Hierarchical Stochastic Model in Bayesian Inference for Engineering Applications: Theoretical Implications and Efficient Approximation

Hierarchical Bayesian models (HBMs) have been increasingly used for various engineering applications. We classify two types of HBM found in the literature as hierarchical prior model (HPM) and hierarchical stochastic model (HSM). Then, we focus on studying the theoretical implications of the HSM. Using examples of polynomial functions, we show that the HSM is capable of separating different types of uncertainties in a system and quantifying uncertainty of reduced order models under the Bayesian model class selection framework. To tackle the huge computational cost for analyzing HSM, we propose an efficient approximation scheme based on importance sampling (IS) and empirical interpolation method (EIM). We illustrate our method using two engineering examples—a molecular dynamics simulation for Krypton and a pharmacokinetic/pharmacodynamics (PKPD) model for cancer drug.

\[ y = f(x, \theta) + \varepsilon \]  

(1)

1 Introduction

Hierarchical Bayesian model (HBM) is a powerful modeling tool extended from the classical Bayesian framework. It has been used to solve many difficult practical problems, such as modeling heterogeneous data, calibrating system with multiple objectives, and inducing sparsity in a model. Its applications span across various fields, such as social science, economics, physics, medical science, and civil engineering [1–4].

In a classical Bayesian setting, users define a stochastic model with parameter \( \theta \) for a forward problem that predicts a quantity of interest and a prior distribution of \( \theta \). When data are available, the Bayes’ Theorem is used to solve the inverse problem by finding the posterior distribution of \( \theta \). For a complex system in reality, we further parameterize the stochastic model and the prior with hyperparameters. This extra level of parameters provides extra flexibility to a model, but generally requires more data to reach a well-posed problem. HBM refers to model classes with such a hierarchy (multiple levels) of parameters. Here, we define two types of HBM commonly found in the literature: a hierarchical prior model (HPM) that further parameterizes the prior and a hierarchical stochastic model (HSM) that further parameterizes the stochastic model (as shown in the likelihood function when evaluated at a given data).

Hierarchical prior model is a well-studied subject in the machine learning and compressive sensing community in the context of sparse Bayesian learning (SBL). Its applications include imaging [5,6], gene selection [7], and Bayesian compressive sensing [8,9]. It is an effective tool to solve ill-posed regression problems by adding extra constraints in the prior distribution under the Bayesian framework. The well-known Bayesian optimization method can also be interpreted using this model structure. On the other hand, the origin of HSM can be traced back to the multilevel regression, and Congdon [13] summarizes recent developments on this topic. Since most of the computational demand in Bayesian analysis comes from evaluating the stochastic model, HSM has a significantly larger computational cost than HPM. Although recently, there is increasing amount of HSM applications, the diversity of HSM uses is not comparable to HPM. However, recent developments in parallel computing open new opportunities to apply HSM to more complicated problems. In this paper, we discuss important theoretical implications of HSM and demonstrate an efficient approximation scheme with practical applications.

Despite the frequent use of the terminology “HBM” in the literature, the distinction between what we called the HSM and the HPM is often omitted. The overlapped usage of HBM has even caused confusion. For example, Guha et al. [14] studied the HPM. Nevertheless, the authors used [15] as a reference, in which the type of HBM mentioned in Ref. [15] is actually the HSM. Hence, in this paper, we begin with a comparison between HPM and HSM in Sec. 2. Then, we turn our focus back to HSM and study its theoretical implications illustrated by a simple polynomial regression in Sec. 3. To tackle the high computational cost for analyzing HSM in practice, we propose an efficient approximation based on the idea of importance sampling (IS) and empirical interpolation method (EIM) [16] in Sec. 4. Section 5 includes two realistic examples using our approximation methods. Finally, we give some concluding remarks in Sec. 6.

2 Classification of Hierarchical Bayesian Model: Hierarchical Prior Model Versus Hierarchical Stochastic Model

For the ease of illustration, we begin with defining the notations that will be repeatedly used in this paper. To predict a quantity of interest \( y \), one may have a stochastic forward model \( F(x, \theta, \varepsilon) \) with input \( x \), model parameters \( \theta \), and the stochasticity part \( \varepsilon \). This function defines the distribution \( p(y|x, \theta) \). One of the most common example of \( F(\cdot) \) is

\[ y = f(x, \theta) + \varepsilon \]  

(1)
where $e_i$ is chosen to follow a zero mean and $\sigma_i$ standard deviation Gaussian distribution, $N(0, \sigma_i)$, for computational convenience. Given a set of input and output data pairs $D = \{(x_i, y_i) | i = 1, \ldots, N_D\}$, the posterior distribution of $\theta$ is inferred by the Bayes’ theorem

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \tag{2}$$

where the likelihood $p(D|\theta) \triangleq p(\hat{y}_1, \ldots, \hat{y}_{N_D}|\hat{x}_1, \ldots, \hat{x}_{N_D}, \theta)$ is the probability of observing the data values $D$ from the predictive model, the prior $p(\theta)$ is the initial belief of the values of $\theta$, and the evidence (or marginal likelihood) $p(D) \triangleq p(\hat{y}_1, \ldots, \hat{y}_{N_D}|\hat{x}_1, \ldots, \hat{x}_{N_D})$ is a critical term used in model selection [20,21]. Then, a robust posterior prediction for an unobserved input–output pair $(x_0, y_0)$ can be obtained by

$$p(y_0|x_0, D) = \int p(y_0|x_0, \theta)p(\theta|D) \, d\theta \tag{3}$$

Hierarchical Bayesian model adds an extra level of hyperparameters $\psi$ to this classical Bayesian model. This new hierarchy can be added to either the prior (HPM) or the stochastic model/likelihood (HSM), which will result in completely different model classes suitable for different problems. The major distinction of the two models comes from different structures of information dependencies between all the stochastic variables. The representations shown in Fig. 1 are very useful to display and study such relations [19]. The arrows denote the directions of the forward models, which imply that Bayes’ Theorem is needed for inference in the opposite direction. Any nodes without an incoming arrow require a prior distribution.

For HPM, an extra node is added to the starting nodes in order to parameterize the prior. This operation does not change the information dependency between the existing parameters. The posterior joint distribution of $\theta$ and $\psi$ is

$$p(\theta, \psi|D) = \frac{p(D|\theta)p(\psi)p(\theta|\psi)}{p(D)} \tag{4}$$

If we integrate out $\psi$ in the analysis, this is equivalent to recovering the classical Bayesian setting from HPM by choosing $p(\psi) = \int p(\psi)p(\theta|\psi) \, d\psi$, because HPM does not affect the stochastic model part. Hence, the hyperparameters in HPM behave as latent variables. In most applications of HPM, we perform optimization of $\psi$ by maximizing $p(D|\psi)$, instead of the robust treatment of $\psi$. For example, in the context of SBL, zero-mean Gaussian priors with different variances are chosen for the coefficients in a linear regression problem. This prior is called the automatic relevance determination prior that is proven to induce sparsity [20,21]. A commonly used algorithm in SBL, called the relevance vector machine, is based on this optimization setup [22,23]. One can interpret this procedure as a Bayesian model selection problem in the continuous space.

For HSM, the hyperparameters affect the basic structure of the stochastic model. As shown in Fig. 1(b), each prediction is defined by a separate set of model parameters, and these parameters are correlated through the hyperparameters $\psi$. Here, the original model parameters behave as latent variables, and $\psi$ becomes the essential variables for future predictions $y_0$. The posterior distribution of $\psi$ is

$$p(\psi|D) = \frac{p(D|\psi)p(\psi)}{\int p(D|\psi)p(\psi) \, d\psi} \tag{5}$$

where $p(D|\psi) = \prod_{i=1}^{N_D} p(D_i|\psi)$

and $p(D_i|\psi) = \int p(D_i|\theta)p(\theta|\psi) \, d\theta$.

Here, we denote the full data set as $D = \{D_i | i = 1, \ldots, N_D\}$, where $D_i$ represents the observed data values for $y_i$. The choice of $p(\theta|\psi)$ is flexible. For example, a statistical model is chosen in Refs. [15] and [13], and $\psi$ is called the hyperparameter vector. The statistical model can be seen as a common prior for all $\theta$. This explicit modeling for the prior separates different uncertainties by isolating out the uncertainty of $\theta$, across multiple groups of predictions. It can give better predictions and also be related to causal inference [10]. Another example of such a hierarchical structure is found in a state-space model, where $\theta_i$ is the state variable and $p(\theta_i|\psi)$ represents the underlying stochastic process. Therefore, $p(\theta_i|\psi)$ can also be a theoretically informed model, instead of a purely empirical or statistical model.

In the current literatures, HPM is mainly used as a tool for optimal prior selection, while HSM is for explicit modeling of the uncertainty of model parameters across multiple groups of predictions for the data. However, HSM has the potential for more sophisticated modeling depending on the choice of $p(\theta|\psi)$, though it comes with a significantly larger computational cost due to the extra integral shown in Eq. (5). We note that since HPM and HSM are two exclusive types of HBM; in theory, they can coexist in a complex HBM and may have more than one level of hyperparameters. It is the limit of computational power that constrains the usage of HBM to be only one level of hyperparameters. In this paper, we study the important theoretical implications of HSM and present an efficient and flexible approximation method for performing Bayesian analysis with HSM. We note that a similar comparison can be found in Ref. [24], but they present it purely from the perspective of a likelihood function of $\psi$, instead of the HBM.

3 Theoretical Implications of Hierarchical Stochastic Model

Different assumptions on the uncertainties in a system can be represented using different hierarchical and nonhierarchical models. In this section, we verify that the effect of Occam’s Razor in

![Fig. 1 Graphical representations of the two HBMs: (a) hierarchical prior model and (b) hierarchical stochastic model](image-url)
the Bayesian model selection framework is applicable to both types of models. Moreover, we study the theoretical implications of using a HSM in three important aspects:

1. Separation of different types of uncertainties in a system (Sec. 3.1).
2. Identification of correlation between predictions or data (Sec. 3.1.2).
3. Uncertainty quantification of reduced order models (Sec. 3.2).

For computational accuracy, we demonstrate our results using polynomial regression because of the availability of many analytical solutions during the Bayesian analyses.

3.1 Separating Different Types of Uncertainties Using Hierarchical Stochastic Model. In this study, we demonstrate how HSM can capture different types of uncertainties using models that are embedded with different sources of stochasticity. We generate synthetic data based on two uncertainty models: (1) zero-mean Gaussian error \( e_i \) added to the function \( f(x, \theta) \), which can represent measurement noise in practice and (2) zero-mean Gaussian error \( e_\theta \) added to the model parameters \( \theta \), which can represent inherent model uncertainty due to, for example, insufficient knowledge of theoretically informed models or environmental variations across multiple experiments. We use a linear function \( f(x, \theta) = \theta x \) (\( \theta \) and \( x \) are scalars) and all data points are generated with independent \( e_i \) and \( e_\theta \). Three types of synthetic data are considered based on \( \theta \), a fixed value of the model parameter:

1. Additive error data, \( D_1: y = f(x, \theta) + e_i, e_i \sim N(0, \sigma_i^2) \).
2. Embedded error data, \( D_{2a}: y = f(x, \theta + e_\theta), e_\theta \sim N(0, \sigma_\theta^2) \).
3. Mixed error data, \( D_{2b}: y = f(x, \theta + e_\theta) + e_i, e_i \sim N(0, \sigma_i^2) \) and \( e_\theta \sim N(0, \sigma_\theta^2) \).

where \( N(z|\mu, \sigma^2) \) denotes a Gaussian distribution for \( z \) with mean \( \mu \) and variance \( \sigma^2 \). We set \( \theta = 1, \sigma_\theta = 0.5, \) and \( \sigma_i = 0.2 \). Each data set contains 1000 data points independently generated from a uniformly distributed \( x \) value between preset bounds and random errors \( e_i \) and \( e_\theta \) from the corresponding Gaussian distributions.

We generate all three types of data twice, once with \( x \) between 0 and 1 and once with \( x \) between 0.4 and 1 (see Figs. 2 and 3, respectively). Note that for the case with \( x \) between 0.4 and 1, the three data sets look a lot more similar to each other visually, which will be reflected in the difficulty of model selection, as will be explained later.

Given the three types of data, we perform Bayesian model class selection on four different stochastic model classes:

1. Nonhierarchical model, \( M_{1u} \): This model class includes a uniform prior for \( \theta \) between –1 and 3, and a Gaussian likelihood for \( \theta \), i.e., \( p(\{x_i, y_i\} | \theta, \sigma_j) = N(y_i | f(x_i, \theta), \sigma_j^2) \) for any \( i \). Here, \( \sigma_j \) is also treated as a parameter to be inferred. We use a uniform prior for \( \sigma_j \) between 0.001 and 1.
2. Hierarchical prior model, \( M_{1h} \): This model class has the same likelihood as \( M_{1u} \) and a hierarchical prior for \( \theta \). The prior is distributed as \( N(0, \mu_\theta, \sigma_\theta^2) \), i.e., the hyperparameters \( \theta = \{ \mu_\theta, \sigma_\theta \} \). We use uniform priors for \( \mu_\theta \) between –1 and 3 and \( \sigma_\theta \) between 0.001 and 1. The prior for \( \sigma_\theta \) follows the one in \( M_{1u} \).
3. Zero noise HSM, \( M_{2u} \): This is an HSM with \( p(\theta|\psi) \) modeled as the same Gaussian prior used for \( \theta \) in \( M_{1u} \) for all \( i \). No additive error is assumed in this model. This implies that the likelihood function of \( \theta_i \) is a delta function centered...
The accuracy of estimating the similarity between $D_1$ and $M_{1a}$ is slightly larger than the prior in $M_{1b}$. This is because their only difference is the prior of $\sigma_1$. We use Monte Carlo simulation to draw posterior samples of the hyperparameters for the "naive" estimation of the evidence. 10K samples are used to obtain accurate estimates with small variances.

### 3.1.1 Results and Discussion

Tables 1 and 2 summarize the results of the Bayesian inference and model selection for all six data sets (three with $x$ between 0 and 1, and three with $x$ between 0.4 and 1). The Bayesian model selection framework is capable of selecting the corresponding model used to generate the data. All models have a relatively good estimate for $M_{1b}$, $M_{2a}$, and $M_{2b}$. We note that the results of $M_{1a}$ and $M_{1b}$ are extremely similar. This is because their only difference is the prior of $\theta$, as explained in Sec. 2. In this case, the marginalized prior $p(\theta) = E[|D]|$ is only slightly larger than the prior in $M_{1a}$. Hence, the evidence of both models is almost the same in all cases.

Furthermore, we observe that in the case of $D_{2b}$ with $x$ between 0.4 and 1, $M_{2a}$ and $M_{2b}$ have a similar posterior model probability. Intuitively, the data with small $x$ values are more sensitive to additive noise because the noise-to-signal ratio is much higher than in the case of larger $x$ values. When such data are not available, it is challenging to distinguish between $D_{2a}$ and $D_{2b}$. This can be proved visually by observing the similarity between $D_{2a}$ and $D_{2b}$ in Fig. 3 compared to the one in Fig. 2.

| Data Model | $E[\theta|D]$ | Std[$\theta|D]$ | $E[\sigma_1|D]$ | Std[$\sigma_1|D]$ | ln(Evid.) | $P(M_1'|D)$ |
|------------|----------------|----------------|-----------------|-----------------|-----------|-------------|
| $D_1$      | $M_{1a}$       | 1.012 0.011     | 0.2030 0.0046   | 167.0 0.570     |           |             |
|            | $M_{1b}$       | 1.012 0.011     | 0.2033 0.0049   | 166.7 0.424     |           |             |
|            | $M_{2a}$       | 0.996 2.000     | 0.0000 0.0000   | -255.67 0.000   |           |             |
|            | $M_{2b}$       | 1.017 0.133     | 0.1976 0.0008   | 162.4 0.006     |           |             |
|            | $M_{2c}$       | 1.000 0.000     | 0.2000 0.2000   |               |           |             |
| $D_{2a}$   | $M_{1a}$       | 1.002 0.016     | 0.2080 0.0065   | -158.2 0.000    |           |             |
|            | $M_{1b}$       | 1.002 0.016     | 0.2814 0.0067   | -158.4 0.000    |           |             |
|            | $M_{2a}$       | 0.978 0.525     | 0.0000 0.0000   | 285.3 1.000     |           |             |
|            | $M_{2b}$       | 1.042 0.572     | 0.0021 0.0000   | 264.6 0.000     |           |             |
|            | $M_{2c}$       | 1.000 0.500     | 0.0000 0.2000   |               |           |             |
| $D_{2b}$   | $M_{1a}$       | 0.998 0.019     | 0.3334 0.0073   | -328.6 0.000    |           |             |
|            | $M_{1b}$       | 0.998 0.019     | 0.3330 0.0071   | -328.7 0.000    |           |             |
|            | $M_{2a}$       | 0.967 2.001     | 0.0000 0.0000   | -1682.3 0.000   |           |             |
|            | $M_{2b}$       | 1.027 0.453     | 0.1956 0.0041   | -266.6 1.000    |           |             |
|            | $M_{2c}$       | 1.000 0.500     | 0.2000 0.2000   |               |           |             |

Note: The row labeled "Ref." shows the actual values for the corresponding estimates.

Our results suggest using $M_{2b}$ for model calibration, when no prior knowledge indicates that additive error is irrelevant. This HSM can appropriately separate the two types of uncertainties in our case study into the additive and nonadditive parts based on the observed data. If computational power allows, Bayesian model class selection should always be performed among a set of different candidate models, as illustrated in this study.

### 3.1.2 Effect of Grouping

In Sec. 3.1, all synthetic data are generated independently based on different stochastic models. In practice, we may group some of the predictions to be modeled under one set of parameter values $\theta_i$, where $i$ is the index for each group of predictions. For example, if an experiment is repeated in different laboratories, we may model the data from the same laboratory using the same set of model parameters, assuming that the data share the same environmental factors during the experiment. When the prediction grouping is not known, we may want to know the most plausible grouping, which defines a unique HSM.

We use the same linear function $f(x, \theta) = ax + b$ as before, but generate a new set of data $D$. First of all, five random samples $\theta^{(i)}$ for $i = 1, \ldots, 5$ are drawn from $N(\theta_i, \sigma_\theta^2)$, then, 11 data points are generated for each $\theta^{(i)}$ to form a data set $D_i$ by using the stochastic forward model $y = f(x, \theta^{(i)}) + \varepsilon_i$, where $\varepsilon_i$ is random error drawn from $N(\varepsilon, \sigma_\varepsilon^2)$ independently for each data point. In this study, $\mu_\varepsilon = 1$, $\sigma_\theta = 0.5$, and $\sigma_\varepsilon = 0.1$. Figure 4(a) shows the five data sets used in the section. The sample mean and standard deviation of $\theta^{(i)}$ are 1.0191 and 0.4079, respectively.

To study the effect of grouping predictions in the HSM, we set up six different HSMs based on different information dependency between the predictions. We perform Bayesian model class selection to find the most plausible HSM to describe the data. The evidence is calculated similar to $M_{2b}$ in Sec. 3.1. Details of the calculation can be found in the Supplementary material, which is available under "Supplemental Data" tab for this paper on the ASME Digital Collection. The six data groupings are the following:

1. Actual grouping, $M_{1}':$ group predictions according to Fig. 4(a).
2. Constant $x$ grouping, $M_{1}':$ predictions with the same $x$ value are grouped together. This case represents grouping data according to the wrong variable. Figure 4(b) shows the resulting grouping.
3. One-half error grouping, $M_{1}':$ starting from $M_{1}'$, the predictions for each data set are further divided into two groups by the centerline of the data points in the set. This centerline is defined by the mean of $\theta$ inferred from each data set.
point in the set. Figure 4(c) shows the resulting data grouping.

(4) One-fourth error grouping, $M_{04}$: starting from $M_{03}$, the predictions for each data set are again further divided into two sets by the centerline of the data points in the set, as done in $M_{03}$.

(5) Single prediction, $M_{05}$: each prediction is treated as an independent data set (same as Sec. 3.1.1).

(6) Random grouping, $M_{06}$: predictions for the data points are randomly collected into five groups. Figure 4(d) shows the resulting grouping.

Table 3 summarizes the results of the Bayesian inference and model class selection for all six HSM. All models have a good estimate for $\mu_l$ (corresponds to $E[l|D]$). However, $M_{02}$ and $M_{06}$, representing two completely wrong groupings, have significantly lower estimates for $\sigma_l$ (given by $\text{Std}[l|D]$), but larger estimates for the mean of $\sigma_l$ (given by $E[\sigma_l|D]$). Hence, they have a very low log-evidence value. $M_{04}$ has the closest estimates of $\mu_l$, $\sigma_l$, and $\sigma_r$; but it is not the most probable model class among the six. $M_{04}$ is the most probable model class even though it has a low value of the estimates for both $\sigma_l$ and $\sigma_r$. We note that when the total number of data points is fixed, the more groups there are, the smaller the average number of predictions in a group is. This affects the likelihood of $\psi$, which equals the product of $p(D_i|\psi)$ for each data set $D_i$ (see Eq. (5)). When the number of $p(D_i|\psi)$ increases, the value of each evidence term may decrease. This decrease is due to a less peaked $p(D_i|\psi)$ as the number of data points in the set decreases. The final evidence $p(D|M_{0k})$ for a given data grouping $M_{0k}$ is a tradeoff between these two factors. Starting from $M_{01}$ being modified to $M_{03}$, if we repeat the process many times, eventually we will reach $M_{06}$. Therefore, $M_{01}$, $M_{03}$, $M_{04}$, and $M_{05}$ can be considered as a sequence of similar data grouping methods. In the end, the tradeoff suggests that

#### Table 3 Results for testing different HSMs

| $p(D_i|\psi)$ | $\text{Std}[l|D]$ | $E[\sigma_l|D]$ | $\text{Std}[\sigma_l|D]$ | $\text{In(Evid.)}$ | $P(M_{0k}|D)$ |
|---------------|------------------|----------------|------------------------|-----------------|-------------|
| $M_{01}$      | 1.050            | 0.528          | 0.091                  | 0.009           | 36.18       | 0.0012 |
| $M_{02}$      | 1.051            | 0.061          | 0.230                  | 0.023           | -3.42       | 0.0000 |
| $M_{03}$      | 1.036            | 0.473          | 0.066                  | 0.007           | 42.88       | 0.9986 |
| $M_{04}$      | 1.024            | 0.432          | 0.061                  | 0.008           | 33.91       | 0.0001 |
| $M_{05}$      | 1.053            | 0.112          | 0.228                  | 0.023           | -3.14       | 0.0000 |
| $M_{06}$      | 1.053            | 0.112          | 0.228                  | 0.023           | -3.14       | 0.0000 |

Ref. 1.000 0.500 0.100 — — —

Fig. 4 Different groupings of data sets. Different colors and marker combinations represent different data sets. The dashed lines denote the actual function without any error: (a) actual grouping, (b) grouping across $x$, (c) over grouping, and (d) random grouping.
$\mathcal{M}_1$ is the most probable grouping. This implies that model class selection based on the evidence may not necessarily lead to the actual grouping if it exists. Overall, it tries to minimize uncertainties in all parameters. On the other hand, if we choose the actual grouping, the uncertainty quantification will be accurate.

### 3.2 Uncertainty Quantification for Reduced Order Models Using Hierarchical Stochastic Model

A reduced order model simplifies the information obtained in the original model. This loss of information can be treated as a source of uncertainty in the Bayesian framework. The structure of this type of uncertainty can be defined if the mapping between the original model and the reduced order model is completely known. In most cases, however, the mapping is not well-defined or too complicated to be studied directly. Therefore, there is no clear way to model such kind of loss of information. In this section, we compare the performance of using different model classes in Sec. 3.1 to represent the reduced order model uncertainty. This study is based on fitting second and third degree polynomial data with linear functions.

#### 3.2.1 Problem Setup for Polynomials

We consider a total of 16 sets of data: eight sets from a quadratic function $f(x) = x^2$ and eight sets from a cubic function $f(x) = x^3$. For each type of function, we generate 20, 50, 100, and 200 data points twice (with and without noise), i.e., a total of eight data sets. The $x$ values of the data points are generated randomly in the interval $[-1, 1]$. Additive Gaussian error with standard deviation $\sigma_y = 0.1$ is chosen for the noise. Figure 5 shows all 16 data sets used in this study.

We perform Bayesian model class selection and posterior robust prediction using three models described in Sec. 3.1: $\mathcal{M}_{1a}$ (denoted as $\mathcal{M}_1$ in this section), $\mathcal{M}_{2a}$, and $\mathcal{M}_{2b}$. Supplementary material, which is available under “Supplemental Data” tab for this paper on the ASME Digital Collection, includes all derivations for the analytical expressions used to estimate the log-evidence values and posterior robust predictions $p(y|D, \mathcal{M}_k)$ for a given model $\mathcal{M}_k$. We evaluate the posterior probability of observing each grid point $(\hat{x}, \hat{y})$ on a two-dimensional (2D) fine grid of $(x, y)$ using the analytical expressions in order to construct the distribution of the posterior robust prediction.

#### 3.2.2 Results and Discussion

Table 4 summarizes the results of model class selection for this study. $\mathcal{M}_{2b}$ is the most probable model class for data with additive noise and $\mathcal{M}_{2a}$ is the most probable model class for data without any additive noise. $\mathcal{M}_1$ is

![Fig. 5 Eight data sets from $y = x^2$ ($P^2$) and eight data sets from $y = x^3$ ($P^3$). The noises are additive Gaussian error with standard deviation $\sigma_y = 0.1$.](image)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Quadratic (no noise)</th>
<th>Quadratic (with noise)</th>
<th>Cubic (no noise)</th>
<th>Cubic (with noise)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>$\ln(\text{Evid.})$</td>
<td>$P(\mathcal{M}_k</td>
<td>D)$</td>
<td>$\ln(\text{Evid.})$</td>
</tr>
<tr>
<td>20</td>
<td>$\mathcal{M}_1$</td>
<td>-10.1</td>
<td>0.0000</td>
<td>-11.4</td>
</tr>
<tr>
<td></td>
<td>$\mathcal{M}_{1a}$</td>
<td>8.6</td>
<td>0.9873</td>
<td>-15.4</td>
</tr>
<tr>
<td></td>
<td>$\mathcal{M}_{2a}$</td>
<td>4.3</td>
<td>0.0127</td>
<td>-6.8</td>
</tr>
<tr>
<td>50</td>
<td>$\mathcal{M}_1$</td>
<td>-31.2</td>
<td>0.0000</td>
<td>-33.4</td>
</tr>
<tr>
<td></td>
<td>$\mathcal{M}_{1a}$</td>
<td>-6.2</td>
<td>0.9920</td>
<td>-1014.3</td>
</tr>
<tr>
<td></td>
<td>$\mathcal{M}_{2a}$</td>
<td>-11.0</td>
<td>0.0080</td>
<td>-27.1</td>
</tr>
<tr>
<td>100</td>
<td>$\mathcal{M}_1$</td>
<td>-67.1</td>
<td>0.0000</td>
<td>-66.2</td>
</tr>
<tr>
<td></td>
<td>$\mathcal{M}_{1a}$</td>
<td>-8.6</td>
<td>0.9685</td>
<td>-61.4</td>
</tr>
<tr>
<td></td>
<td>$\mathcal{M}_{2a}$</td>
<td>-12.0</td>
<td>0.0315</td>
<td>-34.1</td>
</tr>
<tr>
<td>200</td>
<td>$\mathcal{M}_1$</td>
<td>-101.9</td>
<td>0.0000</td>
<td>-102.3</td>
</tr>
<tr>
<td></td>
<td>$\mathcal{M}_{1a}$</td>
<td>57.7</td>
<td>0.9864</td>
<td>-272.5</td>
</tr>
<tr>
<td></td>
<td>$\mathcal{M}_{2a}$</td>
<td>53.4</td>
<td>0.0136</td>
<td>-10.3</td>
</tr>
</tbody>
</table>
the least significant model in most cases because the error caused by the reduced order model is very nonlinear. We do not expect that the additive noise alone is sufficient to explain the error. In this study, the sample size (number of data points) does not play a significant role in the results.

Figure 6 shows one set of the prediction results (the results are not sensitive to the change of the number of sample size). We observe that for cases without additive noise, predictions from $M_{2a}$ and $M_{2b}$ are almost the same. Because $M_{2a}$ is a simpler model than $M_{2b}$ (less parameters), Bayesian model class selection prefers $M_{2a}$ and gives it a higher log-evidence value. For cases with additive noise, the uncertainty of prediction from $M_{2a}$ is significantly larger than $M_{2b}$. This is because $h$ is very sensitive to noise for $x$ values close to zero when a model does not have any additive error component. As a result, Bayesian model class selection prefers $M_{2b}$ for its ability to better fit the data on average.

These results suggest that the HSM with additive error is preferred as the first test for uncertainty quantification of reduced order model. The flexibility of such a model class to handle two types of uncertainty simultaneously (additive error and embedded error in the model parameters) results in a higher chance of discovering the underlying uncertainty structure of a reduced order model.

4 Efficient Approximation of Hierarchical Stochastic Model

The posterior distribution $p(\Psi|D)$ plays an important role in the HSM. Equation (5) shows that the calculation of the likelihood $p(D|\Psi)$ involves evaluations of multiple integrals, which correspond to the evidences for each data set $D_i$ conditional on the hyperparameters $\Psi$. This leads to an extremely large computational cost for sampling from $p(\Psi|D)$, as well as estimating the evidence of the model, $p(D)$. Current approaches include using conjugate pairs for analytical results [13], approximating the integrals with Laplace asymptotic approximation (LAA) [25] or using some advanced Markov chain Monte Carlo techniques [26]. These methods are still restrictive for either studying many complex systems or the efficiency is not very scalable to handle extra data sets.

In this section, we present an efficient approximation method based on a special use of IS. Most current research efforts focus on the HSM that adds only one extra level to the classical Bayesian model. Our method can be intuitively and efficiently extended to any complex HSMs with more levels of parameters. Moreover, we can lay out a standard procedure for fully analyzing any HSMs based on this method.

For ease of illustration, we demonstrate our method based on the commonly used stochastic model shown in Eq. (1). In practice, due to different assumptions on the additive noise, there may be more complicated HSM than the one shown in Fig. 1(b). Here, we consider two possible alternatives (see Fig. 7): (1) independent additive error parameters—applicable to heterogeneous data with different likelihoods [27] and (2) common additive error parameters—applicable to data sets that are expected to have the same measurement errors [25].

Here, we introduce our method to the basic HSM shown in Fig. 1(b), which represents the cases that either $\sigma$ is known or it is uncertain and included with the other uncertain parameters $\theta$. We denote this model class as $M_{11}$. In the Supplementary material, which is available under “Supplemental Data” tab for this paper on the ASME Digital Collection, we present the extension of the
idea to the two alternative HSMs shown in Fig. 7, which we denote as $M_{HS1}$ and $M_{HS3}$. In $M_{HS1}$, the posterior distribution, the likelihood, and the evidence for $\Psi$ are

$$p(\Psi|D, M_{HS1}) = \frac{p(D|\Psi, M_{HS1})p(\Psi|M_{HS1})}{p(D|M_{HS1})}$$

$$p(D|\Psi, M_{HS1}) = \prod_{i=1}^{N_S} p(D_i|\Psi, M_{HS1})$$

$$= \prod_{i=1}^{N_S} \left[ p(D_i|\theta_i, M_{HS1})p(\theta_i|\Psi, M_{HS1}) \right] d\theta_i$$

$$p(D|M_{HS1}) = \int p(D|\Psi, M_{HS1})p(\Psi|M_{HS1}) d\Psi$$

Sampling from $p(\Psi|D, M_{HS1})$ requires repeated evaluations of the likelihood $p(D|\Psi, M_{HS1})$ and thus evaluations of the integrals with different values of $\Psi$. Our idea is to approximate the integrals by IS with a special choice of the proposal PDF that can significantly reduce the total computational cost. In many cases, the likelihood $p(D_i|\theta_i, M_{HS1})$, which is related to the evaluation of $f(x, \Theta)$, dominates the computational effort.

Our method begins by performing Bayesian inference for each data set $D_i$ using the same likelihood $p(D_i|\theta_i, M_{HS1})$, but a different prior (choice of prior is discussed later). To be more specific, we draw samples $\{\theta_i^{(j)}\}_{j=1}^{N_{ij}}$ from the posterior distribution $p(\theta_i|D_i, M_i)$, where $M_i$ denotes this specific stochastic model class

$$p(\theta_i|D_i, M_i) = \frac{p(D_i|\theta_i, M_i)p(\theta_i|M_i)}{p(D_i|M_i)}$$

where $p(D_i|\theta_i, M_i) = p(D_i|\theta_i, M_{HS1})$ (9)

Then, we can approximate $p(D_i|\Psi, M_{HS1})$ based on IS with proposal distribution $q_i(\theta_i) = p(\theta_i|D_i, M_i)$

$$p(D_i|\Psi, M_{HS1}) \approx \frac{1}{N_{ij}} \sum_{j=1}^{N_{ij}} p(D_i|\theta_i^{(j)}, M_{HS1})p\left(\theta_i^{(j)}|\Psi, M_{HS1}\right)$$

$$= \frac{p(D_i|M_i)}{N_{ij}} \sum_{j=1}^{N_{ij}} p\left(\theta_i^{(j)}|M_i\right)$$

where $\theta_i^{(j)} \sim p(\theta_i|D_i, M_i)$

As a result, we only need to perform classical Bayesian inference once for each data set $D_i$ (draw posterior samples and estimate the evidence $p(D_i|M_i)$). Then, the hierarchical analysis comes as a postprocessing of the results. A very good tool for such a Bayesian inference is the transitional Markov chain Monte Carlo method, where the evidence comes as a by-product of drawing posterior samples and the algorithm is inherently parallel [28]. The likelihood $p(D|\Psi, M_{HS1})$ can then be estimated by

$$p(D|\Psi, M_{HS1}) \approx \prod_{i=1}^{N_S} \left[ \frac{p(D_i|M_i)}{N_{ij}} \sum_{j=1}^{N_{ij}} p\left(\theta_i^{(j)}|\Psi, M_{HS1}\right) \right]$$

where $\theta_i^{(j)} \sim p(\theta_i|D_i, M_i)$

In other words, the evidence ratio between $M_{HS1}$ with given hyperparameters and $M_i$ can be approximated by the mean of the prior ratio between the two models over the posterior $\theta_i$ samples of $M_i$.

This approximation suffers the same problem as IS, i.e., the variance of the estimate depends strongly on the closeness of the integrand and the proposal distribution. In our case, they are exactly the same if $p(\theta_i|M_i) = p(\theta_i|\Psi, M_{HS1})$. Therefore, we should choose $p(\theta_i|M_i)$ to minimize its difference to $p(\theta_i|\Psi, M_{HS1})$ for all $\Psi$ weighted by the prior $p(\Psi|M_{HS1})$. If we use the Kullback–Leibler divergence $D_{KL}(p|q)$ as a measure of difference between two distribution $p(x)$ and $q(x)$, this implies the choice

$$p(\theta_i|M_i) = \text{argmin} D_{KL}(p(\theta_i|\Psi, M_{HS1})||p(\theta_i|M_i))$$

$$= \text{argmin} \int D_{KL}(p(x)||q(x)) dx$$

\[ 011006-8 / \text{Vol. 5, MARCH 2019} \] Transactions of the ASME
information is useful to compare with the HSM analysis to give further insight into the system of interest. Furthermore, our method is very efficient for introducing extra data sets because we do not need to rerun the whole problem. Instead, the overhead is only the classical Bayesian inferences for the extra data sets and a postprocessing step that is not computationally intensive.

5 Illustrative Examples

We test our method using two examples: molecular dynamics simulation of Krypton and pharmacokinetic/pharmacodynamic (PKPD) modeling of a cancer drug. Both examples are taken from the previous studies. Their results are used to validate the efficiency and accuracy of our method.

5.1 Molecular Dynamics: Krypton. Wu et al. [25] proposed using the HSM (called the HBM in the paper) to calibrate the Lennard-Jones (LJ) potential \( \psi \) of Krypton based on viscosity data \( D \) reported from nine different laboratories. In that paper, the authors approximated the integrals in Eq. (7) by LAA. Since the joint posterior distribution of \( \psi \) and \( \sigma \) for each data set is close to a Gaussian distribution, LAA is a good approximation for this particular case. We revisit the problem with the same data and the same model class setup as described in Ref. [25]. Instead of the LAA approximation, we use our proposed method to estimate the joint posterior distribution of \( \psi \) and \( \sigma \) and the posterior robust prediction.

Following the assumptions in Ref. [25], we use the HSM with common additive error. First, we build the EIM basis functions based on the method shown in the Supplementary material, which is available under “Supplemental Data” tab for this paper on the ASME Digital Collection. We begin with a training set of \( \psi \) and \( \sigma \) from a coarse grid of 256 points within the boundaries \( 100 \leq \psi \leq 400 \) and \( 0.2 \leq \sigma \leq 0.5 \). A fine grid of 172 points for \( \sigma \) is used between 0.0001 and 0.5, uniformly distributed in log scale. We obtain a total of 50 basis functions for each data set to achieve an error less than 0.001%. Figure 8 shows the results of the EIM approximation for the likelihood function. The grayscale contour shows the actual likelihood values for this experiment. Right plot: maximum error of the EIM estimate for each experiment as a function of the number of basis. The error is normalized by the maximum value of the actual function.

\[
\begin{align*}
\theta & \sim \mathcal{N}(\mu, \Sigma) \\
\psi, \sigma & \sim \mathcal{N}(\hat{\psi}, \hat{\sigma})
\end{align*}
\]

Figure 9 shows the results of the HSM analysis using our proposed methods. We observe that our results are consistent with the results reported in Ref. [25] both in terms of the posterior distributions of the parameters and the posterior-robust prediction. This provides a verification of the accuracy of our method.

5.2 Pharmacokinetic/Pharmacodynamic Model: Cancer Drug. Finley et al. [30] used the classical Bayesian model to study a PKPD model for the antivascular endothelial growth factor (antiVEGF) cancer therapeutic agent, aflibercept. Multiple models with a different number of parameters were calibrated using two types of clinical data: the plasma concentrations of free aflibercept and bound aflibercept. The data can be separated into six groups, each corresponding to a different dosage of the drug. We apply the HSM structure in Fig. 1(b) to the basic model used in Ref. [30], which has three PK model parameters and five prediction error parameters. The prediction of the two types of data using these parameters is calculated by first running the system to reach steady-state and then imposing the corresponding drug dosages. The three model parameters \( \{k_{C}, k_{f}, k_{I} \} \) represent the secretion rate of anti VEGF in the blood, normal tissue, and tumor compartment, respectively. The five prediction error parameters \( \{\varepsilon_{SS}, \varepsilon_{RA}, \varepsilon_{FA}, \varepsilon_{F}, \varepsilon_{I} \} \) represent standard deviations of zero-mean Gaussian distributions for five different types of prediction errors: \( \varepsilon_{SS} \) — error on the steady-state prediction, \( \varepsilon_{RA} \) — error on the bound aflibercept prediction scaled by time, \( \varepsilon_{FA} \) — error on the free aflibercept prediction scaled by time, and \( \varepsilon_{F} \) — error on the free aflibercept prediction without any scaling. We choose independent Gaussian prior distributions for all of the eight parameters in log-10 scale.

Figure 10 and Table 5 compare the posterior distribution of the classical Bayesian model used in Ref. [30] with the posterior distribution of the basic HSM presented in this paper. We observe...
that fitting all data at once (the classical Bayesian model) leads to a significantly larger posterior variance for the model parameters, while the HSM leads to a larger posterior variance for the prediction error parameters. This implies that the knowledge about the model parameters is more transferable across different dosages than the knowledge about the prediction error parameters. We note that the steady-state prediction is not affected by the change of dosage. Indeed, we observe that the inferred value of $r_{\text{SS}}$ in the HSM is similar to the value for the classical Bayesian model. The slight increase of the coefficient of variation (CV) of $\sigma_{\text{SS}}$ for the HSM case is expected because the data used for inference in the HSM are divided into groups of smaller data sets. Moreover, the scaled standard deviations $r_{\text{BAf}}$ and $r_{\text{FAs}}$ are larger, and the fixed standard deviations $r_{\text{BAf}}$ and $r_{\text{FAs}}$ are smaller in the HSM. This is also expected because in the classical Bayesian model, predictions for different dosages may have different levels

<table>
<thead>
<tr>
<th>Table 5 Posterior sample mean and CV for parameters of the PK model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>------------------</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>$k_{\text{EC}}$</td>
</tr>
<tr>
<td>$k_{\text{N}}$</td>
</tr>
<tr>
<td>$k_{T}$</td>
</tr>
<tr>
<td>$\sigma_{\text{SS}}$</td>
</tr>
<tr>
<td>$\sigma_{\text{BAf}}$</td>
</tr>
<tr>
<td>$\sigma_{\text{FAs}}$</td>
</tr>
<tr>
<td>$\sigma_{\text{BAf}}$</td>
</tr>
<tr>
<td>$\sigma_{\text{FAs}}$</td>
</tr>
</tbody>
</table>

Note: All samples are drawn in log-space. The mean is converted to the linear scale. The CV is calculated based on the samples in log-scale.
of sensitivity to the errors scaled by time. Hence, the classical Bayesian model tends to have a larger $\sigma_{p,M}$ and $\sigma_{p,W}$ than $\sigma_{p,B}$ and $\sigma_{p,F}$. As a result, the posterior robust prediction based on the HSM (Fig. 11) has completely different prediction errors for both bound aflibercept and free aflibercept compared to the one in Ref. [30].

6 Conclusion

We demonstrate the benefits of using the hierarchical Bayesian framework in complex systems. Because of the double usage of the term HBM in the literature, we first explain the distinction between two very different HBMs: the HPM and the HSM. Then, we focus on studying the HSM, which has many interesting theoretical implications, as well as many computational challenges in practice. Based on examples of polynomial functions, we suggest that the HSM is capable of explicitly separating different types of uncertainties in a system and can be an effective tool for modeling uncertainty of reduced order models. In order to apply HSM to practical problems, we propose an efficient approximation method to tackle the high computational cost associated with using HSMs. Our method is a “bottom-up” approach that begins by drawing posterior samples of the model parameters for each data set. Then, we use a postprocessing step to move up the hierarchy of the parameters for drawing posterior samples. Building on the basic HSM, we demonstrate our method using two alternative HSMs that have a more complicated hierarchical structure. Finally, we validate our method using two illustrative examples based on the previous studies: a molecular dynamics simulation of Krypton and a PKPD model of a cancer drug.

The HSM is a convenient and effective tool to build the stochastic model/likelihood for a complicated system. It also opens up new types of analyses and results in different conclusions as compared to classical Bayesian inference. Our approximation method provides a standard procedure for analyzing hierarchical models. Beginning with an analysis for each individual data set, our method allows us to move up the hierarchy of the parameters efficiently to potentially extract more in-depth information about the system of interest. Moreover, it is very efficient for sequentially received data sets because our estimation scheme is a computationally fast postprocessing step. So far, we have mainly used a purely statistical (empirical) model for $p(0|\psi)$. In future work, we plan to employ other choices for modeling $p(0|\psi)$, which may motivate the development of new modeling tools.

Acknowledgment

We acknowledge support from ETH Zurich and computational time at the Swiss National Supercomputing Center CSCS under Project No. s448. P. K. and P. A. acknowledge support from the European Research Council (Advanced Investigator Award No. 341117).

References


