Structural Safety 32 (2010) 338–346

Contents lists available at ScienceDirect

Structural Safety

journal homepage: www.elsevier.com/locate/strusafe

Recent developments of Bayesian model class selection and applications in civil engineering

Ka-Veng Yuen *,[1]

Department of Civil and Environmental Engineering, University of Macau, Macao, China

ARTICLE INFO

Article history:
Available online 10 April 2010

Keywords:
Bayesian inference
Empirical relationship
Model class selection
Model updating
Regression
System identification

ABSTRACT

Bayesian model class selection has attracted substantial interest in recent years for selecting the most plausible/suitable class of models based on system input–output data. The Bayesian approach provides a quantitative expression of a principle of model parsimony or of Ockham's razor which in engineering applications can be stated as simpler models are to be preferred over unnecessarily complicated ones. In this paper, some recent developments are reviewed. Linear and nonlinear regression problems are considered in detail. Bayesian model class selection is particularly useful for regression problems since the regression formula order is difficult to be determined solely by physics due to its empirical nature. Applications are presented in different areas of civil engineering, including artificial neural network for damage detection and seismic attenuation empirical relationship.

1. Introduction

The usual approach in system identification is to find the best/optimal model in a prescribed class of models, e.g., class of shear building models with uncertain inter-story stiffnesses or class of quadratic polynomials with uncertain coefficients to be identified. This problem is commonly referred to as *parametric identification*. The more general problem of model class selection has not been well explored in system identification. It is obvious that a more complicated model class often fits the data better than one which has fewer adjustable uncertain parameters. Therefore, if the optimal model class is chosen by minimizing a norm of the fitting error between the output data and the corresponding predictions by the optimal model in each model class, the choice will tend to the ones with more effective free parameters. This approach is therefore likely to lead to over-fitting the data. When an over-fitted model is used for future prediction, it leads to poor results because the model parameters depend too much on the detail of the data and the measurement noise has an important role in the data fitting. Therefore, in model class selection, it is necessary to penalize a complicated model but the quantification of this penalty is not a trivial task. This point was recognized by Jeffreys who did pioneering work on the application of Bayesian methods [12]. In the present context, the selected class of models should agree closely with the observed behavior of the system but otherwise be as simple as possible. In recent years, there has been a reappreciation of the work of Jeffreys on the application of Bayesian methods. In particular, the Bayesian approach to model class selection has been further developed by showing that the *evidence* for each model class provided by the data automatically enforces a quantitative expression of a principle of model parsimony or of Ockham's razor [10,20,21]. There is no need to introduce ad hoc penalty as in some of the earlier work on model class selection.

There are many types of uncertainty involved in civil engineering problems, such as materials, excitation, modeling, emission, and traffic loading. Therefore, there is great demand on explicit treatment of these uncertainties in civil engineering applications and Bayesian analysis is suitable for the need. However, Bayesian applications in civil engineering are still in the early stage so there is great potential for exploration [24]. In this paper, some recent developments and civil engineering applications of Bayesian model class selection are reviewed. In the next section, the Bayesian model class selection method is revisited. It is illustrated for globally identifiable case with asymptotic expansion and for general case with the transitional Markov chain Monte Carlo (TMCMC) method [7]. Special cases of linear and nonlinear regression are considered in detail in Section 3. Bayesian model class selection is particularly useful for the selection of regression formula since the structure of the empirical relationship is difficult to be determined by the physics of the phenomenon alone. In Sections 4 and 5, applications are presented for two areas in civil engineering, including artificial neural network structure design and seismic attenuation empirical relationship. Other applications in the modeling of particulate matter concentration [11] and soil compressibility index empirical relationships [22] have also been studied.

* Tel.: +86 (853) 83974960; fax: +86 (853) 28838314.
E-mail address: kvyuen@umac.mo
[1] Associate professor and head.

0167-4730/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.strusafe.2010.03.011

2. Bayesian model class selection

Use $\mathcal{D}$ to denote some input–output measurements from a physical phenomenon/system. The goal is to use $\mathcal{D}$ to select the most plausible/suitable class of models representing the phenomenon/system among N_C prescribed model classes $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_{N_C}$. Since probability may be interpreted as a measure of plausibility based on specified information [9], the plausibility of a model class conditional on the data $\mathcal{D}$ can be obtained by using the Bayes' theorem:

$$P(\mathcal{C}_j|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{C}_j)P(\mathcal{C}_j)}{p(\mathcal{D})}, \quad j = 1, 2, \ldots, N_C \tag{1}$$

where $p(\mathcal{D}) = \sum_{j=1}^{N_C} p(\mathcal{D}|\mathcal{C}_j)P(\mathcal{C}_j)$ by the theorem of total probability and $P(\mathcal{C}_j)$ expresses the user's judgement on the initial plausibility of the model classes with $\sum_{j=1}^{N_C} P(\mathcal{C}_j) = 1$. The factor $p(\mathcal{D}|\mathcal{C}_j)$ is called the *evidence* for the model class $\mathcal{C}_j$ provided by the data $\mathcal{D}$. The most plausible model class is the one that maximizes $p(\mathcal{D}|\mathcal{C}_j)P(\mathcal{C}_j)$.

The evidence for $\mathcal{C}_j$ provided by the data $\mathcal{D}$ can be obtained from the theorem of total probability:

$$p(\mathcal{D}|\mathcal{C}_j) = \int_{\boldsymbol{\Theta}} p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{C}_j)p(\boldsymbol{\theta}|\mathcal{C}_j)\, d\boldsymbol{\theta}, \quad j = 1, 2, \ldots, N_C \tag{2}$$

where $\boldsymbol{\theta}$ is the parameter vector in the parameter space $\boldsymbol{\Theta} \subset \mathbb{R}^{N_j}$ and it defines each model in $\mathcal{C}_j$. The parameter vector $\boldsymbol{\theta}$ and the parameter space $\boldsymbol{\Theta}$ depend on the model class $\mathcal{C}_j$ even though it is not explicitly reflected in the symbol. This is done for the purpose of simplifying the notation only. The prior PDF $p(\boldsymbol{\theta}|\mathcal{C}_j)$ is specified by the user based on engineering judgement and $p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{C}_j)$ is the likelihood function that represents the contribution of the data in the updating process. Some examples of the establishment of the likelihood function for structural dynamics problems can be found in [14,28,26,25,29,27].

2.1. Globally identifiable case

In *globally identifiable* cases [2], the posterior/updated PDF, $p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{C}_j)$, for the uncertain parameters given a large amount of data may be approximated accurately as Gaussian, so the evidence, $p(\mathcal{D}|\mathcal{C}_j)$, can be approximated by asymptotic approximation [3]:

$$p(\mathcal{D}|\mathcal{C}_j) \approx p(\mathcal{D}|\boldsymbol{\theta}^\star, \mathcal{C}_j)p(\boldsymbol{\theta}^\star|\mathcal{C}_j)(2\pi)^{\frac{N_j}{2}}|\mathcal{H}_j(\boldsymbol{\theta}^\star)|^{-\frac{1}{2}}, \quad j = 1, 2, \ldots, N_C \tag{3}$$

where N_j is the number of uncertain parameters for the model class $\mathcal{C}_j$. The optimal/updated parameter vector $\boldsymbol{\theta}^\star$ is the most probable value, that maximizes $p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{C}_j)$ in the interior of $\boldsymbol{\Theta}$, and $\mathcal{H}_j(\boldsymbol{\theta}^\star)$ is the Hessian matrix of the objective function $-\ln[p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{C}_j)p(\boldsymbol{\theta}|\mathcal{C}_j)]$ with respect to $\boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta}^\star$. The maximum likelihood value, $p(\mathcal{D}|\boldsymbol{\theta}^\star, \mathcal{C}_j)$, in Eq. (3) will be higher for those model classes that make the probability of the data $\mathcal{D}$ higher, or equivalently, those which give better fitting to the data. As mentioned earlier, this likelihood function favors model classes with more effective uncertain parameters.

Except for the maximum likelihood value in Eq. (3), the product of the remaining factors is called the *Ockham factor* [10,20]:

$$\mathcal{O}_j = p(\boldsymbol{\theta}^\star|\mathcal{C}_j)(2\pi)^{\frac{N_j}{2}}|\mathcal{H}_j(\boldsymbol{\theta}^\star)|^{-\frac{1}{2}} \tag{4}$$

It represents a penalty against complicated parameterization. It was shown that the Ockham factor decreases exponentially with the number of uncertain parameters in the model class and the log Ockham factor is given by [3]:

$$\ln \mathcal{O}_j = -\frac{1}{2}N_j \ln N + \mathcal{R}_j \tag{5}$$

where the remainder $\mathcal{R}_j$ depends primarily on the choice of prior PDF and is of order 1 for large N. If the number of data points N in $\mathcal{D}$ is large, the likelihood function will be the dominant one in Eq. (3) because it increases exponentially with N, while the Ockham factor behaves as N^{-1} [23].

The Ockham factor can be further elaborated. It follows from Bayes' theorem that the exact relationship is given by

$$p(\mathcal{D}|\mathcal{C}_j) = p(\mathcal{D}|\boldsymbol{\theta}^\star, \mathcal{C}_j)p(\boldsymbol{\theta}^\star|\mathcal{C}_j)/p(\boldsymbol{\theta}^\star|\mathcal{D}, \mathcal{C}_j) \tag{6}$$

By comparing this equation with Eq. (3), the Ockham factor is approximately equal to the ratio $p(\boldsymbol{\theta}^\star|\mathcal{C}_j)/p(\boldsymbol{\theta}^\star|\mathcal{D}, \mathcal{C}_j)$ which is always less than unity if the data provides any information about the model parameters. For large N, the negative logarithm of this ratio is an asymptotic approximation of the information about $\boldsymbol{\theta}$ provided by data $\mathcal{D}$ [16]. Therefore, the log Ockham factor removes the amount of information about $\boldsymbol{\theta}$ provided by $\mathcal{D}$ from the log likelihood, $\ln p(\mathcal{D}|\boldsymbol{\theta}^\star, \mathcal{C}_j)$, to give the log evidence, $\ln p(\mathcal{D}|\mathcal{C}_j)$.

2.2. General case

For the general case where the posterior PDF may not be approximated by Gaussian distribution, the asymptotic expansion in Eq. (3) is not valid. In general, the log evidence is equal to the difference of two integrals [8]:

$$\ln p(\mathcal{D}|\mathcal{C}_j) = \int_{\boldsymbol{\Theta}} [\ln p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{C}_j)]p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{C}_j)\, d\boldsymbol{\theta} - \int_{\boldsymbol{\Theta}} \left[\ln \frac{p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{C}_j)}{p(\boldsymbol{\theta}|\mathcal{C}_j)}\right] p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{C}_j)\, d\boldsymbol{\theta} \tag{7}$$

The first integral is a measure of the average log goodness-of-fit of the model class $\mathcal{C}_j$ and the second is the relative entropy between the prior and posterior PDFs [17]. The latter quantity is the information gained about the parameters by the data $\mathcal{D}$. Therefore, Eq. (7) states that the log evidence of a model class is equal to the average log goodness-of-fit, penalized by the information gained on the model parameters. Small uncertainty of the model parameters is the consequence of high model output sensitivity due to changes in the model parameters. However, this does not necessary give a good model class for future prediction because the model output may be too sensitive to parametric and modeling error. On the other hand, a robust model class for future prediction should have good data fitting capability and small output difference due to perturbation of the model parameters.

For locally identifiable case or unidentifiable case, the asymptotic expansion in Eq. (3) is not applicable and the transitional Markov chain Monte Carlo algorithm [7] can be used for evidence computation. It is a method for sampling the posterior PDF of a model class and it was developed under the concept of an adaptive Markov chain Monte Carlo simulation procedure [1]. First, a sequence of un-normalized intermediate PDFs, $\mathcal{P}_s(\boldsymbol{\theta})$, $s = 0, 1, \ldots, s_0$, is constructed:

$$\mathcal{P}_s(\boldsymbol{\theta}) \propto p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{C}_j)^{\beta_s} p(\boldsymbol{\theta}|\mathcal{C}_j) \tag{8}$$

where β_s increases monotonically with s such that $\beta_0 = 0$ and $\beta_{s_0} = 1$. Samples of the model parameters are generated according to each intermediate PDF, $\mathcal{P}_s(\boldsymbol{\theta})$, $s = 0, 1, \ldots, s_0$.

The resampling weighting for each sample is the ratio of the target PDFs for the sth and $(s-1)$th levels, evaluated at $\boldsymbol{\theta}_n^{(s-1)}$:

$$w(\boldsymbol{\theta}_n^{(s-1)}) = \frac{p(\mathcal{D}|\boldsymbol{\theta}_n^{(s-1)}, \mathcal{C}_j)^{\beta_s} p(\boldsymbol{\theta}_n^{(s-1)}|\mathcal{C}_j)}{p(\mathcal{D}|\boldsymbol{\theta}_n^{(s-1)}, \mathcal{C}_j)^{\beta_{s-1}} p(\boldsymbol{\theta}_n^{(s-1)}|\mathcal{C}_j)} = p(\mathcal{D}|\boldsymbol{\theta}_n^{(s-1)}, \mathcal{C}_j)^{\beta_s - \beta_{s-1}} \tag{9}$$

Finally, the evidence $p(\mathcal{D}|\mathcal{C}_j)$ can be estimated by the sample average of the resampling weightings:

$$p(\mathcal{D}|\mathcal{C}_j) = \frac{\int_\Theta p(\mathcal{D}|\boldsymbol{\theta},\mathcal{C}_j)p(\boldsymbol{\theta}|\mathcal{C}_j)\,d\boldsymbol{\theta}}{\int_\Theta p(\boldsymbol{\theta}|\mathcal{C}_j)\,d\boldsymbol{\theta}} = \prod_{s=1}^{s_0}\overline{W}_s \tag{10}$$

where the quantities $\overline{W}_s,\ s = 1,2,\ldots,s_0$, are estimated by

$$\overline{W}_s \approx \frac{1}{N}\sum_{n=1}^{N} w\left(\boldsymbol{\theta}_n^{(s-1)}\right) \tag{11}$$

For details, please refer to Ching and Chen [7].

3. Regression problems

3.1. Linear regression problems

A general linear regression formula of a quantity of concern, Q, can be written as

$$Q(\mathbf{x};\mathbf{b},\mathcal{C}_j) = \sum_{l=1}^{N_b} b_l x_l \tag{12}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_{N_b}]^T$ is a known vector including the measured variables used in establishing the regression formula; and $\mathbf{b} = [b_1, b_2, \ldots, b_{N_b}]^T \in \mathbb{R}^{N_b}$ is the vector of the unknown coefficients to be identified. The measurement of Q can be written as

$$y = Q(\mathbf{x};\mathbf{b},\mathcal{C}_j) + \epsilon \tag{13}$$

where ϵ is a Gaussian random variable with zero mean and variance σ_ϵ^2. It is used to represent the measurement noise and modeling error. The uncertain parameters in $\boldsymbol{\theta}$ include the coefficients b_l's and the prediction-error variance σ_ϵ^2 so $N_j = N_b + 1$.

The data $\mathcal{D}$ includes the measurement of $\mathbf{x}$ and the corresponding values for y. By assuming that the prediction errors in different records are statistically independent, the likelihood function is obtained:

$$p(\mathcal{D}|\boldsymbol{\theta},\mathcal{C}_j) = (2\pi)^{-\frac{N}{2}}\sigma_\epsilon^{-N}\exp\left[-\frac{N}{2\sigma_\epsilon^2}J_g(\mathbf{b}|\mathcal{D},\mathcal{C}_j)\right] \tag{14}$$

where N is the total number of measured records. The goodness-of-fit function $J_g(\mathbf{b}|\mathcal{D},\mathcal{C}_j)$ represents the level of data fitting, which is given by

$$J_g(\mathbf{b}|\mathcal{D},\mathcal{C}_j) = \frac{1}{N}\sum_{n=1}^{N}\left[y(n) - \sum_{l=1}^{N_b} b_l x_l(n)\right]^2 \tag{15}$$

A smaller value of this function implies better fitting to the data. In the case of a uniform prior PDF of the coefficients, the optimal coefficient vector $\mathbf{b}^\star$ can be obtained by minimizing $J_g(\mathbf{b}|\mathcal{D},\mathcal{C}_j)$. This is done by solving the linear algebraic equation: $\frac{\partial J_g(\mathbf{b}|\mathcal{D},\mathcal{C}_j)}{\partial \mathbf{b}} = \mathbf{0}$, and the updated coefficient vector is readily obtained:

$$\mathbf{b}^\star = \mathbf{A}^{-1}\begin{bmatrix} \sum_{n=1}^{N} x_1(n)y(n) \\ \sum_{n=1}^{N} x_2(n)y(n) \\ \vdots \\ \sum_{n=1}^{N} x_{N_b}(n)y(n) \end{bmatrix} \tag{16}$$

where $\mathbf{A}$ is an $N_b \times N_b$ symmetric matrix and its (l,l') component is $A^{(l,l')} = \sum_{n=1}^{N} x_l(n)x_{l'}(n)$. Furthermore, the updated fitting-error variance $\sigma_\epsilon^{2\star}$ can be obtained by maximizing the likelihood function $p(\mathcal{D}|\boldsymbol{\theta},\mathcal{C}_j)$:

$$\sigma_\epsilon^{2\star} = \min_{\mathbf{b}} J_g(\mathbf{b}|\mathcal{D},\mathcal{C}_j) = J_g(\mathbf{b}^\star|\mathcal{D},\mathcal{C}_j) \tag{17}$$

It is simply the variance of the fitting error.

For large N, the posterior PDF $p(\boldsymbol{\theta}|\mathcal{D},\mathcal{C}_j)$ is approximately Gaussian centered at the optimal parameter $\boldsymbol{\theta}^\star = \left[\mathbf{b}^{\star T}, \sigma_\epsilon^{2\star}\right]^T$. Therefore, the uncertainty of the parameter estimates can be represented by the covariance matrix given by $\boldsymbol{\Sigma}_\theta = \mathcal{H}_j(\boldsymbol{\theta}^\star)^{-1}$, where the Hessian matrix $\mathcal{H}_j(\boldsymbol{\theta}^\star)$ is given by

$$\mathcal{H}_j(\boldsymbol{\theta}^\star) = \frac{1}{\sigma_\epsilon^{2\star}}\begin{bmatrix} \mathbf{A} & \mathbf{0}_{N_b\times 1} \\ \mathbf{0}_{1\times N_b} & 2N \end{bmatrix} \tag{18}$$

The diagonal elements of the covariance matrix $\boldsymbol{\Sigma}_\theta$ are the marginal variance of the corresponding element of $\boldsymbol{\theta}$ and the quantification of the uncertainty of the model parameters can be used for the uncertainty analysis of the prediction. Finally, by using the asymptotic expansion in Eq. (3), the evidence $p(\mathcal{D}|\mathcal{C}_j)$ can be computed:

$$p(\mathcal{D}|\mathcal{C}_j) = p(\boldsymbol{\theta}^\star|\mathcal{C}_j)\exp\left(-\frac{N}{2}\right)\frac{\sigma_\epsilon^{\star N_j - N}}{\sqrt{2(2\pi)^{N-N_j}N|\mathbf{A}|}} \tag{19}$$

3.2. Nonlinear regression problems

Instead of the linear regression formula in Eq. (12), the class of nonlinear regression formulae is considered in this section. There is generally no closed-form solution of the updated model parameters in contrast to the linear case. The following efficient algorithm may be used to search for the updated parameters. First of all, a nonlinear regression formula of a target quantity, Q, takes the following general form:

$$Q(\mathbf{x};\mathbf{b},\mathbf{c},\mathcal{C}_j) = f(\mathbf{c},\mathbf{x}) + \sum_{l=1}^{N_b} b_l f_l(\mathbf{c},\mathbf{x}) \tag{20}$$

where f and f_l's are the functions specified by the user but with unknown coefficients in $\mathbf{c}$. The measurement of Q is given by

$$y = Q(\mathbf{x};\mathbf{b},\mathbf{c},\mathcal{C}_j) + \epsilon \tag{21}$$

where the fitting/prediction error ϵ is again modeled as zero-mean Gaussian random variable with variance σ_ϵ^2. Similar to Eq. (14), the likelihood function is given by

$$p(\mathcal{D}|\boldsymbol{\theta},\mathcal{C}_j) = (2\pi)^{-\frac{N}{2}}\sigma_\epsilon^{-N}\exp\left[-\frac{N}{2\sigma_\epsilon^2}J_g(\mathbf{b},\mathbf{c}|\mathcal{D},\mathcal{C}_j)\right] \tag{22}$$

where the goodness-of-fit function is

$$J_g(\mathbf{b},\mathbf{c}|\mathcal{D},\mathcal{C}_j) = \frac{1}{N}\sum_{n=1}^{N}\left[y(n) - f(\mathbf{c},\mathbf{x}(n)) - \sum_{l=1}^{N_b} b_l f_l(\mathbf{c},\mathbf{x}(n))\right]^2 \tag{23}$$

As in the linear case, a uniform prior for all the uncertain parameters is taken. Given a particular parameter vector $\mathbf{c}'$ for $\mathbf{c}$, one can compute the conditional optimal value for $\mathbf{b}$ in a similar fashion as Eq. (16):

$$\mathbf{b}^\star(\mathbf{c}') = \mathbf{A}^{-1}\begin{bmatrix} \sum_{n=1}^{N} f_1(\mathbf{c},\mathbf{x}(n))[y(n) - f(\mathbf{c}',\mathbf{x}(n))] \\ \sum_{n=1}^{N} f_2(\mathbf{c},\mathbf{x}(n))[y(n) - f(\mathbf{c}',\mathbf{x}(n))] \\ \vdots \\ \sum_{n=1}^{N} f_{N_b}(\mathbf{c},\mathbf{x}(n))[y(n) - f(\mathbf{c}',\mathbf{x}(n))] \end{bmatrix} \tag{24}$$

and the conditional optimal value for σ_ϵ^2 is given by

$$\sigma_\epsilon^{2\star}(\mathbf{c}') = \min_{\mathbf{b}} J_g(\mathbf{b},\mathbf{c}'|\mathcal{D},\mathcal{C}_j) = J_g(\mathbf{b}^\star,\mathbf{c}'|\mathcal{D},\mathcal{C}_j) \tag{25}$$

By maximizing numerically the goodness-of-fit function in Eq. (23) with respect to $\mathbf{c}$, the updated model parameters can be obtained: $\boldsymbol{\theta}^\star = \left[\mathbf{b}^{\star T}, \mathbf{c}^{\star T}, \sigma_\epsilon^{2\star}\right]^T$. Therefore, the availability of the closed-form solution of the conditional optimal parameters allows one to reduce the original optimization problem with $N_b + N_c + 1$ parameters to the problem with N_c parameters only.

The evidence can be estimated by the following Monte Carlo simulation algorithm. Note that the optimal parameters in $\mathbf{c}$ are not required. The integrand of the evidence integral in Eq. (2) may have complex topology for nonlinear regression formula. However, it can be rewritten as

$$p(\mathcal{D}|\mathcal{C}_j) = \int_C \mathcal{I}(\mathbf{c}|\mathcal{C}_j) p(\mathbf{c}|\mathcal{C}_j)\, d\mathbf{c} \tag{26}$$

if the prior distribution is separable: $p(\mathbf{b},\mathbf{c},\sigma_\epsilon^2|\mathcal{C}_j) = p(\mathbf{b},\sigma_\epsilon^2|\mathcal{C}_j)p(\mathbf{c}|\mathcal{C}_j)$. In this integral, C is the domain for $\mathbf{c}$ and $\mathcal{I}(\mathbf{c}|\mathcal{C}_j)$ is given by

$$\mathcal{I}(\mathbf{c}|\mathcal{C}_j) = \int_0^\infty \int_B p(\mathcal{D}|\boldsymbol{\theta},\mathcal{C}_j) p(\mathbf{b},\sigma_\epsilon^2|\mathcal{C}_j)\, d\mathbf{b}\, d\sigma_\epsilon^2 \tag{27}$$

This integral can be evaluated in the same way as in Eq. (19):

$$\mathcal{I}(\mathbf{c}|\mathcal{C}_j) = p\left(\mathbf{b}^\star(\mathbf{c}), \sigma_\epsilon^{2\star}(\mathbf{c})|\mathcal{C}_j\right) \exp\left(-\frac{N}{2}\right) \frac{\sigma_\epsilon^{\star N_b+1-N}}{\sqrt{2(2\pi)^{N-N_b-1} N|\mathbf{A}|}} \tag{28}$$

where $\mathbf{b}^\star(\mathbf{c})$ and $\sigma_\epsilon^{2\star}(\mathbf{c})$ are the conditional optimal values given by Eqs. (24) and (25). The dimension of the original integral is reduced to N_c and it can be estimated efficiently by Monte Carlo simulation:

$$p(\mathcal{D}|\mathcal{C}_j) = E_{p(\mathbf{c}|\mathcal{C}_j)}\mathcal{I}(\mathbf{c}|\mathcal{C}_j) = \frac{1}{N_s}\sum_{s=1}^{N_s} \mathcal{I}(\mathbf{c}^{(s)}|\mathcal{C}_j) \tag{29}$$

with independent samples $\mathbf{c}^{(s)},\ s = 1,2,\ldots,N_s$, simulated according to the prior PDF $p(\mathbf{c}|\mathcal{C}_j)$.

4. Application to artificial neural network and damage detection

4.1. Problem description

In this section, the Bayesian model class selection approach is applied to structural health monitoring using Artificial Neural Network (ANN) with dynamic data. ANNs are employed for pattern matching in order to detect damage locations and estimate their severity. It is obvious that the selection of the class of ANN models, i.e., the number of hidden layers and the number of hidden neurons in each layer, has crucial effects on its performance. However, these numbers are usually selected by experience or rule of thumb only. In this section, the Bayesian method is applied to select the most plausible class of ANN models. The damage detection method presented here consists of two phases. The damage locations are identified in the first phase using an ANN with damage signatures as the inputs. In the second phase, the severity of the damage identified in the first phase is estimated by another ANN with modal parameter changes as the inputs. These ANNs are designed by the Bayesian model class selection method.

4.2. Pattern matching and damage signature

The idea of pattern matching is to use some prescribed quantities (e.g., a set of modal parameters) as an indicator (ID) of each possible damage case of the structure. If the IDs for all possible damage cases are computed and matched with the measured ID, the damage case corresponding to the best fitting ID is considered as the most probable damage case. A very large database is required to store all the IDs as the number of IDs is huge in practice. Specifically, one needs to consider different number of damage locations and different damage extent. The required computational time and storage space make this approach prohibitive. For instance, let's consider a five-story building with five possible damage locations and five damage levels for each possible damage location. There are 20, 160, 640, 1280, and 1024 possible damage patterns for one to five damage locations, respectively. The total number of damage patterns is 3125 (=5^5, including the undamaged case) for such a simple structure. This number increases drastically with the number of possible damage locations and levels. A possible solution is to proceed in two phases. The damage locations are identified in the first phase, while the corresponding damage extent will be estimated afterwards.

Structural damage induces changes in modal parameters. Damage signature is defined as the ratio of the change of eigenvector of some selected modes to the change of the eigenvalue of a reference mode [18]:

$$\widehat{\boldsymbol{\Psi}} = \frac{1}{\Delta\widehat{\Omega}_r^2}\Delta\widehat{\boldsymbol{\Phi}} \tag{30}$$

where Δ denotes the difference of a quantity between the undamaged and the possibly damaged structure; $\widehat{\Omega}_r$ is the measured natural frequency of a reference mode; $\widehat{\boldsymbol{\Phi}}$ is the vector comprised of the measured mode shapes for all selected modes: $\widehat{\boldsymbol{\Phi}} = \left\{\hat{\boldsymbol{\phi}}_1^T, \hat{\boldsymbol{\phi}}_2^T, \ldots, \hat{\boldsymbol{\phi}}_{N_m}^T\right\}^T$, where $\hat{\boldsymbol{\phi}}_m$ is the measured mode shape of the mth selected mode ($m = 1,2,\ldots,N_m$) and N_m is the total number of selected modes. Damage signatures depend only on the damage locations and the relative damage extent but not the absolute extent. With the finite-element models of the structure, the eigenvalues and eigenvectors of the undamaged and damaged structures can be computed. These theoretical damage signatures for different damage cases are readily obtained in a similar manner as Eq. (30):

$$\boldsymbol{\Psi}(n) = \frac{1}{\Delta\Omega_r^2(n)}\Delta\boldsymbol{\Phi}(n) \quad \text{for } n = 1,2,\ldots,N \tag{31}$$

where n is the index of damage cases; and N is the total number of possible damage cases. Then, the measured damage signatures in Eq. (30) are matched with the theoretical damage signatures in Eq. (31). By using the damage signature, the number of possible patterns for locating damage is significantly reduced. For the five-story building, the number of damage patterns is reduced to 5, 10, 10, 5 and 1 for damage cases for one to five damage locations, respectively. The total number of patterns (including the undamaged and damaged cases) is reduced from 3125 to 32.

4.3. Artificial neural networks based damage detection

In the first phase, the ANN is trained by using the theoretical damage signatures for all possible damage cases for a given damage level as inputs and the corresponding damage location index vector as targets. The damage location index vector is defined as $\mathbf{L} = \{L_1, L_2 \ldots, L_{N_L}\}$, where N_L is the number of all possible damage locations, and L_k is the damage location index for the kth possible damage location. An index of value 1 (0) implies that the corresponding location is damaged (undamaged). Due to the generalization property of ANNs, the trained ANN can be used to estimate/approximate the damage location index vector of the damaged structure by feeding the measured damage signature to the trained ANN. In general, the more input–target pairs employed in the training process, the higher the generalization ability of the trained ANN. After identifying the possible damage locations in the first phase, the damage extent of each identified damage location can be estimated using another ANN in the second phase. Changes of modal parameters and the corresponding damage extent index vector will be treated as input–target training pairs for the ANN in the second phase. This damage extent index vector is defined as $\mathbf{E} = \{E_1, E_2 \ldots, E_{N_E}\}$, where N_E is the number of damage locations identified in the first phase. In general, N_E is much smaller than the number of possible damage locations N_L. E_k is the damage extent index for the kth damage location identified in the first phase.

The value of E_k, in the range from 0 to 1, indicates the percentage reduction of stiffness at the corresponding damage location. Since it is necessary to consider only the possible damage locations detected in the first phase, the total number of training pairs for the ANN in the second phase is reduced to a reasonable level.

4.4. Selection of ANN models with suitable complexity

ANN design here is referred to the determination of the number of hidden layers and the number of neurons in each hidden layer [30,19]. Here, N_I and N_O are the number of the neurons in the input and output layers, respectively. Let $\mathcal{D}$ denote the set of input–target data for ANN training. Different classes of ANN models correspond to models with different number of hidden layers and/or different numbers of neurons in the hidden layers.

Let $\boldsymbol{\theta} \in \mathbb{R}^{N_f}$ denote the unknown parameter vector (including the weights, biases and prediction-error parameter) of an ANN. The likelihood function is given by

$$p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{C}_j) = (2\pi)^{-\frac{NN_O}{2}} \sigma_\epsilon^{-NN_O} \exp\left[-\frac{NN_O}{2\sigma_\epsilon^2} J_g(\boldsymbol{\theta}|\mathcal{D}, \mathcal{C}_j)\right] \tag{32}$$

where σ_ϵ is the standard deviation of the target error; and N is the total number of training pairs, which is taken as the total number of possible damage cases. The goodness-of-fit function $J_g(\boldsymbol{\theta}|\mathcal{D}, \mathcal{C}_j)$ in Eq. (32) is given by

$$J_g(\boldsymbol{\theta}|\mathcal{D}, \mathcal{C}_j) = \frac{1}{NN_O} \sum_{n=1}^{N} \|\widehat{O}(n) - O(n; \boldsymbol{\theta}, \mathcal{C}_j)\|^2 \tag{33}$$

where $O(n; \boldsymbol{\theta}, \mathcal{C}_j)$ is the ANN output of the nth input for a given set of ANN parameters $\boldsymbol{\theta}$ and ANN model class $\mathcal{C}_j$; $\widehat{O}(n)$ is the target of the ANN for the nth input (the output O is the damage location index vector $\mathbf{L}$ for the first phase and it is the damage extent index vector $\mathbf{E}$ for the second phase); and $\|\cdot\|$ denotes the Euclidean norm of a vector. In the case of a uniform prior, the optimal parameter vector $\boldsymbol{\theta}^\star$ is equivalently obtained by minimizing $J_g(\boldsymbol{\theta}|\mathcal{D}, \mathcal{C}_j)$. Furthermore, the optimal fitting-error variance $\sigma_\epsilon^{2\star}$ can be obtained by maximizing the likelihood function $p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{C}_j)$ and it is given by $\sigma_\epsilon^{2\star} = \min_{\boldsymbol{\theta}} J_g(\boldsymbol{\theta}|\mathcal{D}, \mathcal{C}_j) = J_g(\boldsymbol{\theta}^\star|\mathcal{D}, \mathcal{C}_j)$.

4.5. Applications to damage detection of a five-story building

A five-story building is used to demonstrate the Bayesian ANN design and damage detection method. The structure is classically damped with 2.0% damping ratio for all modes and it has uniform distribution of the floor mass and story stiffness. Its natural frequencies are: 1.43, 4.18, 6.59, 8.47 and 9.66 Hz.

4.5.1. Phase I: detection of damage locations

Damage is defined as the reduction of inter-story stiffness. The damage location index vector consists of five elements corresponding to the five stories ($N_L = 5$). The number of input–target pairs for training is $N = 32$, as explained in Section 4.2. Based on the shear building model, natural frequencies and mode shapes of the undamaged and damaged structures can be calculated. To simulate damage in a story, 50% reduction of the inter-story stiffness is introduced. Here, only the first mode is used to compute the damage signature and the ANN has five input neurons and five output neurons ($N_I = N_O = 5$). In this example, only one hidden layer ANNs are considered so the ANN design problem is reduced to the decision of the number of hidden neurons, n_1. A rule of thumb is used for comparison: $n_1 = (N_I + N_O)/2 + \alpha$, where α is 1 or 2 [15]. In this case, n_1 can be taken as 6 or 7. Here, four candidate classes of ANN models, with n_1 = 5, 6, 7 and 8, are considered. The values of the fitting-error parameter $\sigma_\epsilon^\star$ of these model classes are 0.060, 0.027, 0.023, and 0.019 but the class of ANNs with six neurons has dominant plausibility, which is virtually 1.0. The model classes with more hidden neurons have relatively smaller values of the fitting-error parameter $\sigma_\epsilon^\star$, implying better fitting of the input–target training pairs, but not necessarily higher plausibility. It shows clearly that it is not appropriate to select a model class by comparing the values of the fitting-error parameter alone.

In order to examine the capability of the trained ANN, measured damage signatures for different damage cases are obtained. Five percent root-mean-square discrete white noise is added to the simulated acceleration time histories of the structure subjected to wide band excitation. The fundamental frequency and mode shape can then be estimated from the measured frequency spectra. The trained ANN is first examined for 20% damage in a single story, which is different from the 50% assumed in the training process. The outputs (damage location index vector $\mathbf{L}$) of the ANN are summarized in Table 1. Damage Cases 1–5 represent the damage locations in the first to fifth story, respectively. A damage location index with value close to unity indicates that the corresponding story is damaged. For an undamaged location, the corresponding damage location index is close to zero. The ANN approximated damage location index vectors indicate successfully the correct damage locations.

Next, the trained ANN is tested by different damage extent in the second and third story of the building. Nine damage cases are considered. In all cases, the damage of the second story is taken to be 50% while the damage of the third story is considered from 10% to 90% with step size 10%. The ANN outputs are shown in Table 2. The Damage Case x_2/x_3 refers to $x_2\%$ and $x_3\%$ damage of the second and the third story, respectively. It is clearly seen that the ANN approximated damage location index vector can indicate both damage locations when they have similar damage extent (e.g., Damage Cases 50/40, 50/50 and 50/60). Otherwise, only the location with larger damage extent can be detected, e.g., Damage Cases 50/10 and 50/90.

4.5.2. Phase II: quantification of damage extent

The changes of modal frequencies and mode shapes are used as ANN inputs in the second phase. Again, only the first mode is employed and the measured mode shape is normalized such that the element corresponding to the top floor is unity. Since the unity element contains no information, it is not employed as an ANN input. Therefore, the total number of ANN inputs is 5 ($N_I = 5$) including four components of the mode shape and one modal frequency. Two cases are considered to illustrate the procedures of the second phase. The number of ANN outputs is equal to the number of damage locations detected in the first phase.

The first case considers 20% damage in the second story and this corresponds to the Damage Case 2 in Table 1. Based on the finite-element model, changes of modal parameters due to damage of the second floor with 10%,20%,...,70% reduction of the inter-story stiffness are calculated. The modal parameter changes are used as the ANN inputs, while the corresponding percentage reduction of stiffness is the ANN output. The seven sets of input–target pairs

Table 1
ANN outputs for single-damage cases with 20% reduction of inter-story stiffness.

Story	Damage case				
	1	2	3	4	5
1	**1.00**	−0.01	−0.01	−0.01	−0.01
2	0.07	**0.99**	0.16	0.18	0.20
3	0.11	0.06	**1.43**	0.39	0.56
4	0.03	−0.02	0.18	**1.26**	0.25
5	0.01	−0.04	−0.02	−0.01	**1.08**

Table 2
ANN outputs for the case with damage at two different stories.

Damage case	Story				
	1	2	3	4	5
50/10	−0.01	**1.00**	**0.22**	−0.01	−0.01
50/20	−0.00	**1.01**	**0.44**	−0.00	−0.00
50/30	−0.00	**1.01**	**0.68**	0.01	0.00
50/40	0.00	**1.01**	**0.91**	0.02	−0.00
50/50	0.00	**0.99**	**1.07**	0.02	−0.01
50/60	0.00	**0.93**	**1.17**	0.03	−0.02
50/70	0.00	**0.80**	**1.22**	0.04	−0.02
50/80	−0.00	**0.56**	**1.21**	0.03	−0.02
50/90	−0.01	**0.25**	**1.09**	−0.03	−0.03

are employed for training an ANN to estimate the damage extent ($N = 7$). There are five ANN inputs and one ANN output. Based on the Kermanshahi's rule of thumb, the number of hidden neurons may be 4 or 5. However, by the Bayesian approach, the most plausible model class is the one with one hidden neuron ($n_1 = 1$) and its plausibility is close to unity. This is different from the suggestion by the Kermanshahi's rule of thumb, which is based on the numbers of ANN inputs and outputs only. However, the rule neglects other important factors, such as the number of input–target training pairs and the complexity of the function to be approximated. The ANN with only one hidden neuron is used in the second phase to detect the damage extent of the second story. The measured modal parameters of both the undamaged and damaged structures can be obtained from the simulated noisy dynamic data. The changes of modal parameters due to the damage can then be calculated and fed into the trained ANN. The ANN output is 19.68%, which is close to the correct damage extent (20%).

The second case considers 50% and 30% damage of the second and third story, respectively. This corresponds to Damage Case 50/30 in Table 2. Similar to the single-damage case, damage levels from 10% to 70% with step size 10% are considered for the input–target training pairs. Therefore, 49 input–target pairs are employed in the training process ($N = 49$). It turned out that the most plausible class of models is the one with three hidden neurons ($n_1 = 3$) and the plausibility is close to unity. Therefore, it is employed to estimate the damage extent of the two damage locations. The changes in modal parameters are fed to the trained ANN and the outputs are 50.29% and 29.54% for the second and third story, respectively. Again, the damage detection methodology successfully estimates the damage locations and severity.

5. Application to seismic attenuation relationship

5.1. Problem description

Prediction of peak ground acceleration (PGA) has received great attention in the society of civil engineering and engineering seismology for decades. Significant amount of work can be found in predicting the PGA using the magnitude of earthquake, station-to-hypocenter distance, and the properties of the site foundation [13,5,4]. In particular, the Boore–Joyner–Fumal seismic attenuation formula is well-known for estimation of the PGA [4]:

$$\log_{10}\mathrm{PGA} = b_1 + b_2(M - M_0) + b_3(M - M_0)^2 + b_4 r + b_5 \log_{10} r + b_6 G_B + b_7 G_C \tag{34}$$

where M is the moment magnitude of an earthquake; M_0 is a shifting constant and $M_0 = 6$ was used in Boore et al. [4]; r is the observation station-to-hypocenter distance (in kilometer); G_B and G_C are site foundation classification variables: $G_B = 1$ for class B and 0 otherwise, and $G_C = 1$ for class C and 0 otherwise (the classification of foundation will be discussed further in Section 5.3.1).

The predictive model in Eq. (34) is empirical. One may wonder if a more reliable model can be obtained by adding/erasing terms or by considering a different functional form. A good predictive model class should have good capability on fitting the strong-motion records and at the same time be insensitive to observation and modeling error. In this section, the regression formula given in Boore et al. [4] is examined. Thirty-two model classes are constructed by including subsets of terms in the regression formula in Eq. (34). The Bayesian model class selection approach is applied to select among these 32 model classes. A database of 249 strong-motion records from the China Earthquake Data Center [6] is utilized for this study.

5.2. Selection of the predictive model class

In order to examine the suitability of the predictive model class in Eq. (34) and to propose the most suitable one, model class candidates are constructed. Here, predictive model classes are considered in a similar functional form of Eq. (34) but to include different combinations of terms in different model classes. First of all, the constant b_1 is necessary to serve as a scaling factor so all model classes contain b_1. As will be discussed later in Section 5.3.1 that only strong-motion records with moment magnitude $M > 3.5$ are utilized, $M_0 = 3.5$ is taken in this study. For the site properties, both terms ($b_6 G_B$ and $b_7 G_C$) are either included or not included at the same time. As a result, there are totally $2^5 = 32$ model class candidates, namely $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_{32}$. To summarize, all the candidate model classes include the constant b_1 and different model classes include different combinations of the following five groups: $b_2(M - M_0)$, $b_3(M - M_0)^2$, $b_4 r$, $b_5 \log_{10} r$, and $b_6 G_B + b_7 G_C$.

Let $\mathcal{D}$ denote the data that includes the PGA and the corresponding earthquake magnitude, observation station-to-hypocenter distance and site foundation properties. A predictive formula in any model class $\mathcal{C}_j$, which includes subsets of terms of Eq. (34), can be written as

$$\log_{10}\mathrm{PGA}(\mathbf{x}; \mathbf{b}, \mathcal{C}_j) = \sum_{l=1}^{N_b} b_l x_l \tag{35}$$

This falls into the category of linear regression problem in Section 3.1. The variables $x_1, x_2, \ldots, x_{N_b}$ in this case are the corresponding data, i.e., 1, $M - M_0$, $(M - M_0)^2$, r, $\log_{10} r$, G_B and G_C. Note that the first $N_j - 1$ components of the uncertain parameter vector are the uncertain formula coefficients b_l and θ_{N_j} is σ_ϵ^2. In this study, a flat prior PDF is used for not biasing the results prior to the data. The goodness-of-fit function $J_g(\mathbf{b}|\mathcal{D}, \mathcal{C}_j)$ in Eq. (15) is given by

$$J_g(\mathbf{b}|\mathcal{D}, \mathcal{C}_j) = \frac{1}{N} \sum_{n=1}^{N} [\log_{10}\mathrm{PGA}_o(n) - \log_{10}\mathrm{PGA}(\mathbf{x}(n); \mathbf{b}, \mathcal{C}_j)]^2 \tag{36}$$

where $\mathrm{PGA}(\mathbf{x}(n); \mathbf{b}, \mathcal{C}_j)$ is the model predicted PGA and $\mathrm{PGA}_o(n)$ is its corresponding observed value. Since a flat prior PDF is used, the closed-form solution of the optimal coefficient vector $\mathbf{b}^\star$ can be obtained by Eq. (16) with $y(n) = \log_{10}\mathrm{PGA}_o(n)$. Similarly, the most plausible value of the predictive/fitting-error variance is given by $\sigma_\epsilon^{2\star} = \min_{\mathbf{b}} J_g(\mathbf{b}|\mathcal{D}, \mathcal{C}_j) = J_g(\mathbf{b}^\star|\mathcal{D}, \mathcal{C}_j)$. It is attempted to use $\mathcal{D}$ to select the most suitable predictive model class among the aforementioned model class candidates $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_{32}$. The prior plausibility of the model classes is taken as uniform in this study: $P(\mathcal{C}_j) = 1/32$, $j = 1, 2, \ldots, 32$. Finally, the evidence $p(\mathcal{D}|\mathcal{C}_j)$ can be computed by Eq. (19).

5.3. Analysis with strong ground motion measurements

5.3.1. Description of the database

A database of strong-motion records is obtained from the China Earthquake Data Center [6]. In this study, the horizontal components of the PGA records are considered and only the records with $M > 3.5$ are utilized so $M_0 = 3.5$ in Eq. (34). There are totally 94 and 155 records observed from 19 and 13 stations in the Tangshan and Xinjiang region, respectively. The classification of the site foundation depends on its stiffness and the averaged shear velocity over the upper 30 m is used as its measure [4]. Since only the soil/rock types are given by the data center, the site class is defined in a slightly different way from the original definition by Boore et al. [4]. Specifically, granite, sandstone, bedrock, siltstone, and conglomerate are classified as class A. Alluvium, diluvium, and weathering conglomerate are included in class B. Soft soil, clay and subclay are classified as class C.

The range of $\log_{10}$PGA in the data lies in the interval of $[-1,3]$. The prior PDF for b_1 in Eq. (34) is taken to be a uniform distribution in this interval. For the other parameters (if they are included in a predictive model class), they are also taken as uniform distribution in order to let the data to infer the parameter values. By considering the previous work in [13,5,4], the range of the uniform distribution is taken as $[-1,1]$ for b_2, b_3, b_5, b_6 and b_7. For b_4, it was observed that it is much smaller from previous study. This is also expected to happen in this case because r is much larger than $\log_{10}r$ in the range of the data. Specifically, the range for its uniform distribution is taken to be $[-0.01, 0.01]$. Note that as long as the range is large enough, the values of the bounds do not affect the identification results for the parameter vector $\boldsymbol{\theta}$.

5.3.2. Tangshan region

Tangshan city is located in the North China Plain. Table 3 shows the results of the predictive model class selection using the horizontal records from this region. Only the top four, last four and the full model classes are shown in table. The first column shows the ranking of each predictive model class. A smaller number of the ranking corresponds to a higher plausibility of the model class. The second column shows the parameters being included in that model class, e.g., '1 2 3 5' denotes a model class with four terms: $\log_{10}\text{PGA} = b_1 + b_2(M-3.5) + b_3(M-3.5)^2 + b_5\log_{10}r$. The third and fourth columns show the maximum likelihood value and the corresponding standard deviation of the fitting error. The fifth column shows the value of the log Ockham factor, $\ln \mathcal{O}_j$, that indicates the robustness of the model class, and the last column shows the plausibility of each model class. The full model class with all the seven terms has the largest maximum likelihood value. This is intuitive because the full model class has the largest solution space so it is capable to fit the dataset at least as well as any other model classes with less free parameters. However, its robustness is not as good as the other model classes and its log Ockham factor is -21.27, which is the smallest among all the model class candidates. Table 4 shows the optimal parameters of each model class. The numbers in parenthesis denote the standard deviation of that parameter calculated using the Bayesian approach. In order to balance the data fitting capability and robustness, a relatively simple model class is selected and its optimal model is

$$\log_{10}\text{PGA} = 1.9 + 0.76(M-3.5) - 0.19(M-3.5)^2 - 0.86\log_{10}r \tag{37}$$

where PGA is in cm/s^2. The term $-0.86\log_{10}r$ indicates that the PGA decreases with an increasing site-to-hypocenter distance. The PGA decreases by 45% ($\approx (1-2^{-0.86})\times 100\%$) when r is doubled without change of other variables. This predictive model class has plausibility over 0.7 by the Bayesian model selection approach. The terms b_6G_B and b_7G_C are not included in this model. By observing the model classes that include these two terms, one can see that the optimal values of these coefficients are of similar order of its standard deviation of the estimates, implying that the data does not provide significant evidence for such terms to be included. Another noteworthy point is on the model class $\log_{10}\text{PGA} = b_1 + b_2(M-M_0)$, i.e., the 29th model class. The optimal value for b_2 is negative and it seems to imply that the larger the moment magnitude of an earthquake, the smaller the PGA. However, this is only due to the non-uniform distribution of M and r in the dataset. Specifically, the correlation coefficient between the M and r is 0.64 in this dataset. In other words, large value of M in a record often associates

Table 3
Model class selection results (Tangshan).

Ranking	Parameters	$p(\mathcal{D}\|\boldsymbol{\theta}^\star,\mathcal{C}_j)$	$\sigma_\epsilon^\star$	$\ln \mathcal{O}_j$	$P(\mathcal{C}_j\|\mathcal{D})$
1	1 2 3 5	5.54E−09	0.30	−14.56	7.04E−01
2	1 2 4 5	7.65E−10	0.30	−14.05	1.61E−01
3	1 2 3 4 5	5.72E−09	0.30	−16.43	1.13E−01
4	1 2 3 5 6 7	1.15E−08	0.29	−19.41	1.14E−02
7	1 2 3 4 5 6 7	1.15E−08	0.29	−21.27	1.78E−03
29	1 2	5.55E−23	0.42	−8.77	2.32E−12
30	1 6 7	8.42E−23	0.42	−10.33	7.36E−13
31	1 3 6 7	1.54E−21	0.40	−14.21	2.78E−13
32	1 2 6 7	1.04E−22	0.41	−12.92	6.81E−14

Table 5
Model class selection results (Xinjiang).

Ranking	Parameters	$p(\mathcal{D}\|\boldsymbol{\theta}^\star,\mathcal{C}_j)$	$\sigma_\epsilon^\star$	$\ln \mathcal{O}_j$	$P(\mathcal{C}_j\|\mathcal{D})$
1	1 2 5 6 7	4.47E−05	0.26	−17.92	2.23E−01
2	1 2 4 5 6 7	5.30E−04	0.25	−20.48	2.04E−01
3	1 3 5 6 7	1.39E−04	0.26	−19.16	2.00E−01
4	1 3 4 5 6 7	1.61E−03	0.25	−21.74	1.76E−01
8	1 2 3 4 5 6 7	2.31E−03	0.25	−23.87	3.00E−02
29	1 6 7	4.73E−15	0.30	−12.06	8.27E−09
30	1 5 6 7	9.16E−15	0.30	−13.94	2.45E−09
31	1 4 6 7	6.58E−15	0.30	−14.91	6.62E−10
32	1 4 5 6 7	9.29E−15	0.30	−16.28	2.38E−10

Table 4
Optimal parameters of each predictive model class (Tangshan).

Ranking	b_1	b_2 $(M-3.5)$	b_3 $((M-3.5)^2)$	b_4 (r)	b_5 $(\log_{10}r)$	b_6 (G_B)	b_7 (G_C)
1	1.9 (0.17)	0.76 (0.13)	−0.19 (0.043)	–	−0.86 (0.14)	–	–
2	1.8 (0.19)	0.34 (0.06)	–	−0.0025 (0.00066)	−0.66 (0.17)	–	–
3	1.9 (0.19)	0.73 (0.20)	−0.17 (0.083)	−0.00031 (0.0012)	−0.83 (0.19)	–	–
4	1.8 (0.18)	0.76 (0.13)	−0.19 (0.043)	–	−0.85 (0.13)	0.081 (0.07)	0.0033 (0.08)
7	1.8 (0.19)	0.75 (0.20)	−0.18 (0.083)	−0.00011 (0.0012)	−0.83 (0.19)	0.08 (0.07)	0.0031 (0.08)
29	1.2 (0.06)	−0.03 (0.06)	–	–	–	–	–
30	1.1 (0.06)	–	–	–	–	0.10 (0.10)	0.026 (0.11)
31	1.2 (0.06)	–	−0.046 (0.019)	–	–	0.13 (0.10)	0.014 (0.11)
32	1.2 (0.08)	−0.039 (0.06)	–	–	–	0.11 (0.10)	0.024 (0.11)

Table 6
Optimal parameters of each predictive model class (Xinjiang)

Ranking	b_1	b_2 $(M-3.5)$	b_3 $((M-3.5)^2)$	b_4 (r)	b_5 $(\log_{10} r)$	b_6 (G_B)	b_7 (G_C)
1	1.9 (0.17)	0.22 (0.03)	–	–	−0.4 (0.11)	0.24 (0.08)	0.06 (0.10)
2	2.4 (0.25)	0.24 (0.03)	–	0.0015 (0.00068)	−0.74 (0.19)	0.28 (0.08)	0.11 (0.10)
3	2.1 (0.17)	–	0.067 (0.009)	–	−0.4 (0.11)	0.24 (0.08)	0.043 (0.10)
4	2.5 (0.25)	–	0.073 (0.009)	0.0015 (0.00067)	−0.75 (0.19)	0.27 (0.08)	0.092 (0.10)
8	2.5 (0.25)	0.082 (0.10)	0.05 (0.029)	0.0015 (0.00067)	-0.76 (0.19)	0.27 (0.08)	0.097 (0.10)
29	1.7 (0.09)	–	–	–	–	0.093 (0.09)	0.12 (0.10)
30	1.9 (0.20)	–	–	–	−0.14 (0.12)	0.10 (0.09)	0.16 (0.11)
31	1.7 (0.09)	–	–	−0.00038 (0.00046)	–	0.093 (0.09)	0.13 (0.10)
32	2.0 (0.29)	–	–	0.00013 (0.00076)	−0.17 (0.21)	0.10 (0.09)	0.16 (0.11)

with large value of r, that reduces the PGA. Furthermore, the correlation coefficient between the $\log_{10}$PGA and M is −0.0511 and hence the optimal coefficient b_2 is negative. Therefore, a model class with too few free parameters may cause under-fitting to the data.

5.3.3. Xinjiang region

Xinjiang is in the North West of China. In the same fashion as in Tables 3–6 show the predictive model class selection results and the optimal parameters of each model class. From Table 6, the optimal predictive model for the PGA is given by

$$\log_{10}\text{PGA} = 1.9 + 0.22(M-3.5) - 0.4\log_{10} r + 0.24G_B + 0.06G_C \qquad (38)$$

Note that the first four model classes possess similar plausibility, implying that the Bayesian model selection method does not have strong preference on the optimal model class. This is in contrast to the previous case in the Tangshan region, in which the plausibility of the optimal model class is over 0.7. In this case with the data of Xinjiang, a plausibility-weighted predictive formula can be obtained:

$$\log_{10}\text{PGA} = 2.21 + 0.122(M-3.5) + 0.0327(M-3.5)^2 + 0.00071r - 0.56\log_{10} r + 0.26G_B + 0.076G_C \qquad (39)$$

and the associated uncertainty of the parameters $b_1, b_2, \ldots, b_7$ are: 0.21, 0.02, 0.006, 0.00047, 0.15, 0.08 and 0.097, respectively. Of course, one may consider to include five or even more model classes but the results will be virtually the same since the plausibilities of the fifth and below model classes are small. One important point is that this multi-mode predictive formula includes all seven terms but it is not the 'optimal model' in the full model class (8th model class in Table 6) as their parameter values are not the same. This multi-model predictive formula (along with the posterior uncertainty of the parameters) does not have the most powerful data fitting capability but it possesses higher level of robustness than the 'optimal model' in the full model class.

6. Concluding remarks

Bayesian model class selection has attracted substantial interest in recent years for selecting the most plausible/suitable class of models among some specified model classes, based on system measurements. This paper introduced some recent developments of Bayesian model class selection and applications in civil engineering. Asymptotic expansion and Monte Carlo method can be used to compute the evidence of a model class. Special cases of linear and nonlinear regression formula are considered and Bayesian model class selection is particularly useful in this case.

The method was applied to the design of ANN structure. The lack of a practical and rigorous ANN design method is one of the major obstacles in the applications of ANN in structural health monitoring in practice. Although some guidelines were proposed in the literature, they depend highly on user's judgment. The Bayesian ANN design method can be used to quantify the optimality of different classes of ANN models based on the training data. It is also applied to the selection of seismic attenuation empirical formula. A database with 249 strong-motion records in Tangshan and Xinjiang of China is utilized for the analysis. It turns out that the most plausible model class is not the full relationship even though it gives the smallest fitting error. If several predictive model classes possess similar plausibility given the dataset, one can consider a multi-model predictive formula as in the case of the Xinjiang region. The Bayesian approach allows to obtain not only the optimal parameters in a model class but also the associated uncertainty of the parameter estimates. The quantified uncertainty can be further used for the reliability of the prediction. Through the applications in this paper, it is demonstrated that Bayesian identification and model class selection have high potential for applications in different fields of civil engineering since there are many types of uncertainty necessary to be considered. Other applications in the modeling and prediction of air quality and soil compressibility index empirical relationship have been studied but they are not introduced here due to page limit.

Acknowledgements

The author would like to express his deepest gratitude to Prof. Beck for his advice during the PhD study. This paper is dedicated to him for his achievements and inspiration in the direction of Bayesian analysis. This work was supported by the research committee of University of Macau under research grants RG077/07-08S/YKV/FST and RG077/07-08S/09R/YKV/FST. The generous support is gratefully acknowledged.

References

[1] Beck JL, Au SK. Bayesian updating of structural models and reliability using Markov chain Monte Carlo simulation. J Eng Mech (ASCE) 2002;128(4):380–91.
[2] Beck JL, Katafygiotis LS. Updating models and their uncertainties. I: Bayesian statistical framework. J Eng Mech (ASCE) 1998;124(4):455–61.
[3] Beck JL, Yuen K-V. Model selection using response measurements: Bayesian probabilistic approach. J Eng Mech (ASCE) 2004;130(2):192–203.
[4] Boore DM, Joyner WB, Fumal TE. Estimation of response spectra and peak accelerations from Western North American earthquakes: an interim report. Technical report open-file report 93-509, US geological survey; 1993.
[5] Brillinger DR, Preisler HK. Further analysis of the Joyner–Boore attenuation data. Bull Seismolog Soc Am 1985;75(2):611–4.
[6] China Earthquake Data Center. <http://smsd-iem.net/eqkview.asp>.
[7] Ching J, Chen YC. Transitional Markov chain Monte Carlo method for Bayesian model updating, model class selection and model averaging. J Eng Mech (ASCE) 2007;133(7):816–32.
[8] Ching J, Muto M, Beck JL. Structural model updating and health monitoring with incomplete modal data using Gibbs sampler. Comput Aided Civil Infrastruct Eng 2006;21(4):242–57.
[9] Cox RT. The algebra of probable inference. Baltimore: Johns Hopkins Press; 1961.
[10] Gull SF. Bayesian inductive inference and maximum entropy. In: Skilling J, editor. Maximum entropy and Bayesian methods. Boston: Kluwer Academic Publisher; 1988. p. 53–74.
[11] Hoi K-I, Yuen K-V, Mok K-M. Prediction of daily averaged PM10 concentrations by statistical time-varying model. Atmos Environ 2009;43(16):2579–81.
[12] Jeffreys H. Theory of probability. 3rd ed. Oxford Clarendon Press; 1961.

[13] Joyner WB, Boore DM. Peak horizontal acceleration and velocity from strong-motion records including records from the 1979 Imperial Valley, California, earthquake. Bull Seismolog Soc Am 1981;71(6):2011–38.
[14] Katafygiotis LS, Yuen K-V. Bayesian spectral density approach for modal updating using ambient data. Earthquake Eng Struct Dyn 2001;30(8):1103–23.
[15] Kermanshahi B. Design and application of neural networks. Shokodo (Tokyo); 1999 [chapter 3].
[16] Kullback S. Information theory and statistics. Mineola (NY): Dover, Publications Inc.; 1968.
[17] Kullback S, Leibler RA. On information and sufficiency. Ann Math Stat 1951;22(1):79–86.
[18] Lam HF, Ko JM, Wong CW. Localization of damaged structural connections based on experimental modal and sensitivity analysis. J Sound Vib 1998;210(1):91–115.
[19] Lam HF, Yuen K-V, Beck JL. Structural health monitoring via measured Ritz vectors utilizing artificial neural networks. Comput Aided Civil Infrastruct Eng 2006;21(4):232–41.
[20] Mackay DJC. Bayesian interpolation. Neural Comput 1992;4(3):415–47.
[21] Sivia DS. Data analysis: a Bayesian tutorial. Oxford Science Publications; 1996.
[22] Yan WM, Yuen KV, Yoon GL. Bayesian probabilistic approach for the correlations of compressibility index for marine clays. J Geotechn Geoenviron Eng (ASCE) 2009;135(12):1932–40.
[23] Yuen K.-V. PhD thesis: model selection, identification and robust control for dynamical systems. Technical report EERL 2002-03, California Institute of Technology, Pasadena (USA); 2002.
[24] Yuen K-V. Bayesian methods for structural dynamics and civil engineering. John Wiley & Sons Ltd.; 2010.
[25] Yuen K-V, Beck JL. Updating properties of nonlinear dynamical systems with uncertain input. J Eng Mech (ASCE) 2003;129(1):9–20.
[26] Yuen K-V, Beck JL, Katafygiotis LS. Probabilistic approach for modal identification using non-stationary noisy response measurements only. Earthquake Eng Struct Dyn 2002;31(4):1007–23.
[27] Yuen K-V, Katafygiotis L. Substructure identification and health monitoring using response measurement only. Comput Aided Civil Infrastruct Eng 2006;21(4):280–91.
[28] Yuen K-V, Katafygiotis LS. Bayesian time-domain approach for modal updating using ambient data. Probabilist Eng Mech 2001;16(3):219–31.
[29] Yuen K-V, Katafygiotis LS. Model updating using response measurements without knowledge of the input spectrum. Earthquake Eng Struct Dyn 2005;34(2):167–87.
[30] Yuen K-V, Lam HF. On the complexity of artificial neural networks for smart structures monitoring. Eng Struct 2006;28(7):977–84.