

An Adaptive Copula-Based Approach for Model Bias Characterization

Hao Pan

Zhimin Xi
University of Michigan

Ren-Jye Yang
Ford Motor Co

ABSTRACT

A copula-based approach for model bias characterization was previously proposed [18] aiming at improving prediction accuracy compared to other model characterization approaches such as regression and Gaussian Process. This paper proposes an adaptive copula-based approach for model bias identification to enhance the available methodology. The main idea is to use cluster analysis to preprocess data, then apply the copula-based approach using information from each cluster. The final prediction accumulates predictions obtained from each cluster. Two case studies will be used to demonstrate the superiority of the adaptive copula-based approach over its predecessor.

CITATION: Pan, H., Xi, Z., and Yang, R., "An Adaptive Copula-Based Approach for Model Bias Characterization," *SAE Int. J. Mater. Manf.* 8(2):2015, doi:10.4271/2015-01-0455.

1. INTRODUCTION

Computer models that simulate various engineering problems can aid virtual prototyping, reduce product development cycle and cut down the cost of performing the physical tests. However, one may ask the following question: is the model good? Since computer models are, at their best, approximations of real world phenomena based on various assumptions, numerous model validation methodologies have been developed to formally assess the goodness of the computer models. A common definition of model validation is that it is the process of determining the degree to which a computer model is an accurate representation of the real world phenomenon from the perspective of the intended uses of the model [1,2,3]. In essence, this process is to characterize the difference between the computer model results and the true process outputs. The cause of such difference can often come from two sources, observation error and model bias, which are defined commonly in [4-5]. This paper will focus on model bias characterization.

There are various approaches on characterizing model bias in the model validation community [6,7,8,9]. 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 [6]. 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 [7] was proposed to identify an optimal design

when the 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, but 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 [8] was proposed to increase confidence of the optimal design. 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.

In [18], a model bias characterization technique utilizing copula was proposed. It was demonstrated that the proposed approach made significant improvements in model prediction accuracy. This paper further strengthens this approach by incorporating cluster analysis to the existing framework. Data are first preprocessed by applying cluster analysis to group data points based on their similarity. Then copula-based model bias characterization will be applied to each group (or cluster) and model prediction will be made for each cluster. The final model prediction will be a weighted sum of predictions from every cluster. The paper is organized as follows: [Section 2](#) introduces several key concepts including copula modeling and cluster analysis; [Section 3](#) describes the proposed adaptive copula-based model bias characterization approach. [Section 4](#) demonstrates

the application of the proposed copula-based model bias characterization approach in two case studies and compares the results with those obtained by using the regression-based approach; [Section 5](#) concludes the whole paper.

2. COPULA-BASED MODEL BIAS CHARACTERIZATION

2.1. Model Bias

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

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

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. $Y - \varepsilon$ can often be thought of as data obtained from physical tests (test data), while $\hat{Y}(\mathbf{P}, \mathbf{X}, \Phi) + \delta$ can be treated as computer model outputs (CAE data)

2.2. Brief Introduction of Copula

A copula is a general way in statistics to formulate a multivariate distribution, particularly for a bivariate distribution, with various statistical dependence patterns. Formally, a copula is a joint distribution function of standard uniform random variables. According to Sklar's theorem [10], there exists an n -dimensional copula C such that for all T in a real random space,

$$F(T_1, \dots, T_N) = C(F_1(T_1), \dots, F_N(T_N)) \quad (2)$$

where F is an N -dimensional distribution function with marginal functions $F_1, \dots, F_N$. To this date, most copulas only deal with bivariate data due to the fact that there is a lack of practical n -dimensional generalization of the coupling parameter [11,12]. For multivariate data, a usual approach is to analyze the data pair-by-pair using two-dimensional copulas.

2.2.1. Copula Modeling

Copula is used to model the distribution of model bias. The most commonly employed methods to select the best copula are based on a maximum likelihood approach [13,14,15], 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 characterizing the model bias at a few design configurations.

Recently a Bayesian copula approach [11] 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 [11]. 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. 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.

$$H_k: \text{The data come from Copula } C_k, \quad k = 1, \dots, Q \quad (3)$$

The objective is to find the copula with the highest $P_r(H_k|D)$ 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

$$\begin{aligned} Pr(H_k|D) &= \frac{Pr(D|H_k)Pr(H_k)}{Pr(D)} \\ &= \int_{-\infty}^{\infty} \frac{Pr(D|H_k, \tau)Pr(H_k|\tau)Pr(\tau)d\tau}{Pr(D)} \end{aligned} \quad (4)$$

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 $Pr_r(D|H_k, \tau)$ 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 [16]. For more details of the Bayesian copula approach, please refer to the reference document [11].

2.2.2. Identify Model Bias from Multiple 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 $\hat{Y} = a$, that is,

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

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.

1. Generate model bias samples (e.g. $\delta_{N1}, \delta_{N2}, \dots, \delta_{Nj}$) from the conditional PDFs obtained by using Equ. 6;
2. Identify lower (δ_{lower}) and upper (δ_{upper}) bounds from the model bias samples, where

$$\delta_{lower} = \max[\min(\delta_{N1}), \min(\delta_{N2}), \dots, \min(\delta_{Nj})]$$

$$\delta_{upper} = \min[\max(\delta_{N1}), \max(\delta_{N2}), \dots, \max(\delta_{Nj})]$$
3. keep model bias samples which are located between the lower and upper bounds;
4. obtain an empirical PDF of the model bias from samples in step 3

3. PROPOSED APPROACH

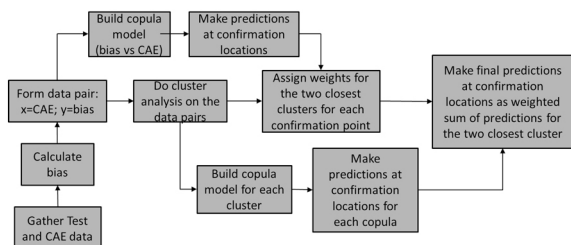


Figure 1. Steps of adaptive copula-based model bias characterization

Important steps in our proposed approach can be summarized as those in the flowchart above (Fig. 1). Test and CAE data are assumed to be available and they will be used as training data to build the clusters and copula models. Bias is then calculated by simply subtracting CAE data from the test data. Data pairs are formulated by treating each of the CAE data as the x coordinate and the corresponding bias as the y coordinate. Cluster analysis is conducted for the data pairs using k-means clustering. And a copula model will be built for each cluster. Weights associated to each cluster will be assigned to each of the confirmation data pairs (again bias vs CAE) based on the distance between the pair and the centroid of each of the clusters. It should be noted here that confirmation test data are often unavailable and as a result the confirmation bias cannot be obtained. To cope with this situation, a prediction using our previously

proposed single copula-based approach [18] will be made and treated as the pseudo confirmation test data. Also using our previous approach, prediction for a specific confirmation point using information (copula sample points) from each cluster will be made. The final prediction is the weighted sum using the weights and the predictions just mentioned.

Compared to our previously proposed approach, the proposed approach here has several enhancements, in the sense that it incorporates cluster analysis to separate data into groups based on their similarity, and uses a weighting scheme to consolidate every piece of information from the clusters, and a sampling approach for calculating expected value at the confirmation locations. All three aspects will be discussed in detail in the following sections.

3.1. Cluster Analysis

3.1.1. k-means Clustering

The k-means clustering is the algorithm used here to preprocess data and group similar data points together. The procedure aims at segmenting the data in such a way that the within-cluster variation is minimized. The clustering process starts by randomly assigning objects to a number of clusters. The objects are then successively reassigned to other clusters to minimize the within-cluster variation, which is basically the (squared) distance from each observation to the center of the associated cluster. If the reallocation of an object to another cluster decreases the within-cluster variation, this object is reassigned to that cluster. With k-means, cluster affiliations can change in the course of the clustering process.

Prior to analysis, we have to decide on the number of clusters. We can tell how many segments are needed, or we may know from previous research what to look for. Based on this information, the algorithm randomly selects a center for each cluster (step 1).

Euclidean distances are computed from the cluster centers to every single object. Each object is then assigned to the cluster center with the shortest distance to it. We now have our initial partitioning of the objects into several clusters.

Based on this initial partition, each cluster's geometric center (i.e., its centroid) is computed (third step).

The distances from each object to the newly located cluster centers are computed and objects are again assigned to a certain cluster on the basis of their minimum distance to other cluster centers. Since the cluster centers' position changed with respect to the initial situation in the first step, this could lead to a different cluster solution.

The k-means procedure now repeats and re-computes the cluster centers of the newly formed clusters, and so on. In other words, steps 3 and 4 are repeated until a predetermined number of iterations are reached, or convergence is achieved (i.e., there is no change in the cluster affiliations).