The Value of Structural Health Monitoring for the
reliable Bridge Management

Zagreb 2-3 March 2017

Bayesian uncertainty quantification and sparse Bayesian learning for model updating in structural health monitoring

James L. Beck[1], **Yong Huang**[2]

DOI: https://doi.org/10.5592/CO/BSHM2017.2.2

[1]George W. Housner Professor of Engineering and Applied Science, California Institute of Technology, 1200 E California Blvd, Pasadena, 91125, USA
[2]Associate Professor of Civil Engineering, Harbin Institute of Technology, 73 Huanghe Road, Harbin, 150090, China

E-mails: [1]jimbeck@caltech.edu; [2]huangyonghere@gmail.com

Abstract. The application of interest in this paper is model updating based on vibration monitoring of an instrumented structure, especially to detect and quantify localized stiffness losses as a proxy for damage. Because of its ability to quantify modeling uncertainty, a Bayesian approach is used in which the relative plausibility of each model in a model class (based on parameterized set of structural models) is quantified by its posterior probability from Bayes' Theorem. In addition, the relative plausibility of each model class within a set of candidate model classes can also be assessed. Computation of this posterior probability from Bayes' Theorem over all candidate model classes automatically applies a quantitative Ockham's razor that trades off a data-fit measure with an information-theoretic measure of model complexity, which penalizes model classes that "over-fit" the data. We present our recent progress in exploring sparse Bayesian learning for structural health monitoring, in which we infer spatially-sparse substructural stiffness reductions in a way that is consistent with the Bayesian Ockham razor. Illustrative results validate the capability of the presented sparse Bayesian learning algorithms for structural health monitoring.

Keywords: Structural Health Monitoring, System Identification, Bayesian Updating, Bayesian model class selection, Probability Logic, Uncertainty Quantification, Bayesian Ockham Razor; Sparse Bayesian Learning, Hierarchical Bayesian Model.

1 Introduction

System identification is the key component in model-based inversions for detection and assessment of damage in structural health monitoring. It uses observed structural response data and prior knowledge to update mathematical models of the behavior of a system such as a bridge or building subject to dynamic excitation. In addition to structural health monitoring, the goals of such data-informed modeling might also include providing a better understanding of the structural system's behavior and allowing more accurate predictions of its future response to specified excitations.

One of the main difficulties is that it is impossible to exactly model the full behavior of a structure by using the limited sensor data and prior knowledge available. Since any model gives an approximation to the real system behavior, there are always modeling uncertainties involved; for example, what values of the model parameters are appropriate and how well does the model predict the real system response? Another difficulty is that for complex system models, single-point parameter estimation often gives non-unique results (e.g. multiple least-squares or maximum likelihood estimates). In order to make more robust predictions, one should track all plausible values of the parameters based on the data and also explicitly treat the uncertain prediction errors (the difference between the response of the real system and that of the system model), as well as possible measurement errors. These issues have motivated numerous researchers to tackle the problem of structural system identification from a Bayesian perspective (e.g. Beck, 2010; Green et al., 2015; Au & Zhang, 2016; Huang et al., 2017b).

In contrast to the point estimates of the parameters used in the conventional deterministic or frequentist probabilistic methods, the Bayesian probabilistic framework uses Bayes' Theorem to quantify the relative plausibility based on the data of all possible values of the model parameters via their posterior PDF (probability density function). This procedure is used to learn about all plausible models for representing the system's behavior where each parameter value specifies a possible model for the system. Since there is always uncertainty in which parameterized model class to choose to represent a system, one can also choose a set of candidate model classes and calculate their posterior probability based on the data by applying Bayes' Theorem at the model class level. An information-theoretic interpretation (Muto & Beck, 2008; Beck, 2010) shows that the posterior probability of each model class depends on the difference between a measure of the average data-fit of the model class and the amount of information extracted from the data by the model class, which penalizes model classes that "over-fit" the data. Comparing the posterior probability of each model class therefore provides a quantitative Ockham's razor (Gull, 1989; Jefferys & Berger, 1992; Mackay, 1992), that is, models should be no more complex than is sufficient to explain the data.

Sparse Bayesian learning (Tipping, 2001a) is a supervised learning framework that is very effective at implementing Ockham's Razor by achieving parsimonious (sparse) representations in the context of regression and classification. It was the basis for the introduction of the relevance vector machine (Tipping, 2000) and sparse principal component analysis (Tipping, 2001b). We give an overview of our recent progress of developing sparse Bayesian learning algorithms for system identification, and present illustrative examples to show the capability of these methods.

2 Bayesian system identification and the Bayesian Ockham Razor

Consider the problem of predicting the output $\mathbf{z}(t)$ to some input $\mathbf{u}(t)$ of a real dynamic system over some time interval, $t \in [0, t_f]$, by using a computational model of the system. We use $\mathbf{u}_n = \mathbf{u}(n\Delta t) \in \mathbb{R}^{N_I}$ and $\mathbf{z}_n = \mathbf{z}(n\Delta t) \in \mathbb{R}^{N_o}$ to denote the real system input and output, respectively, at discrete times $t_n = n\Delta t, n \in \mathbb{Z}^+$, and use $\mathbf{u}_{0:n} = [\mathbf{u}_0^T, \mathbf{u}_1^T, \dots, \mathbf{u}_n^T]^T$ and $\mathbf{z}_{0:n} = [\mathbf{z}_0^T, \mathbf{z}_1^T, \dots, \mathbf{z}_n^T]^T$ to denote the discrete-time histories of the system input and output up to time t_n.

2.1 Stochastic model class

In modeling the I/O (input and output) behavior of a real system, one cannot expect any chosen deterministic model to make perfect predictions and the prediction errors of any such model will be uncertain. This motivates the introduction of a *stochastic* (or *Bayesian*) *model class* $\mathcal{M}$ (Beck, 2010) that consists of a set of *stochastic I/O models* valid for any $n \in \mathbb{Z}^+$ $\{p(\mathbf{z}_{1:n}|\mathbf{u}_{0:n}, \mathbf{w}, \mathcal{M}): \mathbf{w} \in \mathbf{W} \subset \mathbb{R}^{N_p}\}$ (also called *stochastic forward models*) for a system, together with a chosen *prior probability distribution* $p(\mathbf{w}|\mathcal{M})$ over this set that quantifies the initial relative plausibility of each I/O probability model corresponding to each value of the parameter vector $\mathbf{w}$. Any deterministic I/O model of a system that involves uncertain parameters can be used to construct such a model class for the system by *stochastic embedding* (Beck, 2010) in which the *Principle of Maximum Information Entropy* plays an important role (Jaynes 1983; Jaynes 2003) (see (1) in the next sub-section).

Remark 2.1: Probability as a logic provides a rigorous foundation for the Bayesian approach. Probability in *probability logic* is interpreted as the degree of plausibility of a statement on the basis of the specified conditioning information (Cox, 1946,1961; Jaynes, 1957,2003; Beck, 2010). This allows the uncertainty in predictions to be quantified due to our incomplete information because of our limited capacity to collect or understand the relevant information. The probability logic axioms apply to incorporating not only parametric uncertainty (uncertainty about which model in a proposed set should be used to represent the structure's I/O behavior) but also non-parametric uncertainty due to the existence of prediction errors because of the approximate nature of any structural model. This is in contrast to the relative frequency interpretation of probability in Kolmogorov's axioms, which is restricted to "inherently random" physical variables.

2.2 Bayesian updating for a given model class

If sensor data $\boldsymbol{\mathcal{D}}_N = \{\hat{\mathbf{u}}_{0:N}, \hat{\mathbf{y}}_{1:N}\}$ are available where $\hat{\mathbf{y}}_{1:N}$ and $\hat{\mathbf{u}}_{0:N}$ are the measured time histories of the system output and the corresponding measured system input (if available), respectively, sampled at time interval Δt, then a model can be developed to *predict* the *measured* system output $\mathbf{y}_n$ at each time t_n by using:

$$\mathbf{y}_n = \mathbf{z}_n + \mathbf{m}_n = \mathbf{q}_n(\hat{\mathbf{u}}_{0:n}, \mathbf{w}) + \mathbf{e}_n + \mathbf{m}_n \tag{1}$$

where $\mathbf{m}_n$ and $\mathbf{e}_n$ denote the measurement noise and output prediction error at time t_n and the system output equation $\mathbf{z}_n = \mathbf{q}_n(\hat{\mathbf{u}}_{0:n}, \mathbf{w}) + \mathbf{e}_n$ is used where $\mathbf{q}_n$ is the corresponding output of a parameterized deterministic model that can be based on theoretical principles (e.g., a FEM model). A probability model can be chosen for the I/O behavior by selecting a PDF for $\mathbf{e}_{1:n}$ that maximizes Shannon's entropy (a measure of uncertainty) subject to some prior constraints. This procedure is called *stochastic embedding* of the parameterized deterministic model in Beck (2010). A probability model can also be chosen for the measurement error $\{\mathbf{m}_n\}$ based on a separate study of the sensors, where $\{\mathbf{m}_n\}$ is taken independent of the prediction errors $\{\mathbf{e}_n\}$. This leads to a probability model $p(\mathbf{y}_{1:N}|\hat{\mathbf{u}}_{0:N}, \mathbf{w}, \mathcal{M})$ for predicting the sensor output $\mathbf{y}_{1:N}$. In many applications, $\mathbf{m}_n$ is negligible compared with $\mathbf{e}_n$ and so it can be dropped, that is, the difference between the measured system output $\mathbf{y}_n$ and the actual output $\mathbf{z}_n$ is ignored but not the difference between the real system and model outputs, $\mathbf{z}_n$ and $\mathbf{q}_n$.

The data $\boldsymbol{\mathcal{D}}_N$ can be used to update the relative plausibility of each stochastic I/O model $p(\mathbf{y}_{1:n}|\hat{\mathbf{u}}_{0:N}, \mathbf{w}, \mathcal{M})$, $\mathbf{w} \in \mathbf{W} \subset \mathbb{R}^{N_p}$, defined by the stochastic model class $\mathcal{M}$, by computing the *posterior* PDF $p(\mathbf{w}|\boldsymbol{\mathcal{D}}_N, \mathcal{M})$ from *Bayes' Theorem*:

$$p(\mathbf{w}|\boldsymbol{\mathcal{D}}_N, \mathcal{M}) = p(\boldsymbol{\mathcal{D}}_N|\mathbf{w}, \mathcal{M})p(\mathbf{w}|\mathcal{M})/p(\boldsymbol{\mathcal{D}}_N|\mathcal{M}) = c^{-1}p(\boldsymbol{\mathcal{D}}_N|\mathbf{w}, \mathcal{M})p(\mathbf{w}|\mathcal{M}) \tag{2}$$

where $c = p(\boldsymbol{\mathcal{D}}_N|\mathcal{M})$ is the normalizing constant, which is called the *evidence* or *marginal likelihood* for the model class $\mathcal{M}$ given by data $\boldsymbol{\mathcal{D}}_N$; $p(\boldsymbol{\mathcal{D}}_N|\mathbf{w}, \mathcal{M})$, as a function of $\mathbf{w}$, is the *likelihood function* which expresses the probability of getting data $\boldsymbol{\mathcal{D}}_N$ based on the PDF $p(\mathbf{y}_{1:N}|\hat{\mathbf{u}}_{0:N}, \mathbf{w}, \mathcal{M})$ by substituting the measured output data $\hat{\mathbf{y}}_{1:N}$ for $\mathbf{y}_{1:N}$. Note that a model class can be used to perform both prior (initial) and posterior (updated using system sensor data) robust predictive analyses, which can be used during design and operation, respectively, of a structure, based purely on the probability logic axioms (Papadimitriou et al., 2001; Beck & Taflanidis, 2013).

2.3 Bayesian updating for multiple model classes

If $\mathbf{M}$ denotes the proposition that specifies a *discrete* set of candidate model classes $\{\mathcal{M}_m: m = 1,2, \dots, N_M\}$ that is being considered for a system, together with a prior probability distribution $p(\mathcal{M}_m|\mathbf{M})$ over this discrete set, then the posterior PDF $\mathrm{p}(\mathbf{w}|\boldsymbol{\mathcal{D}}_N, \mathbf{M})$ based on $\mathbf{M}$ is given by the Total Probability Theorem:

$$p(\mathbf{w}|\boldsymbol{\mathcal{D}}_N, \mathbf{M}) = \textstyle\sum_{m=1}^{M} p(\mathbf{w}|\boldsymbol{\mathcal{D}}_N, \mathcal{M}_m)P(\mathcal{M}_m|\boldsymbol{\mathcal{D}}_N, \mathbf{M}) \tag{3}$$

where the posterior PDF for each model class $\mathcal{M}_m$ in (3), which comes from (2), is weighted by the posterior probability $P(\mathcal{M}_m|\boldsymbol{\mathcal{D}}_N, \mathbf{M})$ computed from Bayes' Theorem at the model class level:

$$P(\mathcal{M}_m|\boldsymbol{\mathcal{D}}_N, \mathbf{M}) = p(\boldsymbol{\mathcal{D}}_N|\mathcal{M}_m)P(\mathcal{M}_m|\mathbf{M})/p(\boldsymbol{\mathcal{D}}_N|\mathbf{M}) \tag{4}$$

Here, $p(\boldsymbol{\mathcal{D}}_N|\mathcal{M}_m)$ is the *evidence* for $\mathcal{M}_m$ provided by the data $\boldsymbol{\mathcal{D}}_N$ (additional conditioning on $\mathbf{M}$ is irrelevant), which is given by the Total Probability Theorem:

$$p(\boldsymbol{\mathcal{D}}_N|\mathcal{M}_m) = \int p(\boldsymbol{\mathcal{D}}_N|\mathbf{w}, \mathcal{M}_m)p(\mathbf{w}|\mathcal{M}_m)d\mathbf{w} \tag{5}$$

A uniform prior probability distribution can be chosen for the candidate model classes, that is, $P(\mathcal{M}_m|\boldsymbol{M}) = 1/N_M$, if the model classes are considered equally plausible a priori.

The calculation of the posterior probability $P(\mathcal{M}_m|\boldsymbol{\mathcal{D}}_N, \mathbf{M})$ in (4) provides a procedure for *Bayesian model class selection* (or comparison, or assessment), where the computation of the multi-dimensional integral in (5) for the evidence function is vital. If there is no analytical solution for (5), Laplace's approximation method can be used when the model class is globally identifiable based on the available data $\boldsymbol{\mathcal{D}}_N$ (e.g. Beck & Yuen 2004, Beck 2010). When the chosen class of models is unidentifiable or locally identifiable based on the data $\boldsymbol{\mathcal{D}}_N$ so that there are multiple MLEs (*maximum likelihood estimates*) (Beck & Katafygiotis, 1998), only stochastic simulation methods are practical to calculate the model class evidence, such as the TMCMC method (Ching & Chen, 2007), the stationarity method in Cheung & Beck (2010) or the Approximate Bayesian Computation method (Chiachio et al., 2014; Vakilzadeh et al., 2017*)*.

2.4 Bayesian Ockham Razor

Comparing the posterior probability of each candidate model class by (4) automatically implements an elegant and powerful version of Ockham's (Occam's) Razor, known as the *Bayesian Ockham Razor*. The essence of Ockham's Razor has long been advocated for data-based model identification, that is, a simpler model should be preferred over a more complex model if it leads to comparable agreement with the data. However, until recently, the approximate complexity measure for a model did not have a rigorous formulation. Two early attempts are AIC (Akaike, 1974) and BIC (Schwarz, 1978), which trade-off a data-fit measure with a measure of "complexity" proportional to the number of uncertain parameters N_p. Using these simplified criteria for model assessment requires caution, however, because their penalty term for model class complexity depends only on N_p and ignores the effect of the prior distribution.

A recent interesting information-theoretic interpretation (Muto & Beck, 2008; Beck, 2010) shows that the evidence $p(\boldsymbol{\mathcal{D}}_N|\mathcal{M}_m)$ in (5) explicitly builds in a trade-off between a data-fit measure for the model class and an information-theoretic measure of its complexity that quantifies the amount of information that the model class extracts from the data $\boldsymbol{\mathcal{D}}_N$. This result is based on using (2) in the expression for the normalization of the posterior PDF:

$$\begin{aligned}\log[p(\boldsymbol{\mathcal{D}}_N|\mathcal{M}_m)] &= \int \log[p(\boldsymbol{\mathcal{D}}_N|\mathcal{M}_m)]p(\mathbf{w}|\boldsymbol{\mathcal{D}}_N,\mathcal{M}_m)d\mathbf{w} \\ &= \int \log[p(\boldsymbol{\mathcal{D}}_N|\mathbf{w},\mathcal{M}_m)p(\mathbf{w}|\mathcal{M}_m)/p(\mathbf{w}|\boldsymbol{\mathcal{D}}_N,\mathcal{M}_m)]p(\mathbf{w}|\boldsymbol{\mathcal{D}}_N,\mathcal{M}_m)d\mathbf{w} \\ &= \int \log[p(\boldsymbol{\mathcal{D}}_N|\mathbf{w},\mathcal{M}_m)]p(\mathbf{w}|\boldsymbol{\mathcal{D}}_N,\mathcal{M}_m)\,d\mathbf{w} - \int \log[p(\mathbf{w}|\boldsymbol{\mathcal{D}}_N,\mathcal{M}_m)/p(\mathbf{w}|\mathcal{M}_m)]p(\mathbf{w}|\boldsymbol{\mathcal{D}}_N,\mathcal{M}_m)d\mathbf{w} \quad (6) \\ &= \mathbf{E}\big[\log\big(p(\boldsymbol{\mathcal{D}}_N|\mathbf{w},\mathcal{M}_m)\big)\big] - \mathbf{E}\big[\log[p(\mathbf{w}|\boldsymbol{\mathcal{D}}_N,\mathcal{M}_m)/p(\mathbf{w}|\mathcal{M}_m)]\big]\end{aligned}$$

where the expectations $\mathbf{E}[\cdot]$ are taken with respect to the posterior $p(\mathbf{w}|\boldsymbol{\mathcal{D}}_N,\mathcal{M}_m)$. The first term is the posterior mean of the log likelihood function, which is a measure of the average data-fit of the model class $\mathcal{M}_m$, and the second term is the Kullback-Leibler information, or relative entropy of the posterior relative to the prior, which is a measure of the model complexity (the amount of information gain about $\mathcal{M}_m$ from the data $\boldsymbol{\mathcal{D}}_N$) and is always non-negative. This information-theoretic result was first given by Beck & Yuen (2004) for the case of globally identifiable models and then extended to the general case by Ching et al. (2005) where the model may be unidentifiable. The merit of (6) is that it shows rigorously, without introducing ad-hoc concepts, that the log evidence for $\mathcal{M}_m$ explicitly builds in a trade-off between the data-fit of the model class and its information-theoretic complexity. This is important in structural health monitoring applications, since too complex models often lead to over-fitting of the data and the subsequent response predictions may then be unreliable since they depend too much on the details of the specific data, e.g., measurement noise and environmental effects.

3 General formulation of sparse Bayesian learning

3.1 Input-output model specification

Given a set of I/O data $\boldsymbol{\mathcal{D}} = \{\hat{\mathbf{u}}, \hat{\mathbf{y}}\}$, suppose that the model prediction of the output is $\mathbf{y} = \mathbf{f}(\hat{\mathbf{u}}) + \mathbf{e} + \mathbf{m} \in \mathbb{R}^{N_o}$ involving a deterministic function $\mathbf{f}$ of the input vector $\hat{\mathbf{u}}$, along with uncertain prediction error $\mathbf{e}$ and measurement noise $\mathbf{m}$. Assume that the function $\mathbf{f}$ is chosen as a weighted sum of N_p basis functions $\{\boldsymbol{\Theta}_j(\hat{\mathbf{u}})\}_{j=1}^{N_p}$:

$$\mathbf{f}(\hat{\mathbf{u}}) = \textstyle\sum_{j=1}^{N_p} w_j\, \boldsymbol{\Theta}_j(\hat{\mathbf{u}}) = \boldsymbol{\Theta}(\hat{\mathbf{u}})\mathbf{w} \quad (7)$$

where $\boldsymbol{\Theta}$ is an $N_o \times N_p$ matrix with the basis functions $\{\boldsymbol{\Theta}_j\}$ as columns. Analysis of this model is facilitated by the adjustable parameters (or weights) $\mathbf{w} \in \mathbb{R}^{N_p}$ appearing linearly. The objective here is to infer values of the parameters $\{w_j\}_{j=1}^{N_p}$ such that $\boldsymbol{\Theta}(\hat{\mathbf{u}})\mathbf{w}$ is a 'good' approximation of $\mathbf{f}(\hat{\mathbf{u}})$ and the parameter vector $\mathbf{w}$ is sparse. There has been significant recent interest (e.g., Tropp, 2004; Hastie et al, 2015) in the notion of sparse learning algorithms which promote significant numbers of the parameter components w_n to be zero as a means of providing model regularization during inverse problems. These methods have been applied for compressive sensing (Candès, 2006; Donoho, 2006; Huang et al., 2011; Huang et al., 2014; Huang et al., 2016).

3.2 Sparse Bayesian learning model

Sparse Bayesian learning (SBL) encodes a preference for sparser parameter vectors by making a special choice for the prior distribution for the parameter vector $\mathbf{w}$ that is known as the *automatic relevance determination* (ARD) prior (Mackay, 1992; Tipping, 2001a):

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \textstyle\prod_{j=1}^{N_p} p(w_j|\alpha_j) = \prod_{j=1}^{N_p} \mathcal{N}(w_j|0,\alpha_j^{-1}) = \prod_{j=1}^{N_p}\left[(2\pi)^{-1/2}\alpha_j{}^{1/2}\exp\left\{-\frac{1}{2}\alpha_j w_j^2\right\}\right] \quad (8)$$

where the hyperparameter α_j is the prior precision (inverse variance) for w_j. An individual hyperparameter α_j is associated independently with each weight w_j, thereby moderating the strength of the Gaussian prior. Note that an infinite value of α_j implies that the corresponding coefficient w_j has an insignificant prior contribution to the modeling of the measurements $\mathbf{y}$, because it produces essentially a Dirac delta-function at zero for the prior, and so the posterior.

By using *the principle of maximum information entropy* (Jaynes, 1983) and incorporating the first two moments as constraints, the combination of the prediction error and measurement noise $\mathbf{e}$ is modeled as a zero-mean Gaussian vector with covariance matrix $\beta^{-1}\mathbf{I}_{N_o}$, which gives a Gaussian predictive PDF:

$$p(\mathbf{y}|\mathbf{w},\beta) = (2\pi\beta^{-1})^{-\frac{N_o}{2}}\exp\left(-\frac{\beta}{2}\|\mathbf{y} - \boldsymbol{\Theta}(\hat{\mathbf{u}})\mathbf{w}\|_2^2\right) = \textstyle\prod_{j=1}^{N_p}\mathcal{N}(\mathbf{y}|\boldsymbol{\Theta}(\hat{\mathbf{u}})\mathbf{w},\beta^{-1}\mathbf{I}_{N_o}) \quad (9)$$

By substituting the data $\hat{\mathbf{y}}$ for $\mathbf{y}$, (9) gives a Gaussian likelihood function that measures how well the model for specified parameters $\mathbf{w}$ and β predicts the measurements $\hat{\mathbf{y}}$. A stochastic model class $\mathcal{M}(\boldsymbol{\alpha},\beta)$ is then defined by the I/O predictive model in (9) and the prior PDF on $\mathbf{w}$ given by (8).

3.3 Bayesian updating for given model class $\mathcal{M}(\boldsymbol{\alpha}, \boldsymbol{\beta})$

The posterior distribution $p(\mathbf{w}|\hat{\mathbf{y}}, \boldsymbol{\alpha}, \beta)$ over the weight parameters given by model class $\mathcal{M}(\boldsymbol{\alpha}, \beta)$ is computed based on Bayes' theorem:

$$p(\mathbf{w}|\hat{\mathbf{y}}, \boldsymbol{\alpha}, \beta) = p(\hat{\mathbf{y}}|\mathbf{w}, \beta) p(\mathbf{w}|\boldsymbol{\alpha}) / p(\hat{\mathbf{y}}|\boldsymbol{\alpha}, \beta) \tag{10}$$

where $p(\hat{\mathbf{y}}|\boldsymbol{\alpha}, \beta) = \int p(\hat{\mathbf{y}}|\mathbf{w}, \beta) p(\mathbf{w}|\boldsymbol{\alpha}) dw$ is the *evidence* of the model class $\mathcal{M}(\boldsymbol{\alpha}, \beta)$. Since both the prior and likelihood for **w** are Gaussian and the likelihood mean $\boldsymbol{\Theta}(\hat{\mathbf{u}})\mathbf{w}$ is linear in **w**, the posterior PDF can be expressed analytically as a multivariate Gaussian distribution:

$$p(\mathbf{w}|\hat{\mathbf{y}}, \boldsymbol{\alpha}, \beta) = \mathcal{N}(\mathbf{w}|(\boldsymbol{\Theta}^T\boldsymbol{\Theta} + \beta^{-1}\mathbf{A})^{-1}\boldsymbol{\Theta}^T\hat{\mathbf{y}}, (\beta\boldsymbol{\Theta}^T\boldsymbol{\Theta} + \mathbf{A})^{-1}) \tag{11}$$

where $\mathbf{A} = \mathrm{diag}\left(\alpha_j, \dots, \alpha_{N_p}\right)$.

3.4 Hyperparameter learning by evidence maximization

A continuous set of candidate model classes $\mathcal{M}(\boldsymbol{\alpha}, \beta)$ is defined in Subsection 3.2**,** and the robust posterior PDF $p(\mathbf{w}|\hat{\mathbf{y}})$ can be computed by integrating out the posterior uncertainty in $\boldsymbol{\alpha}$ and β as below. We assume that at the posterior $p(\boldsymbol{\alpha}, \beta|\hat{\mathbf{y}})$ is highly peaked at $\{\tilde{\boldsymbol{\alpha}}, \tilde{\beta}\}$ (the MAP (maximum a posteriori) value of $\{\boldsymbol{\alpha}, \beta\}$). We then treat $\{\boldsymbol{\alpha}, \beta\}$ as a 'nuisance' parameter vector and integrate it out by applying Laplace's asymptotic approximation (Beck & Katafygiotis, 1998):

$$p(\mathbf{w}|\hat{\mathbf{y}}) = \int p(\mathbf{w}|\hat{\mathbf{y}}, \boldsymbol{\alpha}, \beta) p(\boldsymbol{\alpha}, \beta|\hat{\mathbf{y}}) d\boldsymbol{\alpha} d\beta \approx p\left(\mathbf{w}|\hat{\mathbf{y}}, \tilde{\boldsymbol{\alpha}}, \tilde{\beta}\right). \tag{12}$$

where: $$\left\{\tilde{\boldsymbol{\alpha}}, \tilde{\beta}\right\} = \arg\max_{[\boldsymbol{\alpha},\beta]} p(\boldsymbol{\alpha}, \beta|\hat{\mathbf{y}}) \; = \arg\max_{[\boldsymbol{\alpha},\beta]} \{p(\hat{\mathbf{y}}|\boldsymbol{\alpha}, \beta) p(\boldsymbol{\alpha}) p(\beta)\} \tag{13}$$

If we assign flat, non-informative prior PDFs for $\boldsymbol{\alpha}$ and β, we equivalently just need to maximize the evidence function $p(\hat{\mathbf{y}}|\boldsymbol{\alpha}, \beta)$. The optimization of $\{\boldsymbol{\alpha}, \beta\}$ is the procedure of *Bayesian model class selection* (Beck & Yuen, 2004) from a continuous set of model classes $\mathcal{M}(\boldsymbol{\alpha}, \beta)$. For larger amounts of data (larger N_o), accurate predictions are expected that are typically highly sparse because the maximization in (13) causes many hyperparameters α_j to approach infinity during the learning process. This is the Bayesian Ockham razor (Gull, 1988; Jefferys & Berger, 1992; Mackay, 1992) at work: the maximization of the evidence function $p(\hat{\mathbf{y}}|\boldsymbol{\alpha}, \beta)$ automatically involves a trade-off between the average data-fit of the model class $\mathcal{M}(\boldsymbol{\alpha}, \beta)$ and model sparseness (more sparseness corresponds to less model complexity), as we discussed in Section 2.4.

3.5 Robust predictions

Having found the MAP estimates $\left\{\tilde{\boldsymbol{\alpha}}, \tilde{\beta}\right\}$, our approximation to the robust predictive distribution of the system response **y** for a given input $\hat{\mathbf{u}}$ would be:

$$\begin{aligned} p(\mathbf{y}|\hat{\mathbf{y}}) = \int p(\mathbf{y}, \mathbf{w}, \boldsymbol{\alpha}, \beta|\hat{\mathbf{y}})\, d\mathbf{w} d\boldsymbol{\alpha} d\beta &= \int p(\mathbf{y}|\mathbf{w}, \boldsymbol{\alpha}, \beta)\, p(\mathbf{w}|\hat{\mathbf{y}}, \boldsymbol{\alpha}, \beta) p(\boldsymbol{\alpha}, \beta|\hat{\mathbf{y}}) d\mathbf{w} d\boldsymbol{\alpha} d\beta \\ &\approx \int p\left(\mathbf{y}|\mathbf{w}, \tilde{\boldsymbol{\alpha}}, \tilde{\beta}\right) p\left(\mathbf{w}|\hat{\mathbf{y}}, \tilde{\boldsymbol{\alpha}}, \tilde{\beta}\right) d\mathbf{w} \end{aligned} \tag{14}$$

This robust predictive PDF takes into account all posterior plausible values of the model parameter vector **w**.

Remark 3.1: A *hierarchical Bayesian model* (Gelman et al., 2013) is involved if we define hyper-priors over the prior precision parameter vector $\boldsymbol{\alpha}$ and prediction error precision parameter β**.** It is typical to assign gamma distributions over $\boldsymbol{\alpha}$ and β (Tipping. 2001a); however, the inverse gamma hyper-prior over $\boldsymbol{\alpha}$ produces a more sparse solution (Babacan, 2010).

Remark 3.2: The learning of the prior precision parameter $\boldsymbol{\alpha}$ is vital to reduce the posterior uncertainties by generating sparse models of **w**, which leads to higher confidence in the predictions. The treatment of the prediction error precision β also affects the algorithm performance significantly, especially when the original model is only approximately sparse (Huang et al., 2016), which is common for structural health monitoring signals.

Remark 3.3: We have found the SBL algorithm suffers from a robustness problem: there are local maxima for (13) that may trap the hyperparameter optimization if the number of measurements N_o is much smaller than the number of model parameters N_p, leading to non-robust Bayesian updating results (Huang et al., 2014). Several robustness enhancement algorithms (Huang et al., 2014; Huang et al., 2016) have been developed by employing different strategies, with the goal of increasing signal reconstruction accuracy in compressive sensing for structural health monitoring signals.

4 Recent progress in applying sparse Bayesian learning to system identification in structural health monitoring

4.1 Hierarchical Bayesian model class

Suppose that we have a vector of identified natural frequencies $\hat{\boldsymbol{\omega}}^2 \in \mathbb{R}^{N_s N_m \times 1}$ (N_s and N_m are the number of modal identifications performed and number of extracted modes for each identification) and mode shapes $\hat{\boldsymbol{\psi}} \in \mathbb{R}^{N_s N_m N_o \times 1}$($N_o$ is the number of measured degrees of freedom). Since the measured degrees of freedom (DOFs) are usually a smaller subset of the DOFs of an appropriate structural model, we introduce the system natural frequencies $\boldsymbol{\omega}^2 \in \mathbb{R}^{N_m \times 1}$ and system mode shapes $\boldsymbol{\phi} \in \mathbb{R}^{N_d N_m \times 1}$ (N_d is number of DOFs of the structural model) to represent the actual underlying modal parameters of the assumed linear dynamics of the structural system at all DOFs corresponding to those of the structural model.

We choose a set of parameterized linear structural models with classical damping to produce normal modes of vibration where each model has the same known mass matrix $\mathbf{M} \in \mathbb{R}^{N_d \times N_d}$ inferred from structural drawings. Taking an appropriate substructuring (perhaps focusing on likely damage locations), we decompose the uncertain stiffness matrix $\mathbf{K} \in \mathbb{R}^{N_d \times N_d}$ as a linear combination of $(N_\theta + 1)$ substructure stiffness matrices $\mathbf{K}_{j,}\, j = 0, 1, \dots N_\theta$:

$$\mathbf{K}(\boldsymbol{\theta}) = \mathbf{K}_0 + \sum_{j=1}^{N_\theta} \theta_j \mathbf{K}_j \tag{15}$$

where $\mathbf{K}_j \in \mathbb{R}^{N_d \times N_d}, j = 1, \dots, N_\theta,$ is the prior choice of the j^{th} substructure stiffness matrix and the corresponding stiffness scaling parameter θ_j is a factor that allows modification of the nominal j^{th}substructure stiffness so it is more consistent with the real structure behavior. The stiffness matrices $\mathbf{K}_j$ could come from a finite-element model of the structure, then it would be appropriate to choose all $\theta_j = 1$ to give the most probable value a priori for the parameter vector $\boldsymbol{\theta} \in \mathbb{R}^{N_\theta}$**.** For damage detection purposes, we will exploit the fact that the onset of stiffness reductions is typically in a small number of locations in the absence of structural collapse, and so the potential change in $\boldsymbol{\theta}$ compared with that of a reference calibration stage is expected to be a *sparse vector* with relatively few non-zero components.

The following joint prior PDF for system parameters $\boldsymbol{\omega}^2$ and $\boldsymbol{\phi}$ and stiffness scaling parameters $\boldsymbol{\theta}$ is chosen (Huang & Beck, 2015a):

$$p(\boldsymbol{\omega}^2, \boldsymbol{\phi}, \boldsymbol{\theta}|\beta) \propto (2\pi/\beta)^{-N_m N_d/2} \exp\left\{-\frac{\beta}{2}\sum_{m=1}^{N_m} \left\|\left(\mathbf{K}(\boldsymbol{\theta}) - \omega_m^2\, \mathbf{M}\right)\boldsymbol{\phi}_m\right\|^2\right\} \tag{16}$$

where the finite value of the equation-error precision parameter β in (16) provides a soft constraint for the eigen-equation and it allows for the explicit control of how closely the system and model modal parameters agree. Note that we can decompose the joint prior PDF $p(\boldsymbol{\omega}^2, \boldsymbol{\phi}, \boldsymbol{\theta}|\beta)$ into the product of a conditional PDF for any one of the parameter vectors and a marginal PDF for the other two parameter vectors. Although the modal parameters are a nonlinear function of the stiffness parameters, we employ a trick to produce a series of coupled linear−in−the−parameter problems.

We choose the unique MAP value $\hat{\boldsymbol{\theta}}_u$ from applying Bayesian updating using a large amount of time-domain vibration data from the calibration state as pseudo-data to define the likelihood function for $\boldsymbol{\theta}$ as:

$$p\left(\hat{\boldsymbol{\theta}}_u|\boldsymbol{\theta}, \boldsymbol{\alpha}\right) = \prod_{i=1}^{N_\theta} \mathcal{N}\left(\hat{\theta}_{u,i}|\theta_i, \alpha_i^{-1}\right) \tag{17}$$

Although the conventional strategy in SBL is to use an ARD Gaussian prior PDF (Tipping, 2001a) to model sparseness, here we incorporate the ARD concept in the likelihood function, along with the prior on $\boldsymbol{\theta}$ in (16). Gaussian likelihood functions $p(\hat{\boldsymbol{\omega}}^2|\boldsymbol{\omega}^2, \boldsymbol{\rho})$ and $p\left(\hat{\boldsymbol{\psi}}|\boldsymbol{\phi}, \boldsymbol{\eta}\right)$ are also defined for system parameters $\boldsymbol{\omega}^2$ and $\boldsymbol{\phi}$ with precision parameters $\boldsymbol{\rho}$ and $\boldsymbol{\eta}$, respectively. In addition, we model our prior uncertainty in the equation error precision β by an exponential hyper-prior $p(\beta|b_0)$ with rate parameter b_0. The proposed modeling constitutes a multi-stage hierarchical model as shown in Figure 1. The bidirectional arrow in the graph of the hierarchical Bayesian model represents the information dependence between structural modal parameters $\boldsymbol{\omega}^2$and $\boldsymbol{\phi}$, which comes from the joint prior $p(\boldsymbol{\omega}^2, \boldsymbol{\phi}|\beta)$.

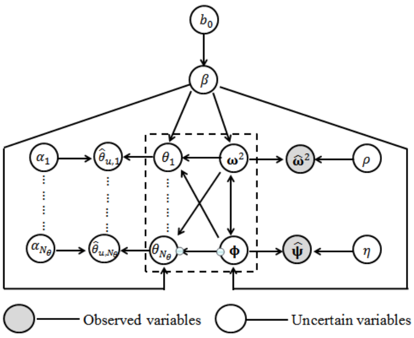


Fig 1. Acyclic graph representing the information flow in the hierarchical Bayesian model for offline SBL algorithms.

4.2 Fast sparse Bayesian learning algorithm

To facilitate the goal of presenting a fast algorithm to perform SBL, we focus on an analytical derivation of the posterior PDF of the stiffness scaling parameter $\boldsymbol{\theta}$ and collect all uncertain parameters except $\boldsymbol{\theta}$ in the vector $\boldsymbol{\delta} = [(\boldsymbol{\omega}^2)^T, \boldsymbol{\rho}^T, \boldsymbol{\phi}^T, \boldsymbol{\eta}, \boldsymbol{\alpha}^T, \beta, b_0]^T$ as 'nuisance' parameters, which are treated by using Laplace's approximation method (their posterior uncertainties are effectively ignored). The stochastic model class $\mathcal{M}(\boldsymbol{\delta})$ for the structural model is defined by the likelihood functions $p(\hat{\boldsymbol{\theta}}_u|\boldsymbol{\theta},\boldsymbol{\alpha}), p(\hat{\boldsymbol{\omega}}^2|\boldsymbol{\omega}^2,\boldsymbol{\rho})$ and $p(\hat{\boldsymbol{\psi}}|\boldsymbol{\phi},\boldsymbol{\eta})$ and the joint priors given by the product of $p(\boldsymbol{\omega}^2,\boldsymbol{\phi},\boldsymbol{\theta}|\beta)$ and $p(\beta|b_0)$. Based on this defined stochastic model class $\mathcal{M}(\boldsymbol{\delta})$, one can use the available modal data $\hat{\boldsymbol{\omega}}^2$ and $\hat{\boldsymbol{\psi}}$ and *pseudo-data* $\hat{\boldsymbol{\theta}}_u$ to update the structural model parameters $\boldsymbol{\theta}$ for system identification purposes. We assume that the posterior $p(\boldsymbol{\delta}|\hat{\boldsymbol{\omega}}^2,\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\theta}}_u)$ is highly peaked at $\tilde{\boldsymbol{\delta}}$ (the MAP value of $\boldsymbol{\delta}$). We then use Laplace's asymptotic approximation (Beck & Katafygiotis, 1998):

$$p(\boldsymbol{\theta}|\,\hat{\boldsymbol{\omega}}^2,\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\theta}}_u) = \int p(\boldsymbol{\theta}|\boldsymbol{\delta},\hat{\boldsymbol{\omega}}^2,\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\theta}}_u)p(\boldsymbol{\delta}|\hat{\boldsymbol{\omega}}^2,\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\theta}}_u)\,d\boldsymbol{\delta} \approx p(\boldsymbol{\theta}|\,\tilde{\boldsymbol{\delta}},\hat{\boldsymbol{\omega}}^2,\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\theta}}_u) \tag{18}$$

where $p(\boldsymbol{\theta}|\boldsymbol{\delta},\hat{\boldsymbol{\omega}}^2,\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\theta}}_u)$ is the posterior PDF for a given model class $\mathcal{M}(\boldsymbol{\delta})$, $\tilde{\boldsymbol{\delta}} = \arg\max p(\boldsymbol{\delta}|\hat{\boldsymbol{\omega}}^2,\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\theta}}_u)$, and

$$p(\boldsymbol{\delta}|\hat{\boldsymbol{\omega}}^2,\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\theta}}_u) \propto p(\hat{\boldsymbol{\omega}}^2,\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\theta}}_u|\boldsymbol{\delta})p(\boldsymbol{\delta})$$

$$= \int p(\hat{\boldsymbol{\theta}}_u|\boldsymbol{\theta},\boldsymbol{\alpha})\,p(\hat{\boldsymbol{\omega}}^2|\boldsymbol{\omega}^2,\boldsymbol{\rho})p(\hat{\boldsymbol{\psi}}|\boldsymbol{\phi},\eta)p(\boldsymbol{\theta}|\boldsymbol{\omega}^2,\boldsymbol{\phi},\beta)p(\boldsymbol{\omega}^2,\boldsymbol{\phi}|\beta)p(\boldsymbol{\rho}|\boldsymbol{\tau})p(\beta|b_0)d\boldsymbol{\theta} \tag{19}$$

where $p(\hat{\boldsymbol{\omega}}^2,\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\theta}}_u|\boldsymbol{\delta})$ is the evidence function for the model class $\mathcal{M}(\boldsymbol{\delta})$. The full posterior uncertainty in $\boldsymbol{\theta}$ is explicitly incorporated when finding the MAP estimates of all parameters in $\boldsymbol{\delta}$, although it is a nontrivial task. The full details of the fast SBL algorithm are given in Huang et al. (2017a).

Remark 4.1: The maximization of evidence in (19) is effectively implementing the Bayesian Ockham Razor by assigning lower probabilities to a structural model whose parameter vector $\boldsymbol{\theta}$ has too large or too small differences from $\hat{\boldsymbol{\theta}}_u$ identified from the calibration state (that is, the model extracts relatively more or less information, respectively, from the system modal parameters $\boldsymbol{\omega}^2$ and $\boldsymbol{\phi}$, and so from the "measured" modal data $\hat{\boldsymbol{\omega}}^2$ and $\hat{\boldsymbol{\psi}}$, which can be seen from the hierarchical model in Figure 1). This process suppresses the occurrence of false and missed alarms for stiffness reductions.

Remark 4.2: It was found that the trade-off stated in Section 2.4 is sensitive to the selection of the equation-error precision parameter β. This motivated us to develop a more sophisticated method, described in the next sub-section, to provide a fuller treatment of the posterior uncertainties, including marginalizing over the posterior uncertainty of β analytically to get a more robust solution.

Remark 4.3: In the fast SBL algorithm, the pseudo-data $\hat{\boldsymbol{\theta}}_u$ is used based on the assumption that it is a unique MAP estimate at the calibration state due to the large amount of time-domain vibration data and identified modal parameters that can be collected. In the next subsection, we relax this assumption by explicitly considering the posterior uncertainty of $\boldsymbol{\theta}_u$ at the calibration stage in case there is not sufficient data to get a posterior on $\boldsymbol{\theta}_u$ that is highly peaked at $\hat{\boldsymbol{\theta}}_u$.

4.3 Sparse Bayesian learning algorithm using Gibbs sampling

The goal of the algorithm presented here is to provide a fuller treatment of the posterior uncertainty by employing MCMC simulation methods, so that the Laplace approximations in the fast SBL algorithm that involve the system modal parameters $\{\boldsymbol{\omega}^2,\boldsymbol{\phi}\}$ and the equation-error precision parameter β can be avoided. Based on the hierarchical model presented in Figure 1, the posterior PDF $p(\boldsymbol{\omega}^2,\boldsymbol{\phi},\boldsymbol{\theta}|\hat{\boldsymbol{\omega}}^2,\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\theta}}_u)$ can be calculated by marginalizing over the parameters $\beta, \eta, \rho, \boldsymbol{\alpha}$ and b_0 in the full posterior PDF from Bayes' theorem as follows:

$$p(\boldsymbol{\omega}^2, \boldsymbol{\phi}, \boldsymbol{\theta} | \widehat{\boldsymbol{\omega}}^2, \widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\theta}}_u) = p(\widehat{\boldsymbol{\omega}}^2 | \boldsymbol{\omega}^2) p(\widehat{\boldsymbol{\psi}} | \boldsymbol{\phi}) p(\widehat{\boldsymbol{\theta}}_u | \boldsymbol{\theta}) p(\boldsymbol{\omega}^2, \boldsymbol{\phi}, \boldsymbol{\theta}) / p(\widehat{\boldsymbol{\omega}}^2, \widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\theta}}_u)$$

$$= \int p(\widehat{\boldsymbol{\omega}}^2 | \boldsymbol{\omega}^2, \rho) p(\widehat{\boldsymbol{\psi}} | \boldsymbol{\phi}, \eta) p(\widehat{\boldsymbol{\theta}}_u | \boldsymbol{\theta}, \boldsymbol{\alpha}) p(\boldsymbol{\omega}^2, \boldsymbol{\phi}, \boldsymbol{\theta} | \beta) p(\beta | b_0) p(\rho, \eta, \boldsymbol{\alpha}, b_0) d\beta d\rho d\eta d\boldsymbol{\alpha} db_0 / p(\widehat{\boldsymbol{\omega}}^2, \widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\theta}}_u) \quad (20)$$

The resulting expression is intractable because the high-dimensional normalizing integral $p(\widehat{\boldsymbol{\omega}}^2, \widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\theta}}_u)$ cannot be computed analytically. Instead, we implement Gibbs Sampling to draw posterior samples from $p(\boldsymbol{\phi}, \boldsymbol{\omega}^2, \boldsymbol{\theta} | \widehat{\boldsymbol{\omega}}^2, \widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\theta}}_u)$ by decomposing the whole model parameter vector into the three groups $\{\boldsymbol{\phi}, \boldsymbol{\omega}^2, \boldsymbol{\theta}\}$ and repeatedly sampling from one parameter group conditional on the other two groups and the available data. We can derive the generic form $p(\mathbf{w}_1 | \hat{\mathbf{y}}, \mathbf{w}_2, \mathbf{w}_3) = \int p(\mathbf{w}_1 | \hat{\mathbf{y}}, \mathbf{w}_2, \mathbf{w}_3, \beta) p(\beta | \hat{\mathbf{y}}, \mathbf{w}_2, \mathbf{w}_3) \, d\beta$ of the conditional posterior PDFs $p(\boldsymbol{\phi} | \hat{\mathbf{y}}, \boldsymbol{\omega}^2, \boldsymbol{\theta})$, $p(\boldsymbol{\omega}^2 | \hat{\mathbf{y}}, \boldsymbol{\phi}, \boldsymbol{\theta})$ and $p(\boldsymbol{\theta} | \hat{\mathbf{y}}, \boldsymbol{\phi}, \boldsymbol{\omega}^2)$ by marginalizing over their corresponding nuisance parameters using Laplace approximations (similar to (18) and (19)). The reader is referred to Huang & Beck (2015b) and Huang et al. (2017b) for detailed information of the SBL algorithm using Gibbs Sampling, including the derivation of the generic form of the conditional posterior PDF $p(\mathbf{w}_1 | \hat{\mathbf{y}}, \mathbf{w}_2, \mathbf{w}_3)$ and the pseudo-codes.

If the Markov chain created by the GS algorithms is ergodic, samples from the marginal posterior distributions $p(\boldsymbol{\theta} | \widehat{\boldsymbol{\omega}}^2, \widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\theta}}_u), p(\boldsymbol{\omega}^2 | \widehat{\boldsymbol{\omega}}^2, \widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\theta}}_u)$ and $p(\boldsymbol{\phi} | \widehat{\boldsymbol{\omega}}^2, \widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\theta}}_u)$ are readily obtained by simply examining the GS samples $\boldsymbol{\theta}^{(n)}$ $(\boldsymbol{\omega}^2)^{(n)}$ and $\boldsymbol{\phi}^{(n)}$, respectively, for larger iteration numbers beyond the burn-in period. Using samples from the marginal posterior PDF $p(\boldsymbol{\theta}_u | \widehat{\boldsymbol{\psi}}_u, \widehat{\boldsymbol{\omega}}_u^2)$ at the calibration stage, we are able to effectively take into account the uncertainty of $\boldsymbol{\theta}_\mathrm{u}$ during the monitoring stage by replacing the MAP value $\widehat{\boldsymbol{\theta}}_u$ with uncertain $\boldsymbol{\theta}_u$, and then drawing samples from the posterior PDF $p(\boldsymbol{\theta} | \widehat{\boldsymbol{\omega}}_d^2, \widehat{\boldsymbol{\psi}}_d, \widehat{\boldsymbol{\omega}}_u^2, \widehat{\boldsymbol{\psi}}_u)$ for the monitoring stage, which is conditional on modal data from both the monitoring and calibration stages.

Remark 4.4: The analytical derivation of the generic conditional posterior PDF $p(\mathbf{w}_1 | \hat{\mathbf{y}}, \mathbf{w}_2, \mathbf{w}_3)$ is important for the effectiveness of this Gibbs Sampling algorithm, which leads to a very desirable feature that it is applicable to linear Bayesian model updating problems of arbitrarily high dimensions, in contrast with other MCMC algorithms.

Remark 4.5: In the Gibbs Sampling Algorithm, by marginalizing over β directly to remove it from the posterior distributions, we get the Student-t conditional PDFs that can be sampled in each step of the algorithm. The Student-t PDFs have heavier tails than the Gaussian PDFs sampled in Algorithm 1 and so the algorithm is more robust to noise and outliers.

Remark 4.6: For the updating of the stiffness scaling parameters $\boldsymbol{\theta}$ and system modal parameters $\boldsymbol{\omega}^2$ and $\boldsymbol{\phi}$, the corresponding model classes $\mathcal{M}(\boldsymbol{\gamma}, b_0), \mathcal{M}(\upsilon, b_0)$ and$\mathcal{M}(\tau, b_0)$are investigated, as seen from the hierarchical Bayesian model in Figure 1. The application of Bayes' Theorem at the model class level automatically penalizes models of $\boldsymbol{\theta}$ ($\boldsymbol{\omega}^2$or $\boldsymbol{\phi}$) that "under-fit" or "over-fit" the associated data $\widehat{\boldsymbol{\theta}}_u$ ($\widehat{\boldsymbol{\omega}}^2$or $\widehat{\boldsymbol{\psi}}$), therefore obtaining reliable updating results for the three parameter vectors, which is the Bayesian Ockham Razor (Beck, 2010) at work.

4.4 Illustrative results of the sparse Bayesian learning algorithms

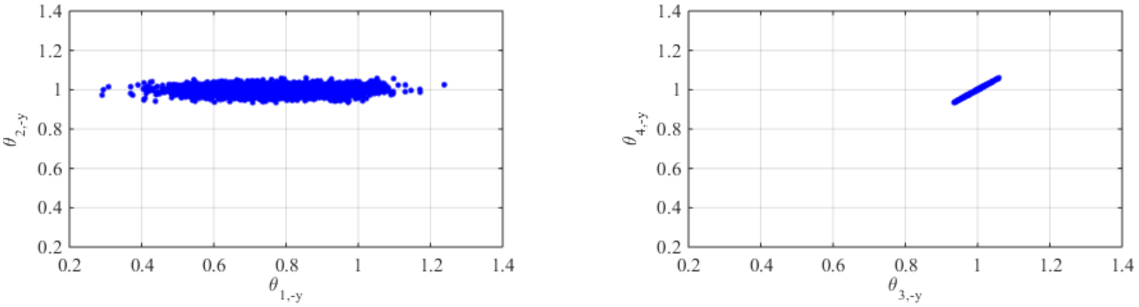
The proposed methodologies are applied to the brace damage patterns in the IASC-ASCE experimental Phase II benchmark problem (Dyke et al., 2003; Ching & Beck, 2003). The benchmark structure is a four-story, two-bay by two-bay steel braced-frame. Three damage configurations (Configs. 4,5,6) and one calibration (undamaged) configuration are investigated in this study. The stiffness scaling parameter vector $\boldsymbol{\theta}$ has 16 components, one for each of the four faces of each of the four stories. The true ratio values for $\theta_{1,-y}$ and $\theta_{4,-y}$ for Config. 4, and $\theta_{1,-y}$ for Config. 5, are 77.4% and the true ratio value $\theta_{1,-y}$ for Config. 6 is 54.9% of the values for the calibration configuration.

In Figure 2, all the samples generated from the Gibbs Sampling algorithm, excluding those in the burn-in period (4000 samples), are plotted in the $\{\theta_{1,-y}, \theta_{2,-y}\}$ and $\{\theta_{3,-y}, \theta_{4,-y}\}$ spaces for Config. 5. They show that the stiffness reduction corresponding to $\theta_{1,-y}$ is correctly identified and quantified as far as the sample means are concerned. Smaller posterior uncertainties can be observed in the stiffness scaling parameters for undamaged substructures, which is a benefit of the procedure of continuous model class selection by learning of the hyperparameters in the SBL formulation.

Figure 3 compares the probability that any stiffness parameter θ_j of a substructure has decreased by more than a prescribed fraction f estimated using the computed posterior PDFs (fast algorithm) or posterior samples (Gibbs Sampling algorithm). It is seen that the two algorithms generate similar results for Config.

4; however, for Config. 5 the posterior uncertainty of the stiffness parameters for the undamaged substructures are smaller for the Gibbs Sampling algorithm. For Config. 6, the occurrence of false damage detections is more unlikely for the Gibbs Sampling algorithm, presumably due to the robust treatment of the equation-error precision parameter β and stiffness parameter vector at the calibration state by a fuller model uncertainty quantification.

Remark 4.7: Much more computing resources are required for the Gibbs Sampling algorithm than the fast algorithm, which is a sacrifice for better posterior uncertainty quantification. Therefore, the choice between these two methods in real applications is a trade-off between the computation time and the level of uncertainty quantification and identification accuracy that the user is willing to accept.



a

b

Fig. 2. Post burn-in samples for some posterior stiffness parameters for the Config. 5 scenario, plotted in: a – $\{\theta_{1,-y}, \theta_{2,-y}\}$; b –$\{\theta_{3,-y}, \theta_{4,-y}\}$ spaces.

5 Concluding remarks

Probability logic combined with a Bayesian approach provides a rigorous framework to quantify modeling uncertainty in model updating in structural health monitoring. It allows plausible reasoning about structural behavior based on incomplete information. A key concept is a stochastic system model class which defines the fundamental probability models that allow robust stochastic structural analyses to be performed. Such a model class can be constructed by stochastic embedding of any deterministic model of the structure's input-output behavior. One distinguishing aspect of the proposed Bayesian framework is marginalization of posterior PDFs, where instead of seeking to estimate all 'nuisance' parameters in the models, we attempt to integrate them out. This allows us to assess the relative plausibility of each model within a set of candidate model classes chosen to represent the uncertain structural behavior. Applying Bayes' Theorem at the model class level automatically penalizes models that are too simple ("under-fit" the data) and too complex ("over-fit" the data), which is the Bayesian Ockham Razor at work. The quantitative implementation of Ockham's Razor is a natural consequence of applying Bayesian updating at the model class level.

Sparse Bayesian learning is an effective strategy to incorporate sparseness during model updating by automatically implementing Ockham's Razor. This alleviates ill-conditioning and ill-posedness in the inverse problem in system identification. Recently developed sparse Bayesian learning algorithms for model updating and system identification have been briefly reviewed and illustrated using identified modal data. A promising performance of the algorithms has been shown by the illustrative results.

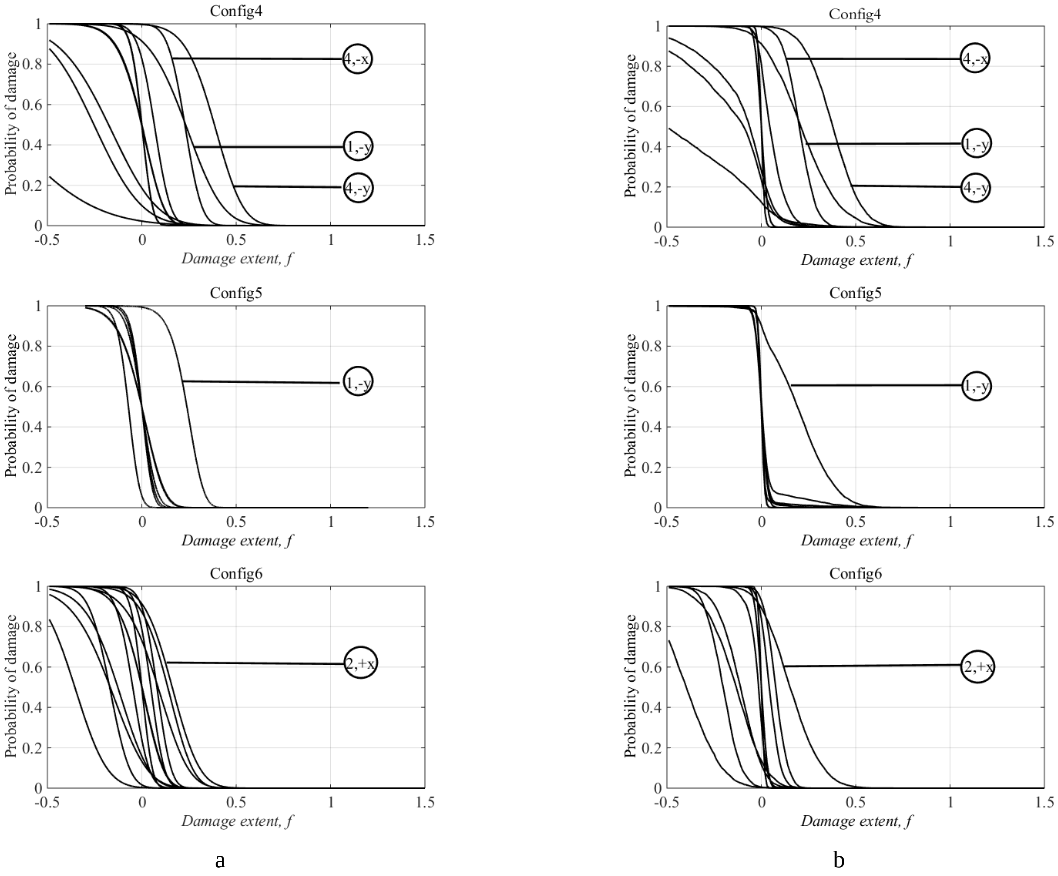


Fig. 3. Estimated damage probability curves for each substructure by running: a –Fast algorithm in Subsection 5.1.2; b – Gibbs Sampling algorithm in Subsection 4.3 using 4,000 post burn-in samples.

Acknowledgement

The second author was supported by the George W. Housner Earthquake Engineering Research fund at the California Institute of Technology and a grant from the National Natural Science Foundation of China (No. 51308161). This support is gratefully acknowledged.

References

Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19: 716-723, 1974.

Siu-Kui Au and Fengliang Zhang, Fundamental two-stage formulation for Bayesian system identification, Part I: General theory, *Mechanical Systems and Signal Processing*, 66–67: 31–42. 2016.

S. Derin Babacan, Rafael Molina and Aggelos K. Katsaggelos. Bayesian compressive sensing using Laplace priors. *IEEE Transactions on Signal Processing*, 19(1): 53-64, 2010.

James O. Berger. *Statistical decision theory and Bayesian analysis*. Springer, second edition,1985.

James L. Beck and Lambros S. Katafygiotis. Updating models and their uncertainties. I: Bayesian statistical framework. *Journal of Engineering Mechanics*, 124: 455-461, 1998.

James L. Beck and Ka-Veng Yuen. Model selection using response measurements: a Bayesian probabilistic approach. *Journal of Engineering Mechanics*, 130: 192-203, 2004.

James L. Beck. Bayesian system identification based on probability logic. *Structural Control and Health Monitoring*, 17: 825-847, 2010.

James L. Beck and Alexandros Taflanidis. Prior and posterior robust stochastic predictions for dynamical systems using probability logic. *International Journal for Uncertainty Quantification*, 3: 271-288, 2013.

Emmanuel J. Candès. Compressive sampling. In *Proceedings of the International Congress of Mathematicians*, Madrid, Spain, 2006.

Sai Hung Cheung and James L. Beck. Calculation of the posterior probability for Bayesian model class assessment and averaging from posterior samples based on dynamic system data. *Computer-Aided Civil and Infrastructure Engineering*, 25: 304-321, 2010.

Manuel Chiachio, James L. Beck, Juan Chiachio and Rus Guillermo. Approximate Bayesian computation by subset simulation. *SIAM Journal on Scientific Computing*, 36 (3): A1339-A1358, 2014.

Jianye Ching and James L. Beck. Two-step Bayesian structure health monitoring approach for IASC-ASCE phase II simulated and experimental benchmark studies, Technical Report EERL 2003-02, Earthquake Engineering Research Laboratory, California Institute of Technology, Pasadena, CA, 2003.

Jianye Ching, Matthew Muto and James L. Beck. Bayesian linear structural model updating using Gibbs sampler with modal data. In *Proceedings of the 9th International Conference on Structural Safety and Reliability*, Rome, Italy, 2005.

Jianye Ching and Yichu Chen. Transitional Markov Chain Monte Carlo method for Bayesian model updating, model class selection and model averaging. *Journal of Engineering Mechanics*, 133: 816-832,2007.

Richard T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14(1): 1–13, 1946.

Richard T. Cox. *The Algebra of Probable Inference*. Johns Hopkins Press: Baltimore, MD, 1961.

David Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4): 1289-1306, 2006.

Shirley J. Dyke, Dionisio Bernal, James L. Beck and Carlos Ventura. Experimental Phase II of the Structural Health Monitoring Benchmark Problem. In *Proceedings of 16th Eng. Mechanics Conf.*, ASCE, Reston, VA, 2003.

Stephen F. Gull. *Bayesian inductive inference and maximum entropy*. In G. J. Erickson & C. R. Smith (eds), Maximum Entropy and Bayesian Methods, 53-74. Dordrecht, Nether-lands: Kluwer Academic Publishers, 1988.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson Aki Vehtari and Donald B. Rubin. *Bayesian data analysis*, third edition, Chapman & Hall/CRC, 2013.

Peter L. Green, Elizabeth J. Cross and Keith Worden. Bayesian system identification of dynamical systems using highly informative training data, *Mechanical Systems and Signal Processing*, 56–57: 109–122, 2015.

Trevor Hastie, Robert Tibshirani and Martin Wainwright. *Statistical Learning with Sparseness: The Lasso and Generalizations*. Chapman and Hall/CRC, 2015.

Yong Huang, James L. Beck, Stephen Wu and Hui Li. Robust Diagnostics for Bayesian Compressive Sensing Technique in Structural Health Monitoring, In *Proceedings of the 8th International Workshop on Structural Health Monitoring*, Stanford, CA, 2011

Yong Huang, James L. Beck, Stephen Wu and Hui Li. Robust Bayesian compressive sensing for signals in structural health monitoring. *Computer-Aided Civil and Infrastructure Engineering*, 29(3):160–179, 2014.

Yong Huang and James L. Beck, Hierarchical Sparse Bayesian Learning for Structural Health Monitoring with incomplete Modal Data. *International Journal for Uncertainty Quantification*, 5(2): 139-169, 2015a.

Yong Huang and James L. Beck. Sparse Bayesian learning with Gibbs Sampling for structural health monitoring with noisy incomplete modal data, In *Proceedings of the 12th International Conference on Applications of Statistics and Probability in Civil Engineering*, Vancouver, Canada, 2015b.

Yong Huang, James L. Beck, Stephen Wu and Hui Li. Bayesian compressive sensing for approximately sparse signals and application to structural health monitoring signals for data loss recovery. *Probabilistic Engineering Mechanics*, 46: 62–79, 2016.

Yong Huang, James L. Beck and Hui Li. Hierarchical sparse Bayesian learning for structural damage detection： Theory, computation and application. *Structural Safety*, 64, 37-53, 2017a.

Yong Huang, James L. Beck and Hui Li. Bayesian system identification based on hierarchical sparse Bayesian learning and Gibbs Sampling with application to structural damage assessment. *Computer Methods in Applied Mechanics and Engineering*, in press, 2017b.

Edwin T.Jaynes. *Papers on Probability, Statistics and Statistical Physics*. R.D. Rosenkrantz (ed.). Dordrecht, Holland: D. Reidel Publishing, 1983.

Edwin T.Jaynes. *Probability theory as logic*. In P.F. Fougere (ed.), Maximum Entropy and Bayesian Methods. Dordrecht, Netherlands: Kluwer Academic Publishers, 1990.

Edwin T.Jaynes. *Probability Theory: The Logic of Science*. Cambridge, UK: Cambridge University Press, 2003.

William H. Jefferys and James O. Berger. Ockham's Razor and Bayesian Analysis. *American Scientist*, 80: 64-72, 1992.

David J.C. Mackay. Bayesian Methods for Adaptive Models. PhD Thesis in Computation and Neural Systems. California Institute of Technology, Pasadena, CA, 1992.

Matthew Muto and James L. Beck. Bayesian updating and model class selection for hysteretic structural models using stochastic simulation. *Journal of Vibration and Control*, 14: 7-34, 2008.

Costas Papadimitriou, James L. Beck and Lambros S. Katafygiotis. Updating robust reliability using structural test data. *Probabilistic Engineering Mechanics*, 16: 103-113,2001.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6: 461-464, 1978.

Michael E. Tipping. *The Relevance Vector Machine*. Advances in Neural Information Processing Systems 12, pages 652-658. MIT Press, 2000.

Michael E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1: 211-244, 2001a.

Michael E. Tipping. Sparse kernel principal component analysis. In *Advances in Neural Information Processing Systems*, volume 13, pages 633–639, Cambridge, MA, 2001b.

Joel A. Tropp. Topics in Sparse Approximation. Ph.D. dissertation, Computational and Applied Mathematics, University Texas at Austin, Austin, TX, 2004.

Majid K. Vakilzadeh, Yong Huang, James L. Beck and Thomas Abrahamsson. Approximate Bayesian Computation by Subset Simulation using hierarchical state-space models. *Mechanical Systems and Signal Processing*, 84, Part B: 2–20, 2017.

Ka-Veng Yuen, James L. Beck and Lambros S. Katafygiotis. Efficient model updating and health monitoring methodology using incomplete modal data without mode matching. *Structural Control and Health Monitoring*, 13 (1): 91–107, 2006.