

Bayesian Probabilistic PCA Approach for Model Validation of Dynamic Systems

Xiaomo Jiang

General Electric Company

Ren-Jye Yang, Saeed Barbat and Para Weerappuli

Ford Motor Company

Copyright © 2009 SAE International

ABSTRACT

In the automobile industry, the reliability and predictive capabilities of computer models for a dynamic system need to be assessed quantitatively. Quantitative validation allows engineers to assess and improve model reliability and quality objectively and ultimately lead to potential reduction in the number of prototypes built and tests. A good metric, which is essential in model validation, requires considering uncertainties in both testing and computer modeling. In addition, it needs to be able to compare multiple responses simultaneously, as multiple quantities are often encountered at different spatial and temporal points of a dynamic system. In this paper, a state-of-the-art validation technology is developed for multivariate complex dynamic systems by exploiting a probabilistic principal component analysis method and Bayesian statistics approach. The probabilistic principal component analysis approach is developed to address multivariate correlation, data uncertainty, and dimensionality reduction. The multivariate Bayesian hypothesis testing method is exploited to quantitatively assess the quality of the computer models for dynamic systems. The proposed method is illustrated with a rear seat child restraint system example with a Hybrid III 3-year old child dummy model, consisting of sixteen sets of test data each having nine response variables.

INTRODUCTION

Computer modeling and simulations play an important role in reducing vehicle prototype tests and shortening

product development time. Its ultimate goal is achieving zero prototypes and replacing the physical prototype testing with computer simulations. To achieve this goal, the quality, reliability and predictive capabilities of the computer models for various vehicle dynamic systems with multiple responses must be quantitatively and systematically assessed with objectivity.

Model validation is the process of comparing model outputs with experimental observations in order to assess the validity or predictive capabilities of computer models. Developing quantitative methods for model validation under uncertainty has attracted considerable research interest in recent years. For example, the fundamental concepts and methodologies for validation of large-scale computer models are being investigated by several government agencies such as the United States Department of Defense (1996), American Institute of Aeronautics and Astronautics (1998), American Society of Mechanical Engineers (2006), and Advanced Simulation and Computing (ASC) program of the United States Department of Energy (2008). Detailed discussions of model verification and validation concepts can be found in many research articles (see e.g., Balci 1997; Roache 1998; Trucano *et al.* 2001; Oberkampf and Trucano 2002 & 2006; Easterling 2003; Gu and Yang 2003; Oberkampf *et al.* 2003 & 2004; Babuska and Oden 2004; Cafeo and Thacker 2004; Chen *et al.* 2004 & 2008; Sargent 2004; Campbell 2004; Bayarri *et al.* 2005; Oberkampf and Barone 2006; Wehrwein and Mourelatos 2006; Schwer 2007; Sun and Vlahopoulos 2007; Jiang and Mahadevan 2007, 2008a, b&c).

Within the context of the model validation for dynamic systems, however, time histories or functional data between model prediction and experimental measurement need to be compared effectively. Subjective engineering judgments based on graphical plots were often used to assess how good the model is (Oberkampf and Trucano 2002 & 2006; Oberkampf and Barone 2006). However, the quality of the computer model cannot be assessed quantitatively via graphical comparison. In addition, many critical issues, such as data correlation between multiple time series, uncertainties in both test and model prediction, and confidence of the model, are ignored. In particular, multiple responses are usually computed and observed at different spatial and temporal points of the dynamic system. These responses need to be compared simultaneously in the process of model validation. In the vehicle crash tests, for example, 212 sensors may be used in the World Side Impact Dummy (WorldSID) test (Cesari *et al.* 2001; Ruan *et al.* 2006) to measure head acceleration, pelvis forces and many other responses in the upper and lower extremities. All these sensor data are in the form of time histories with different measurement units. In such a situation, multiple univariate comparisons without considering correlation among them might yield conflicting inferences about the model quality in the validation process.

Statistical hypothesis testing is one approach to provide quantitative model validation under uncertainty [see Oberkampf and Barone (2006) for a comprehensive state-of-the-art review]. Recently, a few researchers (e.g., Hemez and Doebling 2000; Doebling *et al.* 2001; Paez and Urbina 2002) proposed statistical hypothesis testing-based techniques for assessment of structural dynamic responses to determine whether the computer output matches with the real system output, considering data uncertainty. Classical hypothesis testing is a well-developed statistical method for rejecting a model based on an error statistic (see e.g., Paez and Urbina 2002; Hills and Trucano 2001 and 2002; Hills and Leslie 2003; Easterling 2003; Rutherford and Dowding 2003; Chen *et al.* 2004; Dowding *et al.* 2004; Oberkampf and Barone 2006). In general, this method consists of two main elements: feature extraction and quantitative assessment. The objectives of feature extraction (or metrics selection) are to reduce the dimensionality of the data used for model assessment, and to improve the efficiency and accuracy of model validation. Numerous features have been used in model validation of dynamic systems, such as relative error, principal component decomposition, and frequency response functions. Details of various features extracted from time history data were reported by Hemez and Doebling (2000). More recently, Sarin *et al.* (2008) and Schwer (2007) evaluated two categories of metrics for model validation of dynamic systems, namely, Sprague and Geers metric and Knowles and Gear metric. In these metrics, the

magnitude and phase features are calculated based on the time integration of the wave forms. All these methods, however, aim to quantitatively compare two curves (i.e., model prediction and test data) for a single response. They can not handle multivariate correlation directly, which may not be effective for multivariate model validation of complex dynamic systems, such as vehicle crash tests.

Recently, Bayesian methods have been developed to determine the predictive capabilities of computer models (see e.g., Kennedy and O'Hagan 2001; Zhang and Mahadevan 2003; Chen *et al.* 2004&2008; Mahadevan and Rebba 2005; Rebba 2005; Rebba and Mahadevan 2006; Jiang and Mahadevan 2007, 2008a, b & c). Rebba (2005) compared the Bayesian methods with other statistical validation metrics and approaches in terms of accuracy and adequacy requirements, ease of implementation, etc. One important difference between the Bayesian and classical hypothesis testing approaches is that, the Bayesian approach focuses on model acceptance whereas classical hypothesis testing focuses on model rejection. In particular, Bayesian hypothesis testing-based methods (Zhang and Mahadevan 2003) and Bayesian risk-based decision making methodology (Jiang and Mahadevan 2007&2008a) were explored for model assessment. These metrics were investigated with various model validation problems using limited amounts of experimental data. Details of the validation methods and their applications can be found in Mahadevan and Rebba (2005) and Jiang and Mahadevan (2007, 2008a, b & c).

In this study, a quantitative multivariate validation method is proposed. It exploits two advanced state-of-the-art statistical techniques: Probabilistic Principal Component Analysis (PPCA) and multivariate Bayesian hypothesis testing. The PPCA approach is developed to address three issues: multivariate correlation, data uncertainty, and dimensionality reduction. A multivariate Bayesian hypothesis testing method is developed to quantitatively assess the CAE model for dynamic systems, considering data uncertainty.

PROBABILISTIC PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) (Hotelling 1933) is a well-established statistical method for dimensionality reduction and has been widely applied in data compression, image processing, exploratory data analysis, pattern recognition, and time series prediction (Jolliffe 2002). PCA involves a matrix analysis technique called eigenvalue decomposition. The decomposition produces eigenvalues and eigenvectors representing the amount of variation accounted for by the principal

component and the weights for the original variables, respectively. Its main objective is to transform a set of correlated high dimensional variables to a set of uncorrelated lower dimensional variables (principal components). The important property of PCA is that the principal component projection minimizes the squared reconstruction error in dimensionality reduction. As the PCA is not based on a probabilistic model, it cannot handle data uncertainty. Tipping and Bishop (1999) proposed a probabilistic principal component analysis method to address this issue. In this study, the PPCA approach is developed to (1) reduce data dimensionality, (2) address the multivariate correlation, and (3) consider data uncertainty.

The PPCA is derived from a Gaussian latent variable model which is closely related to statistical factor analysis. The factor analysis is a mathematical technique widely used to reduce the number of variables (dimensionality reduction), while identifying the underlying factors that explain the correlations among multiple variables (Cooley and Lohnes 1971, Morrison 1976). For the convenience of formulation, let $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_N]^T$ represent the $N \times D$ data matrix (either model prediction or experimental measurement in the context of model validation) with $\mathbf{y}_i \in \mathbb{R}^D$, which represents D observable variables each containing N data points. Let $\Phi = [\theta_1, \dots, \theta_N]^T$ be the $N \times d$ data matrix with $\theta_i \in \mathbb{R}^d$ ($d \leq D$) representing d latent variables (factors) that cannot be observed, each containing the corresponding N positions in the latent space. The latent variable model relates the correlated data matrix $\mathbf{Y}$ to the corresponding uncorrelated latent variable matrix Φ , expressed as

$$\mathbf{y}_i = \mathbf{W}\theta_i + \boldsymbol{\mu} + \boldsymbol{\varepsilon}_i, i = 1, 2, \dots, N, \quad (1)$$

where the $D \times d$ weight matrix $\mathbf{W}$ describes the relationship between the two sets of variables $\mathbf{y}_i$ and θ_i , the parameter vector $\boldsymbol{\mu}$ consists of D mean values obtained from the data matrix $\mathbf{Y}$, i.e. $\boldsymbol{\mu} = (1/N) \sum_{i=1}^N \mathbf{y}_i$, and the D -dimensional vector $\boldsymbol{\varepsilon}_i$ represents the error or noise in each variable $\mathbf{y}_i$, usually assumed to consist of independently distributed Gaussian variables with zero mean and unknown variance ψ .

Tipping and Bishop (1999) derived a probabilistic principal component analysis from the statistical factor analysis, with an isotropic noise covariance $\sigma^2 \mathbf{I}$ assumed for the variance ψ . In particular, it was shown that, with the Gaussian distribution assumption for the latent variables, the maximum likelihood estimator for $\mathbf{W}$ spans the principal subspace of the data even when the

σ^2 is non-zero. The use of the isotropic noise model $\sigma^2 \mathbf{I}$ makes the probabilistic PCA technically distinct from the classical factor analysis. The former is covariant under rotation of the original data axes, while the latter is covariant under component-wise rescaling. In addition, the principal axes in the PPCA are in the incremental order (Tipping and Bishop 1999), which can not be realized by the factor analysis.

Note that, within the context of model validation, the test or model prediction may be repeated, or each response of interest may be measured or simulated for more than one time. In such situation, the measurement or prediction error corresponding to each variable can be quantified by statistical data analysis, yielding an additional error vector $\boldsymbol{\varepsilon}_i^*$. The additional error is also assumed to be independently distributed Gaussian variables with zero mean and variance Λ , i.e., $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \Lambda)$, in which Λ is a diagonal matrix, each diagonal element representing the data uncertainty of the corresponding variable. As such, the data matrix $\mathbf{Y}$ in the subsequent analysis becomes the time-dependent mean value of the data for each variable.

The latent variables θ_i in Eq. (1) are conventionally defined to be independently distributed Gaussian variables with zero mean and unit variance, i.e. $\theta_i \sim N(0, \mathbf{I})$. From Eq. (1), the observable variable $\mathbf{y}_i$ can be written in the Gaussian distribution form as

$$\mathbf{y}_i / (\theta_i, \mathbf{W}, \psi) \sim N(\mathbf{W}\theta_i + \boldsymbol{\mu}, \psi), \quad (2)$$

where $\psi = \Lambda + \sigma^2 \mathbf{I}$ combines the measurement or prediction error Λ unique to the response and the variability σ^2 unique to θ_i (the isotropic noise covariance).

It should be pointed out that the latent variables θ_i in the PPCA are intended to explain the correlations between observed variables $\mathbf{y}_i$, while the error variables $\boldsymbol{\varepsilon}_i$ represents the variability unique to θ_i . This is different from the standard PCA which treats covariance and variance identically. The variance of the d -dimensional data matrix is calculated by

$$\Sigma_{ML}^{-1} = \mathbf{I} + \mathbf{W}_{ML}^T \psi_{ML}^{-1} \mathbf{W}_{ML}, \quad (3)$$

where $\psi_{ML} = \Lambda + \sigma_{ML}^2 \mathbf{I}$, and σ_{ML}^2 and $\mathbf{W}_{ML}$ are the maximum likelihood estimates for σ^2 and $\mathbf{W}$ respectively. The unadjusted data in the lower dimensional latent space, $\Phi^* = [\theta_1^*, \dots, \theta_N^*]^T$, is obtained as: