are estimated as $\sum_{i=1}^{3} k_i \sin\theta_i$ and $\sum_{i=1}^{3} c_i \sin\theta_i$ respectively; Eq. (3.14) is used to predict the acceleration of the mass. The statistics of k_i and c_i will be the same as for k and c (Table 3.3) to indicate that the same type of joints are used in the 3-legged system; however there is variability from joint to joint. BNs will be constructed for the response predicted by the single leg and three-legged joint system showing all the relations among the different variables given in Table 3.3.

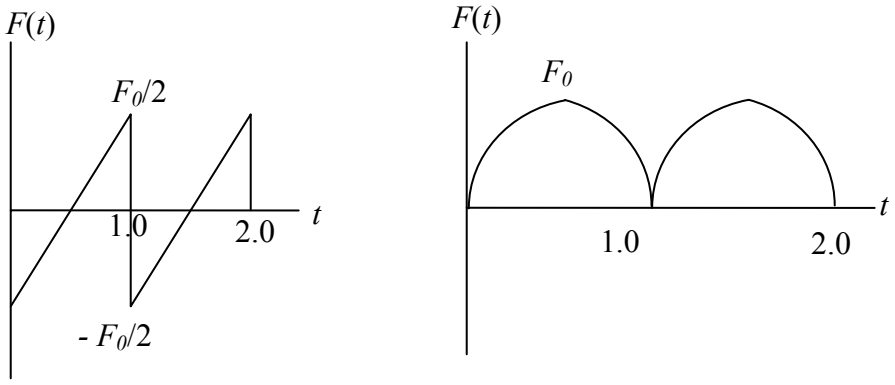


Fig. 3.11 Pulse Loading

a) Triangular load b) Parabolic load

The arbitrary loadings considered are (a) A triangular impulse load starting from zero load and reaching a peak force of F_0 =100 within 1 sec (b) an inverted parabolic loading that starts at zero and reaches peak value of F_0 = 100 at 0.5 sec and goes down to zero at 1 sec. The details of implementation for the various cases listed in Table 3.2 are discussed next. The validation data (12 points) needed for sinusoidal loading on single leg joint is obtained by simulation only to demonstrate the methodology. Thus the 12 measured accelerations are $\boldsymbol{z}$ = {21.639, 22.940, 24.940, 21.696, 24.816, 25.704, 23.163, 22.250, 20.816, 23.354, 22.813, 23.661}. A BN for the validated model is shown in Fig. 3.12.

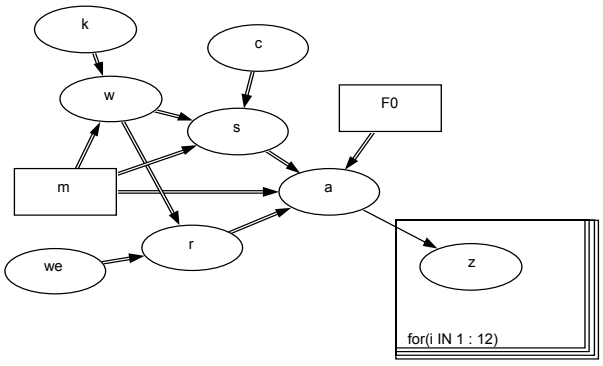


Fig. 3.12 BN for single-leg joint validation

Case 1: The steps involved in validating the maximum acceleration for single-leg joint under arbitrary loading are:

- Update the distribution of acceleration under sinusoidal loading using validation data and hence update the statistics of the parameters k, c, Ω etc.

- Compute the Bayesian validation metric for response under sinusoidal loading as the ratio of posterior and prior densities at an acceleration value predicted for a particular set of values for k, c, Ω etc.
- Using the posterior and prior statistics for k, c etc, calculate the density function of maximum acceleration for the single leg joint under arbitrary loadings (two types mentioned above).
- Compute the validation metric for response under arbitrary loading (application domain) as the ratio of posterior and prior densities at an acceleration value predicted for a particular set of values for k, c, Ω etc.

For the first three cases, the model predictions are made at mean values, $k = 1000$, $c = 7$, $\Omega = 5$. The maximum response under arbitrary loading is computed using a numerical analysis technique (Newmark method).

Case 2: The steps involved in validating the maximum acceleration for three-legged joint under sinusoidal loading are:

- The first two steps are similar to those described in case 1.
- Using the posterior and prior statistics for k, c etc, calculate the density function of maximum acceleration for the 3-legged joint under sinusoidal loading. Note however that k and c will be effective parameters that include the effect of random joint inclination θ.
- Validation metric for the 3-legged joint (application domain) response under sinusoidal loading is the ratio of posterior and prior densities at an acceleration value predicted for a particular set of values for effective parameters k and c, Ω etc.

Case 3: The steps involved in validating the maximum acceleration for three-legged joint under arbitrary loading are:

- The first two steps are similar to those described in case 1.
- Using the new and old statistics for k, c etc, calculate the density function of maximum acceleration for the 3-legged joint under arbitrary loading (two types). Note however that k and c will be effective parameters that include the effect of random joint inclination θ.
- Validation metric for the 3-legged joint (application domain) response under these arbitrary loadings is the ratio of posterior and prior densities at an acceleration value predicted for a particular set of values for effective parameters k and c, Ω etc.

The results obtained in each case are summarized in Table 3.4. The variable B refers to the validation metric in the each domain. The ratio B (of the posterior to prior densities), is always evaluated the mean value of the model prediction in this example.

Table 3.4. Summary of validation and extrapolation results for the 4 cases

Case	***B* in Validation Domain**	***B* in Application Domain**
1	1.82	1.04 (parabolic pulse) 1.03 (triangular pulse)
2	1.82	1.62
3	1.82	1.1 (parabolic pulse) 1.1 (triangular pulse)

A Bayes factor for the decision variable close to 1.0 indicates that validation data are not informative for assessing the model in the application domain. In general, the value of B in the extrapolation domain is lower than in the validation, as expected, since

there should be less confidence in the extrapolation than in the domain where data is available.

3.5.2 Energy dissipation model

Consider another case which studies the energy loss due to friction in bolted lap joints under sinusoidal and arbitrary loadings. Here, the parameters that represent the material and geometric properties are quite different from those described in Section 3.5.1. Thus the following discussion must be viewed as totally different, independent of the spring-mass analogies described so far. Here we consider a four parameter Iwan model (Iwan, 1966) to study the accuracy of the mathematical model in predicting the energy loss due to friction in a lap joint. The purpose of the mathematical model is to predict the dissipation energy D released per cycle at the joint when subjected to impact harmonic (sinusoidal) force amplitude of F_0.

$$D = r^{\chi+3} \frac{4F_S\phi_{\max}(\chi+1)}{\left(\beta+\frac{\chi+1}{\chi+2}\right)(\chi+2)(\chi+3)} \tag{3.15}$$

where the term r is given by solving the following equation below

$$\frac{F_0}{F_S} = r\frac{(\beta+1)-\left(\frac{r^{\chi+1}}{\chi+2}\right)}{\beta+\left(\frac{\chi+1}{\chi+2}\right)} \tag{3.16}$$

where the four parameters R, S, χ and ϕ_{max} are quantified from the experiments and whose statistics are given by Urbina *et al* (2003). β, r and F_s are intermediate variables.

This model has been validated using sufficient data sets and is ready for use in the energy loss prediction under sinusoidal loadings. Now the task is to assess its predictive capabilities under arbitrary loading. Note that this r in Eq. (3.16) is not same as the frequency ratio defined earlier. Again, BNs are employed to extrapolate inference from the validation domain (harmonic loading) to the application domain (arbitrary loading conditions). The computational model given by Eq. (3.17) is valid for sinusoidal or harmonic excitations. Any arbitrary loading may be represented using Fourier series expansion as

$$F_b(t) = a_0 + \sum_{n=1}^{\infty}\left(a_n \cos\left(n\Omega t\right) + b_n \sin\left(n\Omega t\right)\right) \tag{3.17}$$

Thus decomposing the total force function into several sinusoidal force components, the total energy dissipated in the joint can be estimated as the summation of energies dissipated under each of the sine or cosine component. Thus the variable F_0 in Eq. (3.17) is replaced with a_0, a_n, b_n etc for n = 1, 2, 3,… and energy D is computed in parts.

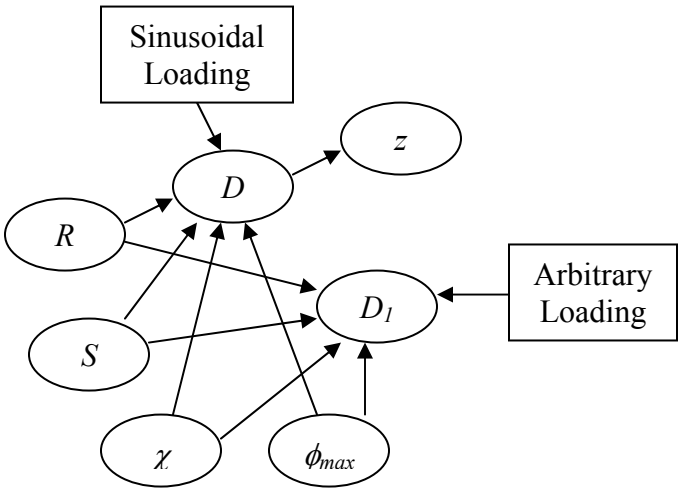


Fig. 3.13 BN for the energy dissipation problem

The model prediction in the application domain is the summation of those energy values. It is also obvious that energy loss under cosine and sine loads will be identical.

The steps involved in validating the model are

- Update the distribution of energy D under sinusoidal loading (magnitude of 320 lb is chosen) using validation data z (12 data points available) and hence update the statistics of the parameters R, S, etc.
- Validation metric for D under sinusoidal loading is the ratio of posterior and prior densities at an energy value predicted for a particular set of values for R, S, etc.
- Using the new and old statistics for R, S etc, calculate the density function of energy loss D_1 for the joint under arbitrary loadings (two types mentioned above). Again, it should be noted that energy D_1 is computed as a sum of energies dissipated by Fourier components of the impulse/ shock force.
- Validation metric for D_1 under arbitrary loading (application domain) is the ratio of posterior and prior densities at an energy value predicted for a particular set of values for R, S, etc.

The triangular and parabolic impulse loads are represented by Fourier series in Eqs. (3.18) and (3.19) respectively.

$$F(t) = \frac{F_0}{2} - F_0 \sum_{n=1}^{\infty} \frac{1}{n\pi} \sin\left(2\pi n t\right) \tag{3.18}$$

$$F(t) = 4F_0 \left[\frac{1}{6} - \sum_{n=1}^{\infty} \frac{1}{n^2 \pi^2} \cos\left(2\pi n t\right) \right] \tag{3.19}$$

For the sake of illustration, the first 5 components of the Fourier series are used to estimate the energy dissipated in the bolted lap-joint and Table 3.5 shows those force components. Thus for triangular shock load, the total energy is computed as

$$D_1 = \mathrm{D}(0.5F_0) + D(0.318F_0) + D(0.159F_0) + D(0.106F_0) + D(0.079F_0) \quad (3.20)$$

where $F_0 = 320$. Similarly, for parabolic shock load, the total energy is computed as

$$D_1 = \mathrm{D}(0.666F_0) + D(0.405F_0) + D(0.101F_0) + D(0.045F_0) + \mathrm{D}(0.025F_0) \quad (3.21)$$

Table 3.5 Fourier components for impulse load

Load	F_1	F_2	F_3	F_4	F_5
Triangular	0.5 F_0	0.318 F_0	0.159 F_0	0.106 F_0	0.079 F_0
Parabolic	0.666F_0	0.405 F_0	0.101 F_0	0.045 F_0	0.025 F_0

The results obtained in each case are summarized below in Table 3.6. The variable B refers to the validation metric in the each domain. The ratio B is always evaluated the mean value of the model prediction in this example.

Table 3.6 Summary of validation and extrapolation results for the 4 cases

***B* in Valid. Domain**	***B* in Appl. domain**
5.02	2.68 (parabolic pulse)
	1.44 (triangular pulse)

3.5.3 Heat flow problem

Consider a transient one dimensional heat flow problem (Hills and Leslie, 2003). The computational model is then time-dependent and so is the target application. Also, the new predictive model has two more additional random input variables in it.

$$T(x,t) = (1-x)\alpha_1 + x\alpha_2 + \sum_{n=1}^{\infty} A_n \exp\left[-\frac{k}{\rho C_p} n^2\pi^2 t\right] \sin(n\pi x) \qquad (3.22)$$

where $A_n = -\frac{2}{n\pi}\left[\alpha_1 - \alpha_2(-1)^n\right]$. The decision variable for the target application is the heat flux defined as $d(x, t) = -k\, dT/dx$ i.e.,

$$d(x,t) = -k\left[\alpha_2 - \alpha_1 + \sum_{n=1}^{\infty} A_n n\pi \exp\left[-\frac{k}{\rho C_p} n^2\pi^2 t\right] \cos(n\pi x)\right] \qquad (3.23)$$

Also, the variable k and ρC_p follow the same statistical distribution N (1, 0.1). Model output corresponding to $t = \infty$ gives the steady state response as obtained in Example 1. In this example, the model predictive capabilities are tested at time t = 0.25 and at a location x = 0.25. Model prediction is made for a set of mean input random variables and experimental data was measured with Gaussian error $\varepsilon_{exp} \sim N(0, 0.5)$. Since both $d(x, t)$ and $T(x, t)$ depend on x, t, α_1, α_2, k and ρC_p, the relations are represented using a Bayesian network as shown in Fig. 3.14.

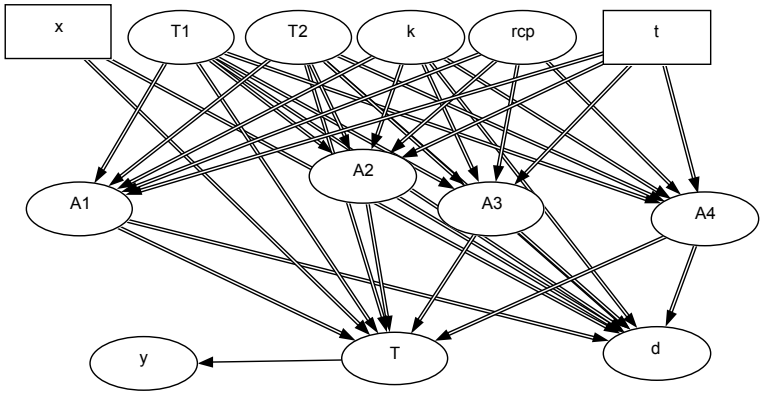


Fig. 3.14. Bayes network for transient heat flow problem

Table 3.7 shows the Bayes factors calculated for the computational model prediction using different observed values of temperature (y) and the corresponding inferences on the target application. For a fixed model prediction (using a fixed set of inputs), the experimental observation value (only single measurement in each case) varied from, being far from prediction to close to the prediction value.

Table 3.7. Validation inference extrapolation for transient heat flow problem

T	y	d	B_T	B_d
12.4	10.16	-9.695	0	0.91
12.4	11.03	-9.695	0.55	1.00
12.4	11.81	-9.695	1.73	1.07
12.4	12.45	-9.695	3.28	1.09

A plot of confidence measure in T (0.25, 0.5) versus confidence measure in d (0.25, 0.5) is shown in Fig. 3.15.

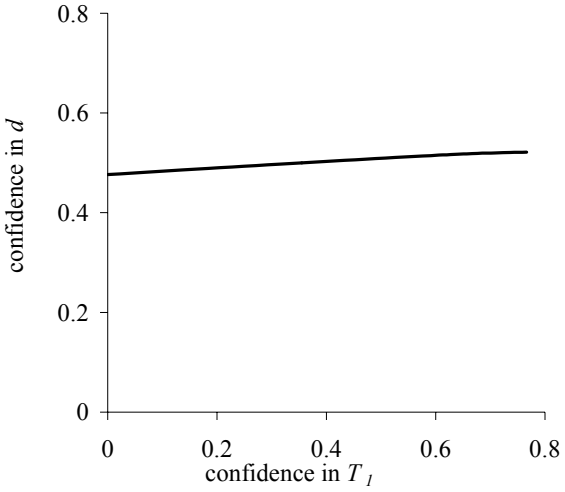


Fig. 3.15. Plot relating confidence in decision variable to validation information

A very flat plot in Fig. 3.15 indicates that when transient conditions are considered, the target application is in fact less sensitive to the validation information. This could be the result of additional parameters introduced in the model and mode complexity.

3.5.4 Extrapolation of stress prediction from nominal to tail loading

A mechanical component in an application is a square plate structure with a circular hole in the center. The plate is subjected to distributed loading along the two straight edges. Finite element (FE) modeling may be used to predict any response quantity of interest related to this plate. The FE model of a quarter the structure is used due to the symmetry, as shown in Fig. 3.16, and the vertical displacement of tip A under the loading is of interest.

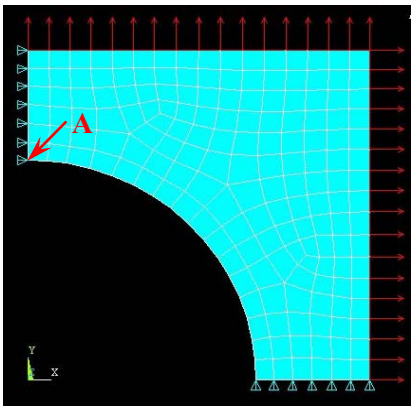


Fig. 3.16 FE model of the plate

The plate has dimensions of 24”x 24” x 1” and the curved edge has a radius of 8”. The Young’s modulus E and the Poisson ratio is v are Gaussian random variables with statistics N (10000, 2000) psi and N (0.2 0.025) respectively. The plate is subjected to uniform loading of equal magnitudes along its edges. For the purpose of analysis, the loading on each edge is assumed Gaussian with statistics $w \sim N$ (500, 50) kips. Elastic

small deflection theory was used in the analysis to determine the displacement of tip A. Appropriate boundary conditions were applied along the other straight edge portions of the plate.

Since the input loading and material properties are random, the model response is also a random quantity. One can estimate the statistical distribution of model response by running the FE code several times using randomly sampled values of the input loading (w, E, v) each time. To avoid this computationally intensive exercise, a stochastic response surface (Tatang *et al*, 1997) using polynomial chaos expansion (Ghanem and Spanos, 1991) was used in this example to represent the tip displacement as a function of the distributed loads along the edges. Although we considered the Poisson's ratio v as a random variable, analysis of variance showed that v has insignificant contribution to the variance of y and hence v is omitted in the response surface. Thus the model output is a function of E and w.

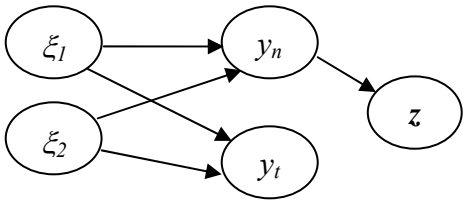


Fig. 3.17 Bayes network for the plate problem: Nominal input to tail input

The stochastic response surface with $R^2 = 0.999$ is

$$y = 1.5306 - 0.3326\ \xi_1 + 0.1544\ \xi_2 + 0.0666\ (\xi_1^2 - 1) - 0.03329\ \xi_1\xi_2 \qquad (3.24)$$

where ξ_1 and ξ_2 are independent standard normal variables. Here ξ_1 and ξ_2 are related to the physical variables E and w using the relation $E = 10000 + 2000\xi_1$ and $w = 500 + 50\xi_2$.

Thus the model response (vertical displacement at tip A) for any values of E and w can be obtained by first transforming each of those values into standard normal space and then substituting them in Eq. (3.24).

Suppose we validate this model in a test setup at its mean input values (w = 500, E = 10,000) whereas in the actual application, the plate experiences larger loads (w = 750, E =10,000). This is Case 2 in Section 3.2.2, where the validation and decision variables are identical but evaluated at mean and tail loads respectively. Suppose the displacement data (5 samples) corresponding to the mean load input is $\boldsymbol{z}$ = {1.215, 1.563, 1.618, 1.962, 1.294}.

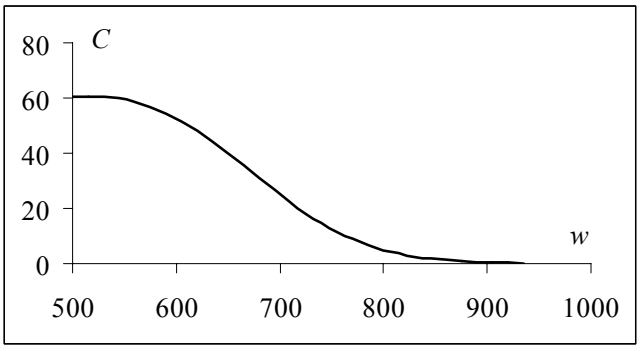


Fig. 3.18 Confidence in prediction at non-nominal loads

The Bayesian network depicting the relations between various quantities is given in Fig. 3.17. Both y_n (nominal) and y_t (tail) are exactly the same functions of ξ_1 and ξ_2 but the response values are evaluated at different inputs. The validation metric B_{yn} at the mean input value is found to be 1.52 which corresponds to 60.3% confidence (i.e., $B_{yn}/(B_{yn}+1)$). The confidence in the model prediction for any other input value (say, from its tail

region) can be calculated, as explained in Section 2.2.2, by evaluating $B_{yt} = \frac{f(y \mid z)}{f(y)}$ at this new input value (from the tail region) first and then by computing C_h using the relation $C_h = B_h / (B_h + 1)$. Given the experimental data at nominal loading, the confidence in the model prediction at different load values (equal magnitude on all edges) is estimated and shown in Fig. 3.18. At $w = 750$, $B_h = 0.142$ and $C = 12.46\%$. As we collect more data at higher load values, one should expect the confidence curve to move to the right, indicating increasing confidence at higher loads. With the current information, the confidence drops below 50% at $w = 612$ lb. Thus the proposed methodology can also be used to determine the limits of extrapolation.

Different loading conditions

Suppose the plate is subject to uniform loading w of equal magnitude along its edges in the validation domain and point load P in the application domain. It is assumed that the load P acts at the midpoint along the edge of the quarter plate. For the purpose of analysis, the loading on each edge is assumed Gaussian with statistics $w \sim N$ (500, 50) kips and $P \sim N$ (6000, 1200) kips. Further the material properties are random variables as well with distributions $E \sim N$ (10,000; 1000) psi while $v \sim N$ (0.2, 0.025). A linear elastic, small deflection theory was used in the analysis to determine the displacement of tip 'A' shown in Fig. 3.16. Since the input loading is random, the model response in both domains will also be a random quantity. The FE model is the only common link between the two domains. The BN for this problem is shown in Fig. 3.19 and the common independent variables are E and v.

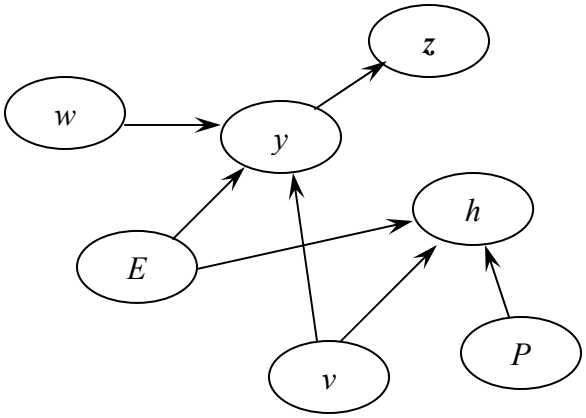


Fig. 3.19 Bayes network for the plate problem: Different loading conditions

Model y represents the response under distributed loading while h represents response under point loading. The stochastic response surface is constructed for h in terms of E and P only since the variable v is found out to have no significant effect on h.

The stochastic response surface with $R^2 = 0.999$ is

$$h = 0.5623 - 0.1222\ \xi_1 + 0.1134\ \xi_3 + 0.02447\ (\xi_1^2 - 1) - 0.02445\ \xi_1 \xi_3 \qquad (3.25)$$

where ξ_1 and ξ_3 are standard normal variables related to E and P using the relations $10000 + 2000\ \xi_1$ and $6000 + 1200\ \xi_3$ respectively. Suppose the data $\boldsymbol{z}$ is used to update the response in the validation domain y, then the linking variable E and hence the decision variable h are updated through the Bayes network. The Bayes factors for y and h evaluated at the mean values of E, w and P are estimated to be 1.52 and 1.1 respectively. This is Case 2 in Section 3.2.2 where the input conditions are physically different in the validation and extrapolation domains.

3.5.5 Multivariate extrapolation

A thermal decomposition model of a polyurethane foam was developed by Hobbs *et al* (1999). The model predicts the foam decomposition front location as a function of time and the computational model involves solving a series of partial differential equations using numerical methods. Each of those codes corresponds to different chemical and physical processes. The boundary condition consists of a uniform rate of heating maintaining a constant temperature at one edge of the foam. The rate of decomposition depends on several model parameters such as material properties (density, specific heat, and emissivity), chemical properties (bond population, heat of reaction etc) of the foam, and activation energies (that affect the chemical bond breaking rates). Thus the model prediction is a function of 25 input parameters. Further, the uncertainty of those parameters is characterized using statistical distributions. The statistics of 16 activation energy parameters were estimated from 18 experiments (Hills *et al*, 2004). Although the histograms of each of those parameters did not have symmetry and are bimodal in some cases, the activation energies were assumed to be Gaussian for the sake of analysis. The remaining 9 parameters relating to the material and chemical properties of the foam are assumed to be Gaussian as well from a previous analysis (Dowding *et al*, 2004). The details of the statistics and correlation have been omitted in this examples as they are found in Hills *et al* (2004).

The uncertainty in the model output can be represented using a statistical distribution whose statistics can be obtained in two ways; in the first approach, the 18 sets of input parameters from 18 experiments can be used directly to calculate the model output 18 times. This method however is not so useful for ‘making new predictions’ for a

set of input values other than those that have already been used or to predict the response at a different time period. Alternatively, an approximate mathematical model such as a response surface can be built to predict the response quantity of interest (in this case, location of decomposition front) as a function of the random input parameters and time. This uncertainty in the inputs can be propagated to the output through this approximate mathematical model repeatedly to obtain the output statistics as well as to make predictions for future use without accessing the full suit of codes. This reduces the computational effort and saves time for later uses of the model for design. Before the approximate model is set for use in an application, it needs to be validated at least for the range in which the test data is available. Using the validation inference, the confidence in the model output for a new set of input parameters outside the validation domain has to be computed. This example thus serves as a case study for the multivariate extrapolation discussed in Section 3.3.

The approximate mathematical model is based on a first-order Taylor series expansion, constructed as a function of the random parameters, around the mean values of the parameters as

$$x(\boldsymbol{\alpha},t) = x\left(\boldsymbol{\mu}_{\boldsymbol{\alpha}},t\right) + \sum_{i=1}^{25} \left.\frac{\partial x(\boldsymbol{\alpha},t)}{\partial \alpha_i}\right|_{\mu_{\alpha_i}} \left(\alpha_i - \mu_{\alpha_i}\right) \tag{3.26}$$

where $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \ldots \alpha_{25}\}$ is the vector of input random parameters and $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$ corresponding mean vector. In this example, $x(\boldsymbol{\mu}_{\boldsymbol{\alpha}}, t)$ is assumed to be Gaussian; thus the model output at any time instant is Gaussian as well from Eq. (3.26). To validate this model that was built around the mean parameter vector, experiments were conducted to measure the actual location of the decomposition front. For a given heating rate with a temperature of 600

^{0}C, measurements were taken at discrete, irregular time intervals and the response observed corresponds to the mean input model parameters.

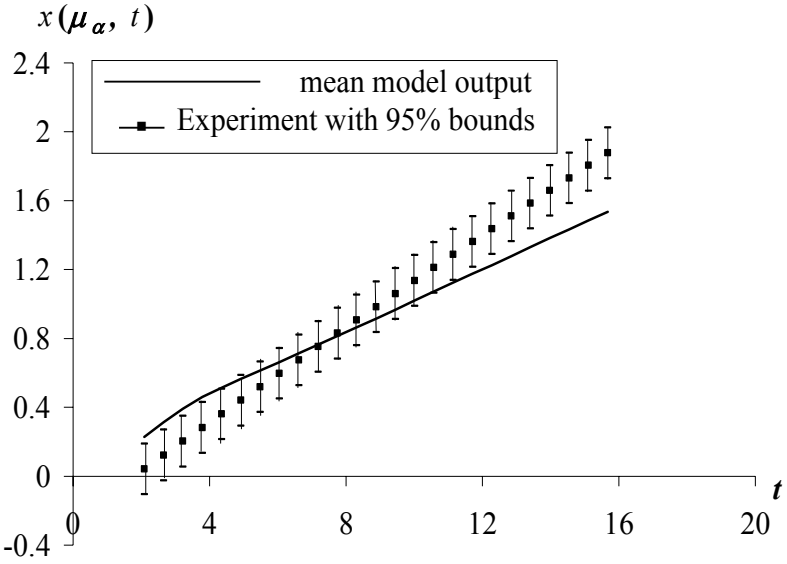


Fig. 3.20. Model prediction versus experiment

The readings are available at 110 different time instants but for illustration purposes, only the readings from the first 25 time instants were used in this example. It is just a coincidence that the number of random input parameters used in the model is same as the number of time locations at which the response is measured. Although the response measured at each time period is correlated, no additional information on the correlation among the measurements $\boldsymbol{y}$ is available, and thus the measurement uncertainty is assumed from previous experience to be $\boldsymbol{\Sigma} = \text{cov}(\boldsymbol{y}) = 0.075^2\,\mathbf{I}$, where $\mathbf{I}$ is the identity matrix of dimensions 25 × 25. Also, $\boldsymbol{\Sigma}$ is substituted for $\mathbf{V}$ in Eq. (2.9) to calculate the likelihood. Fig. 3.20 shows the plot of mean model prediction $x(\mu_\alpha, t)$ versus the observed response

***y*.** Model validation in this case involves determining whether the mean model output vector is statistically close enough to the experimental observation vector.

Suppose the prior prediction $x(\boldsymbol{\mu}_{\boldsymbol{\alpha}}, t_j)$ for j = 1 to 25, is normal with mean vector $\boldsymbol{\eta} = x(\boldsymbol{\mu}_{\boldsymbol{\alpha}}, t)$ and covariance matrix $\boldsymbol{\Lambda}$. The covariance of the model output is derived from the covariance of the input parameters using the relation $\boldsymbol{\Lambda} = \nabla_{\boldsymbol{\alpha}}\boldsymbol{x}(\boldsymbol{\alpha}, \boldsymbol{t}).\mathrm{cov}(\boldsymbol{\alpha}).\nabla_{\boldsymbol{\alpha}}\boldsymbol{x}(\boldsymbol{\alpha}, \boldsymbol{t})$. Having observed the data $\boldsymbol{y}$ with Gaussian measurement uncertainty having zero mean and covariance structure $\boldsymbol{\Sigma} = \mathrm{cov}(\boldsymbol{y}) = 0.075^2\,\mathbf{I}$, the posterior joint density for $\boldsymbol{x}$ will be multivariate normal as well. The posterior mean and covariance matrix for the model output variables are given by

$$\boldsymbol{\eta}_p = (\boldsymbol{\Lambda}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}[\boldsymbol{\Lambda}^{-1}\boldsymbol{\eta} + \boldsymbol{\Sigma}^{-1}\boldsymbol{y}]$$

$$\boldsymbol{\Lambda}_p = (\boldsymbol{\Lambda}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1} \tag{3.27}$$

Using Eq. (2.9), the aggregate validation metric is computed at the mean value $\boldsymbol{\eta}$ as

$$B_y = \left.\frac{f\left(x(\boldsymbol{\alpha},\boldsymbol{t})\,|\,\boldsymbol{y}\right)}{f\left(x(\boldsymbol{\alpha},\boldsymbol{t})\right)}\right|_{\boldsymbol{\eta}} = \frac{|\boldsymbol{\Lambda}|^{\frac{1}{2}}}{|\boldsymbol{\Lambda}_p|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}\left(\boldsymbol{\eta}-\boldsymbol{\eta}_p\right)^T\boldsymbol{\Lambda}_p^{-1}\left(\boldsymbol{\eta}-\boldsymbol{\eta}_p\right)\right] \tag{3.28}$$

The Bayes factor for any prediction other than the mean can be obtained using the experimental data and prior statistics of mean model output as described in Section 3.3. Suppose we need to assess the confidence in an arbitrary model output at $\boldsymbol{x}$, the Bayes factor is calculated as

$$B_h = \left.\frac{f\left(x(\boldsymbol{\alpha},\boldsymbol{t})\,|\,\boldsymbol{y}\right)}{f\left(x(\boldsymbol{\alpha},\boldsymbol{t})\right)}\right|_{\boldsymbol{x}} = \frac{|\boldsymbol{\Lambda}|^{\frac{1}{2}}}{|\boldsymbol{\Lambda}_p|^{\frac{1}{2}}}\frac{\exp\left[-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\eta}_p\right)^T\boldsymbol{\Lambda}_p^{-1}\left(\boldsymbol{x}-\boldsymbol{\eta}_p\right)\right]}{\exp\left[-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\eta}\right)^T\boldsymbol{\Lambda}\left(\boldsymbol{x}-\boldsymbol{\eta}\right)\right]} \tag{3.29}$$

Thus B_y is found to be 0.215 and B_h evaluated at $x(\boldsymbol{\alpha} + 0.1\boldsymbol{\sigma}, \boldsymbol{t})$ is estimated as 0.0018. A value of 0.215 indicates that the model prediction at mean is not close enough to the

mean observation vector and only 17% confidence exists in the mean model output based on the available data. The confidence at $x(\boldsymbol{\alpha} + 0.1\boldsymbol{\sigma}, \boldsymbol{t})$ is even smaller.

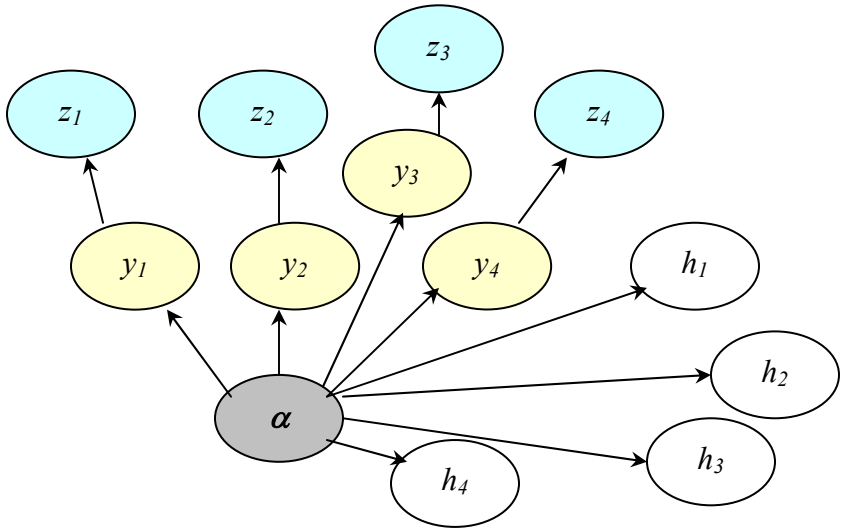


Fig. 3.21. BN for Multivariate Extrapolation

When conjugate priors are not available, a Markov Chain Monte Carlo simulation technique like Gibbs sampling (Spiegelhalter *et al*, 2002) may be employed to calculate the metric given in Eq. (3.29). In such cases, multiple nodes for y_1, y_2, y_3.. and h_1, h_2, h_3 etc are defined as shown in Fig. 3.21. Several reasons can be attributed for this apparent model and data discrepancy. Even graphically, the mean model output did not fall within the 95% confidence intervals for the measurement most of the time, which should give a preliminary indication of model inadequacy.

(1) Model form error: The first-order Taylor series may be inadequate in representing the model prediction for mean inputs in Eq. (3.27) and even the sensitivities $\nabla_{\boldsymbol{\alpha}}\boldsymbol{x}(\boldsymbol{\alpha}, \boldsymbol{t})$ calculated using perturbation methods may not have been accurate.

(2) UQ for model input and output: Another major source of error could be the uncertainty characterization of the input model parameters from limited data. With the errors in the estimation of sensitivities, the covariance matrix of the model output variables is also affected.

(3) Experimental error: With the lack of complete information on measurement uncertainty, the experimentally observed response at different time periods are assumed to be independent in the definition of $\boldsymbol{\Sigma}$ while the computational model output had perfect correlation structure defined using $\boldsymbol{\Lambda}$. This metric is affected by such deficiencies.

Computational aspects: The mean model output being nearly a linear function of time results in a highly skewed covariance matrix $\boldsymbol{\Lambda}_p$, which affects the accuracy of the metric defined in Eqns. (3.29) and (3.30). Suppose we ignore the correlation among the several model output variables, the aggregate metrics B_y and B_h have been computed under the independence assumption and are found to be 240367 and 1927 respectively. The region in the multidimensional space where the model outputs have more than 50% probability of being correct cannot be estimated easily, unlike the univariate case. It was also observed that slight deviations in the experimental result from the model prediction resulted in large changes in the validation metric value (details not shown in this example). Thus multivariate tests could be more stringent compared to marginal comparisons.

3.5.6 Analytical methods for Bayesian analysis

Consider the finite element plate problem shown in Section 3.5.4. Suppose we do not wish to construct a stochastic response surface for the stress prediction y as given in

Eq. 3.24 and still wish to compute B_y and B_h, saddlepoint approximations described in Section 3.4 can be applied for that purpose. The Bayesian network for this example is shown in Fig. 3.19. The goal is to extrapolation inferences across different loading conditions. Once the data $\boldsymbol{z}$ is used to update the response in the validation domain y, the linking variables E and v and hence the decision variable h can also be updated. The prior and posterior densities of y at some model prediction value y_0, can be estimated using Eq. (3.8) as

$$f_y(y)\Big|_{y_0} = \frac{1}{(2\pi)^{\frac{3}{2}}}\left(\frac{\left|\boldsymbol{\Omega}(y_0)\right|}{|\boldsymbol{\Omega}|\left|\left(\nabla y(E,v,w)\right)^T \boldsymbol{\Omega}(y_0)\left(\nabla y(E,v,w)\right)\right|}\right)^{1/2} \frac{f\left(\hat{E}(y_0),\hat{v}(y_0),\hat{w}(y_0)\right)}{f\left(\hat{E},\hat{v},\hat{w}\right)}$$

$$f_y(y\,|\,\boldsymbol{z})\Big|_{y_0} = \frac{1}{(2\pi)^{\frac{3}{2}}}\left(\frac{\left|\boldsymbol{\Omega}(y_0)\right|}{|\boldsymbol{\Omega}|\left|\left(\nabla y(E,v,w)\right)^T \boldsymbol{\Omega}(y_0)\left(\nabla y(E,v,w)\right)\right|}\right)^{1/2} \frac{f\left(\hat{E}(y_0),\hat{v}(y_0),\hat{w}(y_0)\,|\,\boldsymbol{z}\right)}{f\left(\hat{E},\hat{v},\hat{w}\,|\,\boldsymbol{z}\right)} \tag{3.30}$$

The posterior density $f\left(\hat{E},\hat{v},\hat{w}\,|\,\boldsymbol{z}\right)$ shown in Eq. (3.31) can be evaluated using the method described in Section 2.2 and Eq. (3.6). When the data z = 0.013 has been observed with a Gaussian measurement uncertainty ($\sigma_{exp}^2 = 0.0025$), the likelihood $f(z \mid y)$ can be taken as Gaussian as well with mean y and variance σ_{exp}^2 for substituting in Eq. (3.30). The Bayes factors for y and h evaluated at the mean values of E, v, w and P were estimated to be 21.4 and 1.21 respectively. If the response quantities of interest in the validation and application domains are not sensitive to the linking variables, the posterior densities of those common variables do not affect the decision variable as well, in which case the Bayes factor for the decision variable will be close to 1.0. In such cases,

inferences cannot be drawn effectively for the application domain based on the validation data.

Thus the inferences from one loading condition to the other have been propagated through the BN concept proposed in this example and using approximate methods to reduce the number of FE code evaluations. A typical Gibbs sampling procedure that calls the FE code directly would have required nearly 50,000 function evaluations. (A response surface constructed to replace the FE code would also need considerable number of function evaluations depending on the nature of the problem. Also, there is no prior guarantee that all FE models can be represented by second or third order response surfaces with sufficient accuracy). The saddlepoint Laplace approximation required 56 function evaluations for each type of loading (uniform and point loads), thus far the most efficient.

3.6 Summary

Bayesian methodology helps to propagate inferences from the validation domain to the target application domain through the Bayes network approach. Two cases of extrapolation were considered: Extrapolation of validation inferences from one response quantity (for which data is available) to a different response quantity (for which data is absent), and from one input condition to another. The second case included two situations: the input variables in validation and application domains are physically different, and the inputs in the two domains come from two regions of a distribution.

From the numerical examples, it is seen that the sensitivity analysis (second-order variance-based, especially relevant in Bayesian methodology) must be conducted for the

system-level model before component level tests are conducted. The numerical examples can also be used to demonstrate the proposed extrapolation methodology when the underlying physics changes i.e., materials can have elasto-plastic behavior in the application domain. In these particular problems, we have relatively adequate knowledge on the behavior of physical systems. However, this is difficult for systems where the effects of physics change are unknown. Estimating the confidence bounds in the multivariate case is still a numerical challenge which needs further work.

Saddlepoint-based Laplace approximations were used in this study to carry out marginal and conditional density estimation. Although the accuracy of such approximate methods has been investigated previously in the literature, extensive study on their use for model validation remains to be done. Adaptive rejection sampling (ARS) and nonparametric methods have been briefly discussed in this chapter. ARS still has very limited use for even a small problem like the plate model with a hole and hence can be used for parametric analytical model updating purposes only. We conclude that the Laplace approximation methods are promising for validation and extrapolation applications involving very large scale models. Future work in this direction involves application of these techniques to the case of correlated input variables and accuracy estimation studies.

CHAPTER IV

ERROR ESTIMATION IN V&V

4.1 Motivation

The complex phenomena involved in engineering systems are increasingly being sought to be modeled and simulated using numerical methods. Several computational methods and techniques have been developed to accomplish this objective. There is a need to assess the accuracy of these simulations by comparing computational predictions with experimental test data. Conducting full-scale physical experiments, however, could be uneconomical and time consuming. Also, computational models incorporate many assumptions and approximations. Therefore, they need to be subjected to rigorous and efficient verification and validation (V & V) before they can be applied to practical problems with confidence. While chapters 2 and 3 dealt with the issue of validation, this chapter discusses about verification of computational models.

Verification refers to the assessment of accuracy of the solution with respect to known solutions. The aim of the verification process is to identify, quantify and reduce the errors in the computational model (AIAA, 1998). Total uncertainty in computational analysis is understood to arise from a full range of modeling and simulation activities which can be broadly classified as variabilities, errors, and uncertainties. The computational activities which comprise uncertainty quantification can be broadly classified as nondeterministic analysis (assessment and propagation of uncertainty) and

numerical error estimation. Also, measurement error should be included in both inputs and outputs in the validation metric.

Non-deterministic analysis methods have mostly been concerned with propagating variabilities in model parameters (usually defined in terms of probability distributions) through one or more models with the goal of estimating some statistics of interest on the predicted quantities of the models. These methods include Monte Carlo simulation (Iman & Conover, 1982; McKay *et al*, 1979; Deodatis *et al*, 1995), first-order and second-order reliability methods (Hasofer & Lind, 1974; Hohenbichler *et al*, 1987), stochastic finite element methods (Yamazaki & Shinozuka, 1988; Ghanem & Spanos 1991) and response surface methods (Schueller *et al*, 1989; Myers & Montgomery, 1995). A collocation-based stochastic finite element method will be pursued in this study for its efficiency in non-deterministic analysis and ability to quantify errors in the modeling and simulation process.

Thus the current chapter develops methods to quantify and assess the relative influence of errors in numerical modeling vs. measurement error in validation experiments. One of the errors in numerical solution is discretization error. This error is the result of using a discretization method with a finite number of degrees of freedom to solve the set of differential equations. These errors lead to a bias in the computed solution with respect to the true solution of the continuous differential equations. A detailed investigation of discretization error estimation in non-deterministic analysis will be presented. In this study, the non-deterministic analysis is performed using a Stochastic Response Surface Method (SRSM) in which the output response surface is represented by polynomial chaos expansion. There is truncation error in SRSM due to the finite number

of terms in the response surface and this error should be quantified. Besides errors in numerical solution, measurement errors in input variables and their effect on the prediction, and errors in the measurement of output variables during validation experiments will be included in the model validation framework.

4.2 Errors in numerical solution

When continuum models (such as partial differential equations) are used to represent a physical phenomenon, and approximate methods are used to solve those equations, numerical errors are introduced in the solution. The different types of numerical error can combine linearly or nonlinearly but the scope of this chapter is to quantify those errors. Since the errors are derived from the model solution which in turn depends on the uncertain inputs, numerical errors may also be treated as uncertain. This section develops various error estimation and uncertainty quantification methods that are needed for V&V process.

4.2.1 Discretization error (ε_h)

When continuum structures are analyzed through discretized models, the predictions from such models contain numerical errors. Various measures or error estimators have been developed to minimize the discretization error and to adaptively refine the deterministic model. Initial studies in error estimation focused on the convergence and stability of the solution and not specifically the quantification of error. Babushka and Rheinbolt (1978) introduced techniques to approximate the error in energy or energy norm and formed a basis for the error estimation. Elemental residual methods

and interpolation estimates were developed for *a priori* error estimation (Demkowicz *et al*, 1984) in the field of computational fluid dynamics. Extrapolation techniques have been used to estimate the global estimates for the *h*- version of finite element method (Szabo, 1986). Recovery-based methods, wherein the given solution is compared with the solution by a smoothened model, were developed by Zienkiewicz and Zhu (1987). A super-convergent patch recovery-based error estimator was also developed by Zienkiewicz and Zhu (1992). Also, bounds for the global-error estimates and methods for local error estimates (referred to as goal-oriented approach) have been developed (Ainsworth & Oden, 1993, 1997; Babuska *et al*, 1994; Dow, 1999).

The subject of *a posteriori* error estimation is now well established as a result of the above studies, and the error estimates are being investigated for application to mesh refinement problems involving elliptic, parabolic and hyperbolic partial differential equations. The robustness, consistency, stability, and convergence of some these error estimators and indicators around singularity locations still require study. Among the error estimators (e.g., Ainsworth & Oden, 1993, 1997; Babuska *et al*, 1994; Dow, 1999) which have been developed in deterministic finite analysis as well as in classical methods, four easily computable error estimators were extended by Rebba (2002) to numerical analysis with stochasticity. Most of the error estimators have been found to be only useful for adaptive mesh refinement, but not for quantifying the actual error. The actual error is best described by Richardson extrapolation and its ease of computation has attracted the V&V research community (Roache, 2002). Thus, an error estimator based on Richardson extrapolation (Richards, 1997) is considered here for the sake of illustration. However,

the proposed model validation methodology in this study is quite general, and can be implemented with any appropriate error estimator.

Richardson Extrapolation

In the Richardson extrapolation, the error due to grid size is given by

$$\varepsilon_h = \frac{y_1 - y_2}{r^p - 1} \tag{4.1}$$

where the grid refinement ratio $r = \frac{h_2}{h_1}$ and y_1, y_2 are the solutions with the two mesh sizes: coarse and fine. The order of convergence p can be obtained from the equation:

$$p = \ln\left(\frac{y_3 - y_2}{y_2 - y_1}\right) / \ln(r) \tag{4.2}$$

where y_3 is the solution with the finest grid size. The grid refinement ratio is assumed to be constant, i.e., $r = \frac{h_2}{h_1} = \frac{h_3}{h_2}$. The rate of convergence p was computed from Eq. (4.2), using the mean values of the responses y_1, y_2, and y_3 at three different mesh sizes h_1, h_2, and h_3, and using r.

Note that for a particular realization of the random variables, discretization error is by itself deterministic, but in non-deterministic analysis, its randomness arises due to randomness in the input variables. Since the random response is a function of random input variables, the error in the computation of this response is also random and a function of random input variables. Recognition of this fact has led to several studies (Alvin, 2000; Babuska & Chatzipantelidis, 2002), attempting to quantify the discretization error in non-deterministic analysis. As pointed out by Alvin, the dependence of the error estimate on the values of the input parameters of the model

should be account for. Many equations in Babuska and Chatzipantelidis's (2002) clearly show the dependence of the error estimate on the input random parameter.

4.2.2 Errors due to element selection and shape function order (ε_p)

Elemental errors arise during the formulation of the elements and are usually reduced by improving the model prior to the analysis (*a priori* error analysis). These errors may occur due to selection of lower order shape functions and/or due to the approximations made in the geometry of the element. The interpolation polynomial functions used to compute the displacements might introduce error. For example, the use of linear strain elements as opposed to the non-linear elements adds to elemental errors. Practical error estimators have not been developed for quantifying these model errors. Mesh refinement may reduce these errors to some extent if not completely eliminate them. One may use Richardson extrapolation formula for deriving error estimates in a *p*-version finite element method. However this method could be computationally not so efficient to implement for large scale FE models.

4.2.3 Errors due to stochastic analysis

Errors in stochastic analysis are method-dependent, i.e. sampling error occurs in Monte Carlo methods and truncation error occurs in series expansion-based methods such as spectral stochastic finite element method and response surfaces. For the response surfaces, truncation error is usually treated as a Gaussian random variable with zero mean and constant variance. This error variance can be minimized by increasing the order of polynomial used in the response surface or more sample points are selected to fit the

regression models. When Monte Carlo simulation is used to estimate a parameter and if σ is its sample variance, then the error due to Monte Carlo sampling follows a Gaussian distribution with zero mean and a variance of σ^2/n.

4.2.4 Stochastic distribution of discretization error

If the physical, model and data uncertainties are modeled through probabilistic analysis, then the response is not a single value but follows a statistical distribution. Various methods are available to carry out probabilistic analysis to quantify the uncertainty in the output variables, given the statistical distribution of the input variables. Available uncertainty propagation models can be classified into three categories (Haldar & Mahadevan, 2000): (a) analytical methods, (b) sampling based methods, and (c) response surface methods. The choice of method depends on the nature of model used for predicting the output, and the needs with respect to accuracy and efficiency. In this study, a response surface approach is pursued to estimate the distribution of the discretization error. The statistical distribution of the error can then be easily obtained by simulating the input random variables in the response surface model.

Stochastic Response Surface Method (*SRSM*)

A polynomial chaos-based response surface is used, which is found to have superior convergence characteristics than traditional response surface models (Rebba, 2002). The response surface is constructed by approximating both the input and output random variables through series expansions of standard random variables ξ_i. For example, a normal random variable can be expressed in terms of its parameters as $\mu + \sigma\xi$ where ξ is a standard normal variable. A uniform random variable bounded between

a and b is expressed as $a + \frac{b-a}{2}\left(1 + erf\left(\frac{\xi}{\sqrt{2}}\right)\right)$. Similarly, a lognormal random variable with parameters λ and δ can be expressed as exp ($\lambda + \delta\xi$). The output response surface is expressed through a polynomial chaos expansion by:

$$y = a_o + \sum_{i_1=1}^{n} a_{i_1}\Gamma_1(\xi_{1_1}) + \sum_{i_1=1}^{n}\sum_{i_2=1}^{i_1} a_{i_1 i_2}\Gamma_2(\xi_{i_1},\xi_{i_2}) + \sum_{i_1=1}^{n}\sum_{i_2=1}^{i_1}\sum_{i_3=1}^{i_2} a_{i_1 i_2 i_3}\Gamma_3(\xi_{i_1},\xi_{i_2},\xi_{i_3}) + ...(4.3)$$

where y is the output and $\Gamma_p(\xi_{i_1},...,\xi_{i_p})$ are multi-dimensional Hermite polynomials of degree p given by

$$\Gamma_p(\xi_{i_1},...,\xi_{i_p}) = (-1)^p e^{\frac{1}{2}\xi^T\xi}\frac{\partial^p}{\partial\xi_{i_1}...\partial\xi_{i_p}} e^{-\frac{1}{2}\xi^T\xi} \qquad (4.4)$$

where ξ is a vector of independent standard normal variables $\{\xi_{i_k}\}_{k=1}^{p}$. The response surface in Eq. (4.3) is referred to here as a stochastic response surface, to distinguish it from conventional response surfaces. The series could be truncated to a finite number of terms. The accuracy of the computational model depends on the order of the expansion. Additional transformations are necessary if the variables are correlated.

The unknown coefficients may be estimated by various methods such as the Galerkin method or the collocation method (Isukapalli & Georgopoulos, 1999). The latter is used in this study, where the model outputs are computed at a set of collocation points. These collocation points are selected from the roots of the Hermite polynomial of a higher order and are made to capture points from regions of high probability (Tatang, 1997). Response surfaces for model output at different mesh sizes y_1, y_2, y_3 can be

constructed and substituted in Eq. (4.1) to derive a single response surface for eh. Thus statistical distribution of eh can be derived by simulating ξ_1, ξ_2 etc.

4.3 Errors in experimental measurement

In measurement theory (Ang & Tang, 1975), the estimated mean value from the observations is usually assumed to be the true measurement or true value of the underlying variable. The error of the estimated mean value consists of two components: systematic error or bias error, random error. Systematic error depends on the quantity measured, the experimental conditions, and the measurement technique. It may attributed certain well-defined factors whose effects can be determined and thus corrected by a constant bias factor. Random error, which is the other component of measurement error, has a random distribution and can be quantified using statistics. The Student's *t* and Chi-Square distributions in conjunction with the Central Limit Theorem provide a mechanism for determining the required number of observations (Caria, 2000). When a set of observation data is available, the statistical estimate of the mean is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4.5}$$

and the corresponding variance is

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{4.6}$$

The random error about the mean is the standard error of the mean:

$$\sigma_{\bar{x}} = \frac{s_x}{\sqrt{n}} \tag{4.7}$$

Thus the total error in x may be defined as a random variable with Gaussian distribution with zero mean and variance (Barford, 1985) as $\sigma^2 = \sigma_e^2 + \frac{S_x^2}{n}$ in which σ^2_e gives the contribution from systematic error s^2_x/n is random error which can be reduced by increasing the sample size. The above definition of measurement error is valid for both input variables and output variables. However, the measurement error in the input variables will be propagated to the prediction of the output, while the measurement error in the output variables will affect the likelihood function to be used in Bayesian model validation.

4.3.1 Measurement error in the input (ε_d)

If the relationship between input and output is given by

$$u = f(x_1, x_2, ..., x_m) \tag{4.8}$$

then the error in the prediction of the output due to the measurement error in the input variables can be expressed as

$$\varepsilon_d = \Delta u = \sum_{i=1}^{m} \left(\frac{\partial f}{\partial x_i} \right|_{x=\bar{x}} \delta x_i \tag{4.9}$$

in which δx_i is the measurement error in ith input random variable x_i and $\left(\frac{\partial f}{\partial x_i} \right|_{x=\bar{x}}$ is the first order sensitivity coefficient of the model output u with respect to the ith input random variable x_i. Since the measurement error in each input variable can be quantified according to Eq. (4.7), the key to quantifying the error term ε_d is to compute the sensitivity coefficients which are partial derivatives. The partial derivatives may be

obtained either by analytical differentiation or by numerical differentiation (i.e., finite differences). The choice of the method is problem-dependent.

4.3.2 Measurement error in the output (ε_{exp})

Suppose the output response quantity from an experiment is measured as y_{exp}. This result deviates from the true solution due to error in measuring the outcome, denoted here as ε_{exp}.

$$y_{exp} = y_{true} - \varepsilon_{exp} \tag{4.10}$$

The measurement error ε_{exp} is usually assumed to follow a normal distribution with zero mean and a constant standard deviation σ_{exp} (Barford, 1985) that depends on the quantity measured, the experimental conditions, and the measurement technique. (Systematic errors in the test can result in a non-zero mean in the measurement error and hence should be eliminated. Also, the variance in the measured outcome may have resulted from a combination of various factors. Our goal in this study is not to address these various factors; it is assumed that the total variance has already been calculated). The systematic error is deterministic, related to the accuracy and occurs due to bias in the measurement; this error can be eliminated. The random measurement error is difficult to measure from a single experiment but the parameters of its distribution can be determined from repeated observations. The experimental errors are usually assumed to follow a normal distribution due to the following properties (Barford, 1995):

- Positive and negative errors can occur with equal probability (symmetric distribution)
- Small errors are more likely to occur than large errors in a controlled experiment

In general, experimental errors may have zero mean but could still be non-Gaussian.

4.4 Illustration

A simple numerical example is given here to illustrate the proposed stochastic analysis of discretization error. The FEM model (or code) is simply treated as a "black-box" and the size of the problem only changes the computational effort but not the concept. Consider a plate with a hole in the center subjected to uniform distributed loading on the edges. Making use of the symmetry, only a quarter of the plate is analyzed, as shown in Fig. 4.1. The Young's modulus of the plate is assumed to be constant throughout the plate (isotropic) and its value is 10,000 ksi. Also, the plate has unit thickness. A finite element model of the plate is created using the software ANSYS (Version 6.1). The domain is discretized into elemental areas. A linear elastic, plane-stress analysis was performed using ANSYS. Consider two independent input lognormal random variables w_1, w_2 with same mean value of 12 ksi and standard deviation of 2.4 ksi; and one output Von Mises stress σ_v at point A. Two different levels of mesh size are chosen in ANSYS: a coarser mesh with 216 elements and a finer mesh with 486 elements. The ratio of the grid sizes, r, is found to be 0.666. A finite element analysis with a more refined mesh (1102 elements) is carried out to estimate the order of convergence p and it is found to be close to 0.9 as per Eq. (4.2). For this particular example problem, analytical solution is available (Timoshenko & Goodier, 1970) which could also be used to compute p.

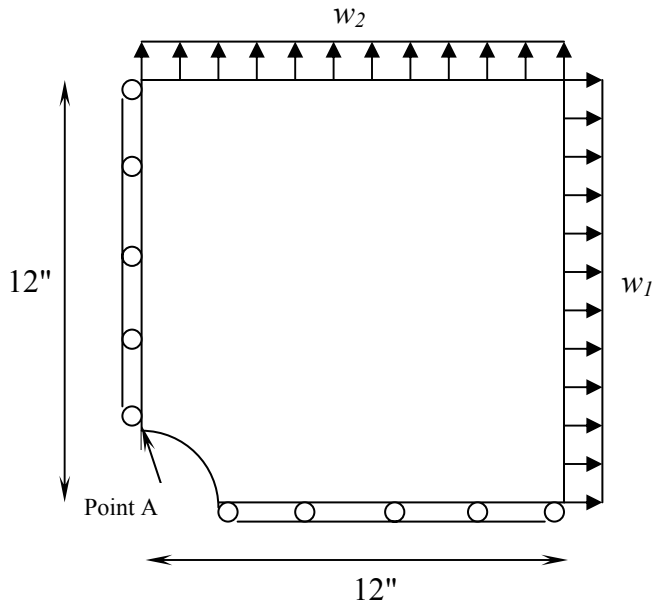
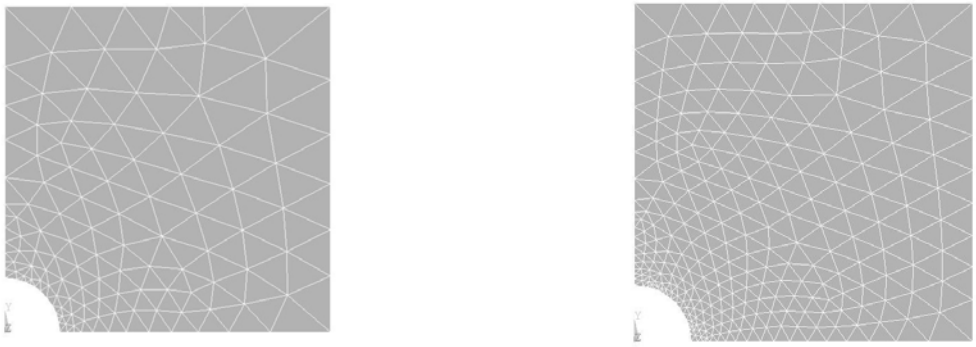


Fig. 4.1. One quarter of a plate with a circular hole at the center

Clearly, if the model output is uncertain, the convergence parameter p also should be a random variable according to Eq. (4.2). In this simple example, the model responses corresponding to the mean input loads $w_1 = 12$ and $w_2 = 12$ have been used to estimate p.



(a) Coarse mesh (216 elements) (b) Fine mesh (486 elements)

Fig. 4.2. Finite element models for the plate

The error estimate is thus calculated as $(\sigma_{v1} - \sigma_{v2})/(0.666^{0.9} - 1)$ where σ_{v1} and σ_{v2} are the Von Mises stresses at point A for the coarse and fine mesh sizes respectively. The input variables are expressed in terms of standard random variables ξ_1, ξ_2 as $(2.465 + 0.198\xi_1)$, $(2.465 + 0.198\xi_2)$ respectively. A second order polynomial chaos expansion for a single output and two input variables is given by

$$\sigma_v = a_o + a_1\xi_1 + a_2\xi_2 + a_3(\xi_1^2 - 1) + a_4(\xi_2^2 - 1) + a_5\xi_1\xi_2 \qquad (4.11)$$

At least 6 sample points are needed to estimate the 6 unknown coefficients in the second order response surface in Eq. (4.11), and 17 samples are needed for a third-order response surface (Isukapalli & Georgopoulos, 1999). These samples are selected at collocation points which are combinations of roots of a Hermite polynomial of one order higher than the order of polynomial expansion. The resulting first, second, and third order response surfaces are computed using multiple linear regression as

$$\sigma_v = 21.6489 + 6.088\xi_1 - 1.833\xi_2 \qquad (4.12a)$$

$$\sigma_v = 21.7985 + 6.165\xi_1 - 1.85\xi_2 + 0.6088(\xi_1^2 - 1) - 0.1743(\xi_2^2 - 1) - 0.0127\xi_1\xi_2 \qquad (4.12b)$$

$$\begin{aligned}\sigma_v = {} & 21.8 + 6.172\xi_1 - 1.856\xi_2 + 0.6144(\xi_1^2 - 1) - 0.1777(\xi_2^2 - 1) - 0.0098\xi_1\xi_2 \\ & + 0.0399(\xi_1^3 - 3\xi_1) - 0.0109(\xi_1^2\xi_2 - \xi_2) + 0.0016(\xi_2^2\xi_1 - \xi_1) - 0.0053(\xi_2^3 - 3\xi_2)\end{aligned} \qquad (4.12c)$$

The residual errors from the regression for the three different response surfaces are 0.356, 0.018 and 0.0002 ksi respectively. Relative to the mean Von-Mises stress, the percentage errors are 1.65%, 0.08% and 0.001% respectively. For practical applications, one may terminate the response surface construction at this stage. Therefore, the 3rd order response surface is used for further computations below.

This same example was also done with normally distributed w_1 and w_2, in which case a first order response surface was found to be adequate i.e, the standard error was

estimated as 0.0008 ksi or 0.004% relative error. Thus, simply changing the distribution from normal to lognormal made it necessary to use a higher order response surface in Eq. (4.12c). Thus higher-order response surfaces may be necessitated by the nature of the randomness in the input variables, even when the physical problem is simple and linear. The third-order response surface for Von mises stress in Eq. (4.12c) is used now to illustrate the stochastic estimation of the discretization error. The response surfaces for Von-Mises stress using two different mesh sizes are given by

$$\begin{aligned}\sigma_{v_1} &= 21.8 + 6.172\xi_1 - 1.856\xi_2 + 0.6144(\xi_1^2 - 1) - 0.1777(\xi_2^2 - 1) - 0.00981\xi_1\xi_2 \\ &+ 0.0399(\xi_1^3 - 3\xi_1) - 0.0109(\xi_1^2\xi_2 - \xi_2) + 0.00167(\xi_2^2\xi_1 - \xi_1) - 0.0053(\xi_2^3 - 3\xi_2) + \varepsilon_1 \\ \sigma_{v_2} &= 22.514 + 6.59\xi_1 - 2.11\xi_2 + 0.6509(\xi_1^2 - 1) - 0.224(\xi_2^2 - 1) - 0.00328\xi_1\xi_2 \\ &+ 0.0433(\xi_1^3 - 3\xi_1) - 0.0067(\xi_1^2\xi_2 - \xi_2) + 0.00037(\xi_2^2\xi_1 - \xi_1) - 0.00072(\xi_2^3 - 3\xi_2) + \varepsilon_2\end{aligned} \tag{4.13}$$

where ε_1 and ε_2 are the residuals, which are found to follow normal distributions with zero mean values and standard deviations of 0.0064 and 0.01 respectively. Since the residuals are negligible compared to the actual response, they are not included in further calculations. The Richardson extrapolation-based error estimator is thus calculated using Eq. (1) as $(f_1 - f_2)/(0.694)$, i.e.,

$$\begin{aligned}\varepsilon_h &= 1.0228 + 0.6018\xi_1 - 0.366\xi_2 + 0.0526(\xi_1^2 - 1) - 0.0676(\xi_2^2 - 1) + 0.0094\xi_1\xi_2 \\ &+ 0.00488(\xi_1^3 - 3\xi_1) + 0.00607(\xi_1^2\xi_2 - \xi_2) - 0.00188(\xi_2^2\xi_1 - \xi_1) + 0.0066(\xi_2^3 - 3\xi_2)\end{aligned} \tag{4.14}$$

The error ε_h is stochastic, whose distribution is obtained from Eq. (4.14) by considering the distributions of the input variables ξ_1 and ξ_2. The error estimator ε_h is found to follow a normal distribution with a mean value of 1.0212 ksi and standard deviation of 0.70777 ksi.

4.5 Summary

Several sources of uncertainties are identified and errors from both simulation and experimental measurement are quantified and included in this study. Once these various errors are quantified, they can be included as additional variables to the model response. Thus the validation metric is affected by these various types of errors. The sensitivities of the Bayes factor with respect to the different sources of error can be quantified, in order to facilitate model refinement after validation. This chapter examined the role of discretization error, error due to stochastic analysis, and measurement errors. The next chapter includes model form uncertainty, reliability analysis error etc in the design.

CHAPTER V

INCLUSION OF MODEL ERRORS IN DESIGN

5.1 Motivation

The development of high performance computers in recent years is leading to an ever increasing reliance on computational models to analyze and design complex engineering systems. However, such simulation models incorporate many assumptions and approximations, thus leading to errors in the prediction. Reliability-based design optimization (RBDO) commonly evaluates the reliability constraints of the physical system through the use of computational models. Before we assess the reliability of the actual physical system, the performance of the simulation model itself needs to be assessed by comparing the model prediction against observations, using specific validation experiments. A rigorous verification and validation process is needed to effectively quantify the uncertainties and errors in the system analysis model, and the model uncertainties and errors should be accounted for in the design optimization.

Uncertainty in engineering analysis arises from three types of sources: (1) Physical or inherent variability: This is commonly represented through random variables in the context of RBDO and generally quantified by probability distributions estimated from observed data; (2) Information uncertainty, due to either limited or qualitative information: In the context of probabilistic modeling, limited data leads to statistical uncertainty, i.e., the uncertainty in the statistical distribution parameters of the random variables identified in the first source. Qualitative information

is handled through epistemic uncertainty methods such as fuzzy sets, possibility theory, evidence theory etc. (3) Model uncertainty and errors, which arise from selection of model form and parameters, assumptions, and approximations at several stages of analysis. In the context of RBDO, this should not only include concerns about the system analysis model, but the reliability estimation method also.

This study only deals with the above sources of uncertainty in a probabilistic context. The reliability analysis in most RBDO studies has been concerned with physical uncertainty to estimate the probability of failure or a reliability index. A few recent studies have also considered the statistical uncertainty mentioned above, which induces scatter in the estimated failure probability. When statistical uncertainties are considered, the reliability estimate is not a single number but follows a probability distribution.

When multiple models are available to describe a physical phenomenon, selecting one of the models for use involves *model uncertainty*, and the resulting model prediction will then contain *model error*. Model selection uncertainty is epistemic in nature, and has been sought to be mitigated through Bayesian model averaging in some studies, but not directly quantified. On the other hand, model error -- the difference between prediction and observation -- can be directly quantified and incorporated in RBDO. Therefore, this study focuses on the quantification and inclusion of model error in RBDO.

The many sources of physical system model error are broadly grouped into two components in this study: model form error and solution approximation (numerical) error. Model form error includes assumptions about system behavior, boundary conditions, model parameters and input variables. When continuum mathematical models are discretized using finite element or finite difference methods, the solution approximation

contains *discretization error*. Response surface approximations have been used as surrogates for large computational models in many design optimization studies, and this introduces additional *truncation error* in the design solution.

The reliability analysis used in RBDO also has errors, which can again be classified as model form and solution approximation errors. Methods such as FORM, SORM etc., commonly used in reliability analysis introduce solution approximation error since they give only approximate estimates of the probability integral. There could also be model form error, e.g., selection of distribution types for the random variables used in the reliability analysis. Model form error may also be introduced due to incorrect or approximate formulation of the limit state function. In addition, there may be statistical uncertainty in the distribution parameters of the random variables due to limited data.

If physical system model error is defined as the difference between test data and model prediction, both of which are treated as random variables in this study, then model error is also a random variable. Once the model error (which includes both model form and numerical errors) statistics are quantified, this study treats this as an additional random variable in RBDO.

Thus the objective of this study is to quantify and include model error in RBDO. First a brief overview of concepts and previous work with respect to uncertainties and errors in physical system analysis, reliability analysis, and reliability-based design is presented. Next a methodology is proposed to quantify model form errors. Two methods are proposed to estimate the statistics of model form error. Numerical examples are provided in the end to illustrate the application of the proposed methodology to mechanical systems design.

5.2 Modeling uncertainties and errors

5.2.1 Model form errors

Model uncertainty arises during the model selection process when we replace physical reality with mathematical models. Earlier studies on this topic have focused on model uncertainty reduction rather than quantification. When there exist several possible models to describe a phenomenon, a Bayesian approach can be used to include all the candidate models by assigning weights (the probability of each model being correct). The model weights may be updated when new observation/data becomes available. This approach has been applied to probability distribution type uncertainty and linear regression model uncertainty problems in statistics (Edwards, 1984; Guedes Soares, 1988; Draper, 1995; Volinsky *et al*, 1997), and was recently used to account for mechanical model uncertainty (Zhang and Mahadevan, 2000; Der Kiureghian, 2001). This method reduces the model form uncertainty and model errors but does not quantify them explicitly. In many practical situations, only one model may be available, in which case Bayesian model averaging is not useful. Whether single or multiple models are used, model error (difference between observation and prediction) is directly observable, and offers a clear approach to account for model uncertainty in design.

Bayesian methods have also been used to update prior model error distributions using the data (Onatski and Williams, 2003). Model selection has also been addressed using a decision-theoretic approach (Radhakrishnan and McAdams, 1995), considering the costs of developing or choosing a highly complicated model versus needed adequacy for the application.

In finite element (FE) analysis, Mehta (1996) and Reid (1998) list the sources of modeling errors, and provide cautions and steps to be implemented to control model form error, but do not quantify it. Hierarchical modeling has been suggested (Kurowski and Szabo, 1997; Oberkampf *et al*, 2002) to improve the current simplistic model with a more complicated model in increasing steps and check if the solution converges to a limit. Hierarchical modeling should be implemented by keeping all other factors (such as mesh size, boundary conditions etc) constant, in order to isolate the effect of model form. Since various types of errors during modeling may cancel each other, overall comparisons with experiments alone can be hazardous for accepting model predictions, and should be accompanied by quantification of various error sources (Kurowski, 2001; Rebba *et al*, 2004).

5.2.2 Solution approximation errors

When continuum mathematical problems are solved through discretized numerical procedures, the predictions from such models contain numerical solution errors. An extensive discussion of the quantification of discretization error and derivation of stochastic distribution of the error has been presented in Chapter 4. Hence further explanation of this error is limited in this section.

5.2.3 Approximations in reliability analysis

Model-based reliability analysis generally uses a demand vs. capacity format, corresponding to a desired performance criterion. Suppose R is the capacity and S is the

demand (both of which are treated as random variables), then a performance function or limit state function g is constructed as

$$g = R - S \tag{5.1}$$

Failure is defined to occur when g is less than zero, and the corresponding failure probability p_f is computed as $P(g < 0)$, knowing the statistics of R and S. R and S could be functions of a vector of basic random variables X, with a joint probability density function $f_X(x)$. Then the failure probability may be estimated as $p_f = \int_{g<0} f_X(x)dx$. The accuracy of this probability estimate is affected by both model form errors (limit state formulation, selection of distributions of the random variables X) and solution approximation errors (e.g., use of FORM, SORM etc.).

Errors in the computational model of the physical system obviously are propagated to the limit state function g in reliability analysis. Also, one might be uncertain about the actual physics behind the limit state and use an empirical model (e.g., fatigue life prediction limit state). Sometimes, a complicated limit state is simplified for the sake of fast evaluation of the reliability index. For the last case, a Model Correction Factor (MCF) method (Ditlevsen and Arnbjerg-Nielsen, 1994) has been suggested to iteratively shift the most probable point (MPP) on the approximate (linear) limit state to the more realistic formulation of the limit state. Several variations of this method have been applied in system-level structural reliability analysis and stochastic process simulation (Ditlevsen and Johannesen, 1999; Franchin *et al*, 2002). A model correction factor can only reduce the *reliability analysis errors* to an unknown extent, but it still

does not quantify the limit state modeling error. Uncertainty in distribution type has been handled by Bayesian averaging as mentioned earlier.

5.2.4 Model uncertainty in Reliability-based design optimization (RBDO)

Many studies on design under uncertainty, in the context of minimizing expected cost subject to reliability constraints, can be found in the literature (Mahadevan, 2004). Reliability-based design optimization techniques have been studied for automotive industry applications (e.g., Du and Chen, 2000; Hoffman *et al*, 2003; Zou, 2004), structural engineering (e.g., Rao, 1984, Royset *et al*, 2001; Faber and Sorensen, 2003), and aerospace systems (e.g., Smith and Mahadevan, 2005). A typical RBDO problem involves (a) minimizing the cost subject to reliability and physical constraints, or (b) maximizing the reliability subject to cost and physical constraints. Weight is used as a surrogate for cost in many RBDO studies. Cost may include manufacturing, operational, maintenance, failure and repair costs (or life cycle costs in general).

A simple, typical RBDO formulation with only component-level reliability constraints is as follows:

$$\text{Minimize } h(\boldsymbol{d}, \boldsymbol{X}) \tag{5.2}$$

$$\text{s.t. } p_{f_i} = P(g_i(\boldsymbol{X}) \leq 0) < p_i \qquad \text{for } i = 1, 2, \ldots, k$$

where $h(.)$ is the objective function (or cost function), $\boldsymbol{d}$ is a set of design variables, $\boldsymbol{X}$ is a set of input random variables and p_i could be i^{th} threshold failure probability. Further each of the design variables $\boldsymbol{d}$ may be bounded. The vector $\boldsymbol{d}$ includes both deterministic design variables and distribution parameters of random design variables. A number of RBDO studies have focused on developing computationally efficient methods to solve

Eq. (5.2). Various nested, decoupled, and single-loop methods are available. The focus of this study is different; we wish to study how model errors in physical system analysis and reliability analysis may be included in the final design ***d***.

RBDO incorporates physical variabilities through the random variables ***X*** but the statistical parameters that describe those probability distributions can be uncertain due to limited data. This additional statistical uncertainty introduces variability in the reliability constraint satisfaction and/or the objective function, and the design must be insensitive (robust) to such variations. In that case, we aim to simultaneously optimize the mean and variance of the performance measure (cost, reliability etc.). This has been referred to as reliability-based robust design (RBRD), and Eq. (2) may be revised as.

$$\text{Min: } E[h(\boldsymbol{d}, \boldsymbol{X})] \text{ and Min: } Var[h(\boldsymbol{d}, \boldsymbol{X})] \tag{5.3}$$

$$\text{s.t. } \mu_{p_{f_i}} < p_i$$

$$\sigma_{p_{f_i}} < p^s{}_i \qquad \text{for } i = 1, 2, \ldots, k$$

where ***d*** is a set of design variables, ***X*** is a set of random variables and, p_i and $p^s{}_i$ represent user-defined limits on the mean and standard deviation of the failure probability estimate for the i^{th} limit state. Thus reliability based robust design (RBRD) uses two objective functions, and several multi-objective optimization methods are available to solve this problem. When the objective function is in the form of a polynomial response surface, Chen *et al* (2004) have proposed an efficient method to analytically compute the mean and variance of the objective function under statistical uncertainty.

A simple weighted sum formulation for RBRD in terms of the reliability index was pursued by Stoebner and Mahadevan (2000), by minimizing the variance of the

reliability index while increasing its mean value. Alternatively, Du *et al* (2003) proposed a percentile formulation to combine the two objectives into a single objective problem.

This section discussed the occurrence of model errors in system analysis, reliability analysis, and reliability-based design. A review of earlier studies shows that most of their concern has been with reducing the modeling errors, not quantifying them or explicitly including them in the design. Next section proposes techniques to quantify model errors so that they can be properly accounted for in reliability-based design optimization.

5.3 Quantification of model errors

As mentioned earlier, the many sources of model error are grouped into two components in this study: (1) behavior assumptions and selection of model form and parameters; and (2) subsequent approximations in numerical implementation. There are other sources of error such as software coding errors and human implementation. It is hoped that they can be eliminated with careful verification, and hence are not included in this discussion. Fig. 5.1 describes the different stages where different types of errors are introduced, all of which contribute to overall model error.

The first step in model-based simulation is to understand the physical concepts involved in a phenomenon. The domain or environment in which the system functions needs to be stated. For example, fluid-structure interaction or solid-solid impact will be modeled differently. Physical mechanisms such as heat transfer, structural dynamics, coupled electro-thermal effects, etc. must be identified. The quantity of interest for the target application and the required inputs to the system must be defined in the conceptual

modeling stage. The next step is to select a mathematical model based on a particular theory, perhaps from a set of possible models. The choice is usually made from past experience and understanding of the system.

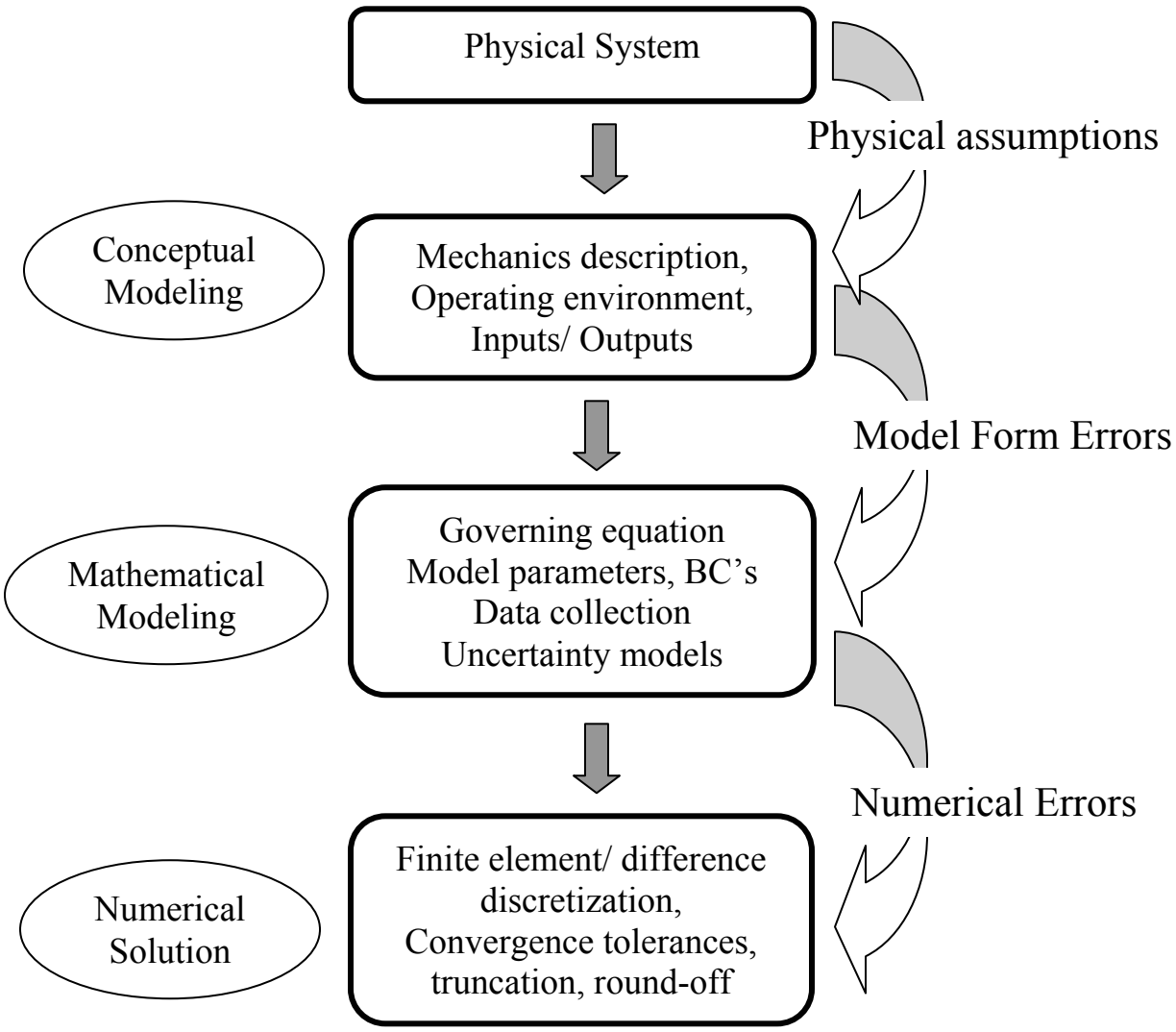


Fig. 5.1 Errors in phases of modeling and simulation

At the model selection stage, we choose a continuum mathematical model (e.g., a partial differential equation with appropriate boundary conditions) to represent the physical phenomenon. Also model parameters (relating to the boundary conditions, for instance), often unknown, are assumed to have some values. At this stage, model form errors will be introduced reflecting the discrepancy between the mathematical model and physical reality. Actually, model form error is introduced even in the conceptual modeling stage, but appears in a quantitative form at the mathematical modeling stage. In the next stage (numerical solution), the continuum model is discretized and solved to compute the system response.

This discretization process introduces numerical errors due to finite mesh or step sizes, convergence tolerances, round-off errors, etc. Errors introduced by response surface approximations are referred to as truncation errors. Refererring to Fig. 5.1, the true physical system response y_{true} is first approximated by a continuum mathematical model resulting in model form error ε_{mf}.

$$y_{true} = y_{cont} + \varepsilon_{mf} \tag{5.4}$$

where y_{cont} is the continuum solution. In the next step, y_{cont} is approximately calculated by a discretized numerical solution y_{pred} (e.g., finite element model), leading to numerical solution error ε_{num}.

$$y_{cont} = y_{pred} + \varepsilon_{num} \tag{5.5}$$

where y_{pred} is the response predicted by the numerical model. Combining the above two equations, we get

$$y_{true} = y_{pred} + \varepsilon_{mf} + \varepsilon_{num} \tag{5.6}$$

This approach is different from the popular notation of total error being represented as square root of sum of squared errors (RSSE) (e.g., Coleman and Stern, 1997) (In such discussions, the term "error" has been actually used to imply variance, and it makes sense to add variances and compute the square root of the sum of variances). RSSE is an indirect indicator of overall model error and does not quantify the actual error.

5.3.1 Quantification of numerical solution errors (ε_{num})

Finite element discretization error may be estimated based on the Richardson extrapolation (Richards, 1997) method. In this method, the error due to grid size h_1 (for a coarse mesh) is given by

$$\varepsilon_{num} = \frac{y_1 - y_2}{r^p - 1} \tag{5.7}$$

where the grid refinement ratio $r = h_2/ h_1$, and y_1 and y_2 are the solutions with coarse and fine meshes respectively. The order of convergence p can be obtained from the relation $p = \ln\left(\frac{y_3 - y_2}{y_2 - y_1}\right) / \ln(r)$ where y_3 is the solution with the finest grid size, and $r = h_2/ h_1 = h_3/ h_2$.

Other numerical errors

The other sources of numerical solution error include bugs in the computer codes, convergence tolerances, truncation errors, singularities corrupting the solution, inappropriate shape functions (in case of finite element, finite difference-based problems) etc. Some of these errors such as response surface error (i.e., truncation error) can be quantified. However, all these errors combine in a nonlinear form that is impossible to

derive. The best way to handle such errors is to minimize them through careful code verification, solution verification, convergence studies etc. If these steps are carried out systematically, it is possible that the miscellaneous numerical errors could become negligible compared to discretization error and model form error.

5.3.2 Model form error (ε_{mf})

This can be quantified only by comparing model prediction to physically observed response. However, the experimental observation y_{obs} is a random variable due to imprecision in the instrument and variation in test conditions. The true value is equal to the observed test result plus the experimental error:

$$y_{true} = y_{obs} + \varepsilon_{exp} \tag{5.8}$$

The experimental error ε_{exp} is usually assumed to follow a normal distribution as described in Chapter 4. In general, experimental errors may have zero mean but could still be non-Gaussian. From Eq. (5.6) and Eq. (5.8), one can write:

$$y_{obs} + \varepsilon_{exp} = y_{pred} + \varepsilon_{mf} + \varepsilon_{num} \tag{5.9}$$

Denoting (y_{obs} - y_{pred}) as ε_{obs} (observed error),

$$\varepsilon_{obs} = \varepsilon_{mf} + \varepsilon_{num} - \varepsilon_{exp} \tag{5.10}$$

Thus model form error can be expressed as

$$\varepsilon_{mf} = \varepsilon_{obs} - \varepsilon_{num} + \varepsilon_{exp} \tag{5.11}$$

5.3.3 Model parameter error

In this study, this error is lumped into model form error and is not treated separately. However, in some applications, it may be necessary to quantify this explicitly. The error (during measurement and/or selection) in the model parameters and input variables will

be propagated to the prediction of the output. If the relationship between input and output is given by $y = u(x_1, x_2, ..x_m)$, then the error in the prediction of the output due to error in the input variables may be approximated using a first-order sensitivity analysis as

$$\varepsilon_d = \Delta y = \sum_{i=1}^{m} \left(\frac{\partial u}{\partial x_i} \right)\Bigg|_{x = \bar{x}} \delta x_i \tag{5.12}$$

in which δx_i is the measurement error in i^{th} input random variable x_i and $\left.\frac{\partial u}{\partial x_i}\right|_{x=\bar{x}}$ is the first order sensitivity coefficient of the model output y with respect to the i^{th} input random variable x_i.

5.3.4 Quantification of statistics of model form error

In Eq. (5.11), ε_{num} is a random variable whose distributions can be estimated from the distributions of model outputs at coarse and fine meshes as explained in Section 5.3.1. The experimental error ε_{exp} is also a random variable but from a single experiment, we do not know the precise value of ε_{exp} and therefore of ε_{mf}. Only the statistics such as mean, variance and if possible the distribution of the random variable ε_{exp} can be estimated based on repeated observations and through prior experience. If all the terms in the right-hand side of Eq. (5.11) are treated as random variables, then the model form error ε_{mf} also becomes a random variable. Two methods are investigated below to quantify its statistics.

Resampling Method

The assumption in the resampling method is that the underlying distribution of the data is parametric and we use the sample data set to generate the underlying population of the parameters (Good, 1999). In the basic bootstrapping method (Efron, 1979), we derive the

distributions of the statistical parameters such as mean $\mu_{\varepsilon_{mf}}$ and standard deviation $\sigma_{\varepsilon_{mf}}$ of the model error by resampling a large number of ε_{obs} values from the existing finite data set. Each time a value for ($\varepsilon_{obs} - \varepsilon_{num}$) is *resampled*, a randomly generated term ε_{exp} is added to it. Thus, many samples of ε_{mf} are obtained, which can be used to compute the statistical parameters of ε_{mf}. Repeating this procedure a number of times provides several sets of samples. These sets can be used to compute the statistics of the distribution parameters of ε_{mf}.

When only a finite number of values for ε_{mf} are available, the histogram constructed for the model form error will be quite coarse as shown in Fig. 5.2, thus not suitable for identifying the distribution of ε_{mf}. (Fig. 5.2 is only for the sake of illustration; hence the scale is irrelevant at this point). However, a smoother histogram can be obtained by filling the gaps in the histogram through an interpolation technique. The new histogram could then be used to derive an approximate continuous distribution for the model form error ε_{mf}. A smoothed bootstrapping method based on the interpolation of the original data (Silverman and Young, 1987) can be used for this purpose. First we resample ε_{mf} from the finite number of samples, and a small random term is added to each resampled value. This random term is again scaled down so that it does not affect the estimation of population statistics.

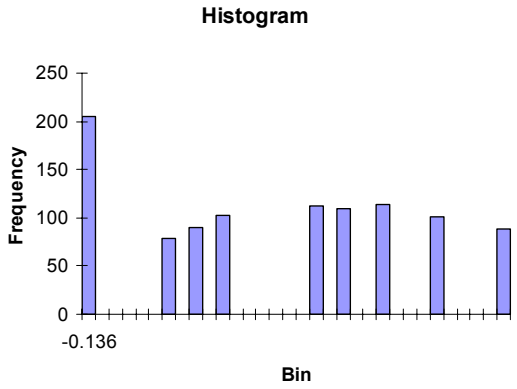


Fig. 5.2 Discrete histogram for bootstrapped samples

Suppose a bootstrap sample set $y_1^*, y_2^*, \ldots, y_n^*$ is generated by drawing n values at random (with replacement) from the original sample $x_1, x_2, \ldots, x_n$. A smoothed bootstrap resample $x_1^*, x_2^*, \ldots, x_n^*$ is obtained by calculating

$$x_i^* = \bar{y}^* + \frac{\left(y_i^* - \bar{y}^* + \delta\varepsilon_i\right)}{\sqrt{1 + \frac{\delta^2}{s^2}}} \tag{5.13}$$

for $i = 1, \ldots, n$, where $\bar{y}^*$ is the mean of the y_i^*, s^2 is the sample variance of the observations y_i, and ε_i are the random errors drawn from $N(0, 1)$. Here, δ known as window size determines the level of smoothing and is usually determined arbitrarily as $s/\sqrt{n}$. We repeat this process a large number of times (say 10,000) to obtain $10000n$ unique samples of x. These large number of realizations are then used to construct a continuous probability density function for x. In a similar fashion, the smoothed samples of ε_{mf} can be obtained to derive its empirical distribution.

Analytical approximation of finite samples

An alternative approach is to use saddlepoint methods (Jensen, 1995) that make use of characteristic functions and their variations to derive approximate empirical probability

distributions. Suppose each i.i.d sample x_i from the data $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ set has some common unknown distribution function $f_X(x)$ with cumulant generating function (CGF) $K(\lambda)$. By definition, CGF for a continuous random variable x is derived from its PDF $f_X(x)$ as

$$K(\lambda) = \log\left(\int e^{\lambda x} f_X(x) dx\right) \tag{5.14}$$

If we have discrete data set, then a simple calculation gives

$$K(\lambda) = \log\left\{\frac{\sum_{i=1}^{n} \exp(\lambda x_i)}{n}\right\} \tag{5.15}$$

Then a parameter λ_x can be defined as a solution to $\frac{dK(\lambda_x)}{d\lambda} = x$ i.e., from Eq. (5.16):

$$x = \frac{\sum_{i=1}^{n} x_i \exp(\lambda x_i)}{\sum_{i=1}^{n} \exp(\lambda x_i)} \tag{5.16}$$

Then the probability density function of x can be written approximately as

$$f_X(x) = \left\{\frac{n}{2\pi K''(\lambda_x)}\right\}^{\frac{1}{2}} \exp\left[n\left\{K(\lambda_x) - x\lambda_x\right\}\right] \tag{5.17}$$

Further, an analytical CDF of x can be computed as (Daniels, 1987)

$$F_X(x) = \Phi(w_x) + \phi(w_x)\left[\frac{1}{w_x} - \frac{1}{z_x}\right] \tag{5.18}$$

where $w_x = sign(\lambda_x)\left[2n\left\{x\lambda_x - K(\lambda_x)\right\}\right]^{\frac{1}{2}}$ and $z_x = \lambda_x\left\{nK''(\lambda_x)\right\}^{\frac{1}{2}}$. Using this idea, an empirical continuous PDF or CDF can be derived for the model form error ε_{mf} from finite

sample size of size n. Obviously, the accuracy of the distributions thus derived using saddlepoint approximations improves with increased sample size.

In this section, errors in various stages of modeling and simulation are identified, and procedures for the quantification of various types of errors (in numerical solution and model form) are developed. When finite data is available, distributions for the model form error can be constructed using saddlepoint approximations or smoothed bootstrapping. Next section develops the method to include the effect of model errors in the design optimization.

5.4 Inclusion of model errors in rbdo

The model errors are proposed to be included in RBDO in two steps. In the first step, the limit state function is modified to reflect the physical system model error. In the second step, the error due to the use of an approximate reliability analysis method is included during the RBDO iterations.

Consider a reliability calculation $p_f = P(g < 0)$, where $g = R - S$. The computational model response S_{model} may be augmented with the overall error in the physical system model ε_{model} to construct the response quantity S to be used in reliability analysis, as

$$S = S_{model} + \varepsilon_{model} \tag{5.19}$$

where ε_{model} is the sum of ε_{mf} and ε_{num}. The statistical distribution of ε_{mf} is estimated using either interpolated resampling or saddlepoint approximation described in Section 5.3, and the statistical distribution of ε_{num} is estimated using the Richardson extrapolation method.

Or in general, ε_g, the overall error in g, may be computed based on the error in calculating S.

The design formulation is the same as shown in Eq. (5.2), with the additional random variable ε_g introduced as follows:

$$\text{Minimize: } h(\boldsymbol{d}, \boldsymbol{X}) \tag{5.20}$$

$$\text{s.t. } p_{f_i} = P(g_i(\boldsymbol{X}) + \varepsilon_g \leq 0) < p_i \qquad \text{for } i = 1, 2, \ldots, k$$

The reliability analysis method (e.g., FORM which is commonly used in RBDO) induces errors as well. This error could be reduced by using Monte Carlo simulation to evaluate the reliability constraints, which might not be feasible if the function evaluation is expensive (e.g., finite element analysis). A more efficient method is proposed here, along the following steps:

1. The RBDO iterations are first conducted using FORM to evaluate the reliability constraints. (This means that several efficient single loop and decoupled RBDO methods can be used).
2. Once the FORM-based RBDO reaches an optimum, the reliability constraints are evaluated using Monte Carlo simulation. This helps to quantify the error $\varepsilon_{FORM} = p_{MC} - p_{FORM}$ in the evaluation of each reliability constraint.
3. The right hand side of each reliability constraint is augmented with the term ε_{FORM} with the appropriate sign. That is, if the Monte Carlo estimate p_{MC} is higher than the FORM estimate p_{FORM}, then the target failure probability on the right hand side of the i^{th} reliability constraint in Eq. (5.20) becomes $p_i - \varepsilon^i_{FORM}$. This means that if FORM is found to underestimate the failure probability, the constraint

appropriately becomes more stringent, and vice versa. Thus the optimization problem becomes

$$\text{Minimize: } h(\boldsymbol{d}, \boldsymbol{X}) \tag{5.21}$$

$$\text{s.t. } p_{f_i} = P(g_i(\boldsymbol{X}) + \varepsilon_{g_i} \le 0) < p_i - \varepsilon^i_{FORM} \qquad \text{for } i = 1, 2, \ldots, k$$

4. FORM-based RBDO is once again carried out with Eq. (5.21).
5. Steps 2 to 4 are repeated until convergence.

After step 2, if FORM is found to overestimate the failure probability, the designer may choose one of two options: either continue with the remaining steps till convergence, or stop at step 2 and accept the conservative design provided by FORM.

Notice that the RBDO formulation in Eq. (5.21) incorporates both types of error: ε_{g_i} represents the error in the physical system model, and ε^i_{FORM} represents the error due to the reliability analysis method. Of course, the estimation of ε^i_{FORM} is based on Monte Carlo simulation, which itself has error. A simple formula for the precision (standard deviation) Monte Carlo simulation is given as (Haldar and Mahadevan, 2000):

$$\varepsilon = \sqrt{\frac{P_f^T (1 - P_f^T)}{N}} \tag{5.22}$$

where N is the number of samples, and P_f^T is the true failure probability, approximated by the estimated probability. However, the Monte Carlo error is reducible, by increasing the number of samples N. Therefore, the Monte Carlo error should first be reduced to a negligible amount before quantifying the FORM error. Several efficient methods such as adaptive importance sampling (Zou *et al*, 2004) are available to reduce the Monte Carlo computational effort while achieving the required accuracy.

5.5 Numerical examples

5.5.1 Gear-shaft assembly

Consider a mechanical drive shaft assembled into a press-fit gear wheel as shown in Fig. 5.3. The objective is to determine the radii of the solid shaft R and the gear wheel R_0 such that the assembly meets the design torque requirements reliably without slipping at the fit interface (Cruse, 1997). The interface length L is known and the interference fit tolerated in this assembly Δ is also deterministic. The maximum torque T that can be transmitted by the assembly (fit) without any slippage can be given in terms of the coefficient of friction η at the fit, interface length L (or gear wheel width in this case), interference fit Δ and the interference pressure p as (Shigley *et al*, 2004)

$$T = 2\pi\eta pLR^2 \tag{5.23}$$

The interface pressure can be derived using the assumption of a thick cylinder for the gear wheel and the shaft as

$$p = \frac{\Delta}{R\left[\frac{1}{E_0}\left(\frac{R_0^2 + R^2}{R_0^2 - R^2} + \nu_0\right) + \frac{1}{E_i}\left(1 - \nu_0\right)\right]} \tag{5.24}$$

where E_0 and E_i are Young's moduli, ν_0 and ν_i are the Poisson ratios of the gear wheel and the drive shaft respectively. The values for the width of the gear wheel L and the interference fit Δ are assumed deterministic here for the sake of simplicity. The material properties and the coefficient of friction have inherent variability and are beyond the designer's control; hence they are assumed to be random variables. The values for the deterministic variables and statistics of the various uncertain parameters are given in Table 5.1.

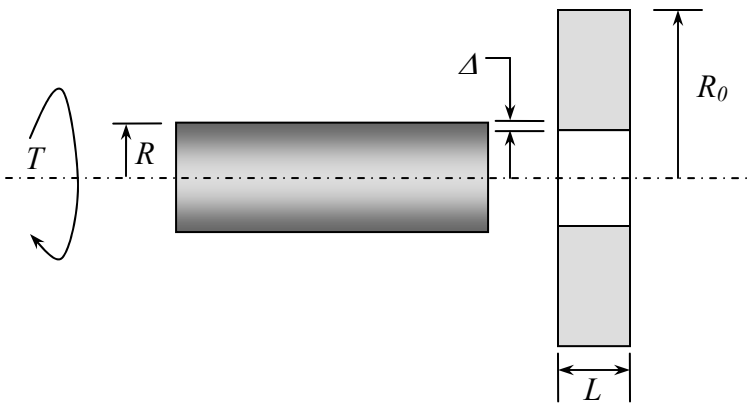


Fig. 5.3 Torque shaft

The initial values for the design variables R (shaft radius) and R_0 (wheel radius), both of which are assumed deterministic in this example, are 8 and 16 respectively. The two variables are bounded as $5 \le R \le 9$ and $10 \le R_0 \le 20$ respectively. (Strictly speaking, they are random; if their variability is significant, we can treat them as random variables and use their mean values as design optimization variables).

Table 5.1. Statistics of variables in assembly design

Variable	**Type**	**Mean**	**Std. Dev**
E_0	Normal	10000 units	200 units
E_i	Normal	8000 units	200 units
v_0	Normal	0.2	0.05
v_i	Normal	0.15	0.05
η	Normal	0.75	0.2
Δ	Deterministic	0.01 units	-
L	Deterministic	4 units	-

Suppose we wish to ensure that the maximum torque transmitted by the assembly fit exceeds a threshold value T_0. The probability of achieving the design requirement needs to be evaluated first. A limit state is defined as $g = T - T_0$ and failure is defined when the torque delivered (T) is less than T_0 i.e., when $g < 0$. The analytical model for the interface pressure given in Eq. (5.24) is derived on the assumption that the pressure across the length L is uniform when the shaft is driven into the gear wheel by force. In reality this assumption may not be true and there may be non-uniform pressure built in the interface. Since the lengths of the shaft and the hub are not the same, stress concentration can occur at each end of the hub (gear wheel). It is also assumed that the two components have only elastic strains in them after the fit is assembled and this assumption is never checked. All these issues introduce model form error. The equations of the mechanical problem are simple and can be solved analytically, no numerical procedure is required; hence numerical solution error ε_{num} is not considered in this example.

Thus accounting for the model form error only, the limit state can be rewritten as $g = T_{pred} + \varepsilon_{mf} - T_0$. Knowing the specific densities of the shaft ρ_i and the gear wheel ρ_0, the total weight of the assembly can be estimated as

$$W = \pi L g_a[\rho_0 R_0^2 + (\rho_i - \rho_0)R^2] \tag{5.25}$$

where g_a is the acceleration due to gravity. In this illustrative example, ρ_0 and ρ_i are assumed to be 7.85 and 7.95 respectively (this introduces model parameter selection error which is lumped in model form error). Also T_0 is assumed to be 1000 units.

Model form error quantification

Suppose 12 different assemblies (to simulate the variability that might occur in the real world) are tested to measure the torque T_{obs} delivered by the fit. The corresponding model

predictions T_{pred} are obtained using the same inputs used in the test setup (In reality however, not all inputs to the computational model can be measured accurately). Using observed values T_{obs} = {4805.943, 4797.649, 3918.362, 4759.615, 5363.197, 7187.641, 6213.017, 5456.729, 5173.763, 5926.158, 4737.321, 4865.384} and predicted values T_{pred} = {4681.648, 4636.136, 3796.127, 4572.260, 5078.693, 6999.631, 6048.302, 5230.336, 4898.267, 5690.531, 4616.021, 4645.201,}, ε_{obs} or the difference between T_{obs} and T_{pred} is calculated for 12 data points. The experimental error in this problem is assumed to be Gaussian with zero mean and constant variance σ^2_{exp} = 100, for the sake of illustration. For this example, the model form error ε_{mf} is found to have a normal distribution with mean 192 and standard deviation of 73 units. Either the interpolated resampling or saddlepoint approximation technique in Section 5.3 may be used to estimate the probability distribution of ε_{mf}, using the twelve data points and model predictions.

Reliability-Based Design Optimization

The RBDO formulation is

$$\text{Min: } W = \pi L g_a[\rho_0 R_0^2 + (\rho_i - \rho_0)R^2]$$

$$\text{s.t: } P(T < T_0) \le p_0 \qquad (5.26)$$

where p_0 is assumed to be 0.002 in this example. Since the torque T transmitted by the mechanical assembly depends on both R and R_0, the probability $P(T < T_0)$ also depends on those respective radii. One can use Monte Carlo simulation or FORM to evaluate the probability constraint in Eq. (5.26). The optimization problem can be solved in three ways, depending on the reliability analysis method – FORM alone, Monte Carlo alone, or applying corrections to the FORM estimate by comparing with the Monte Carlo estimate (using the algorithm developed in Section 5.4). The results of the first two options are

shown in Table 5.2, for two cases – ignoring model error, and including model error. Table 5.2 reports the optimum solution for the design variables R and R_0, the total weight W, and mean torque delivered by the optimum design. Comparison of the results of the first two options shows very close agreement, therefore the third option is not necessary in this problem.

Basic Monte Carlo simulation requires a very large number of samples in this problem (about 10 million, based on Eq. 5.22) to achieve the level of accuracy needed to quantify the FORM analysis error. Therefore, an adaptive importance sampling (AIS) procedure (Zou *et al*, 2004) is used, which achieves similar accuracy within 10,000 samples.

Table 5.2. Optimal design solution for the mechanical assembly

Probability estimation method	**Ignoring model error**			**Including model error**		
	$(R, R_0)_{opt}$	W	**Mean Torque**	$(R, R_0)_{opt}$	W	**Mean Torque**
FORM	(7.43, 12.91)	1314	4290 units	(6.98, 10.69)	854	3360 units
Monte Carlo (AIS)	(7.42, 12.91)	1313	4290 units	(6.98, 10.69)	854	3360 units

Comparing the two columns in Table 5.2 (ignoring model error and including model error), it is seen that the structure weight W is less when model error is included. In this example, the computational model underestimates the torque delivered by the assembly (as evident from positive model form bias), and therefore overestimates the failure probability. In order to meet the reliability requirements, the RBDO algorithm tries to design for a larger torque and hence produces a heavier structure. When the

physical model error is included, the underestimation of the delivered torque is corrected, and a lighter structure meets the design requirements.

Comparison of the optimum solutions across the two rows shows that the use of FORM is adequate to evaluate the reliability constraint in this problem. There is very little difference between the FORM-based and Monte Carlo-based solutions. For the sake of completeness, the probabilities of failure corresponding to the FORM-based solutions are calculated using Monte Carlo simulation (AIS), and are found to be 0.00198 and 0.002 respectively without and with model form error (FORM estimated 0.002 in both cases). Since the results of FORM and Monte Carlo are very close, the third option of quantifying the FORM analysis error and re-solving the RBDO problem is not pursued. The number of limit-state or g evaluations in each of the above methods is shown in Table 5.3.

Table 5.3. Computational efficiency for the assembly design

Method	**Ignoring model error**	**Including model error**
FORM	480	560
Monte Carlo (AIS)	200,000	200,000

Several assumptions regarding the physical behavior of the fit introduced model form error in this example. The main objective of the study, to include the effect of model error on the reliability constraint and hence on the RBDO solution, was achieved and demonstrated in this simple mechanical problem. In the next example, a finite element model is used for the system analysis instead of a closed-form analytical equation, thus creating both model form and numerical solution errors.

5.5.2 Shape optimization of cantilever plate

Consider a cantilever plate with three holes of equal size as shown in Fig. 5.4. The plate has a length L, height h and unit thickness. The structure is subjected to uniform loading w along its span. The design goal is to determine the hole radius r that minimizes the total weight (or area) of the plate such that the probability of vertical displacement at the free end being greater than a threshold level is less than an allowable value. The random variables in this problem are Young's modulus $E \sim N$ (10,000, 200) units and loading $w \sim N$ (100, 20) units. The design variable r is in fact deterministic but in order to construct a stochastic response surface for the displacement in terms of E, w and r, the radius of each hole was varied uniformly in the range 0.25 to 1.25. The Poisson ratio v and height h are assumed to be 0.2 and 4 respectively (deterministic). The structure has an overall length of L = 12 units with the holes equally spaced apart at a distance of 0.25L from each other as shown in Fig. 5.4.

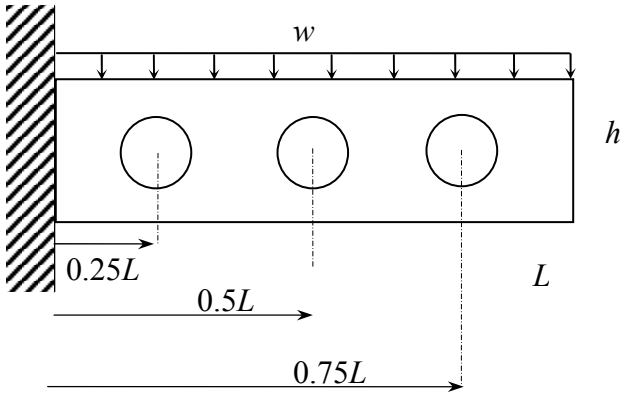


Fig. 5.4 Cantilever plate with three holes

The vertical displacement D at the free end is computed using a finite element code and the threshold displacement D_0 is set to be 9.0 units. Thus the design optimization problem is formulated as

$$\text{Minimize:} \quad A = Lh - 3\pi r^2$$

$$\text{s. t: } P(g < 0) < 0.002 \tag{5.27}$$

where $g = D_0 - D$. In this example, a stochastic response surface (Tatang *et al*, 1997) using polynomial chaos (Ghanem and Spanos, 1991) is constructed for the coarse computational model first. Three levels of hole radius are chosen and for each such configuration, three different mesh sizes were chosen (coarse, fine and finest). The response surface corresponding to a coarse model is given by

$$D_1 = 5.6018 - 0.113\xi_1 + 1.289\xi_2 + 0.5769\xi_3 + 0.2263\xi_3^2 + 0.1145\xi_2\xi_3 + \varepsilon_1 \tag{5.28a}$$

where ξ_1, ξ_2 and ξ_3 are standard normal variables related to E, w and r through the relations $E = 10000 + 200\xi_1$, $w = 100 + 20\xi_2$, $\mu_r = 0.25 + \Phi(\xi_3)$. The residual (or truncation error) ε_1 is observed to follow a normal distribution with zero mean and variance of 0.0121. (This is negligible compared to the mean value of D_1, thus the response surface is accurate). As stated earlier, the reason for varying r as a uniform random variable was to construct a single response function in terms of E, w and r instead of multiple surfaces for different values of r. Once the response surface is constructed, one can derive the probability distribution of D_1 for each r value. Thus Eq. (5.28a) may be rewritten as

$$D_1 = 5.602 - 0.113\xi_1 + 1.29\xi_2 + 0.577\Phi^{-1}(r - .25) + 0.226\left[\Phi^{-1}(r - .25)\right]^2 + 0.1145\xi_2\Phi^{-1}(r - .25) + \varepsilon_1 \tag{5.28b}$$

The response surface in Eq. (5.28b) is then used for computing the statistics of D_1 at any r by generating 10,000 samples of ξ_1, and ξ_2. For instance, D_1 is observed to follow a normal distribution with mean a value of 5.602 units and a standard deviation of 1.295 units when r = 0.75. The response D_1 follows normal distribution with mean 5.315 and standard deviation of 1.218 units when r = 0.5 and so forth. Similar to Eq. (5.28a), response surfaces for finer and finest mesh sizes are constructed as

$$D_2 = 6.3156 - 0.12794\xi_1 + 1.278\xi_2 + 0.9856\xi_3 + 0.3917\xi_3^2 + 0.1967\xi_2\xi_3 + \varepsilon_2 \qquad (5.28c)$$

$$D_3 = 6.4473 - 0.1308\xi_1 + 1.3066\xi_2 + 1.0945\xi_3 + 0.443\xi_3^2 + 0.2185\xi_2\xi_3 + \varepsilon_3 \qquad (5.28d)$$

Again, the residuals (or truncation errors) ε_2, ε_3 and ε had very small mean values and variances compared to their respective mean model predictions. Next, based on the Richardson extrapolation estimate in Eq. (5.7), the discretization error in this plate problem is expressed as

$$\varepsilon_h = 0.8378 - 0.0191\xi_1 + 0.1202\xi_2 + 0.642\xi_3 + 0.2014\xi_3^2 + 0.2173\xi_2\xi_3 + \varepsilon \qquad (5.29)$$

Similar to D_1, the response surface in Eq. (5.29) is used for computing the statistics of ε_h at any r by generating 10,000 samples of ξ_1 and ξ_2. ε_h is observed to follow a normal distribution with mean value of 0.8378 units (indicating positive model bias) and a standard deviation of 0.1217 units when r = 0.75. Thus the statistics of numerical solution error are computed. Note that the above computation is used to estimate two types of error – discretization error ε_h due to the finite element discretization of the continuum structure, and truncation errors ε_1, ε_2, and ε_3 due to the response surface approximations of the finite element model. However the truncation errors were observed to be negligibly small, as mentioned above. Thus only discretization error is considered further.

The next step is to compute the statistics of model form error. Suppose 8 different plates have been tested for displacement over a range of loads (this represents the sample of cantilever plates with possible range of configurations). The corresponding predictions and measurements are obtained as D_{pred} = {6.239, 7.017, 7.334, 6.770, 9.146, 6.512, 5.030, 5.573} and D_{obs} = {6.095, 7.605, 7.692, 7.367, 9.373, 7.028, 5.464, 5.687} respectively. The experimental error is assumed to be Gaussian with zero mean and constant variance $\sigma^2_{exp} = 0.01$. Using this data, the model form error ε_{mf} is found to have a Weibull distribution with 3 parameters: scale, shape and location as (0.6316, 2.2846, 0.8947). The mean of this distribution is 0.34 (indicating positive model bias) and standard deviation is 0.26.

Thus the modeling error in this problem has both components: (1) model form error ε_{mf}, due to the mathematical modeling of the plate behavior; and (2) numerical (finite element) solution of the mathematical model (ε_h). In this example, the mean value of ε_h (0.837) is larger than that of the model form error ε_{mf} (0.1217). The next step is RBDO. Either D_1 alone can be used for design (i.e., ignoring modeling error) or ($D_1 + \varepsilon_{mf} + \varepsilon_h$) can be used for design (i.e., including model error). Both cases are considered in this example. The RBDO results with three options – FORM alone, Monte Carlo (AIS) alone, and quantifying the FORM analysis error and re-solving the RBDO problem using FORM – are summarized in Table 5.4. Table 5.4 reports the optimum solution for the hole radius, corresponding area A (indicates weight), and mean displacement produced by the optimum design. In this problem, the finite element model appears to underestimate the actual displacement and hence overestimate the reliability. The target reliability is then achieved with a lighter structure. When model error is included in the

design, the reliability constraint becomes more stringent, and the RBDO results in a heavier structure (i.e., smaller hole radius), as shown in Table 5.4.

Table 5.4. Optimal design solution for the cantilever problem

Probability estimate in the constraint	**Ignoring model error**			**Including model error**		
	r_{opt}	A	**Mean Displ.**	r_{opt}	A	**Mean Displ.**
FORM	0.595	44.66	5.43 units	0.3695	46.71	5.23 units
Monte Carlo (AIS)	0.595	44.66	5.43 units	0.3742	46.68	5.24 units
FORM estimate $+ \varepsilon_{\text{FORM}}$	-	-	-	0.3742	46.68	5.24 units

Note that both ε_{mf} and ε_h have positive means in this problem (calculated using D_{pred} and D_{obs}, and Eq. 5.29), confirming that the computational model underestimates the displacement. In Table 5.4, both FORM and Monte Carlo simulation gave the same solution when model error is ignored. This is not surprising, since the displacement D (approximated by D_1) computed by the finite element model is found to have a normal distribution, and the limit state function is simply $g = D_0 - D_1$, where D_0 is a constant. Therefore the third method, i.e., estimating FORM error and re-solving the RBDO problem, is unnecessary in this case.

When the discretization error and model form error are included in the RBDO, the limit state function becomes $g = D_0 - (D_1 + \varepsilon_{mf} + \varepsilon_h)$. Both the error terms are non-Gaussian, creating non-linearities in the equivalent normal transformation within FORM, and hence the FORM and Monte Carlo results are different, as shown in Table 5.4. When the FORM solution is evaluated using Monte Carlo simulation, its failure probability is found to be 0.00184, i.e., FORM overestimated the failure probability and produced a

heavier structure. In the third row of Table 5.4, the FORM reliability analysis error is included as per the algorithm in Section 5.4, leading to a lighter structure (same as that found by the use of Monte Carlo all the way, i.e., second row in Table 4). The number of limit-state *g* evaluations in each method is shown in Table 5.5.

Table 5.5. Computational efficiency for the cantilever plate design

Method	**Without model form error**	**With model form error**
FORM	192	240
Monte Carlo (AIS)	120,000	120,000
FORM estimate + $\varepsilon_{\text{FORM}}$	-	40, 960

5.6 Summary

While previous RBDO methods have included the randomness in physical variables, this study proposes a methodology to quantify errors due to system model form, numerical solution approximation, and reliability analysis approximation, and then explicitly include these errors in reliability-based design optimization, in the context of probabilistic analysis. Two different methods were proposed to estimate the statistical distribution of model form error using limited data. Richardson extrapolation-based estimates can be used to quantify finite element discretization errors. An iterative scheme was proposed to include the reliability analysis error in the design using a limited number of Monte Carlo analyses.

This study grouped the many sources of model error into a few broad categories for the purpose of quantification, and further refinement may be pursued to quantify the

contributions of other sources to overall model error. For the sake of RBDO, only overall model error distribution is required for inclusion in the optimization formulation of Eq. (5.20) or (5.21). However, it is desirable to quantify the different sources of error, in order to facilitate trade-off decisions regarding resource allocation for model improvement. The use of response surfaces or simplified closed form analytical expressions is quite common in optimization due to the computational expense, and appropriate truncation errors need to be quantified and included in the design. (The second numerical example in Section 5.5 quantified the truncation error in the response surface, but this error was found to be negligibly small for that particular problem).

The numerical examples in this study were carried out using the classical nested loop RBDO formulation and the number of g evaluations needed in each case was reported in Section 5.5. The focus of this study is not on efficiency, but on the inclusion of various sources of error in the design optimization. Several more efficient RBDO methods (single loop and sequential) have been developed in recent years, and all these methods can be enhanced to incorporate model error. Future work in this direction also needs to include system reliability constraints, and needs to consider additional approximations in calculating system reliability.

CHAPTER VI

CONCLUSIONS AND FUTURE WORK

6.1 Synopsis

Model validation has mostly been a graphical comparison exercise in the past and uncertainties were not rigorously incorporated in making any inferences. While the terminology on validation is now well established, this study is the first of its kind in actually proposing and implementing a quantitative framework for model validation. Several types of validation metrics have been suggested that measure how well the model outputs match with the data when both quantities are uncertain. Statistical hypothesis testing procedures formed the basis for defining most of the metrics. Numerical examples highlighted the different inferences conveyed by point null versus interval-based hypothesis testing formulations. Both classical and Bayesian methods have been explored depending on the amount of prior information available, nature of model prediction and test data. The proposed validation metrics will enable modelers to assess the confidence in their computer model predictions and help make informed decisions regarding need and resource allocation for further data collection.

The validation process estimates the confidence in a model prediction made within a certain validation domain or the region in which test data is available. Often, the decision variable or the target region of application could be different from the response quantity validated. In order to assess the predictive capabilities of the model beyond the test region, we need to extrapolate the inferences across various domains. The term

extrapolation in this study should not be confused with the commonly used time series or spatial extrapolation found in financial, geostatistical fields. The goal in this study was to derive 'validation metrics' for predictions in the application domain, based on those estimated in the validation domain. In this regard, the Bayesian network methodology was found to be promising. The concept was also used for validating system level models where test data may be available only at the component level and Bayesian networks were used to represent the components, subsystems and the full system.

While validation answers the question whether we are solving the right equation, verification on the other hand attempts to answer if we are solving the equation right. Verification, involving quantification and minimization of various errors and uncertainties arising in implementing a computation model for prediction, was implemented. This study focused on finite element models where the mesh size could be a large contributor of numerical error in the prediction. The study explored stochastic response surfaces to estimate uncertainties in the discretization error due to uncertainties in the model inputs. This study proposed a way to assess physical model form errors and reliability analysis errors for use in design under probabilistic constraints. An iterative algorithm was proposed for including model errors and reliability analysis errors in optimization.

6.2 Future work

The validation metrics so far developed in this study deal with continuous random variables only. Future work in this direction includes developing metrics to include discrete variables and/or a combination of discrete and continuous variables for the input

or output. For dynamic response problems, model comparisons may have to be made in the frequency domain and methods need to be developed for characterizing uncertainties in the spectral densities. Use of expensive simulation techniques like Markov Chain Monte Carlo sampling etc., pose computational challenges in model validation. The use of efficient techniques for Bayesian updating, such as saddlepoint approximations introduced in this study, needs to be investigated further for practical applications. Alternative formulations of the extrapolation problem are also expected from a continued research. Most DOE (design of experiments) methods have been developed with the aim of replacing the full model evaluation each time and for uncertainty propagation. Some research is required for developing DOE techniques particularly suited for validation and extrapolation. Also, design of validation experiments should make use of the statistical information available from the computational model.

Validation is quite subjective, in the sense that the formulation of the validation metric depends on the questions we would like to be answered and the decisions one makes at the end of modeling process. If a wrong model is accepted to be valid, the model user will face risk in future applications and design. If we decide to reject a valid model, the model builder has to collect additional data, spend additional resources and time to improve the model. These costs amount to the model developer's risk. A framework for balancing model developer risk vs. code user risk needs to be developed. This framework can further be extended to other stages of a V&V process such as model selection, code verification, design of experiments etc.

REFERENCES

1. Ainsworth, M., and Oden, J. N., "A Unified Approach to aposteriori Error Estimation based on Elemental Residuals," *Numer. Math,* Vol.65, pp: 23-50, 1993.

2. Ainsworth, M., and Oden, J. N., "aposteriori Error Estimation in Finite Element Analysis," *Comput. Methods Appl. Mech. Engg.*, Vol.142, pp: 1-88, 1997.

3. Alvin, K. F., "A Method for Estimating Discretization Error in Non-deterministic Analysis," *AIAA Journal,* Vol. 38 (5), pp: 910-916, 2000.

4. American Institute of Aeronautics and Astronautics, *Guide for the Verification and Validation of Computational Fluid Dynamics Simulations*, AIAA-G-077-1998, Reston, VA, 1998.

5. Ang, A. H-S., and Tang, W. H., *Probability Concepts in Engineering Planning and Design: Volume I: Basic Principles*, John Wiley, New York, 1975.

6. Babuska, I., and Rheinboldt, W. C., "aposteriori Error Estimates for the Finite Element Method," *Int. J. Numer. Methods Engg.*, Vol.12, pp. 1597-1615, 1978.

7. Babuska, I., Strouboulis, T., Upadhpyay, C. S., Gangaraj, S. K., and Copps, K., ''Validation of aposteriori Error Estimators by Numerical Approach,'' *Int. J. Numer. Methods Engg.,* Vol. 37, pp: 1073-1123, 1994.

8. Babuska, I., and Chatzipantelidis, P., "On Solving Elliptic Stochastic Partial Differential Equations," *Comput. Methods Appl. Mech. Engg.*, Vol.191 (37-38), pp: 4093-4122, 2002.

9. Barford, N. C., *Experimental measurements: precision, error, and truth*, Wiley, New York, 1985.

10. Bayarri, M. J., Berger, J. O., Higdon, D., Kennedy, M. C.,Kottas, A., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., Lin, C. H., and Tu, J., "A Framework for Validation of Computer Models," *SIAM Conference on Mathematics for Industry: Challenges and Frontiers*, October 13-15, Toronto, Canada, 2003.

11. Beck, M. B., Mulkey, L. A., and Barnwell, T. O., "Model Validation for Predictive Exposure Assessments," *EPA White paper for Risk Assessment Forum*, 1994.

12. Berger, J. O., and Delampady, M., "Testing Precise Hypotheses," *Statistical Science*, Vol. 2, pp: 317-352, 1987.

13. Berger, J. O., and Pericchi, L. R., "The Intrinsic Bayes Factor for Model Selection and Prediction," *J. Amer. Statist. Assoc*, Vol. 91, pp: 109-122, 1996.

14. Berger, J. O., and Sellke, T., "Testing a Point Null Hypothesis: The Irreconcilability of *p*-values and Evidence," *Journal of American Statistical Association*, Vol. 82 (397), pp: 112-139, 1987.

15. Boehm, B., "Verifying and Validating Software Requirements and Design Specifications," *IEEE Software*, 1984.

16. Box, G. E. P., and Jenkins, G. M., *Time series analysis, forecasting and control*, Holden-day, San Francisco, 1974.

17. Busch, U., and Heimann, D., "Statistical-Dynamical Extrapolation of Nested Regional Climate Simulations," *Climate Research*, Vol. 19, pp: 1-13, 2001.

18. Caria, M., *Measurement analysis: An Introduction to the Statistical Analysis of Laboratory data in Physics, Chemistry and the Life Sciences,* Imperial college press, London, 2000.

19. Castillo, E., Sarabia, J. M., Solares, C., and Gomez, P., "Uncertainty Analysis in Fault Trees and Bayesian Networks using FORM/SORM Methods," *J. of Reliability Engineering and System Safety*, Vol. 65(1), pp: 29-40, 1999.

20. Chen, W., Jin, R., and Sudjianto, A., "Analytical Variance-Based Global Sensitivity Analysis in Simulation-Based Design under Uncertainty," *Proceedings of ASME Design Automation Conference*, *DETC2004-57484*, Salt Lake City, UT, September 28-October 2, 2004.

21. Cohen, J., "The Earth is Round ($p < 0.05$)," *American Psychologist*, Vol. 49, pp: 997-1003, 1994.

22. Coleman, H. W., and Stern, F., "Uncertainties and CFD Code Validation," *J. of Fluids Engg.*, Vol. 119, pp: 795-803, 1997.

23. Cruse, T. A., *Reliability-based Mechanical Design*, Marcel Dekker Inc, New York, 1997.

24. Dahll, G., "Combining Disparate Sources of Information in the Safety Assessment of Software-based Systems," *Nuclear Engg and Design*, Vol. 195, pp: 307-319, 2000.

25. Daniels, H. E., "Tail Probability Approximations," *International Statistical Review*, Vol. 54, pp: 34-48, 1987.

26. Davison, A. C., and Hinkley, D., "Saddlepoint Approximations in Resampling Methods," *Biometrika*, Vol. 78 (3), pp: 417-431, 1988.

27. Defense Modeling and Simulation Office, DoD *Verification, Validation, and Accreditation* (*VV&A*) *Recommended Practices Guide*, Office of the Director of Defense Research and Engr., www.dmso.mil/docslib, Alexandria, VA, April, 1996.

28. Demkowicz, L., Oden, J. T., and Strouboulis, T., "Adaptive Finite Elements for Flow Problems with Moving Boundaries. Part I: Variational Principles and aposteriori Error Estimates," *Comput. Methods Appl. Mech. Engg.*, Vol. 46, pp: 217-251, 1984.

29. Deodatis, G., Popescu, R., and Prevost, J. H., "Simulation of Stochastic Processes and Fields for Monte Carlo Applications: Some Recent Developments," *Proceedings of ASME Design Engineering Technical conference*, Boston, MA, Vol. 3, pp: 955-965, September 17-20, 1995.

30. Der Kiureghian, A., "Analysis of Structural Reliability under Model and Statistical Uncertainties: a Bayesian Approach," *Computational Structural Engineering*, Vol. 1 (2), pp: 81-87, 2001.

31. Devroye, L., and Gyorfi, L., *Nonparametric Density Estimation*, John Wiley & Sons, 1985.

32. Ditlevsen, O. and Arnbjerg-Nielsen, T., "Model Correction Factor Method in Structural Reliability," *Journal of Engineering Mechanics, ASCE*, Vol. 120 (1), pp: 1-10, 1994.

33. Ditlevsen, O., and Johannesen, J.M., "Model Correction Factor Method for System Analysis," *ICASP 8, Sydney December 12-19, 1999, Applications of Statistics and Probability, Civil Engineering Reliability and Risk Analysis* (Eds.: R.E. Melchers and M.G. Stuart), Balkema, Rotterdam, pp: 1011-1018, 2000.

34. Dow, J. O., *A Unified Approach to the Finite Element Method and Error Analysis Procedures*, Academic Press, San Diego, CA, 1999.

35. Dowding, K., Hills, R. G., Leslie, I. H., Pilch, M., Rutherford, B., and Hobbs, M. L., "Case Study for Model Validation: Assessing a Model for Thermal Decomposition of Polyurethane Foam," Report No. SAND2004-3632, Sandia National Laboratories, Albuquerque, NM, 2004.

36. Draper, D., "Assessment and Propagation of Model Uncertainty," *Journal of the Royal Statistical Society Series B*, Vol. 57 (1), pp: 45-97, 1995.

37. Du X., Chen W. "Sequential Optimization and Reliability Assessment Method for Efficient Probabilistic Design", Paper No. DETC2002/DAC-34127, Proceedings of *ASME Design Automation Conference*, Montreal, Canada, 2002.

38. Du. X, Sudjianto, A., and Chen, W., "An Integrated Framework for Optimization using Inverse Reliability Strategy", Paper No. DETC-DAC48706, *ASME Design Automation Conference*, Chicago, IL, 2003.

39. Edwards, G., "A Bayesian Procedure for Drawing Inference from Random Data". *Reliability Engineering*, Vol. 9, pp: 1-17, 1984.

40. Efron, B., "Bootstrap Method: Another Look at the Jackknife," *Ann. Statist*., Vol (7), pp: 1-26, 1979.

41. Efron, B., and Tibshirani, R., *An Introduction to Bootstrap*, No. 57 in Monographs on Statistics and Applied Probability, Chapman & Hall, New York, 1993.

42. Faber, M., and Sørensen, J. D., "Reliability-based Code Calibration-The JCSS Approach," *Proceedings of the 9th International Conference on Applications of Statistics and Probability in Civil Engineering ICASP9,* Vol. 2, pp. 927-935, San Francisco, USA, July 6-9, 2003.

43. Fern, E. F., and Monroe, K. B., "Effect-Size Estimates: Issues and Problems in Interpretation," *Journal of Consumer Research*, Vol. 23 (2), pp: 89-105, 1996.

44. Ferrandiz, J. R., "Bayesian Inference on Mahalanobis Distance: An Alternative Approach to Bayesian Model Testing," In *Bayesian Statistics 2*, Eds: Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M., pp: 645-654, North-Holland, Amsterdam, 1985.

45. Franchin, P., Ditlevsen, O., and Der Kiureghian, A., "Model Correction Factor Method for Reliability Problems Involving Integrals of Non-Gaussian Random Fields," *Probabilistic Engineering Mechanics*, Vol. 17 (2), pp: 109-122, 2002.

46. Friis-Hansen, A., Friis-Hansen, P., and Christensen, C. F., "Reliability Analysis of Upheaval Buckling-updating and Cost Optimization," *Proceedings of 8th ASCE Specialty Conference on Probabilistic Mechanics and Structural Reliability*, Notre Dame, 2000.

47. Ghanem, R. G., and Spanos, P. D., *Stochastic Finite Elements: A Spectral Approach*, Springer-Verlag: New York, 1991.

48. Gilks, W. R., and Wild, P., "Adaptive Rejection Sampling for Gibbs Sampling," *Applied Statistics*, Vol. 41 (2), pp: 337-348, 1992.

49. Gilks, G. R., Richardson, S., and Spiegelhalter, D. J., *Markov Chain Monte Carlo in Practice*, Interdisciplinary Statistics, Chapman & Hall/CRC, London, 1996.

50. Good, P. I., *Resampling Methods: A Practical Guide to Data Analysis*, Springer-Verlag, New York, March 1999.

51. Guedes Soares, C., "Bayesian Prediction of Design Wave Height," In: Thoft-Christensen P, editor. *Reliability and Optimization of Structural Systems '88*, Springer-Verlag, pp: 311-323, 1988.

52. Haldar, A., and Mahadevan, S., *Probability, Reliability and Statistical Methods in Engineering Design*, John Wiley & Sons, New York, 2000.

53. Hasofer, A. M., and Lind, N. C., "An Exact and Invariant First Order Reliability Format," *J. Eng. Mech., ASCE*, Vol. 100, EM1, pp: 111-121, 1974.

54. Hasselman, T. K., and Wathugala, W. G., "A Hierarchical Approach for Model Validation and Uncertainty Quantification," *Proceedings of Fifth World Congress on Computational Mechanics* (*WCCM V*), Vienna, Austria, 2002.

55. Heckerman, D., Geiger, D., and Chickering, D., "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data," *Proceedings of Tenth Conference on Uncertainty in Artificial Intelligence*, Seattle, WA, pp: 293 – 301, 1994.

56. Higdon, D., Kennedy, M. C., Cavendish, J., Cafeo J., and Ryne, R. D., "Combining Field Data and Computer Simulations for Calibration and Prediction," *SIAM Journal on Scientific Computing*, Vol. 26, pp: 448-466, 2004.

57. Hills, R. G., and T. G. Trucano, "Statistical Validation of Engineering and Scientific Models: A Maximum Likelihood Based Metric," Technical Report No. SAND2001-1783, Sandia National Laboratories, Albuquerque, NM, 2001.

58. Hills, R. G., and Leslie, I. H., "Statistical Validation of Engineering and Scientific Models: Validation Experiments to Application," Technical Report No. SAND2003-0706, Sandia National Laboratories, Albuquerque, NM, 2003.

59. Hobbs, M. L., Erickson, K. L., and Tze, C. Y., "Modeling Decomposition of Unconfined Rigid Polyurethane Foam," SAND99-2758, Sandia National Laboratories, Albuquerque, NM, 1999.

60. Hoffman, R. M., Sudjianto, A., Du, X., and Stout J., "Robust Piston Design and Optimization Using Piston Secondary Motion Analysis," *Proceedings of SAE 2003*, Detroit, USA, March, 2003.

61. Hohenbichler, M., Gollwitzer, S., Kruse, W., and Rackwitz, R., "New Light on First- and Second-Order Reliability Methods," *Structural Safety*, Vol. 4, pp: 267-284, 1987.

62. Iman, R. L., and Conover, W. J. "A Distribution Free Approach to Introducing Rank Correlation among Input Variables," *Communication on statistics- simulation and computation* Vol. 11(3), pp: 311-334, 1982.

63. Isukapalli, S. S., and Georgopoulos, P. G., "Computational Methods for the Efficient Sensitivity and Uncertainty Analysis of Models for Environmental and Biological Systems," Technical Report, State Univ. of New Jersey, 1999.

64. Jeffreys, H. J., *Theory of Probability*, Third Edition, Oxford: Clarendon Press, 1961.

65. Jensen, J. L., *Saddlepoint Approximations*, Oxford Science Publications, New York, 1995.

66. Jensen, F. V., *Bayesian Networks and Decision Graphs*, Springer-Verlag, New York, 2001.

67. Johnson, D. H., "Insignificance of Statistical Significance Testing," *Journal of Wildlife Management*, Vol. 63, pp: 763-772, 1999.

68. Kurowski, P., and Szabo, B., "How to Find Errors in the Finite Element Models," *Machine Design*, September 25 issue, 1997.

69. Kirk, R., "Practical Significance: A Concept Whose Time has Come," *Educational and Psychological Measurement*, Vol. 56(5), pp: 746-759, 1996.

70. Kurowski, P., "Easily Made Errors Mar FEA Results," *Machine Design*, September 13 issue, 2001.

71. Lawley, D. N., "Tests of Significance for the Latent Roots of Covariance and Correlation Matrices," *Biometrika*, Vol. 43(1/2), pp: 128-136, 1956.

72. Leonard, T., and Hsu, J. S. J., *Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers*, Cambridge University Press, Cambridge 1999.

73. Livingston, R. J., Diaz, R. J., and White, D. C., "Field Validation of Laboratory-Derived Multispecies Aquatic Test Systems," EPA/600/4-85/039, U.S. Environmental Protection Agency, Environmental Research Laboratory, Gulf Breeze, FL, 56p, 1985.

74. Mahadevan, S., "Design Optimization for Reliability and Robustness," *Proceedings of SAE World Congress 2004,* Detroit, Michigan, USA, March 8-11, 2004.

75. Mahadevan, S., and Rebba, R., "Computational Model Validation under Uncertainty," *Proceedings of ICASP9*, San Francisco, California, USA, July 6-9, 2003.

76. Mahadevan, S., and Rebba, R., "Validation of Reliability Computational Models using Bayes Networks," *J. of Reliability Engineering and System Safety*, Vol. 87 (2), pp. 223-232, 2005.

77. Mahadevan, S., Zhang, R., and Smith, N., "Bayesian Networks for System Reliability Reassessment," *Structural Safety*, Vol. 23, pp. 231-251, 2001.

78. McKay, M. D., Beckman, R. J., and Conover, W. J., "A Comparison of Three Methods for Selecting Values of Input Variables in Analysis of Output from a Computer Code," *Technometrices* Vol. 2, pp: 239-245, 1979.

79. Mehta, U. B., "*Guide to Credible Computer Simulations of Fluid Flows*," *AIAA Journal of Propulsion and Power*, Vol. 12 (5), pp: 940-948, 1996.

80. Myers, R. H., and Montgomery, D. C., *Response Surface Methodology*, John Wiley and Sons, New York, 1995.

81. Nataf, A., "D´etermination des distributions de probalit´es dont les marges sont donn´ees," *Comptes Rendus de l'Academie des Sciences*, Vol. 225, pp: 42–43, 1962.

82. Oberkampf, W. L., and Barone, M. F., "Measures of Agreement between Computation and Experiment: Validation Metrics," *Proceedings of AIAA 34thFluid Dynamics Conferences,* Portland, Oregon, USA, June 28-July 1, 2004.

83. Oberkampf, W. L., Trucano, T. G., and Hirsch, C., "Verification, Validation and Predictive Capability in Computational Engineering and Physics," *Proceedings of*

Foundations '02, Workshop on V&V, Johns Hopkins University, Maryland, October 22-23, 2002.

84. Oberkamf W.L., and Trucano, T. G., "Verification and Validation in Computational Fluid Dynamics," *Progress in Aerospace Sciences*, Vol. 38 (3), pp: 209-272, 2002.

85. Onatski, A., and Williams, N., "Modeling Model Uncertainty," *Journal of European Economic Association*, Vol. 1 (5), pp: 1087-1122, 2003.

86. Pawitan Y. *In All Likelihood: Statistical Modeling and Inference using Likelihood*. New York; Oxford Science Publications; 2001.

87. Pilch, M. M., and Trucano, T. G., "Validation Metrics," Report No. SAND2001-1411P, Sandia National Laboratories, Albuquerque, NM, 2001.

88. Pontius Jr, R. G., and Batchu, K., "Using the Relative Operating Characteristic to Quantify Certainty in Prediction of Location of Land Cover Change in India," *Transactions in GIS*, Vol. 7(4), pp: 467-484, 2003.

89. Pontius Jr, R. G., Agrawal, A., and Huffaker, D., "Estimating the Uncertainty of Land-Cover Extrapolations while Constructing a Raster Map from Tabular Data," *J. of Geographical Systems*, Vol. 5(3), pp: 253-273, 2003.

90. Radhakrishnan, R., and McAdams, D. A., "A Methodology for Model Selection in Engineering Design," *ASME Journal of Mechanical Design*, Vol. 127 (3), pp: 378-387, 2005.

91. Rao, S. S., "Design with Uncertain Parameters and Stochastic Processes," *AIAA Journal*, Vol. 22, pp: 1670-1678, 1984.

92. Rebba, R., "Computational Model Validation under Uncertainty," Master's thesis, Department of Civil and Environmental Engineering, Vanderbilt University, Nashville, TN, USA 2002.

93. Rebba, R., and Mahadevan, S., "Verification and Validation of Simulation Models under Uncertainty," *Proceedings of USNCCM7*, Albuquerque, New Mexico, USA, July 27-21, 2003.

94. Rebba, R., Mahadevan, S., and Huang, S., "Validation and Error Estimation of Computational Models," *Proceedings of Fourth International Conference on Sensitivity Analysis of Model Output* (*SAMO*), Santa Fe, NM, March, 2004.

95. Reid, J. D., "Admissible Modeling Errors or Modeling Simplifications," *Finite Element Analysis and Design*, Vol. 29 (1), pp: 49-63, 1998.

96. Richards, S. A., "Completed Richardson Extrapolation in Space and Time," *Comm. Numer. Methods Engg*, Vol. 13, pp: 573-58, 1997.

97. Roache, P. J., *Verification and Validation in Computational Science and Engineering*, Hermosa publishers, Albuquerque, NM, 2002.

98. Robinson, A., P., and Froese, R. E., "Model Validation using Equivalence Tests," *Ecological Modeling*, Vol. 176 (3-4), pp: 349-358, 2004.

99. Rosenblatt, M., "Remarks on Multivariate Transformation," *Ann. Math. Stat*, Vol. 23(3), pp: 470-472, 1952.

100. Royset, J. O., Der Kiureghian, A. and Polak, E., "Reliability-based Optimal Structural Design by the Decoupling Approach," *Reliability Engineering & System Safety*, Vol. 73, pp: 213-221, 2001.

101. Scott, D., W., *Multivariate Density Estimation: Theory, practice, and Visualization*, John Wiley & Sons, New York, 1992.

102. Schervish, M. J., *Theory of Statistics*, Springer-Verlag, New York, 1995.

103. Schueller, G. I., Bucher, C. G. Bourgund, U., and Ouypornprasert, W., "On Efficient Computational Schemes to Calculate Failure Probabilities", *Probabilistic engineering mechanics*, Vol. 4(1), pp: 10-18, 1989.

104. Segalman, D. J., Paez, T. L., Smallwood, D. O, Sumali, A. H., and Urbina, A., "Status and Integrated Road-map for Joints Modeling Research," Report No. SAND2003-0897, Sandia National Laboratories, Albuquerque, NM, 2003.

105. Silverman, B. W., and Young, G. A., "The Bootstraps: To Smooth or not to Smooth," *Biometrika*, Vol. 74, pp: 469 – 479, 1987.

106. Smallwood, D., Gregory, D., and Coleman, R., "A Three-parameter Constitutive Model for a Joint which Exhibits a Power Law Relationship between Energy Loss and Relative Displacement," *Proceedings of the 72nd Shock and Vibration Symposium*, Destin, Florida November 12-16, 2001.

107. Smith, N. L., and Mahadevan, S., "Integrating System-level and Component-level Designs under Uncertainty," *Journal of Spacecrafts and Rockets*, Vol. 42 (4), pp: 752-760, 2005.

108. Spearman, C., "The Proof and Measurement of Association between Two Things," *American Journal of Psychology*, Vol. 15 (1), pp: 72-101, 1904.

109. Spiegelhalter, D. J., Thomas, A., and Best, N. G., *WinBUGS Version 1.4 User Manual*, Cambridge, UK, MRC Biostatistics Unit; 2002.

110. Srivastava, M. S., and Hui, T. K., "On Assessing Multivariate Normality based on Shapiro-Wilk W Statistics," *Statistics and Probability*, Vol. 12, pp: 61-72, 1987.

111. Srivastava, M. S., *Methods of Multivariate Statistics*, John Wiley, New York 2002.

112. Stoebner, A. M., and Mahadevan, S., "Robustness in Reliability-based Design," *Paper 2000-1508, AIAA/ASME/ASCE/AHS/ASC Structures, 41st SDM Conference and Exhibit*, Atlanta, GA, April 3-6, 2000.

113. Tapia, R. A., *Non-parametric Probability Density Estimation*, Johns Hopkins University Press, Baltimore 1978.

114. Tatang, M. A., Pan, W. W., Prinn, R. G., and McRae, G. J., "An Efficient Method for Parametric Uncertainty Analysis of Numerical Geophysical Model" *Journal of Geophysical Research-Atmospheres*, Vol. 102(D18), pp: 21925-21932, 1997.

115. Thacker, B. H., and Huyse, L. J., "Role of Non-Determinism in Validation of Computational Mechanics Models," *Proceedings of the Fifth World Congress on Computational Mechanics* (*WCCM V*), Vienna, Austria, July 2002.

116. Thompson, B., "Why Multivariate Methods are Usually Vital in Research: Some Basic Concepts," *Biennial meeting of the Southwestern Society for Research in Human Development*, Austin, TX, (ERIC Document Reproduction Service No. ED 367 678), February, 1994a.

117. Tierney, L., and Kadane, J. B., "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of American Statistical Association*, Vol. 81 (393), pp: 82-86, 1986.

118. Tierney, L., Kass, R. E., and Kadane, J. B., "Approximate Marginal Densities of Nonlinear Functions," *Biometrika*, Vol. 76 (3), pp: 425-433, 1989; Corrections: Vol. 78 (1), pp: 233-234, 1991.

119. Timoshenko, S. P., and Goodier, J. N., *Theory of Elasticity,* 2nd Ed. New York, McGraw-Hill, 1951.

120. Trucano, T. G., Easterling, R. G., Dowding, K. J., Paez, T. L., Urbina A., Romero, V. J., Rutherford, B. M., and Hills, R. G., "Description of the Sandia Validation Metrics Project," Technical Report No. SAND2001-1339, Sandia National Laboratories, Albuquerque, NM, 2001.

121. Urbina, A., and Paez, T. L., "Statistical Validation of Structural Dynamics models," *Annual Technical Meeting & Exposition of the Institute of Environmental Sciences and Technology,* Phoenix, AZ, 2001.

122. Urbina, A., Paez, T. L., Hasselman, T. K., Wathugala, G. W., Yap, K., "Assessment of Model Accuracy Relative to Stochastic System Behavior," *Proceedings of 44th AIAA Structures, Structural Dynamics, Materials Conference*, April 7-10, Norfolk, VA, 2003.

123. Volinsky, C. T., Madigan D., Raftery, A. E., and Kronmal, R. A., "Bayesian Model Averaging in Proportional Hazard Models: Assessing the Risk of a Stroke", *Applied Statistics*, Vol. 46(4), pp: 433-448, 1997.

124. Wellek, S., *Testing Statistical Hypotheses of Equivalence*, Chapman & Hall, 2002.

125. Wentzell, P. D., Andrews, D. T., Hamilton, D. C., Faber, K., and Kowalski, B. R., "Maximum Likelihood Principal Component Analysis," *J. Chemomet*., Vol. 11, pp: 339-366, 1997.

126. Wilks, D. S., *Statistical Methods in Atmospheric Sciences: An Introduction,* Academic Press, 1995.

127. Williams, W. H., and Goodman, M. L., "A Simple Method for the Construction of Empirical Confidence Limits for Economic Forecasts," *Journal of the American Statistical Association,* Vol.66 (336), pp: 752-754, 1971.

128. Yamazaki, F., Shinozuka, M., and Dasgupta, G., "Neumann Expansion for Stochastic Finite Element Analysis," *J. Engg. Mech, ASCE*, Vol. 114 (8), pp: 1335-1354, 1988.

129. Zhang, R., and Mahadevan, S., "Model Uncertainty and Bayesian Updating in Reliability-based Inspection," *J. of Structural Safety*, Vol. 22 (2), pp: 145-160, 2000.

130. Young, G. A., and Daniels, H. E., "Bootstrap Bias," *Biometrika*, Vol. 77 (1), pp: 179-185, 1990.

131. Zhang, R., and Mahadevan, S., "Bayesian Methodology for Reliability Model Acceptance," *Reliability Engineering & System Safety*, Vol. 80 (1), pp: 95-103, 2003.

132. Zienkiewicz, O. C., and Zhu, J. Z., ''A Simple Error Estimator and Adaptive Procedure for Practical Engineering Analysis,'' *Int. J. Numer. Methods Engg.*, Vol. 24, pp: 337-357, 1987.

133. Zienkiewicz, O. C., and Zhu, J. Z., "The Super-convergent Patch Recovery and aposteriori Estimates. Part 2: Error Estimates and Adaptivity," *Int. J. Numer. Methods Engg.,* Vol.33, pp: 1335-1382, 1992.

134. Zou, T., "Efficient Methods for Reliability-based design Optimization," Doctoral Dissertation, Department of Civil Engineering, Vanderbilt University, Nashville, USA, 2004.