

On Stochastic Model Interpolation and Extrapolation Methods for Vehicle Design

Zhenfei Zhan, Yan Fu and Ren-Jye Yang
Ford Motor Co

ABSTRACT

Finite Element (FE) models are widely used in automotive for vehicle design. Even with increasing speed of computers, the simulation of high fidelity FE models is still too time-consuming to perform direct design optimization. As a result, response surface models (RSMs) are commonly used as surrogates of the FE models to reduce the turn-around time. However, RSM may introduce additional sources of uncertainty, such as model bias, and so on. The uncertainty and model bias will affect the trustworthiness of design decisions in design processes. This calls for the development of stochastic model interpolation and extrapolation methods that can address the discrepancy between the RSM and the FE results, and provide prediction intervals of model responses under uncertainty. This paper investigates and compares three stochastic methods for model interpolation and extrapolation, and they are: (1) Bayesian inference-based method, (2) Gaussian process modeling-based method, and (3) Copula-based method, and several validation metrics are proposed to evaluate the prediction results. An analytical case study and a vehicle design problem are used to demonstrate the advantages and disadvantages of these methods.

CITATION: Zhan, Z., Fu, Y. and Yang, R., "On Stochastic Model Interpolation and Extrapolation Methods for Vehicle Design," *SAE Int. J. Mater. Manf.* 6(3):2013, doi:10.4271/2013-01-1386.

INTRODUCTION

Computational Aided Engineering (CAE) becomes a vital tool in industries. In automotive industry, Finite Element (FE) models are widely used in vehicle design. Model validation [1,2,3,4,5,6] is the process of comparing CAE model outputs with test measurements in order to assess the validity and predictive capabilities of the CAE model for its intended usage. Successful implementation of model validation will increase the chance of virtual testing and may lead to significant reduction in prototype building and testing of vehicle designs.

However, even with increasing speed of computers, the simulation time of full FE models is still too time-consuming to use it for direct design optimization or robust design. It sometimes takes hours or even days to complete a single FE simulation run. Hence, response surface model (RSM) is commonly used as a surrogate of the FE model to reduce the development time. Various RSM techniques have been developed in recent years including radial basis function (RBF) [7], artificial neural network (ANN), support vector regression (SVR), and Kriging (KRG) [8, 9]. The effectiveness of different RSM techniques varies based on the different modeling criteria, amount of available samples, and the nonlinear nature of the physical phenomena itself [10,

11]. No matter what RSM method is used, there is always some discrepancy between the results of the true FE model and the RSM. This is especially true when the RSMs are used to predict new designs in terms of interpolation and extrapolation. This calls for the development of stochastic model interpolation and extrapolation methods that can address the discrepancy between the RSM and FE models and provide bias-corrected prediction intervals of RSM model responses under uncertainty.

Various statistical inference techniques have been developed [12], and Bayesian methods have been investigated extensively for validation and calibration of complex CAE models [13]. The Bayes network links the computational model from component level to system level in the context of both physics-based extrapolation and statistical extrapolation. Rebba and Mahadevan [14, 15] developed and implemented a Bayesian framework for drawing inferences for predictions in the untested domain using Bayes networks. Jiang and Mahadevan [16] proposed Bayesian hypothesis testing-based methodology for model validation and confidence extrapolation under uncertainty using limited test data. Kim and Kwak [17] introduced the concept of the design space optimization. By starting with a small design space and advancing to a larger space, the method eventually

achieves an optimal design space. Chen et al [18], Li et al [19] and Kakkolaras et al [20] brought up the needs for validating models from the perspective of engineering design, and provide stochastic assessment of the validity of a model. However, these methods did not fully address the needs for bias correction and design interpolation and extrapolation under uncertainty. With these considerations, three interpolation and extrapolation methods are investigated. They are Bayesian inference-based method [21], Gaussian process modeling (GPM)-based method [22], and Copula-based method [23].

An adequate validation metric is required to evaluate the models' predictive capability. Developing quantitative model validation methods has attracted considerable researchers' interest in recent years. Statistical hypothesis testing [3, 4], Sprague and Geers, and Knowles and Gear metrics [5] were proposed to provide quantitative model validation. Bayesian based metrics [24, 25] were developed to determine the predictive capabilities of CAE models. They were applied to various model validation problems and showed significant potential. In addition, reliability based validation methods [14, 15, 26, 27] provides a direct, objective and quantitative assessment of quality of a CAE model. It incorporates both accuracy and adequacy requirement to facilitate applications. Liu, et al. [28] summarized four main types of stochastic validation metrics to consider uncertainties from both physical tests and computer models, namely classical hypothesis testing, Bayes factor, frequentist's metric, and area metric. In this paper, three validation metrics, e.g. area metric, reliability-based metric, and Bayesian confidence metric, are employed to evaluate the effectiveness of the three stochastic interpolation and extrapolation methods under uncertainty.

The rest of the paper is organized as follows: three stochastic interpolation and extrapolation methods are briefly introduced, followed by description of three stochastic model validation metrics. Next, a one dimensional analytical case study and a vehicle frontal impact case study are used to investigate the three stochastic interpolation and extrapolation methods. Finally, some discussions and summary are given in the end.

STOCHASTIC INTERPOLATION & EXTRAPOLATION METHODS

In this section, three stochastic interpolation and extrapolation methods are introduced. For the convenience of formulation, let $\mathbf{x}$ represents design variables, Y_t and Y_m stand for FE model result (test) and RSM model result (model) respectively. The relationship between $Y_t(\mathbf{x})$ and $Y_m(\mathbf{x})$ can often be generalized as follows:

$$Y_t(\mathbf{x}) = Y_m(\mathbf{x}) + \delta(\mathbf{x}) + \varepsilon(\mathbf{x}) \quad (1)$$

where ε is the random variable representing the test error that may depend on $\mathbf{x}$, and δ is the bias of the CAE model. Hence,

it can be drawn that the "reality" or "true value" of output, namely Y_{true} , is equal to the test observation minus the test error, and the CAE model prediction plus the model bias, as shown in Equation 2.

$$Y_{\text{true}}(\mathbf{x}) = Y_t(\mathbf{x}) - \varepsilon(\mathbf{x}) = Y_m(\mathbf{x}) + \delta(\mathbf{x}) \quad (2)$$

Equation 2 serves as the governing equation for all three methods.

Bayesian Inference-Based Model Interpolation and Extrapolation Method

Zhan et al [21] proposed a Bayesian inference-based model interpolation and extrapolation method. This method utilizes the advantages of Bayesian inference and response surface model (RSM). The process starts with design of experiment (DOE) matrix for validation in the design space, followed by performing repeated physical tests and CAE simulations. The difference between test and CAE data are then calculated as evidence for the Bayesian updating of hyper-parameters of the bias distribution. After obtaining the prior information, the posterior distributions of prediction bias hyper-parameters are calculated.

Following Equation 2, assume $\tau(\mathbf{x}) = \delta(\mathbf{x}) + \varepsilon(\mathbf{x})$. It is noted that the prediction bias is an essential step for estimating the value of Y_{true} . For the convenience of formulation and method demonstration, it is reasonable to assume the bias $\tau(\mathbf{x}) = \delta(\mathbf{x}) + \varepsilon(\mathbf{x})$ follows a normal distribution with known variance as:

$$\tau(\mathbf{x}) \sim N(\tau_\mu(\mathbf{x}), \tau_{\sigma^2}(\mathbf{x})) \quad (3)$$

Given the known variance τ_{σ^2} , the mean value τ_μ needs to be inferred with available data. In Bayesian probability theory, if the posterior distributions $p(\theta|\mathbf{x})$ are in the same family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions. For example, the Gaussian family is self-conjugate with respect to a Gaussian likelihood function. If the likelihood function is Gaussian, choosing a Gaussian prior over the mean will ensure that the posterior distribution is also Gaussian. In this study, the prior distribution of τ_μ can be assumed as:

$$\tau_{\mu 0} \sim N(\mu_0, \sigma_0^2) \quad (4)$$

Consider the general problem of inferring a distribution for a parameter θ given some datum or data $\mathbf{x}$. From Bayes' theorem, the posterior distribution is equal to the product of the likelihood function and prior $p(\theta)$, normalized (divided) by the probability of the data $p(\mathbf{x})$ as Equation 5:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta} \quad (5)$$

Given the assumption of normally distributed conjugate prior, let n denote the number of test observations, the posterior distribution of τ_μ can be expressed as

$$\tau_{\mu 1} \sim N(\mu_1, \sigma_1^2) \quad (6)$$

Where

$$\mu_1 = \left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^n x_i}{\sigma^2} \right) / \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right), \quad \sigma_1^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1}.$$

Based on the test and CAE data, the bias distribution at each design point $\mathbf{x}$ is obtained as: $\tau(\mathbf{x}) \sim N(\tau_{\mu 1}(\mathbf{x}), \tau_{\sigma^2}(\mathbf{x}))$. It is noted that this method can also be used for conjugate prior of different distributions and non-conjugate prior as well.

Two RSMs are built for the mean and standard deviation of the prediction bias. After obtaining the CAE simulation result, the prediction intervals (PIs) of the output at new designs are calculated. Confirmation tests at new designs are used to validate the prediction results. If the test results of a new design are within the corresponding PIs, the prediction by interpolation or extrapolation is considered successful. Based on the result, the decision maker then decides to accept or reject the prediction result. If accepted, it is then passed to the downstream engineers for design optimization or robust design. On the other hand, if the prediction is rejected, additional DOE will be added and repeat the process until good quality prediction is achieved [21].

Gaussian Process Modeling-Based Interpolation and Extrapolation Method

A Gaussian random process (or in short, Gaussian process) $f(\bullet)$ is a random process that has two Gaussian properties. It can be represented in the following form:

$$f(\mathbf{x}) \sim GP(m(\mathbf{x}), V(\mathbf{x}, \mathbf{x}')) \quad (7)$$

where $\mathbf{x}$ denotes time or multidimensional space location. $m(\mathbf{x})$ is the *mean function* of the Gaussian process, and $V(\mathbf{x}, \mathbf{x}')$ is the *covariance function* of the Gaussian process. The two Gaussian properties are that 1) the random variable $f(\mathbf{x})$ at location $\mathbf{x}$ follows a normal distribution

$$f(\mathbf{x}) \sim N(m(\mathbf{x}), V(\mathbf{x}, \mathbf{x})) \quad (8)$$

and that 2) a finite collection of random variables $f(\mathbf{x}_1), f(\mathbf{x}_2), \dots, f(\mathbf{x}_n)$ at locations $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ follows a multivariate normal distribution.

A Gaussian process modeling (GPM)-based interpolation and extrapolation method was proposed by Jiang et al [22]. The general idea is to use the basic concept of Gaussian process modeling for modeling bias correction. According to Equation (2), assume that the CAE model $Y_m(\mathbf{x})$ can be replaced by a Gaussian process with hyperparameters β_m, σ_m^2 , and ω_m^2 .

$$Y_m(\mathbf{x}) \sim GP(h_m(\mathbf{x})^T \beta_m, \sigma_m^2 \exp\{-\sum_{k=1}^p \omega_k^m (x_k - x'_k)^2\}) \quad (9)$$

and that the bias function $\delta(\mathbf{x})$ can be modeled as a Gaussian process with hyperparameters $\beta_\delta, \sigma_\delta^2$, and ω_δ^2 .

$$\delta(\mathbf{x}) \sim GP(h_\delta(\mathbf{x})^T \beta_\delta, \sigma_\delta^2 \exp\{-\sum_{k=1}^p \omega_k^\delta (x_k - x'_k)^2\}) \quad (10)$$

Then all the hyper-parameters ($\beta_m, \sigma_m^2, \omega_m^2, \beta_\delta, \sigma_\delta^2$, and ω_δ^2) are decided through maximum likelihood estimation (MLE). The details of the GPM-based interpolation and extrapolation method can be referred to Jiang et al [22].

Copula-Based Interpolation and Extrapolation Method

In probability theory and statistics, copulas are used to describe the dependence between random variables. The cumulative distribution function of a random vector can be written in terms of marginal distribution functions and a copula. The marginal distribution functions describe the marginal distribution of each component of the random vector and the copula describes the dependence structure between the components. Copulas are popular in statistical applications as they allow one to easily model and estimate the distribution of random vectors by estimating marginal and copula separately.

Xi et al [23] proposed a Copula-based interpolation and extrapolation method. The principle is to first characterize the model bias in the design space, then correct the model prediction by adding the characterized model bias. According to the Copula modeling, it is possible to predict the possible model bias for the new design configuration. For example, Copula modeling between Y_m and δ allows us to find the possible model bias $\delta(\mathbf{x})$ for a realization of $Y_m(\mathbf{x})$ at a new design configuration. Mathematically, this is a process to identify the conditional PDF of the model bias δ given of $Y_m(\mathbf{x}) = a$ as shown in Equation 11.

$$C(F_{Y_m}(Y_m), F_\delta(\delta)) | Y_m = a \quad (11)$$

Meanwhile, at a new design configuration the design variable values are known. 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, the intersection of these conditional PDFs for the model bias needs to be identified. 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 the same number of model bias samples (e.g. $\delta_{N,1}, \delta_{N,2}, \dots, \delta_{N,j}$) from total j conditional PDFs in Equation 11, where N is the number of random samples

Step 2. identify the lower bound (δ_{Lower}) and upper bound (δ_{upper}) from the j set of model bias samples, where $\delta_{\text{Lower}} = \max[\min(\delta_{N,1}), \min(\delta_{N,2}), \dots, \min(\delta_{N,j})]$, $\delta_{\text{upper}} = \min[\max(\delta_{N,1}), \max(\delta_{N,2}), \dots, \max(\delta_{N,j})]$.

Step 3. obtain the model bias samples located between the lower and upper bounds;

Step 4. calculate the empirical PDF of the model bias from based on the obtained samples in step 3.

The details of the Copula-based interpolation and extrapolation method can be referred to Xi et al [23].

VALIDATION METRICS FOR METHOD COMPARISON

In order to investigate the behavior of different stochastic interpolation and extrapolation methods, this section introduces three validation metrics that can be used to evaluate the prediction results at different discrete design points. For the convenience of formulation, Let $Y_t(x)$ and $Y_{m_corr}(x)$ represent the physical test data and bias corrected CAE model predictions distribution at design x , where $Y_{m_corr}(x)$ is the combination of RSM prediction model $Y_m(x)$ and prediction bias model $\tau(X)$.

Area Metric

In order to measure the agreement of the entire distributions of model prediction and experiments, Ferson et al [2] proposed a validation metric which uses the area between the model prediction distribution $f(Y_{m_corr})$ and the experimental distribution $f(Y_t)$. The quantitative measure of the discrepancy between the two distributions is formulated as

$$d(f(Y_t), f(Y_{m_corr})) = \int_{-\infty}^{+\infty} |f(Y_t(x)) - f(Y_{m_corr}(x))| dx \quad (12)$$

The area metric has two important characteristics. First, the area metric measures the differences between the entire distributions from the experiments and model predictions. Second, the metric can be used when only a few data points from experiments or model predictions are available.

Liu et al [28] summarized the characteristics of the area based validation metrics. The advantages include: (1) the area based metrics directly use the area between the distributions of model predictions and experimental observations as an objective metric to quantify the agreement, so that wider distribution of the model prediction will lead to a larger discrepancy. (2) The area metric generalizes to a deterministic comparison when uncertainty goes to zero and degenerates the comparison of two stochastic quantities to a comparison between two scalar values from predictions and physical observations. (3) One distinctive advantage of these metrics is its capability of integrating sparse physical observations at multiple validation sites to assess the global predictive capability over a specified domain of interest. Several disadvantages of the area based metrics are noted as follows: (1) When there is a lack of sufficient experimental data, using the empirical distribution derived from physical observations may lead to underestimation or overestimation. Thus, the existing area metric does not provide a confidence level of the metric due to the lack of sufficient data. (2) The existing area based metrics only compares marginal distributions; therefore, it is better suited for single response or uncorrelated multiple responses. How to deal with the problem with multiple correlated responses needs to be investigated.

Reliability Based Metric

Rebba et al [15] proposed model reliability based metric that measures the agreement between the model prediction and experiments. It is adopted by Zhan et al [27] to quantitatively assess the CAE model quality of dynamic systems.

Let Φ represents the difference data between the test and CAE data, that is $\Phi = y_t(x) - Y_{m_corr}(x)$. The difference data Φ is the validating feature selected for reliability based validation metric. Model reliability is defined as a simple validation metric $r = p(\epsilon_{rl} < \Phi < \epsilon_{ru})$, which represents the probability that the observed difference is within a small interval, also called the threshold $[\epsilon_{rl}, \epsilon_{ru}]$. The threshold helps to estimate the probability, and the adequacy (confidence) requirement is a , such that we accept the model prediction only when $p(\epsilon_{rl} < \Phi < \epsilon_{ru}) > a$. The physical test and model simulation results can be univariate or multivariate time series. The validation problem is to ask if the probability that the difference data between the test and model simulation results are within certain interval is larger than a predefined level a , that is $p(\epsilon_{rl} < \Phi < \epsilon_{ru}) > a$. Here we assume Φ follows a normal distribution $N(\mu_\Phi, \Sigma_\Phi)$, where μ_Φ is the difference

mean and Σ_{Φ} is the covariance matrix of difference. Then the model reliability is calculated as:

$$r = p(\Phi < \varepsilon_u) - p(\Phi < \varepsilon_l) \quad (13)$$

The obtained model reliability is provided to decision makers to assess the model validity and predictive capacity. If the model is validated with an acceptable reliability, design optimization can then be performed to reduce the vehicle's weight while improving vehicle performances and quality. On the other hand, if not, it can be either rejected or proposed for improvement, and the validation process is repeated.

One of the crucial steps in Reliability based validation metric is to define a proper threshold $[\varepsilon_{rl}, \varepsilon_{ru}]$. In this study, in order to differentiate the capability of different method, a tight threshold is selected as 2% of the maximum absolute of the test data. The reliability metric results are ranging from 0 to 1, 1 means the whole difference distribution are within the allowable interval so that the test and CAE model distributions are considered perfectly match with each other.

Bayesian Confidence Metric

Jiang et al [24] and Zhan et al [25] used interval-based Bayesian hypothesis testing method to quantitatively assess the CAE model quality. Within the context of binary hypothesis testing for multivariate model validation, the formulation of Bayesian interval-based hypotheses is represented as $\mathbf{H}_0: \varepsilon_{bl} \leq \Phi \leq \varepsilon_{bu}$ versus $\mathbf{H}_1: \Phi < \varepsilon_{bl} \text{ or } \Phi > \varepsilon_{bu}$, where $[\varepsilon_{bl}, \varepsilon_{bu}]$ is a predefined threshold interval. Here whether the difference Φ is within an allowable limit $[\varepsilon_{bl}, \varepsilon_{bu}]$ is tested. Assuming that Φ has a probability density function under each hypothesis, i.e., $\Phi|\mathbf{H}_0 \sim f(\Phi|\mathbf{H}_0)$, $\Phi|\mathbf{H}_1 \sim f(\Phi|\mathbf{H}_1)$. In this study, it is assumed that 1) the difference Φ follows a normal distribution $N(\mu_{\Phi}, \Sigma_{\Phi})$ with the covariance matrix Σ_{Φ} , and 2) a prior density function of δ under both null and alternative hypotheses, denoted by $f(\delta)$, is taken to be $N(\rho, \Lambda)$. This selection assumes that the amount of information in the prior is equal to that in the observation, which is consistent with the Fisher information-based method. Its numerical integration should be referred to Jiang [24]. In this study, same interval with reliability based metric are selected as a percentage of the maximum absolute of the test data. The Bayesian confidence metric results are also ranging from 0 to 1.

CASE STUDY & DISCUSSION

Analytical Example

A one dimensional mathematical example [10, 29] is used to study the characteristics of different interpolation and extrapolation methods. The function is:

$$y = (6x - 2)^2 \sin(12x - 4) + \varepsilon \quad (14)$$

where the independent test error $\varepsilon \sim N(0, 0.07^2)$, $x \sim [0, 1]$. A Kriging response surface model is constructed based on 6

samples at $x = [0, 0.22, 0.39, 0.63, 0.86, 1]$, as shown in Figure 1. It is noted that there is significant discrepancy between the true function curve and the Kriging response surface model curve, especially in the domain between $[0, 0.2]$ and $[0.65, 0.85]$ where the behavior of true function is obviously not captured by the RSM curve. Another set of 10 design points at $x = [0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.95]$ with 5 repeated tests at each design are obtained for bias construction. The true model, the Kriging RSM model, and the repeated tests are shown in Figure 1.

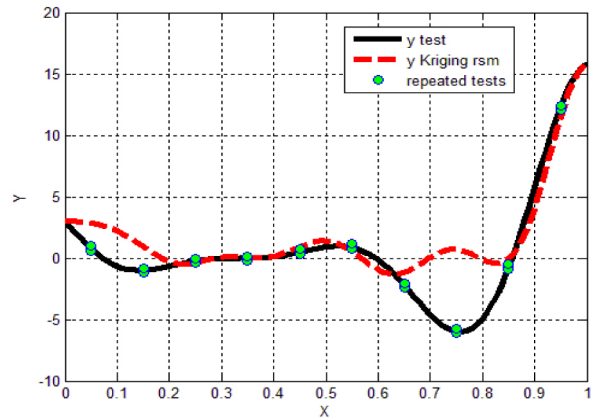


Figure 1. True model, RSM model, and repeated tests

Three stochastic interpolation and extrapolation methods are employed to provide prediction intervals with 99% confidence level, and the results are shown in Figure 2.

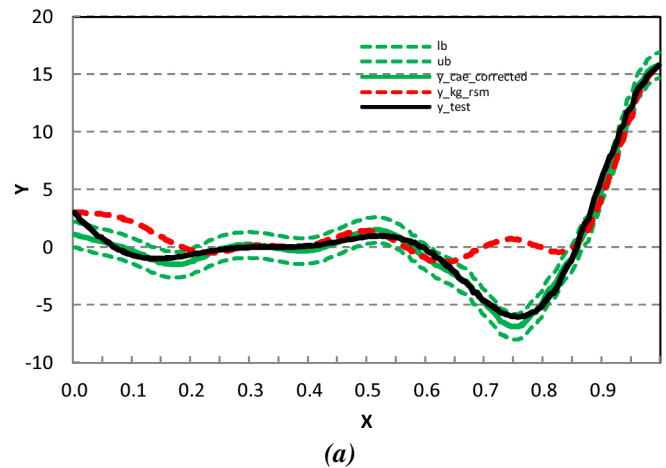


Figure 2. Prediction results of (a) Bayesian inference-based method, (b) Gaussian Process modeling (GPM)-based method, (c) Copula-based Method