

A Nonparametric Bootstrap Approach to Variable-size Local-domain Design Optimization and Computer Model Validation

Dorin Drignei, Zissimos Mourelatos and Vijitashwa Pandey
Oakland Univ.

Michael Kokkolaras
Univ. of Michigan-Ann Arbor

ABSTRACT

Design optimization often relies on computational models, which are subjected to a validation process to ensure their accuracy. Because validation of computer models in the entire design space can be costly, a recent approach was proposed where design optimization and model validation were concurrently performed using a sequential approach with both fixed and variable-size local domains. The variable-size approach used parametric distributions such as Gaussian to quantify the variability in test data and model predictions, and a maximum likelihood estimation to calibrate the prediction model. Also, a parametric bootstrap method was used to size each local domain. In this article, we generalize the variable-size approach, by not assuming any distribution such as Gaussian. A nonparametric bootstrap methodology is instead used to size the local domains. We expect its generality to be useful in applications where distributional assumptions are difficult to verify, or not met at all. A heat conduction problem illustrates the proposed methodology.

CITATION: Drignei, D., Mourelatos, Z., Pandey, V. and Kokkolaras, M., "A Nonparametric Bootstrap Approach to Variable-size Local-domain Design Optimization and Computer Model Validation," *SAE Int. J. Mater. Manf.* 5(1):2012, doi: 10.4271/2012-01-0226.

1. INTRODUCTION

Simulation-based design optimization requires that the model is valid in a wide range of design variables. In addition, there is inherent uncertainty in both the model calibration parameters and the tests that are conducted to obtain the data to be used in the validation process. If the model is validated for the entire range of its inputs, it will always yield designs that will perform as predicted. However, this may require a large amount of resources to conduct multiple tests in order to quantify test uncertainty at each point of a sufficiently large set in a potentially high-dimensional space of the model inputs.

Due to limited resources, current practice uses computational models in simulation-based design optimization that have been validated only at a relatively small number of points. This approach can compromise local model accuracy for the benefit of using one single global model. [Li et al. \(2010\)](#) presented a trust-region-like design optimization methodology and showed that design optimization using a global model can yield a different and worse "optimal" design relative to the one obtained by using

a model that is calibrated when necessary as the optimization process progresses.

In [Li et al. \(2010\)](#), a sequential methodology was presented to validate a simulation model. Starting with an initial design, an optimization problem is solved within a local domain around that design in which we consider the simulation model to be valid within a given confidence level. This approach can potentially reduce the number of required tests, which are carefully chosen at each design point. In this manner, available resources can be utilized effectively while ensuring the validity of the simulation model in the vicinity of the design candidates generated by the optimization process.

[Drignei et al. \(2011\)](#) improved the methodology in [Li et al. \(2010\)](#) by using a variable-size local domain method for concurrent design optimization and model validation. The size and shape of the local domain at each stage of a sequential design optimization process were determined from a parametric bootstrap methodology under the Gaussian assumption, involving maximum likelihood estimators of unknown model parameters. Validation through calibration (i.e. adjustment of simulation model parameters in order to

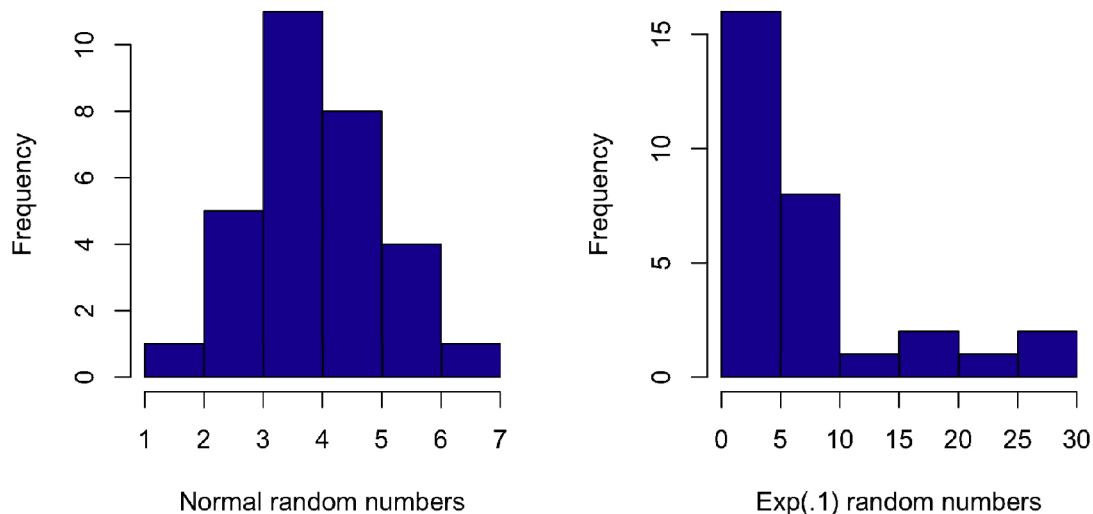


Figure 1. Simulated random numbers

increase the confidence level that there is statistical similarity between predictions and tests) was carried out in the local domain at each stage. The sequential process continues until the local domain does not change from stage to stage during the design optimization process, ensuring convergence to an optimal design.

The assumption of Gaussian data is prevalent in many studies for simplicity reasons. While many data sets exhibit a Gaussian behavior, there are instances where this is not the case. For example, data that are severely skewed, multimodal or include outliers may not be Gaussian. In this paper, we explore an alternative approach to variable-size local-domain for concurrent design optimization and model validation that does not rely on the assumption of Gaussian (or any other) distribution. For that, we use a nonparametric bootstrap statistical method.

An outline of this paper is as follows. In Section 2 we discuss and contrast the parametric and nonparametric bootstrap methods. In Section 3 we outline the proposed approach to local design optimization and model validation using nonparametric bootstrap. In Section 4 we use this method to design a porous material insulator using heat conduction in a solid.

2. PARAMETRIC VERSUS NONPARAMETRIC BOOTSTRAP

This section provides a basic description of the parametric and nonparametric bootstrap methods highlighting their differences and application domains. Suppose we have a sample of observations that are generated from a distribution whose functional form is known but its parameters are not (e.g. Gaussian distribution with unknown mean and standard deviation). The parametric bootstrap provides a method for statistical inference of the unknown parameters such as estimating the confidence intervals of the mean value.

Using an available sample (i.e. a finite number of data), the parameters of the distribution are first estimated. Then the distribution with the estimated parameters is sampled to obtain B samples of n random numbers each. Each of the B samples is used to compute the statistic of interest S_b , $b = 1, \dots, B$. Then a 90% confidence interval for the parameter of interest (e.g. mean value) is obtained from the 5% and 95% percentiles of S_b , $b = 1, \dots, B$. For example, a sample of 30 random numbers is generated from a Gaussian distribution with a mean value of 4. The left panel of Figure 1 shows their histogram. The estimated mean and standard deviation of this Gaussian distribution are $\bar{x} = 3.79$ and $s = 1.07$, respectively. To illustrate the parametric bootstrap, we generated $B=100$ samples of size $n = 30$ from the Gaussian distribution with these estimated parameters and obtained $B=100$ sample means S_b . The 90% parametric bootstrap confidence interval was then calculated as (3.45, 4.08). The exact 90%

confidence interval $\bar{x} \pm 1.645 \frac{s}{\sqrt{30}}$ is (3.46, 4.11). We note that the parametric bootstrap and the exact confidence intervals agree here, and they both include the true mean of 4.

The right panel of Figure 1 shows a histogram of 30 simulated exponential random numbers with parameter $\lambda = 0.1$. Thus, the mean value of the exponential random variable

that generated these numbers is equal to $\frac{1}{\lambda} = 10$. We

computed the exact 90% confidence interval $\bar{x} \pm 1.645 \frac{s}{\sqrt{30}}$ for this mean value as (4.84, 9.58) under the wrong assumption that these random numbers were generated from a Gaussian random variable. Clearly, this interval doesn't include the mean value of 10, emphasizing that statistical inference methods can be wrong if the Gaussian distribution is used for data whose distribution deviates significantly from Gaussian.

When one is not sure about the distribution of available data, nonparametric methods for statistical inference should be used. One such method called nonparametric bootstrap (Higgins, 2004) is based on the concept of re-sampling. For example, if 3.10, 17.98, 1.09, 6.63, 2.64 is a sequence of five random numbers from the exponential distribution with parameter $\lambda = 0.1$, the sequence 3.10, 17.98, 1.09, 3.10, 17.98 is obtained by re-sampling with replacement the first sequence. Notice that the second sample has repeated values since the sampling is done with replacement. In nonparametric bootstrap, one obtains B re-samples of the original sample with replacement, and then calculates the statistic of interest S_b (e.g. the sample mean) for each sample b . The distribution of S_b is finally used to perform statistical inference. The 5% and 95% quantiles of S_b , $b = 1, \dots, B$ can be used as the lower and upper bounds of a 90% confidence interval for the parameter of interest. For the example in the right panel of Figure 1, the 90% nonparametric bootstrap confidence interval based on $B=100$ is (4.65, 10.10), which includes the mean value of 10.

3. PROPOSED NONPARAMETRIC BOOTSTRAP APPROACH

In our previous research (Li et al., 2010, Drignei et al., 2011) we have used a concurrent optimization and model calibration approach using sequential local domains. The approach calibrates the prediction model sequentially within each local domain ensuring validation of the design candidates as they are generated during optimization. For a two-dimensional problem for example, the design optimization starts with an initial design $\mathbf{d}=(d_1, d_2)$ and determines the size of a local domain around $\mathbf{d}$, by calibrating the simulation model at $\mathbf{d}$ so that the model is statistically similar to available test data at $\mathbf{d}$ for a certain confidence level. After the size of the local domain is determined, optimization is performed to obtain the optimal design within the local domain, completing the first stage of the sequential approach. A second stage is then initiated using the optimal design of the first stage as the initial design of the second stage, and repeating the process by first determining the size of the new local domain centered at the new initial design and then performing optimization within the new local domain. This process is assumed converged if the optimal design is within the current local domain and its Euclidean distance from the previous optimum is less than a predetermined absolute error.

Let $y^r(\mathbf{d}, \mathbf{p}, \tau)$ be the real (superscript r) and unknown response for design variables $\mathbf{d} \in \mathcal{R}^{n_d}$ and design parameters $\mathbf{p} \in \mathcal{R}^{n_p}$. The argument τ indicates that the response is, in general, time dependent. Let $y^t(\mathbf{d}, \mathbf{p}, \tau)$ be the test response (superscript t), and $y^m(\mathbf{d}, \mathbf{p}, \mathbf{c}(\mathbf{d}, \mathbf{p}), \tau)$ be the response from the simulation (prediction) model (superscript m), where $\mathbf{c} \in \mathcal{R}^{n_c}$ is the vector of model calibration parameters. A bold

letter indicates a vector. If $\mathbf{x}=(\mathbf{d}, \mathbf{p}) \in \mathcal{R}^{n_d+n_p}$, we have (Mahadevan and Rebba, 2005):

$$y^t(\mathbf{x}, \tau) = y^m(\mathbf{x}, \mathbf{c}, \tau) + \varepsilon(\mathbf{x}, \tau), \quad (1)$$

where $\varepsilon(\mathbf{x}, \tau)$ is a random quantity representing the experimental (test) error, of mean vector zero and covariance matrix Γ . At each design point $\mathbf{x}$, time τ is discretized at a finite number of time points and the resulting vector of all values is denoted by $\boldsymbol{\varepsilon}(\mathbf{x})$. Similarly, the vectors of responses and predictions at the same finite number of time points, are $\mathbf{y}^t(\mathbf{x})$ and $\mathbf{y}^m(\mathbf{x}, \mathbf{c})$, respectively. For simplicity, Equation (1) does not include the model bias explicitly although it can be easily included as explained in Drignei et al. (2011). The covariance matrix Γ may depend on statistical parameters $\boldsymbol{\theta}$ which must be estimated. If $\boldsymbol{\phi} = (\mathbf{c}, \boldsymbol{\theta})$ includes the calibration parameters $\mathbf{c}$ and/or the statistical parameters $\boldsymbol{\theta}$, then their estimated values can be obtained by minimizing the following generalized least-squares function

$$L(\boldsymbol{\phi} | \boldsymbol{\varepsilon}(\mathbf{x})) = \boldsymbol{\varepsilon}(\mathbf{x})^T \Gamma^{-1} \boldsymbol{\varepsilon}(\mathbf{x}). \quad (2)$$

For simplicity we choose Γ as the identity matrix multiplied by a scale parameter σ^2 . Therefore, in this paper we have $\boldsymbol{\theta} = \sigma^2$.

We then use the test data $\mathbf{y}^t(\mathbf{x})$, obtained during the process of validating the simulation model at design $\mathbf{x}$, to determine the size of the local domain around $\mathbf{x}$, within which the simulation model can be assumed valid with a given confidence level.

When the design $\mathbf{x}$ is fixed, the generalized least-squares function of the parameter vector $\boldsymbol{\phi}$ given the available test data, is provided by $L(\boldsymbol{\phi} | \boldsymbol{\varepsilon}(\mathbf{x}))$ in Equation (2). To define the size of the domain around $\mathbf{x}$, we propose to use an extended generalized least-squares function $L(\boldsymbol{\phi}, \mathbf{x} | \boldsymbol{\varepsilon}(\mathbf{x}))$, which has the same form as L in Equation (2), but with $\mathbf{x}$ as an additional statistical parameter. The parameters $(\boldsymbol{\phi}, \mathbf{x})$ are then estimated using a nonparametric bootstrap approach (Higgins 2004) based on simulated data. The contribution of this work is in using a nonparametric bootstrap approach instead of a parametric approach as in Drignei, et al. (2011). The method is nonparametric because it does not assume a specific parametric distribution such as Gaussian for the test data.

Specifically, we generate data $\bar{\boldsymbol{\varepsilon}}(\mathbf{x})$ by sampling with replacement the error $\boldsymbol{\varepsilon}(\mathbf{x}) = \mathbf{y}^t(\mathbf{x}) - \mathbf{y}^m(\mathbf{x}, \mathbf{c})$ to obtain simulated test data $\bar{\mathbf{y}}^t(\mathbf{x})$ as

$$\bar{\mathbf{y}}^t(\mathbf{x}) = \mathbf{y}^m(\mathbf{x}, \hat{\mathbf{c}}) + \bar{\boldsymbol{\varepsilon}}(\mathbf{x}). \quad (3)$$

The bar in Equation (3) indicates simulated (not actual) data. We then re-estimate the statistical model (i.e. re-