

An Ensemble Approach for Model Bias Prediction

Zhimin Xi
Univ. of Michigan

Yan Fu and Ren-Jye Yang
Ford Motor Co

ABSTRACT

Model validation is a 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. In reliability based design, the intended use of the model is to identify an optimal design with the minimum cost function while satisfying all reliability constraints. It is pivotal that computational models should be validated before conducting the reliability based design. This paper presents an ensemble approach for model bias prediction in order to correct predictions of computational models. The basic idea is to first characterize the model bias of computational models, then correct the model prediction by adding the characterized model bias. The ensemble approach is composed of two prediction mechanisms: 1) response surface of model bias, and 2) Copula modeling of a series of relationships between design variables and the model bias, between model prediction and the model bias. Advantages of both mechanisms are utilized with an accuracy based weighting approach. A vehicle design of front impact example is used to demonstrate the effectiveness of the proposed methodology.

Keywords: model validation, reliability based design, model bias prediction, Copula

CITATION: Xi, Z., Fu, Y. and Yang, R., "An Ensemble Approach for Model Bias Prediction," *SAE Int. J. Mater. Manf.* 6(3):2013, doi:10.4271/2013-01-1387.

1. INTRODUCTION

Computational models are extensively used in design of engineering systems to analyze, predict, and optimize the performance of real physical systems. When models are accurate and reliable, demand to conduct time consuming and costly physical experiments can be reduced, and product/system development time and cost can be significantly decreased. However, models are generally built on the basis of many assumptions and simplifications and therefore model validation becomes a necessary 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 [1,2,3]. Unlike deterministic validation, development of statistically validated models is very difficult because of the uncertainty from model random variables, model bias, and uncontrollable random parameters in experiments.

In reliability based design, the intended use of the model is to identify an optimal design with the minimum cost function while satisfying all reliability constraints. It is pivotal that computational models should be validated before conducting the reliability based design. However, validation

of computational models in the design space is challenging because only limited physical experiments may be available at a few design configurations due to the lack of resources. One approach is to characterize the model bias using a regression model at a specified design configuration and assume this regression model is also applicable to other design configurations [4]. Hence, original model prediction can be corrected by adding the characterized model bias. This approach may produce inaccurate model prediction because it is tricky to correctly identify coefficients of the regression model based on only one design configuration. Alternatively, a design-driven validation approach [5] was proposed to identify an optimal design when model is inaccurate for predicting real system performances. However, model accuracy is only locally improved by requiring more physical experiments in a small region near the 'optimal' design. In this approach, model is validated near the 'optimal' design instead of the design space. Therefore, the identified 'optimal' design may not be the true optimum because the design optimization is conducted in an invalidated design space. Recently, a sequential model validation approach [6] was proposed to increase confidence of the optimal design.



Documents

Document Sharing Hub

Model validation is first conducted at the baseline design; followed by design optimization in a small validated domain to find next design configuration. Model validation is again executed in the identified design configuration, resulting in a new small validated domain. This process is sequentially carried out till the convergence of the design. This approach is computationally expensive because design optimization and model validation are coupled together.

The objective of this paper is to develop an ensemble approach for model bias prediction that is able to improve the model prediction accuracy in the design space under uncertainty. In particular, the model prediction accuracy stands for the mean and standard deviation prediction of the performance of interest. The basic idea is the same as previous approaches [4,5,6] by correcting the model prediction through adding the characterized model bias. The contribution of this paper is to propose an ensemble approach for model bias prediction which is composed of two prediction mechanisms: 1) response surface of the model bias, and 2) Copula modeling of a series of relationships between design variables and the model bias, between model prediction and the model bias. Advantages of both mechanisms are utilized with an accuracy based weighting approach. Rest of the paper is organized as following. Section 2 reviews a model bias prediction mechanism using response surface. Section 3 expounds Copula modeling for model bias prediction. Section 4 proposes a fusion system to fuse the two prediction mechanisms. Section 5 demonstrates the proposed methodology using a vehicle design of front impact example. Finally, conclusion is made in Section 6.

2. REVIEW OF MODEL BIAS PREDICTION USING RESPONSE SURFACE

A general relationship between the model prediction and the test data can be expressed as [7]

$$\hat{Y}(\mathbf{P}, \mathbf{X}, \phi) + \delta = Y - \varepsilon \quad (1)$$

where $\hat{Y}$ and Y are the predicted and the measured system performances, respectively, δ is the model bias, ε is the measurement error, $\mathbf{P}$ is a vector of the deterministic model variable, $\mathbf{X}$ and ϕ are the vectors of the irreducible and reducible model random variables, respectively. The irreducible random variables ($\mathbf{X}$) are characterized using probability density functions (PDFs) with sufficient information. The reducible random variables (ϕ) are derived from the lack of information for describing the uncertainty. For example, parameters, i.e., the mean and the variance of the PDF, or even distribution types are uncertain unless sufficient information is collected. The measurement error ε is mainly affected by the equipment accuracy and human errors. Eq. (1) can be rewritten as Eq. (2) if we ignore the measurement error.

$$\hat{Y}(\mathbf{P}, \mathbf{X}, \phi) + \delta = Y \quad (2)$$

In study of model validation for reliability based design, the objective is to improve the model prediction accuracy in the design space. A two-step calibration procedure was proposed to accurately characterize the model bias in the design space [8].

Step 1. calibrate unknown model variables ϕ through separate calibration experiments;

Step 2. calibrate the model bias using the statistical calibration technique at each design configuration.

Then four steps are used to obtain the model bias in the design space with the aid of the response surface.

Step 1. identify optimal distributions of the model bias using the two-step calibration procedure under several design configurations;

Step 2. construct response surfaces for the central moments (e.g., mean and standard deviation) of the model bias;

Step 3. calculate the central moments of the model bias at new design configurations on the basis of the response surfaces;

Step 4. approximate the distributions of the model bias at new design configurations.

Response surface of the model bias plays a critical role in the process of obtaining corrected model prediction. Its accuracy mainly depends on three factors including: 1) nonlinearity of the model bias in the design space; 2) amount and location of the identified model bias in the design space; and 3) algorithm of the response surface method.

3. COPULA MODELING FOR MODEL BIAS PREDICTION

3.1. Review of Copula

Copula is a general way in statistics to formulate a multivariate distribution with various types of statistical dependence. A Copula is a joint distribution function of standard uniform random variables. According to Sklar's theorem [9], there exists an n -dimensional Copula C such that for all y in a real random space

$$F(y_1, \dots, y_n) = C(F_1(y_1), \dots, F_n(y_n)) \quad (3)$$

where F is an n -dimensional distribution function with marginal functions $F_1, \dots, F_n$. Most Copulas deal with bivariate data due to the lack of practical n -dimensional generalization of the coupling parameter [10-11]. For multivariate data, the usual approach is to analyze the data pair-by-pair using two-dimensional Copulas.

3.2. Copula Modeling

The most commonly employed methods to select the best Copula are based on a maximum likelihood approach [12,13,14], which relies on the estimation of an optimal parameter set. In other words, comparisons are made among Copulas with given optimal parameters. The Copula selection on the basis of the maximum likelihood approach typically demands sufficient bivariate data to ensure the accurate Copula selection. Hence, such an approach may not be suitable in this study because there is typically a lack of validation experiments for charactering the model bias at a few design configurations.

Recently a Bayesian Copula approach [10] was proposed to select the best Copula. It was further shown that this approach provides more reliable identification of true Copulas even with the small amount of samples because, unlike the maximum likelihood approach, the selection of the best Copula is independent of the Copula parameter estimation [10]. Hence, we employ the Bayesian Copula approach for modeling statistical dependence between design variables and the model bias, between model prediction and the model bias, where the design variable indicates the controllable model variable (e.g. $\mathbf{P}$, $\mathbf{X}$ and $(\phi$ in Eq.(2)). Determination of the best Copula is done in two steps: 1) selection of optimal marginal distributions; and 2) determination of an optimal Copula. The marginal distribution is determined using the maximum likelihood approach. Next, we focus on the second step for determining the optimal Copula using the Bayesian approach.

In the Bayesian Copula approach, a set of hypotheses are first made as follows.

H_k : The data come from Copula C_k $k = 1, \dots, Q$

The objective is to find the Copula with the highest $\Pr(H_k | D)$ from a finite set of Copulas, where D represents bivariate data in a standard uniform space. Based on the Bayes' theorem, the probability that data come from the Copula C_k is expressed as

$$\Pr(H_k | D) = \frac{\Pr(D | H_k) \Pr(H_k)}{\Pr(D)} = \frac{\int_{-1}^1 \Pr(D | H_k, \tau) \Pr(H_k | \tau) \Pr(\tau) d\tau}{\Pr(D)} \quad (4)$$

where τ is the Kendall's tau, which is a non-parametric measure of the statistical dependence associated to Copulas. Kendall's tau (τ) belongs to the set of each Copula and the outcome is equally likely. All Copulas are equally probable with respect to a given τ which reflects no preference over the Copulas. The likelihood $\Pr(D | H_k, \tau)$ depends upon τ and can be calculated from the Copula PDF as

$$\Pr(D | H_k, \tau) = \prod_{l=1}^m c_k(u_{1l}, u_{2l} | \tau) \quad (5)$$

where $c_k(\cdot)$ is the PDF of the k^{th} Copula; m is the total number of coupling data; u_{1l} and u_{2l} are the l^{th} realizations of

the statistically dependent bivariate variables. The normalization of $\Pr(D)$ can be computed using the sum rule [15]. For more details of the Bayesian Copula approach, please refer to the reference document [10].

3.3. Model Bias Prediction Using Copulas

According to the Copula modeling, we are able to predict the possible model bias for the new design configuration. For example, Copula modeling between $\hat{Y}$ and δ allows us to find the possible model bias δ for a realization of $\hat{Y}$ (e.g. $\hat{Y}=a$) at a new design configuration. Mathematically, this is a process to identify the conditional PDF of the model bias δ given $\hat{Y}=a$ as shown in Eq. (6).

$$C(F_{\hat{Y}}(\hat{y}), F_{\delta}(\delta)) | \hat{y}=a \quad (6)$$

Meanwhile, we also know the design variable (e.g., $x_1=a_1, x_2=a_2, \dots, x_j=a_j$) at the new design configuration. Thus, the possible realizations of the model bias δ must simultaneously satisfy all the conditional PDFs identified from a series of Copula models. In other words, we need to find the intersection of these conditional PDFs for the model bias. Theoretically, it is very difficult. Therefore, a sampling approach is proposed for the model bias prediction. Four steps are conducted to have the empirical PDF of the model bias.

Step 1. generate model bias samples (e.g. $\delta_{N,1}, \delta_{N,2}, \dots, \delta_{N,j}$) from conditional PDFs obtained in Eq. (6), where N is the number of random samples;

Step 2. identify lower (δ_{lower}) and upper (δ_{upper}) bounds from the model bias samples, where

$$\begin{aligned} \delta_{\text{lower}} &= \max [\min(\delta_{N,1}), \min(\delta_{N,2}), \dots, \min(\delta_{N,j})] \\ \delta_{\text{upper}} &= \min [\max(\delta_{N,1}), \max(\delta_{N,2}), \dots, \max(\delta_{N,j})] \end{aligned}$$

Step 3. keep model bias samples which are located between the lower and upper bounds;

Step 4. obtain an empirical PDF of the model bias from samples in step 3.

4. FUSION OF MODEL BIAS PREDICTION

Combining approximation algorithms into an ensemble has already found its applications in a wide variety of research fields, such as the development of committees of neural networks [16-17], the meta modeling for the design of modern engineering systems [18,19,20], the detection of traffic incidents [21], the quantification of model uncertainty [22], etc. This section proposes a fusion approach for model bias prediction using both the Copula and response surface approaches. In particular, this study employs a k-fold cross validation (CV) approach to estimate the accuracy of each