American Journal of Mathematical and Computational Sciences
2018; 3(1): 50-61
http://www.aascit.org/journal/ajmcs

AASCIT American Association for Science and Technology

Probabilistic Reduced Order Modeling Using a Bayesian Approach

Indika Udagedara[1], Brian Todd Helenbrook[1, 2], Aaron Luttman[3], Jared Catenacci[3]

[1]Department of Mathematics, Clarkson University, New York, United States of America
[2]Department of Mechanical and Aeronautical Engineering, Clarkson University, New York, United States of America
[3]Signal Processing and Applied Mathematics, Nevada National Security Site, Las Vegas, United States of America

Email address

udagedig@clarkson.edu (I. Udagedara), bhelenbr@clarkson.edu (B. T. Helenbrook), LuttmaAB@nv.doe.gov (A. Luttman), CatenaJW@nv.doe.gov (J. Catenacci)

Citation

Indika Udagedara, Brian Todd Helenbrook, Aaron Luttman, Jared Catenacci. Probabilistic Reduced Order Modeling Using a Bayesian Approach. *American Journal of Mathematical and Computational Sciences.* Vol. 3, No. 2, 2018, pp. 50-61.

Received: February 22, 2018; **Accepted**: April 24, 2018; **Published**: May 31, 2018

Abstract: A method for probabilistic reduced order modeling (ROM) is developed for stochastic problems. Probabilistic principal component analysis (PPCA) was modified to generate a basis for the reduced order model from training data, in such a way that it allows the noise in the training data to be estimated and also determines the variance of the latent variables. This variance information is then used as a prior in a new probabilistic data projection approach. Together these techniques give a fully probabilistic method for creating ROMs that allow accurate predictions of noise-free data from data that is dominated by noise.

Keywords: Bayesian Parameter Estimation, Model Selection, Noise Reduction, Probabilistic Principal Component Analysis, Reduced Order Modeling

1. Introduction

The goal of this work is to develop reduced order modeling techniques for problems with stochastic noise. The main idea is to use low-noise training data to generate basis functions that accurately represent the system response, then, given a noisy data measurement, project that data onto the basis to generate an accurate estimate of the noise-free data. In our previous work [1], we showed that noisy Monte Carlo simulation data from radiation transport simulations could be significantly improved by the projection process. Although that work demonstrated the potential of the approach, many of the steps needed to make the process practical were incomplete. Specifically, the formulation in [1] did not properly account for noise in the generation of the basis functions, had no mechanism for selecting the number of basis functions in the ROM, and used L_2 projection of the trial data without mathematical justification. Accurate results were obtained essentially by knowing the correct answer and then choosing the model parameters such that the error was minimized. This approach obviously is not feasible for any practical problem.

In this work, a probabilistic approach is used to remedy these deficiencies. To generate the basis, the stochastic formulation of the principal component analysis (PCA) [2-8], known as probabilistic principal component analysis (PPCA) [9-14], is used. PPCA identifies the noise in the training data, which cannot be estimated with conventional PCA. We have also improved the formulation of PPCA to remove the unnecessary assumption that all latent variables have unit variance. The new formulation provides more physical insight into the eigenvalues of the PPCA and their relation to the variance of the latent variables. This information was necessary for the new projection procedure that was developed.

The probabilistic formulation also provides a method for selecting the number of basis functions to include in the ROM. This can be done by combining PPCA with a Bayesian model selection [15-24] criterion. Here the Bayesian information criterion (BIC) [15, 17, 24-27] is used to identify the optimal number of basis functions to include in the ROM, and it is demonstrated that this approach reliably chooses the number of basis functions that can be identified given that the training data alsoincludes noise.

Lastly, a new approach for projecting the trial data is derived. This approach uses prior information obtained from the PPCA of the training data to improve the projection of the trial data. In our previous work, L_2 projection was used, which basically corresponds to a projection with no prior knowledge of the projection coefficients. In the following, it is

demonstrated that using the training data to define a prior for the projection coefficients leads to significantly improved results when the trial data has more noise than the training data, which is typically the case. Together, these techniques define a method for reduced order modeling of problems with stochastic noise that can be applied to practical scenarios.

The paper is organized as follows. First the framework for the reduced order model generation and the modified PPCA approach is derived. This is then followed by a discussion of the model selection approach (BIC) and the derivation of the method for projecting the trial data. To demonstrate the benefits of this new probabilistic formulation, the ROM procedure is applied to a simple stochastic model problem, and predictions from the ROM are compared to our previous approach and to noise-free data, which was known for the model problem.

2. Reduced Order Modeling Formulation

The ROM is formulated assuming the data is created by a process of the form

$$\vec{y} = \sum_{j=1}^{m} w_j \vec{\varphi}_j + \vec{\mu} + \vec{\varepsilon} = \Phi \vec{w} + \vec{\mu} + \vec{\varepsilon}, \tag{1}$$

where $\vec{y}$ is a single data realization for a particular physical scenario, which could be obtained either from a numerical simulation or an experiment. The dimension of $\vec{y}$ is d, which depends on the physical problem being studied. The $\vec{\varphi}_j,\ j \in [1,m]$, are basis functions that are scenario independent, and the w_j's are latent variables that vary from realization to realization. The w_j's being random variables implies that there is a probability associated with the occurrence of each physical scenario. The mean of the latent variables is assumed to be zero such that $\vec{\mu}$ is the mean of the data over all scenarios, making $\vec{\mu}$ a scenario-independent constant vector. The noise in the process is $\vec{\varepsilon}$, also a random variable, which could represent noise in the experimental measurements or in the stochastic numerical simulation approach used to generate the data. The noise is assumed to be generated by a zero-mean Gaussian process with covariance $\sigma_\varepsilon^2 I$, where I is the identity matrix of dimension d.

Although the data is assumed to be generated by a process of the form given by (1), none of the parameters of the model ($\Phi, \vec{\mu}, m$) are known. The first step of the reduced order modeling process is to generate a set of "training data" that can be used to estimate these parameters. The training data is a set $Y = \{\vec{y}_k\}$, for $k = 1,2,...,n$, of realizations of the process. The scenarios associated with these realizations are chosen randomly according to the probability density function predicting the occurrence of any given scenario. The generation of the data can be through either numerical simulation or experiment, and both are assumed to also include random noise.

2.1. PPCA

To estimate Φ and $\vec{\mu}$ given Y, PPCA is used. Our formulation of PPCA is an improvement on the standard derivation (given in Refs. [10, 11]), where it is assumed that the latent variables, $\vec{w}$, are uncorrelated and follow a Gaussian distribution with unit covariance. In the following, it is also assumed that the latent variables are uncorrelated and follow a Gaussian distribution, but the covariance is not a-priori assumed to be 1. Instead the variances, $\{\sigma_{w_i}^2\}_{i=1}^m$, are estimated as part of the derivation. In the following, a condensed derivation is given that highlights the main differences between the new and original derivation.

Bayes' formula to estimate the unknowns, Φ, σ_ε^2, $\sigma_{w_i}^2$, and $\vec{\mu}$ in the model is

$$p\left(\Phi, \sigma_\varepsilon^2, \sigma_{w_i}^2, \vec{\mu} \mid Y\right) \propto p\left(Y \mid \Phi, \sigma_\varepsilon^2, \sigma_{w_i}^2, \vec{\mu}\right) p\left(\Phi, \sigma_\varepsilon^2, \sigma_{w_i}^2, \vec{\mu}\right), \tag{2}$$

where $p\left(\Phi, \sigma_\varepsilon^2, \sigma_{w_i}^2, \vec{\mu} \mid Y\right)$ is the posterior distribution, $p\left(\Phi, \sigma_\varepsilon^2, \sigma_{w_i}^2, \vec{\mu}\right)$ is the prior distribution and $p\left(Y \mid \Phi, \sigma_\varepsilon^2, \sigma_{w_i}^2, \vec{\mu}\right)$ is the likelihood distribution. PPCA uses a maximum likelihood estimator (MLE) [28-31] to find the unknown parameters assuming no prior knowledge about their values i.e. the prior is assumed to be uniform. The MLE is thus obtained by maximizing the log-likelihood function, $\log p\left(Y \mid \Phi, \sigma_\varepsilon^2, \sigma_{w_i}^2, \vec{\mu}\right)$.

In order to obtain the likelihood function, the probability distribution of an individual realization, $\vec{y}$, conditioned on $\Phi, \sigma_\varepsilon^2, \sigma_{w_i}^2, \vec{\mu}$, first needs to be identified. This distribution, $p(\vec{y} \mid \Phi, \sigma_\varepsilon^2, \sigma_{w_i}^2, \vec{\mu})$, is called the predictive distribution and can be obtained using the following relations

$$p(\vec{y} \mid \Phi, \sigma_\varepsilon^2, \sigma_{w_i}^2, \vec{\mu}) = \int_{-\infty}^{\infty} p(\vec{y}, \vec{w} \mid \Phi, \sigma_\varepsilon^2, \sigma_{w_i}^2, \vec{\mu}) d\vec{w} = \int_{-\infty}^{\infty} p(\vec{y} \mid \vec{w}, \Phi, \sigma_\varepsilon^2, \vec{\mu}) p(\vec{w} \mid \sigma_{w_i}^2) d\vec{w}, \tag{3}$$

where the integral is taken over all components of the vector $\vec{w}$. Assuming the noise in (1) is Gaussian with zero mean, the probability distribution of $\vec{y}$ conditioned on the latent variable, $\vec{w}$, and the parameters σ_ε^2, $\vec{\mu}$, and Φ is given by

$$p\left(\vec{y} \mid \vec{w}, \Phi, \sigma_\varepsilon^2, \vec{\mu}\right) = \mathsf{N}\left(\Phi \vec{w} + \vec{\mu}, \sigma_\varepsilon^2 I\right), \tag{4}$$

where the notation $\mathsf{N}\left(\Phi\vec{w}+\vec{\mu},\sigma_\varepsilon^2 I\right)$ indicates a Gaussian distribution with mean $\Phi\vec{w}+\vec{\mu}$ and co-variance matrix $\sigma_\varepsilon^2 I$.

The latent variables are assumed to be uncorrelated and follow a zero-mean Gaussian distribution with a covariance of $\sigma_{w_i}^2$, thus $p(\vec{w}\,|\,\sigma_{w_i}^2)=\mathsf{N}(\vec{0},\Sigma^2)$, where Σ^2 is an $m\times m$ diagonal matrix with the values $\sigma_{w_i}^2$ on the diagonal. Based on this assumption, one can show that the predictive distribution is Gaussian of the form $\mathsf{N}\left(\vec{\mu},\Phi\Sigma^2\Phi^T+\sigma_\varepsilon^2 I\right)$.

The likelihood distribution, $p\left(Y\,|\,\Phi,\sigma_\varepsilon^2,\Sigma^2,\vec{\mu}\right)$ for the training data, $Y=\{\vec{y}_k\}$, for $k=1,2,...,n$ is the product of the individual predictive distributions. The log likelihood can be shown to be

$$L=\log p\left(Y\,|\,\Phi,\sigma_\varepsilon^2,\Sigma^2,\vec{\mu}\right)=-\frac{dn}{2}\log 2\pi-\frac{n}{2}\log|\,\Phi\Sigma^2\Phi^T+\sigma_\varepsilon^2 I\,|-\frac{1}{2}\sum_{k=1}^{n}\left(\vec{y}_k-\vec{\mu}\right)^T\left(\Phi\Sigma^2\Phi^T+\sigma_\varepsilon^2 I\right)^{-1}\left(\vec{y}_k-\vec{\mu}\right), \tag{5}$$

where $|\cdot|$ denotes the determinant of a matrix.

As the prior is uniform, the most probable values of the posterior distribution can be determined by maximizing L with respect to the unknown parameters, Φ, $\vec{\mu}$, σ_ε^2. In all the following, the subscript MP indicates these most probable values. The maximization process gives the following estimate for the mean $\vec{\mu}$,

$$\vec{\mu}_{MP}=\frac{1}{n}\sum_{k=1}^{n}\vec{y}_k. \tag{6}$$

The estimate of σ_ε^2 is

$$\sigma_{\varepsilon(MP)}^2=\frac{1}{d-m}\sum_{i=m+1}^{d}\lambda_i, \tag{7}$$

where the λ_i's are the eigenvalues of the data covariance matrix, $S=\frac{1}{n}\sum_{k=1}^{n}\left(\vec{y}_k-\vec{\mu}_{MP}\right)\left(\vec{y}_k-\vec{\mu}_{MP}\right)^T$. The maximum likelihood estimate for σ_ε^2 can be interpreted as the average magnitude of the eigenvalues of dimension greater than m. These eigenvalues can only be caused by noise, as there are only m latent variables.

Minimizing (5) with respect to Φ, one obtains

$$\Phi_{MP}\Sigma_{MP}=U\left(\Lambda-\sigma_{\varepsilon(MP)}^2 I\right)^{\frac{1}{2}}R, \tag{8}$$

where U is a $d\times m$ matrix whose columns are given by a complete subset of (orthonormal) eigenvectors of the data covariance matrix S, Λ is the $m\times m$ diagonal matrix consisting of the first m largest eigenvalues of S, and R is an arbitrary $m\times m$ orthonormal matrix. This equation is almost the same as in [10, 11], except for the Σ term. In [10, 11], R was chosen to be the identity matrix, which then determined Φ. The disadvantage of this choice is that the column vectors of Φ then each must have a magnitude determined by the diagonal matrix $\left(\Lambda-\sigma_{\varepsilon(MP)}^2 I\right)^{\frac{1}{2}}$. This scaling of the basis functions is necessary to ensure that the latent variables all have a variance of unity. In the new formulation, we can satisfy (8) by choosing

$$\Phi_{MP}=U \tag{9}$$

and the estimate for Σ as

$$\Sigma_{MP}=\left(\Lambda-\sigma_{\varepsilon(MP)}^2 I\right)^{\frac{1}{2}}. \tag{10}$$

This allows us to have standard orthonormalbasis functions, and also correctly identifies the variance of the latent variables (as we confirm in the example problem below). Manipulating (10), the relation can put in the following form

$$\Lambda=\Sigma_{MP}^2+\sigma_{\varepsilon(MP)}^2 I. \tag{11}$$

This shows that the eigenvalues of the covariance matrix S are the variance of latent variables summed with the variance of the measurement error. The first m eigenvalues consists of both variances, however the eigenvalues greater than m are strictly due to random measurement error.

2.2. Model Selection

In the previous section, the model parameters were estimated assuming that the dimension of the ROM, m, was a known, fixed number. PPCA together with Bayesian model selection criteria can be used to predict the number of basis functions required for the ROM. There are many different Bayesian model selection criteria [17, 18, 23, 24, 32-35], and here we choose the Bayesian information criterion (BIC) [15, 17, 24-27]. BIC selectsthe value of m that minimizes

$$f_{BIC}(m)=-2L_{MP}+\left(m\left(d-1-\frac{m-1}{2}\right)+d+1\right)\log n, \tag{12}$$

where L_{MP} is the maximum value of the likelihood distribution and the term in the outer parentheses in (12) is the number of estimatable parameters in the model, both of which depend on m. The number of estimatable parameters arise from Φ_{MP}, $\vec{\mu}_{MP}$, and $\sigma_{\varepsilon_{MP}}^2$. There are

$(d-1)+(d-2)+(d-3)+...+(d-1-(m-1))$ parameters in Φ_{MP}. Here the $d-1$ comes from the fact that the first basis vector is required to be normalized to have magnitude 1 so when $m=1$, one can only choose $d-1$ independent variables. Because of the requirement of orthogonality the number of free parameters in choosing a basis vector decreases by 1 for each additional basis vector. This results in the number of free parameters in Φ being $m(d-1-(m-1)/2)$. The number of parameters in $\vec{\mu}_{MP}$ and $\sigma^2_{\varepsilon_{MP}}$ are d and 1, respectively, giving the total shown in parentheses above.

L_{MP} is obtained by inserting the maximum likelihood estimates of the parameters (6), (7), (9), and (10) into (5). Following the simplification techniques in [27] but with our maximum likelihood results, this becomes

$$L_{MP} = -\frac{dn}{2}\log(2\pi) - \frac{n}{2}\left(\sum_{j=1}^{m}\log(\lambda_j) + (d-m)\log\left(\frac{1}{d-m}\sum_{j=m+1}^{d}\lambda_j\right) + d\right). \tag{13}$$

To find the most probable value for m, $f_{BIC}(m)$ is calculated as a function of m, $m \in [1,d]$, and the value of m that minimizes the function is chosen.

2.3. Bayesian Projection with Gaussian Prior

The above sections determined the model parameters, Φ, $\vec{\mu}$ and m, which is to say that the reduced order model is now constructed from thetraining data. In this section, a latent variable vector $\vec{w}$ is estimated given a "trial" data vector $\vec{y}$ that is obtained from a new scenario drawn from the distribution of scenario probabilities, i.e. we now project trial data onto the ROM. A trial vector $\vec{y}$ includes noise, ε_T, which is drawn from a zero-mean Gaussian distribution $\mathsf{N}(0, \sigma^2_{\varepsilon_T})$, and the magnitude of this noise, $\sigma^2_{\varepsilon_T}$, is assumed to be different (typically larger) than that of the training data. In our previous work L_2 projection of the trial data was used to estimate the latent variables and no estimate was given for $\sigma^2_{\varepsilon_T}$; here the latent variables are estimated using Bayesian parameter estimation with a Gaussian prior. The estimate of Σ^2_{MP} obtained from the training data is used as the covariance of the prior on the latent variables.

To estimate the probability distribution of $\vec{w}$ and $\sigma^2_{\varepsilon_T}$, conditioned on the observed data, $\vec{y}$, and parameters $\vec{\mu}$, Σ^2, and Φ, Bayes' theorem is used. Applying Bayes' theorem, assuming that Σ^2 and $\sigma^2_{\varepsilon_T}$ are independent, we have

$$\begin{aligned} p\left(\vec{w},\sigma^2_{\varepsilon_T} \mid \vec{y},\Phi,\Sigma^2,\vec{\mu}\right) &\propto p\left(\vec{y} \mid \vec{w},\sigma^2_{\varepsilon_T},\Phi,\Sigma^2,\vec{\mu}\right) p\left(\vec{w},\sigma^2_{\varepsilon_T} \mid \Phi,\Sigma^2,\vec{\mu}\right) \\ &\propto p\left(\vec{y} \mid \vec{w},\sigma^2_{\varepsilon_T},\Phi,\vec{\mu}\right) p\left(\vec{w} \mid \Sigma^2\right) p\left(\sigma^2_{\varepsilon_T}\right). \end{aligned} \tag{14}$$

Assuming the model given by (1) holds for the trial data as well, the probability distribution of $\vec{y}$ is

$$p\left(\vec{y} \mid \vec{w},\Phi,\vec{\mu},\sigma^2_{\varepsilon_T}\right) \propto |\sigma^2_{\varepsilon_T} I|^{-1/2} \exp\left(-\frac{(\vec{y}-\Phi\vec{w}-\vec{\mu})^T(\vec{y}-\Phi\vec{w}-\vec{\mu})}{2\left(\sigma^2_{\varepsilon_T}\right)}\right). \tag{15}$$

As assumed before, the probability of $\vec{w}$ for a given scenario is Gaussian with mean zero and covariance Σ^2

$$p\left(\vec{w} \mid \Sigma^2\right) \propto |\Sigma^2|^{-1/2} \exp\left(-\frac{1}{2}\vec{w}^T\Sigma^{-2}\vec{w}\right), \tag{16}$$

so, assuming a uniform prior distribution for $\sigma^2_{\varepsilon_T}$, the log posterior is obtained from (14), (15), and (16) as

$$\log p\left(\vec{w},\sigma^2_{\varepsilon_T} \mid \vec{y},\Phi,\Sigma^2,\vec{\mu}\right) \propto \frac{1}{2}\log|\sigma^2_{\varepsilon_T} I| + \frac{(\vec{y}-\Phi\vec{w}-\vec{\mu})^T(\vec{y}-\Phi\vec{w}-\vec{\mu})}{2\sigma^2_{\varepsilon_T}} + \frac{1}{2}\log|\Sigma^2| + \frac{1}{2}\vec{w}^T\Sigma^{-2}\vec{w} \tag{17}$$

Setting the derivatives of (17) with respect to $\vec{w}$ and $\sigma^2_{\varepsilon_T}$ to zero to find the maxima gives

$$\vec{w}_{MP} = \left(I + \sigma^2_{\varepsilon_{T_{MP}}}\Sigma^{-2}\right)^{-1}\Phi^T(\vec{y}-\vec{\mu}) \tag{18}$$

and

$$\sigma^2_{\varepsilon_{T_{MP}}} = \frac{1}{d}\left(\vec{y} - \Phi\vec{w}_{MP} - \vec{\mu}\right)^T \left(\vec{y} - \Phi\vec{w}_{MP} - \vec{\mu}\right). \quad (19)$$

Equations (18) and (19) are a system of non-linear equations with unknowns $\vec{w}_{MP}$ and $\sigma^2_{\varepsilon_{T_{MP}}}$. The pair of equations can be solved using a fixed point iteration, where it is first assumed that $\sigma^2_{\varepsilon_{T_{MP}}}$ is zero, then (18) is used to calculate $\vec{w}_{MP}$. Equation (19), which can be interpreted as a calculation of the noise in the data assuming the true data is given by $\Phi\vec{w}_{MP} + \vec{\mu}$, can then be used to calculate $\sigma^2_{\varepsilon_{T_{MP}}}$. This new value of $\sigma^2_{\varepsilon_{T_{MP}}}$ is used in (18) and the process is repeated until (18) and (19) are satisfied to a specified tolerance.

If $\sigma^2_{\varepsilon_{T_{MP}}}$ is small relative to the values of $\sigma^2_{w_i}$, then the matrix in parentheses in (18) is essentially the identity matrix and the L_2 projection result, $\vec{w} = \Phi^T\left(\vec{y} - \vec{\mu}\right)$, is recovered. This is the approach that was used in our previous work and is also the result that would be obtained assuming a uniform prior on $\vec{w}$ instead of a Gaussian prior. However, when the noise in the data is large relative to the variance of a latent variable, i.e. $\sigma^2_{\varepsilon_{T_{MP}}} / \sigma^2_{w_i}$ is large, the Gaussian-prior projection reduces the magnitude of the L_2 projection value of w_i to account for the fact that the noise in the data is causing an estimation for w_i that is larger than the expected variation of that latent variable. We show in the following section that this can significantly improve the accuracy of the projection for these conditions.

3. Demonstration

This section illustrates the above discussed ROM techniques for a model problem where the data is generated using a model of the form given by (1) i.e. the data is generated as a linear combination of a finite number of basis functions and latent variables with a mean vector and added random noise. As all of the parameters of the data generation are known, the ROM process can be validated by comparing the estimated parameters to those used to generate the data.

The q basis functions used to generate the data are discrete sine waves given by

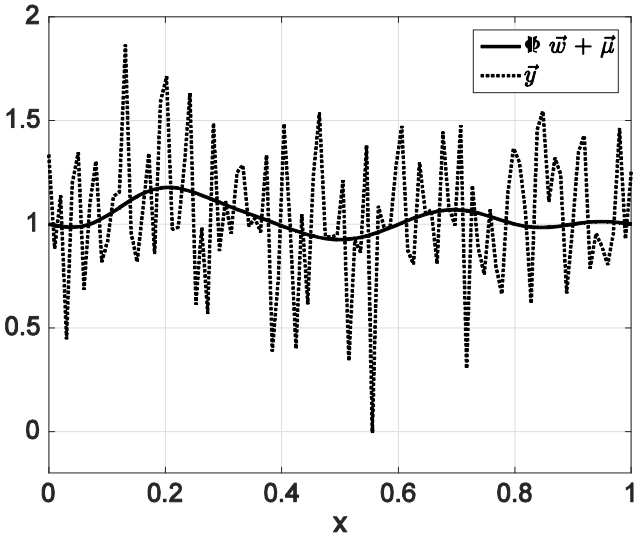
$$\vec{\varphi}_j = \frac{\sin\left(j\pi\vec{x}\right)}{\|\sin\left(j\pi\vec{x}\right)\|} \quad \text{for} \quad j \in [1, q],$$

where $\vec{x} \in \mathsf{R}^d$ is a vector of $d = 100$ uniformly spaced points from the domain [0,1] including endpoints. In the above, the norm $\|.\|$ is the Euclidean vector norm such that $\vec{\varphi}_j \cdot \vec{\varphi}_j = 1$. With this normalization, the peak value of the basis functions is $1/\sqrt{(d/2)} \approx 0.14$.

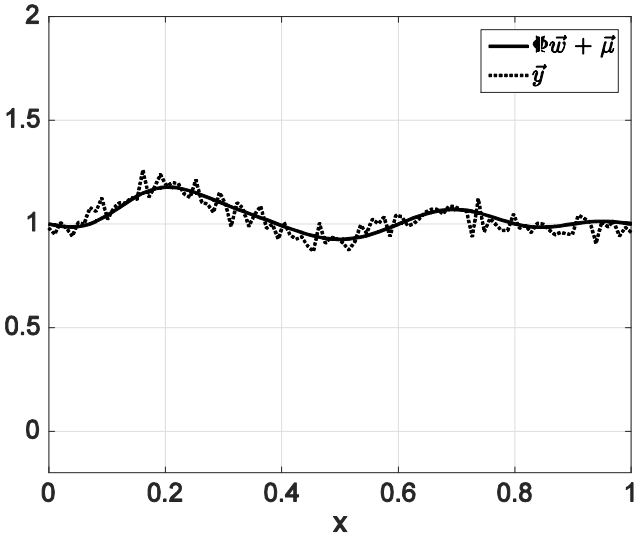
The latent variables, w_j, were sampled from Gaussian distributions with variances of $\sigma^2_{w_j}$, chosen to be

$$\sigma^2_{w_j} = \frac{1}{2^{j-1}} \quad \text{for} \quad j \in [1, q].$$

Unless stated otherwise, $q = 10$ basis functions and latent variables were used to create the data. The mean, $\vec{\mu}$, was a vector of ones.



(a)



(b)

Figure 1. *A single realization from the training data (a)* σ^2_ε *= 1/10 (b)* σ^2_ε *= 1/400..*

The noise vector, $\vec{\varepsilon}$ was also sampled from a Gaussian

distribution, either $\mathsf{N}\left(\vec{0},\sigma_\varepsilon^2 I\right)$ for the training data or $\mathsf{N}\left(\vec{0},\sigma_{\varepsilon_T}^2 I\right)$ for the trial data. The data generation process was repeated $n=10000$ times to generate the training data set Y. To investigate the effect of noise in the training data, two training data sets were studied, one with $\sigma_\varepsilon^2=1/10$ and the other with $\sigma_\varepsilon^2=1/400$. Figure 1 shows a typical realization of a data vector from the two training data sets. From figure 1(a) it can be seen that, in the case of $\sigma_\varepsilon^2=1/10$, the noise in the data is significant with a magnitude on the order $\pm\sqrt{1/10}\approx 0.32$. Similarly in figure 1(b), the noise magnitude is $\pm\sqrt{1/400}=0.05$.

Averaging over all of the data vectors of the data set Y determines the most probable mean vector $\vec{\mu}_{MP}$. This is shown in Figure 2 for both data sets. The fluctuations in the mean should scale as $\left(\sum_i \sigma_{w_i}+\sigma_\varepsilon\right)/\sqrt{n}$, which is equal to 0.0088 and 0.0061 for $\sigma_\varepsilon^2=1/10$ and $1/400$, respectively. This is in good agreement with what is observed in Figure 2. The dominant source of error in determining the mean is not the noise in the data, but rather determining the average outcome of the scenarios. For this reason, the fluctuations in both estimated means are of similar magnitude.

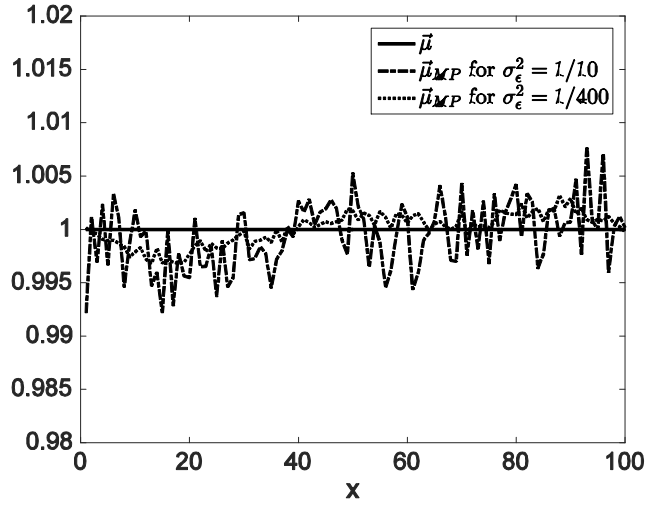


Figure 2. *Mean distribution for the $\sigma_\varepsilon^2=1/10$ and $\sigma_\varepsilon^2=1/400$ training data.*

3.1. PPCA

According to (9) in Section 2.1, the estimates for the most probable basis functions are the dominant subset of m eigenvectors of the data covariance matrix, $S=\frac{1}{n}\sum_{k=1}^{n}\left(\vec{y}_k-\vec{\mu}_{MP}\right)\left(\vec{y}_k-\vec{\mu}_{MP}\right)^T$. Figure 3 shows the eigenvalue spectrum for the two sets of training data with $\sigma_\varepsilon^2=1/10$ and $1/400$. Based on (11), the eigenvalues are expected to decay like $1/2^{i-1}$ because of the $\sigma_{w_i}^2$ term and then plateau at a value of σ_ε^2. The predicted eigenvalues based on (11) are also shown on the plot. The curves agree well indicating that the new formulation of PPCA accurately predicts the dependence of the eigenvalues on both $\sigma_{w_i}^2$ and σ_ε^2.

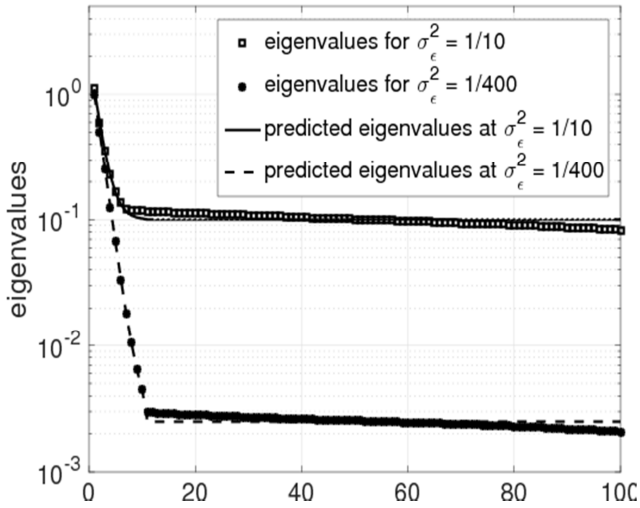


Figure 3. *Eigenvalue spectra of the data covariance matrices created using the $\sigma_\varepsilon^2=1/10$ and $\sigma_\varepsilon^2=1/400$ training data. Predicted spectra based on (11) are also shown.*

Figure 4 shows the first, third, fifth, and seventh basis functions for the $\sigma_\varepsilon^2=1/10$ and $\sigma_\varepsilon^2=1/400$ training data. PPCA is able to extract basis functions that are less affected by noise than the actual data; compare the magnitude of the noise in Figure 4(a) for $\sigma_\varepsilon^2=1/10$ with the noise in Figure 1(a). Even for the 7th mode in the case of $\sigma_\varepsilon^2=1/10$, shown in Figure 4(b), which had $\sigma_{w_7}^2=1/2^6=0.016$ much smaller than σ_ε^2, PPCA is able to roughly obtain the correct functional form. This is primarily because of the large number of training data vectors used ($n=10000$), which allows the PPCA to detect the form of the basis in spite of the numerical noise. This indicates that one can obtain accurate basis functions for the ROM by either reducing the noise in the data or by increasing the number of data vectors in the training data.

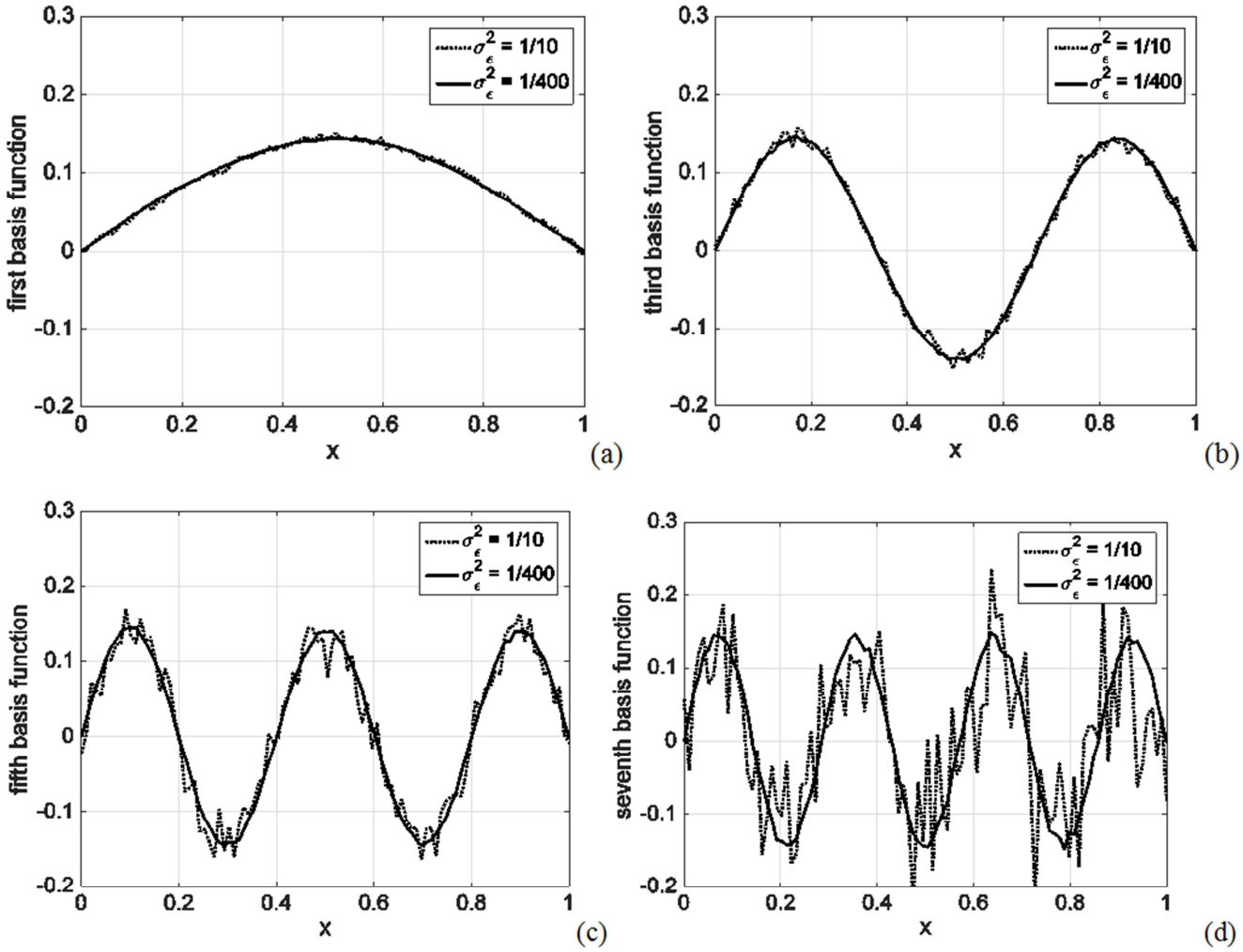
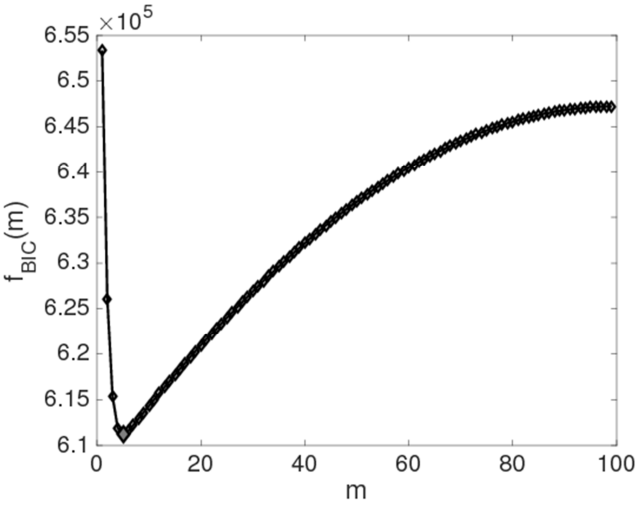


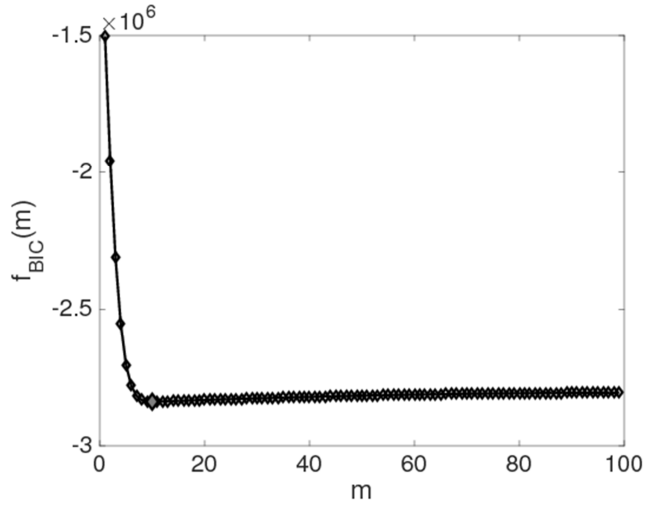
Figure 4. *Eigenvectors 1, 3, 5, 7 generated using training data with* σ_ε^2 *= 1/10 and* σ_ε^2 *= 1/400.*

3.2. Model Selection

As described in Section 2.2, the Bayesian information criterion (BIC) given in (12) is used to determine the dimensionality m. Figure 5 shows $f_{BIC}(m)$ as a function of m for the two training data sets. The minimum of $f_{BIC}(m)$ occurs at $m=5$ and $m=10$ for the $\sigma_\varepsilon^2 = 1/10$ and $\sigma_\varepsilon^2 = 1/400$ training data, respectively. Examining figure 3, it seems that BIC chooses m at the point of the change in decay rate of the eigenvalues. For $\sigma_\varepsilon^2 = 1/10$, this occurs at $m=5$ even though only the first three latent variables have $\sigma_{w_i}^2 > \sigma_\varepsilon^2$. For $\sigma_\varepsilon^2 = 1/400$, this occurs at $m=10$, even though only the first eight latent variables have $\sigma_{w_i} > \sigma_\varepsilon^2$.



(a)



(b)

Figure 5. *BIC as a function of model dimension m (a) for the $\sigma_\varepsilon^2 = 1/10$ training data (minimum is at 5 basis functions) (b) for the $\sigma_\varepsilon^2 = 1/400$ training data (minimum is at 10 basis functions).*

Once the number of latent variables is estimated, (7) can be used to estimate σ_ε^2 and (10) to estimate $\sigma_{w_i}^2$. Table 1 summarizes the estimated and true values of σ_ε^2 and $\sigma_{w_i}^2$ for the two sets of training data. There are only five values of $\sigma_{w_i}^2$ estimated for the training data with $\sigma_\varepsilon^2 = 1/10$, because BIC estimated that $m = 5$ for this data. The first thing to observe from the data is that the estimates of $\sigma_{w_i}^2$ are accurate; the percentage errors are less than 5% in all cases and in most cases the error is less than 2%. This is also true of the estimates of σ_ε^2; the errors are less than 6% for both training data sets. This confirms that the new formulation of PPCA correctly estimates $\sigma_{w_i}^2$.

Another observation from the estimated values of $\sigma_{w_i}^2$ is that the errors do not vary significantly between the σ_ε^2 = 1/400 and 1/10 data sets. This implies that for this particular data set, the main source of error is not the noise in the measurements, but rather the number of data vectors used to create the data set (n). Both data sets used $n = 10,000$ so the accuracy of the estimates of $\sigma_{w_i}^2$ are similar.

Table 1. *Estimated and true values of $\sigma_{w_i}^2$ and σ_ε^2 for the two training data sets.*

	True	σ_ε^2 = 1/400	σ_ε^2 = 1/10
σ_ε^2	-	0.002497	0.100297
$\sigma_{w_1}^2$	1.0000	0.9926	1.0039
$\sigma_{w_2}^2$	0.5000	0.4911	0.4930
$\sigma_{w_3}^2$	0.2500	0.2460	0.2496
$\sigma_{w_4}^2$	0.1250	0.1218	0.1278
$\sigma_{w_5}^2$	0.0625	0.0645	0.0671
$\sigma_{w_6}^2$	0.0313	0.0307	-
$\sigma_{w_7}^2$	0.0156	0.0157	-
$\sigma_{w_8}^2$	0.0078	0.0079	-
$\sigma_{w_9}^2$	0.0039	0.0039	-
$\sigma_{w_{10}}^2$	0.0019	0.0020	-

3.3. Bayesian Projection with Gaussian Prior

The goal of the projection process is to use the developed model to make an accurate estimate of the true signal, $\Phi\vec{w} + \vec{\mu}$, given a "trial" data vector $\vec{y}$ from an unknown scenario ($\vec{w}$) containing random noise $\vec{\varepsilon}$ with unknown magnitude, $\sigma_{\varepsilon_T}^2$. To test the approach described in Section 2.3, two trial data vectors were generated, one with $\sigma_{\varepsilon_T}^2 = 1/5$ and the other with $\sigma_{\varepsilon_T}^2 = 1/200$. The Bayesian projection approach was then used to estimate $\vec{w}$ and $\sigma_{\varepsilon_T}^2$ using (18) and (19). To understand how the training process affects the final results, projections were performed with the 5 and 10 basis function models created above using the $\sigma_\varepsilon^2 = 1/10$ and 1/400 training data respectively. One additional ROM was created using training data with $\sigma_\varepsilon^2 = 1.5\times10^{-12}$ and $m = 50$. BIC selected 42 basis functions for this model.

Projections for four cases were performed. Three cases correspond to trial data with a large noise magnitude ($\sigma_{\varepsilon_T}^2 = 1/5$). For these cases reconstructions were done with the ROMs obtained from the σ_ε^2 = 1/10, 1/400, and 1.5×10^{-12} training data. These cases, allowed us to investigate how the quality of the basis functions and the number of basis functions in the ROM affected the projections. The last case performed a reconstruction using lower noise trial data ($\sigma_{\varepsilon_T}^2 = 1/200$) with the ROM obtained from the σ_ε^2 = 1/400 data. This case was used to determine the effect of the magnitude of noise in the trial data. The combination of low noise in the trial data but high noise in the training data is not of practical interest, because it is assumed that the training data will be of the same or higher quality than the trial data.

Figure 6 shows the projection results. In each figure, the red dotted curve is the ROM projection using a Gaussian prior and the blue dash-dotted curve represents a standard L_2 projection. The gray dash line is the trial data, $\vec{y}$, and the solid black line is the true signal i.e. without the added noise. Figure 6(a) shows the reconstruction of the $\sigma_{\varepsilon_T}^2 = 1/5$ trial data fromthe ROM created using the $\sigma_\varepsilon^2 = 1/10$ training data and Figure 6(b) shows the reconstruction using the ROM

created using the $\sigma_\varepsilon^2 = 1/400$ training data. Both ROMs significantly reduce the noise in the data, but the $\sigma_\varepsilon^2 = 1/400$ ROM produces smoother predictions because of the higher quality eigenvectors obtained with the PPCA. Comparing the L_2 projection to the Gaussian-prior projections shows that the Gaussian prior projections give a closer approximation to the true solution. This is true in both cases, but more so in the $\sigma_\varepsilon^2 = 1/400$ ROM. This is because this ROM has more basis functions, which allows the L_2 projection to better represent the noise in the data and thus increases the deviation from the true signal. This is further verified by Figure 6(c), which shows that with increasing numbers of basis functions in the ROM (42 for this case), the L_2 projection result actually deviates from the true solution whereas the Gaussian projection does not. The projection of the $\sigma_{\varepsilon_T}^2 = 1/200$ data shown in Figure 6(d) shows that as the noise in the trial data is reduced, both projections approach the noise-free signal, but the Gaussian-prior projection is still more accurate and has fewer spurious oscillation that the L_2 projection.

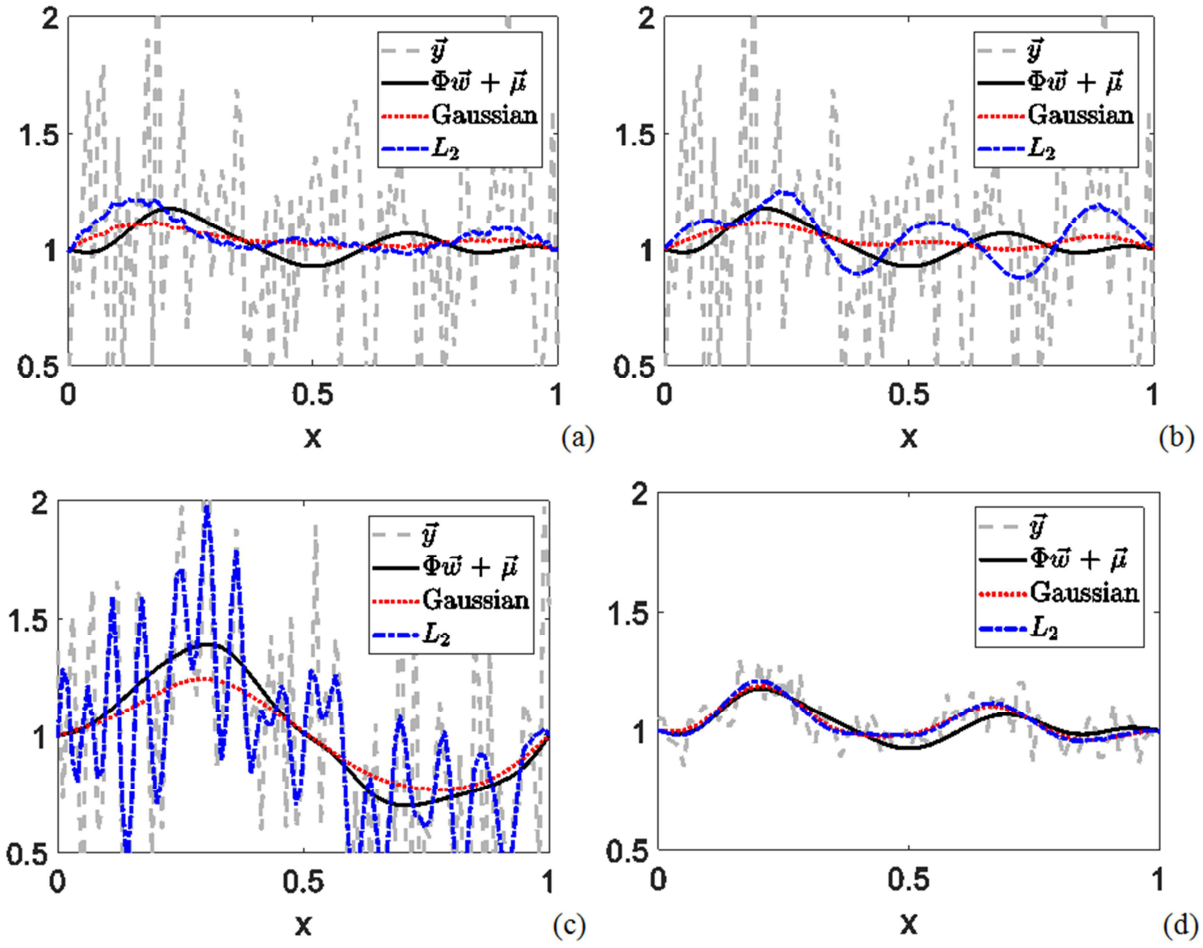


Figure 6. *Realization* $\vec{y}$ *, true solution* $\Phi\vec{w}+\vec{\mu}$ *, Gaussian-prior projection and* L_2 *projection for the cases: (a)* $\sigma_{\varepsilon_T}^2 = 1/5$ *,* $\sigma_\varepsilon^2 = 1/10$ *, (b)* $\sigma_{\varepsilon_T}^2 = 1/5$ *,* $\sigma_\varepsilon^2 = 1/400$ *, (c)* $\sigma_{\varepsilon_T}^2 = 1/5$ *,* $\sigma_\varepsilon^2 = 1.5\times10^{-12}$ *, (d)* $\sigma_{\varepsilon_T}^2 = 1/200$ *,* $\sigma_\varepsilon^2 = 1/400$ *.*

Table 2 shows the true and estimated values for $\sigma_{\varepsilon_T}^2$ for the four trial cases shown in Figure 6. Estimated values are compared for Gaussian and L_2 projection. Note that (19) is used to calculate the $\sigma_{\varepsilon_T}^2$ for both projection approaches. As the number of basis functions in the model increase, the value of $\sigma_{\varepsilon_T}^2$ for the Gaussian-prior projection remains relatively constant while the L_2 projection approach decreases. All of the predicted values using the Gaussian-prior projection are within 3%.

Table 2. *True values and estimated values of the variance of the trial data using Gaussian and L_2 projection.*

Case	True σ_{ε}^2	True $\sigma_{\varepsilon_T}^2$	Predicted $\sigma_{\varepsilon_T}^2$ - Gauss. proj.	Predicted $\sigma_{\varepsilon_T}^2$ - L_2 proj.
a	1/10	1/5	0.2312	0.2294
b	1/400	1/5	0.2264	0.1854
c	1.5×10^{-12}	1/5	0.2067	0.1309
d	1/400	1/200	0.00467	0.00456

To verify that the above identified trends are not particular to the trial data vector examined, for 10000 trial data realizations an error was computed by comparing a given signal, $\vec{y}$, which could be either the trial data vector, the L_2 projection of the trial data, or the Gaussian projection of the trial data, to the true signal. For the above described cases, the error is defined as

$$E = \sqrt{(\vec{y} - \vec{y}_{true})^T(\vec{y} - \vec{y}_{true})} \tag{20}$$

For the trial data, using the form given by eq. 1, the average error is equal to

$$E_{ave} = \int_{-\infty}^{\infty} \sqrt{\vec{\varepsilon}^T\vec{\varepsilon}}\, p(\vec{\varepsilon}) d\vec{\varepsilon},$$

where $p(\vec{\varepsilon})$ is the probability distribution of the noise in the data. The analytical value of the average error for the trial data with $\sigma_{\varepsilon_T}^2 = 1/5$ is 4.4610 and for $\sigma_{\varepsilon_T}^2 = 1/200$ is 0.7053. The error of the projections, which is shown in Figure 7, is normalized with respect to the analytical values of the average error in the trial data. The horizontal red and blue curves in the figure are the mean value of the normalized projection error over the 10000 realizations for the Gaussian-prior projection and the L_2 projection, respectively. For all cases, both projection processes reduced the error. For Gaussian projection, in cases a, b, and c, the error was reduced by a factor of 0.17. The reduction in error for case d was less because there was less noise in the trial data. The Gaussian-prior projection was on average more accurate than the L_2 projection. Consistent with the differences seen in Figure 6(b) and (c), this difference is most significant for the cases with $\sigma_{\varepsilon_T}^2 = 1/5$ using the ROMs with a larger numbers of basis functions i.e. cases b and c. Comparing the mean Gaussian-prior projection error between plots a, b, and c shows that the Gaussian-prior projection errors are relatively insensitive to the number of basis functions in the model.

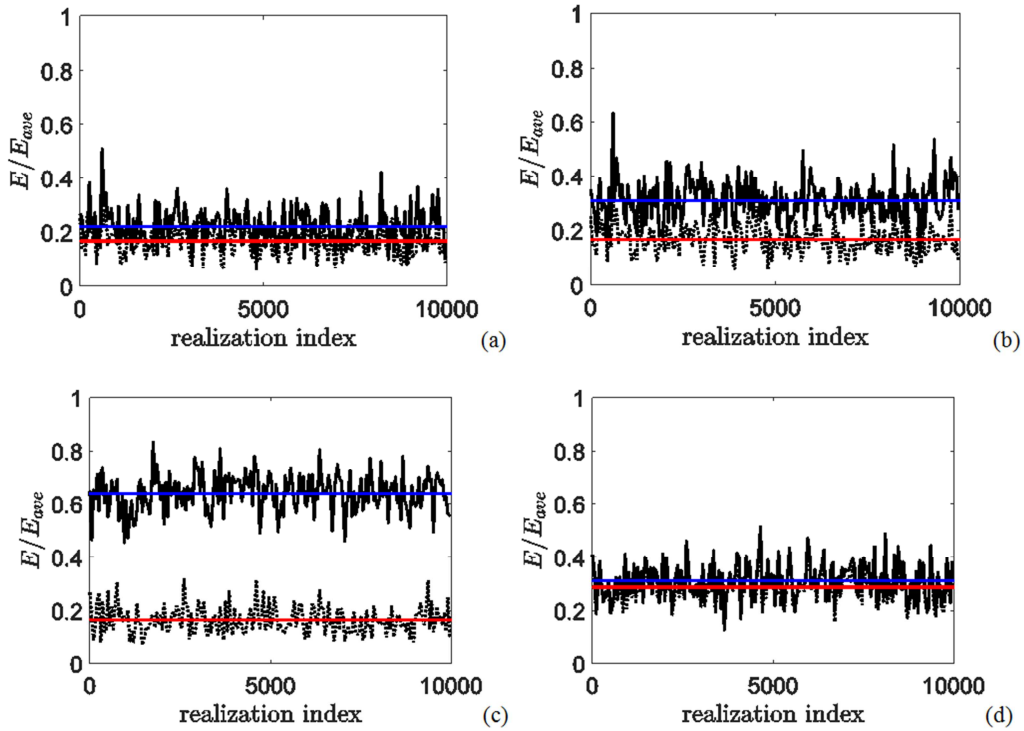


Figure 7. *L_2 norm of the error of Gaussian-prior projections and L_2 projections for 10000 trial realizations.: (a) $\sigma_{\varepsilon_T}^2 = 1/5$, $\sigma_{\varepsilon}^2 = 1/10$, (b) $\sigma_{\varepsilon_T}^2 = 1/5$, $\sigma_{\varepsilon}^2 = 1/400$, (c) $\sigma_{\varepsilon_T}^2 = 1/5$, $\sigma_{\varepsilon}^2 = 1.5\times10^{-12}$, (d) $\sigma_{\varepsilon_T}^2 = 1/200$, $\sigma_{\varepsilon}^2 = 1/400$. These errors are normalized respect to the analytical values of the average error in the trial data. The Gaussian-prior projection errors and the L-2 projection errors are shown as the dash line and the solid black line, respectively. The red line represents the average error for Gaussian-prior projection and the blue line represents the average error for L_2 projection.*

4. Conclusions

In this work, a probabilistic ROM process was developed for stochastic problems. This process allows reduced order modeling techniques to be applied to problems where the defining parameters are sampled from a probability distribution and the measured data has random noise. No ad-hoc assumptions are necessary to derive the model and apply it to trial data. The process provides a comprehensive probabilistic approach for deriving reduced order models of stochastic systems. These reduced order models can then be used to provide rapid, accurate predictions for stochastic problems where repeated analyses of similar scenarios must be performed.

An improved derivation of PPCA was also given that relaxes the unnecessary assumption that the variance of the latent variable is unity. The new approach provides an accurate estimate of this variance and also gives orthonormal basis functions. (In standard PPCA, the basis function are orthogonal but not orthonormal.) These improvements allowed a more intuitive interpretation of the eigenvalues of the PPCA and their relation to the noise in the data and the variance of the latent variables.

The information obtained from the improved PPCA was used to create a Gaussian prior for the latent variables in the on-line part of the ROM process. In the on-line step, Bayesian parameter estimation is used to estimate the latent variables associated with a data vector from a new scenario with an unknown amount of noise. These latent variables are then used to reconstruct the noise-free signal. The model problem showed that the true (noise-free) signal could be accurately reproduced from noisy data using this approach, much more so than with a standard L_2 projection especially when the noise in the data vector is large and there are many basis functions in the model.

Acknowledgements

This manuscript was authored in part by National Security Technologies, LLC, under contract DE-AC52-06NA25946 with the U.S. Department of Energy and supported by the Site-Directed Research and Development program. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The U.S. Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan). DOE/NV/25946--3089.

References

[1] Indika Udagedara, Brian T Helenbrook, Aaron Luttman, and Stephen E Mitchell. Reduced order modeling for accelerated Monte Carlo simulations in radiation transport. *Applied Mathematics and Computation*, 267: 237–251, 2015.

[2] K Peason. On lines and planes of closest fit to systems of point in space. *Philosophical Magazine*, 2: 559–572, 1901.

[3] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2 (1-3): 37–52, 1987.

[4] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2 (4): 433–459, 2010.

[5] Mark Richardson. Principal component analysis. *URL: http://people.maths.ox.ac.uk/richardsonm/SignalProcPCA. pdf (last access: 3. 5. 2013)*, 2009.

[6] Rasmus Bro and Age K Smilde. Principal component analysis. *Analytical Methods*, 6 (9): 2812–2831, 2014.

[7] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[8] Bruce Moore. Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Transactions on Automatic Control*, 26 (1): 17–32, 1981.

[9] Neil Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, 6 (Nov): 1783–1816, 2005.

[10] Michael E Tipping and Christopher M Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11 (2): 443–482, 1999.

[11] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61 (3): 611–622, 1999.

[12] Christian F Beckmann and Stephen M Smith. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, 23 (2): 137–152, 2004.

[13] Fang X Wu. Gene regulatory network modelling: a state-space approach. *International Journal of Data Mining and Bioinformatics*, 2 (1): 1–14, 2008.

[14] Alexander Ilin and Tapani Raiko. Practical approaches to principal component analysis in the presence of missing values. *Journal of Machine Learning Research*, 11 (Jul): 1957–2000, 2010.

[15] Jiahua Chen and Zehua Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95 (3): 759–771, 2008.

[16] Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael PH Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6 (31): 187–202, 2009.

[17] Larry Wasserman. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44 (1): 92–107, 2000.

[18] Kenneth P Burnham and David R Anderson. Multimodel inference understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33 (2): 261–304, 2004.

[19] Walter Zucchini. An introduction to model selection. *Journal of Mathematical Psychology*, 44 (1): 41–61, 2000.

[20] James L Beck and Ka V Yuen. Model selection using response measurements: Bayesian probabilistic approach. *Journal of Engineering Mechanics*, 130 (2): 192–203, 2004.

[21] Adrian E Raftery. Bayesian model selection in structural equation models. *Sage Focus Editions*, 154: 163–163, 1993.

[22] Adrian E Raftery. Bayesian model selection in social research. *Sociological Methodology*, pages 111–163, 1995.

[23] David Posada and Thomas R Buckley. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53 (5): 793–808, 2004.

[24] Jerald B Johnson and Kristian S Omland. Model selection in ecology and evolution. *Trends in Ecology & Evolution*, 19 (2): 101–108, 2004.

[25] Kenneth P Burnham and David Anderson. *Model selection and multi-model inference*. Taylor & Francis, 2003.

[26] Sadanori Konishi, Tomohiro Ando, and Seiya Imoto. Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika*, 91 (1): 27–43, 2004.

[27] Gerda Claeskens, Nils L Hjort, et al. *Model selection and model averaging*, volume 330. Cambridge University Press Cambridge, 2008.

[28] Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222: 309–368, 1922.

[29] FW Scholz. *Maximum likelihood estimation*. Wiley Online Library, 1985.

[30] Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25, 1982.

[31] Bradley Efron and David V Hinkley. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, 65 (3): 457–483, 1978.

[32] Klaas E Stephan, Will D Penny, Jean Daunizeau, Rosalyn J Moran, and Karl J Friston. Bayesian model selection for group studies. *Neuroimage*, 46 (4): 1004–1017, 2009.

[33] Henry D Acquah. Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship. *Journal of Development and Agricultural Economics*, 2 (1): 001–006, 2010.

[34] Joseph E Cavanaugh. Unifying the derivations for the Akaike and corrected Akaike information criteria. *Statistics & Probability Letters*, 33 (2): 201–208, 1997.

[35] Hamparsum Bozdogan. Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology*, 44 (1): 62–91, 2000.