Available online at www.sciencedirect.com
ScienceDirect
Reliability Engineering and System Safety 93 (2008) 1197–1207
RELIABILITY ENGINEERING & SYSTEM SAFETY
www.elsevier.com/locate/ress

Computational methods for model reliability assessment

Ramesh Rebba[a,*], Sankaran Mahadevan[b]

[a]General Motors R&D, Warren, MI 48090, USA
[b]Vanderbilt University, Nashville, TN 37235, USA

Received 29 June 2006; received in revised form 11 June 2007; accepted 5 August 2007
Available online 10 August 2007

Abstract

This paper investigates various statistical approaches for the validation of computational models when both model prediction and experimental observation have uncertainties, and proposes two new methods for this purpose. The first method utilizes hypothesis testing to accept or reject a model at a desired significance level. Interval-based hypothesis testing is found to be more practically useful for model validation than the commonly used point null hypothesis testing. Both classical and Bayesian approaches are investigated. The second and more direct method formulates model validation as a limit state-based reliability estimation problem. Both simulation-based and analytical methods are presented to compute the model reliability for single or multiple comparisons of the model output and observed data. The proposed methods are illustrated and compared using numerical examples.

Keywords: Hypothesis testing; Model validation; *p*-value; Reliability; *t*-test

1. Motivation

Various types of uncertainties and errors are involved 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 variability 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 or more uncertain quantities.

Validation assessments can be made using qualitative graphical methods or quantitative statistical techniques; the focus of this paper is on the latter. Depending on the nature or form of model output and experimental data, model validation may involve comparison of means or variances or even two probability distributions. A validation metric would then provide a quantitative assessment of the agreement between prediction and observation [1]. 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 support whether the degree of confidence with which we accept or reject a model is adequate for the intended model use. This paper investigates the ability of several validation metrics to address both accuracy and adequacy requirements for engineering applications.

Various types of quantitative 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 [2]. Sargent [3] and Balci [4] outlined the general framework for model verification, validation and accreditation by defining various terminologies. Coleman and Stern [5] combined various types of errors and uncertainties arising in computational fluid dynamics 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 were assumed to be independent of

*Corresponding author. Tel.: +1 586 986 1654.
E-mail address: ramesh.rebba@gm.com (R. Rebba).

0951-8320/$ - see front matter
doi:10.1016/j.ress.2007.08.001

Nomenclature

A^2	Anderson Darling statistic
α	significance level
B	Bayes factor
C	posterior probability of null hypothesis
c	reliability requirement
D	dissipation energy
ε	accuracy requirement
H_0	null hypothesis
H_a	alternative hypothesis
LN	lognormal distribution
N	normal distribution
n	sample size
r	model reliability
s^2	sample variance

each other and were combined linearly. The term uncertainty has been used synonymously with standard deviation in that paper. Thus, the total uncertainty in the comparison error or the standard deviation of E is estimated as

$$\sigma_E = \sqrt{\sigma^2_{\mathrm{SPD}} + \sigma^2_{\mathrm{SMA}} + \sigma^2_{\mathrm{SN}} + \sigma^2_{\mathrm{D}}}. \tag{1}$$

The model prediction is said to be inadequate if $|E| < \sigma_E$. The confidence with which we accept or reject a model prediction is not reported with this metric.

Since the metric proposed in Eq. (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 $\mathbf{x}_{\mathrm{model}}$ and $\mathbf{x}_{\mathrm{exp}}$, a validation metric based on the Mahalanobis distance was proposed [6]:

$$d^2 = (\mathbf{x}_{\mathrm{model}} - \mathbf{x}_{\mathrm{exp}})^{\mathrm{T}}(\mathrm{cov}(\mathbf{x}_{\mathrm{model}}) + \mathrm{cov}(\mathbf{x}_{\mathrm{exp}}))^{-1}(\mathbf{x}_{\mathrm{model}} - \mathbf{x}_{\mathrm{exp}}). \tag{2}$$

The model prediction is said to be close to the data when d^2 is less than some critical value $\chi_\alpha{}^2(n)$, where n is the number of data points or predictions and α is the significance level. The corresponding p-value is computed as $P(d^2 > d^2_{\mathrm{obs}})$. The meaning and interpretation of p-value will be discussed later in Section 2. For the distance metric in Eq. (2) to follow a χ^2 distribution, both the model output and field response have to follow normal distribution. When model and field responses are not normal, they both can be transformed into normal space.

Classical hypothesis testing methods have been explored for model validation assessment in recent literatures [7,8]. Zhang and Mahadevan [9] 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 $B = (n+1)(n!/(n-k)!k!)p^k(1-p)^{n-k}$. If $B > 1.0$, we conclude that the data favor the model prediction. Recently the method has been extended to the validation of more generalized model outputs, both univariate and multivariate [10]. The validation metric in that case is the ratio of posterior to prior densities of the model prediction. Other Bayesian approaches for model validation have focused more on calibration of the model using the data and providing posterior probability intervals rather than a direct assessment of the degree of match between prediction and observation [11,12].

Some alternatives to p-values and Bayes factors are also available in the literature. One approach is the use of decision-theoretic utility or loss functions [13]. Denoting d_0 as the decision to accept the null hypothesis that model prediction and data are equal and d_1 as the decision to accept the alternative, one can define the utility function $u(d_i, \theta)$ of choosing d_i when θ is the parameter we wish to validate. Using the Bayesian approach, having observed the data x, the decision d_1 is the optimal decision if and only if $E[u(d_1, \theta) - u(d_0, \theta)|x] > 0$. The difference in the utility functions is usually chosen as a squared loss function or an absolute error metric [14]. Recently, Jiang and Mahadevan [15] have developed a Bayesian decision-making methodology for computational model validation, considering the risk of using the current model, data support for the current model, and cost of acquiring new information to improve the model.

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 [16,17] 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. Spearman's rank correlation [18] is sometimes used as a correlation indicator between data and prediction.

This paper explores two statistical methods for model validation: (a) hypothesis testing and (b) limit state-based reliability analysis. Section 2 discusses point-null testing using classical and Bayesian statistics. The concept of p-value is discussed in detail for its use as a validation metric. In Section 3, a more practical interval-based hypothesis testing is proposed as an alternative to point null testing. In Section 4, a more direct approach that formulates model validation as a reliability analysis problem is proposed. The probability that the difference between model prediction and data is within a given threshold is calculated for univariate and multivariate comparisons. The proposed methodologies are illustrated and compared through numerical examples.