

A Copula-Based Approach for Model Bias Characterization

Zhimin Xi and Pan Hao
University of Michigan

Yan Fu and Ren-Jye Yang
Ford Motor Co.

ABSTRACT

Available methodologies for model bias identification are mainly regression-based approaches, such as Gaussian process, Bayesian inference-based models and so on. Accuracy and efficiency of these methodologies may degrade for characterizing the model bias when more system inputs are considered in the prediction model due to the curse of dimensionality for regression-based approaches. This paper proposes a copula-based approach for model bias identification without suffering the curse of dimensionality. The main idea is to build general statistical relationships between the model bias and the model prediction including all system inputs using copulas so that possible model bias distributions can be effectively identified at any new design configurations of the system. Two engineering case studies whose dimensionalities range from medium to high will be employed to demonstrate the effectiveness of the copula-based approach.

CITATION: Xi, Z., Hao, P., Fu, Y., and Yang, R., "A Copula-Based Approach for Model Bias Characterization," *SAE Int. J. Passeng. Cars - Mech. Syst.* 7(2):2014, doi:10.4271/2014-01-0735.

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.

The purpose of this paper is to propose a model bias characterization technique utilizing copula. Model bias is first characterized in the design space, then the original model prediction is corrected by adding the characterized model bias using copula. The paper is organized as follows: Section 2

introduces copula with the focus on a Bayesian copula approach; [Section 3](#) 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 4](#) concludes the whole paper.

2. COPULA-BASED MODEL BIAS CHARACTERIZATION

2.1. 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,

$$F(T_1, \dots, T_N) = C(F_1(T_1), \dots, F_N(T_N)) \quad (1)$$

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. Proposed Approach for Model Bias Characterization Using Copula

2.2.1. Model Bias

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

$$\hat{Y}(P, X, \Phi) + \delta = Y - \varepsilon \quad (2)$$

where $\hat{Y}$ and Y are the predicted and the measured system performances, respectively, δ is the model bias, ε is the measurement error, P is a vector of the deterministic model variable, X and Φ are the vectors of the irreducible and reducible model random variables, respectively. The irreducible random variables (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 $\hat{Y}(P, X, \Phi) + \delta$ can be treated as computer model outputs (CAE data).

2.2.2. 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 charactering 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 $Pr(H_k|D)$ 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

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

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 $Pr(D|H_k, \tau)$ depends upon τ and can be calculated from the copula PDF as