Fundamental Two-Stage Formulation for Bayesian System Identification, Part I: General Theory

Siu-Kui Au[1] and Feng-Liang Zhang[2*]

[1] Center for Engineering Dynamics and Institute for Risk and Uncertainty, University of Liverpool, UK

[2] Research Institute of Structural Engineering and Disaster Reduction, Tongji University, China

Abstract

Structural system identification is concerned with the determination of structural model parameters (e.g., stiffness, mass) based on measured response data collected from the subject structure. For linear structures, one popular strategy is to adopt a 'two-stage' approach. That is, modal identification (e.g., frequency, mode shape) is performed in Stage I, whose information is used for inferring the structural parameters in Stage II. Different variants of Bayesian two-stage formulations have been proposed in the past. A prediction error model is commonly introduced to build a link between Stage I and Stage II, treating the most probable values of the natural frequencies and mode shapes identified in Stage I as 'data' for Stage II. This type of formulation, which casts a prediction error model through descriptive statistics, involves heuristics that distort the fundamental nature of the Bayesian approach, although it has appeared to be inevitable. In this paper, a fundamental theory is developed for the Bayesian two-stage problem. The posterior distribution of structural parameters is derived rigorously in terms of the information available in the problem, namely the prior distribution of structural parameters, the posterior distribution of modal parameters in Stage I and the distribution of modal parameters conditional on the structural parameters that connects Stages I and II. The theory reveals a fundamental principle that ensures no double-counting of prior information in the two-stage identification process. Mathematical statements are also derived that provide insights into the role of the structural modeling error. Beyond the original structural model identification problem that motivated the work, the developed theory can be applied in more general settings. In the companion paper, examples with synthetic and real experimental data are provided to illustrate the proposed theory.

Keywords: Bayesian; double counting principle; modal identification; prediction error; system identification; two-stage approach

[*] Corresponding author. E-mail: fengliangzhang@hotmail.com, fengliangzhang@tongji.edu.cn; Office phone: +86 21 6598 7352; Office fax: +86 21 6598 2668

1. Introduction

System identification involves making inference about the parameters of a mathematical model based on observed measurements of the real system. Driven by the increasing demand for understanding and using mathematical models of nature and engineered systems consistent with observations, it has become one of the most important problems in modern science and engineering. The Bayesian approach provides a fundamental means for system identification, resolving uncertainties due to the lack of information in the context of probability logic [1][2][3]. The parameters are viewed as uncertain variables and the identification results are cast in terms of their probability distribution after incorporating information from the observed data.

Let $\boldsymbol{\theta}$ be a set of model parameters to be identified from available data D. Bayesian system identification aims at determining the 'posterior distribution' (i.e., given data), $p(\boldsymbol{\theta} \mid D)$. Using Bayes' Theorem,

$$p(\boldsymbol{\theta} \mid D) = p(D)^{-1} p(\boldsymbol{\theta}) p(D \mid \boldsymbol{\theta}) \tag{1}$$

where $p(D)^{-1}$ is a normalizing constant; $p(\boldsymbol{\theta})$ is the 'prior distribution' of $\boldsymbol{\theta}$ (i.e., in the absence of data); and $p(D \mid \boldsymbol{\theta})$ is the 'likelihood function' that gives the distribution of D for a given $\boldsymbol{\theta}$. If the relationship between $\boldsymbol{\theta}$ and D is complicated, the identification problem can be very challenging. For example, it can be difficult in the first place to formulate the likelihood function $p(D \mid \boldsymbol{\theta})$ in an explicit form conducive to analytics or computations. The problem may not be 'globally identifiable', i.e., there is more than one or even an infinite number of most probable values, reflecting the fact that the available data is not sufficient for delineating their plausibility. In this case it is also difficult to extract information (e.g., descriptive statistics) about the posterior distribution of $\boldsymbol{\theta}$ [4].

In view of the difficulty of identifying $\boldsymbol{\theta}$ directly from the data D, a 'two-stage' approach has been suggested to convert the original problem into two sub-problems which are more intuitive. This has been motivated by structural system identification problems, where the target is to identify the structural parameters (e.g., stiffness, mass) from vibration data (e.g., acceleration) measured from the subject structure [5][6][7]. In Stage I the modal properties, i.e., natural frequencies, damping ratios, mode shapes, etc., are first identified. Their identification result is

then used for identifying the structural parameters in Stage II. Although intuitive, formulating the two-stage problem in strict accordance with Bayes' rule is mathematically non-trivial. One needs to express the posterior distribution of $\boldsymbol{\theta}$ in Stage II in terms of the posterior distribution of the modal parameters in Stage I, in an explicit manner consistent with the relationship between the modal properties and the data (Stage I), the relationship between the modal parameters and the structural parameters and the prior information on the structural parameters.

Two-stage formulations with different variants have been proposed, e.g., [8][9][10][11][12]. Applications can be found in, e.g., [13][14][15]. The existing formations, however, involve heuristics in the formulation of the likelihood function $p(D|\boldsymbol{\theta})$ in Stage II. In one popular formulation, in order to link Stage I and II, the most probable value (MPV, a descriptive statistic) of the modal parameters in Stage I is taken as 'data' and modeled to consist of the structural model prediction (which depends on the structural parameters) and a prediction error. The statistical properties of the latter is determined from either ensemble statistics of identification results in Stage I in early developments, or from the posterior statistics in more recent developments [16]. Casting a prediction error model on the most probable modal parameters, which is merely a descriptive statistic characterizing the posterior distribution, has philosophical issues and distorts the fundamental nature of a Bayesian approach. Due to the non-trivial nature of the two-stage problem, however, the heuristic treatment has so far appeared to be unavoidable.

In this work, we develop a general fundamental theory for the Bayesian two-stage problem and apply it to structural system identification based on ambient vibration data. It is presented in two companion papers. In this paper, we derive the equation that fundamentally expresses the posterior distribution of $\boldsymbol{\theta}$ in terms of the posterior distribution of the parameters identified in Stage I. Theoretical issues associated with the formulation are investigated in detail. In the companion paper [17], the general theory is applied to the case of structural model identification using ambient vibration data. Illustrative examples with synthetic and experimental data are presented to verify the method and investigate its applications.

2. Problem context

For clarity we first present the context of the two-stage identification problem. Recall from the introduction that the target is to determine the posterior distribution of $\boldsymbol{\theta}$ from the measured data D. For discussion purposes we refer $\boldsymbol{\theta}$ as the 'structural parameters'. This terminology is motivated from the structural system identification problem. Suppose there is a 'data prediction model' whose set of parameters $\boldsymbol{\alpha}$ can be readily identified from D, in the sense that the likelihood function $p(D|\boldsymbol{\alpha})$ is available in explicit form and conducive to computations. Although not mathematically required, $\boldsymbol{\alpha}$ is often globally identifiable from the data and its posterior statistics (e.g., most probable value and covariance matrix) can be determined efficiently. Both $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ are related to the data D but intuitively the relationship between $\boldsymbol{\alpha}$ and D is more direct and characteristic. Without loss of generality, divide the parameters in $\boldsymbol{\alpha}$ into two groups,

$$\boldsymbol{\alpha} = [\varpi, \upsilon] \tag{2}$$

where ϖ is related to $\boldsymbol{\theta}$ but υ is not (and is possibly null). The relationship between ϖ and $\boldsymbol{\theta}$ is described through a 'structural prediction model' that gives a prediction of ϖ for a given $\boldsymbol{\theta}$ at least probabilistically, in terms of the conditional distribution $p(\varpi|\boldsymbol{\theta})$.

The idea of a two-stage approach is to first identify ϖ from D (Stage I) and then use the identification result to identify $\boldsymbol{\theta}$ (Stage II). The potential advantage is that the result in Stage I is often demanded and hence calculated anyway; and is relatively easy to obtain and check intuitively, providing a quick consolidation of the information in the data that is useful for making inference about $\boldsymbol{\theta}$. Of course, Stage II can still be difficult but it is likely to be easier than the direct approach and its result is easier to check or understand because the relationship between $\boldsymbol{\theta}$ and ϖ (the structural model) can be more tractable than that between $\boldsymbol{\theta}$ and D.

Figure 1 gives a schematic diagram of the two-stage Bayesian identification problem described above. Table 1 summarizes the nomenclature. In the particular context of structural system identification that motivated the approach, $\boldsymbol{\theta}$ is the set of structural parameters (e.g., stiffness, mass) to be identified from the measured vibration data D (e.g., acceleration histories); and $\boldsymbol{\alpha}$ is the set of all modal parameters (e.g., natural frequencies, damping ratios, mode shapes, etc) that

are necessary to give a complete probabilistic description of the data D in terms of the likelihood function $p(D|\boldsymbol{\alpha})$. The relationship between $\boldsymbol{\theta}$ and D depends on the structural model. The structural parameters $\boldsymbol{\theta}$ are not always globally identifiable from the data, e.g., when there are too many structural parameters or too few data. On the other hand, for linear elastic structures the modal parameters $\boldsymbol{\alpha}$ of well-excited modes can be readily identified from vibration data. They are often under frequent demand in practice as they govern resonance characteristics and directly affect the vibration response of structures. Within $\boldsymbol{\alpha}=[\varpi,\upsilon]$, the set ϖ consists of the natural frequencies and mode shapes, which can be predicted by an available finite element model (the structural model) at least probabilistically when $\boldsymbol{\theta}$ is given; υ consists of the damping ratios and other parameters that according to the structural model are not related to $\boldsymbol{\theta}$.

The central task of establishing the theory for the Bayesian two-stage problem is to express the posterior distribution of $\boldsymbol{\theta}$ in Stage II in terms of the posterior distribution of ϖ in Stage I. The main result of this paper is encapsulated in the form of a theorem, presented below for the ease of reading. It shall be proven in Section 4 and its associated issues shall be investigated.

Theorem 1 (Two-stage formulation)

Let $p(D|\varpi,\upsilon)$ be the likelihood function formulated according to the data prediction model in Stage I; $p(\varpi|\boldsymbol{\theta})$ be the conditional distribution of ϖ formulated according to the structural prediction model in Stage II; $p(\varpi,\upsilon|\boldsymbol{\theta})$ be the joint (prior) conditional distribution of (ϖ,υ) for given $\boldsymbol{\theta}$ and $p(\upsilon)$ be the prior distribution of υ. Assume that

$$p(D|\varpi,\upsilon,\boldsymbol{\theta}) = p(D|\varpi,\upsilon) \tag{3}$$

$$p(\varpi,\upsilon|\boldsymbol{\theta}) = p(\varpi|\boldsymbol{\theta})p(\upsilon) \tag{4}$$

Then the posterior distribution of $\boldsymbol{\theta}$ given the data D and prior information is given by

$$p(\boldsymbol{\theta}|D) \propto p(\boldsymbol{\theta})\int p_0(\varpi|D)p(\varpi|\boldsymbol{\theta})d\varpi \tag{5}$$

where

$$p_0(\varpi|D) = \int p_0(\varpi,\upsilon|D)d\upsilon \tag{6}$$

is the marginal posterior distribution of ϖ using (hypothetically) a uniform (i.e., constant) prior distribution for ϖ in Stage I; and

$$p_0(\varpi, \upsilon \mid D) \propto p(D \mid \varpi, \upsilon) p(\upsilon) \tag{7}$$

is the corresponding (hypothetical) joint posterior distribution of (ϖ, υ).

■

Equation (3) says that the information of the structural parameters θ is redundant for the determining the plausibility of the data D when the parameters in Stage I, $\alpha = [\varpi, \upsilon]$, are given. Equation (4) distinguishes two types of parameters in the structural prediction model, allowing for parameters ϖ that depend on θ and parameters υ that do not.

3. Basic formulation

In this section we derive the basic equation that expresses the posterior distribution of θ in terms of the identification result in Stage I. The theory to be developed essentially translates the interfacing information in Figure 1 into Bayesian mathematical forms.

Assume that the probability distribution of the data D can be fully determined for given (ϖ, υ). This gives the 'likelihood function' $p(D \mid \varpi, \upsilon)$, which must be formulated based on the data prediction model. The set ϖ contains the parameters that are related to θ; and υ contains the remaining unrelated ones. Specifically, it is assumed that

$$p(\varpi, \upsilon \mid \theta) = p(\varpi \mid \theta) p(\upsilon) \tag{8}$$

This is equivalent to the following two basic assumptions:

1) ϖ and υ are conditionally independent for a given θ;

2) θ is informative about ϖ but not υ

Equation (8) implies that ϖ and υ are unconditionally independent because, using the theorem of total probability,

$$
\begin{aligned}
p(\varpi, \upsilon) &= \int p(\varpi, \upsilon \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \int p(\varpi \mid \boldsymbol{\theta}) p(\upsilon) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \int p(\varpi \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \times p(\upsilon) \\
&= p(\varpi) p(\upsilon)
\end{aligned}
\tag{9}
$$

since

$$
p(\varpi) = \int p(\varpi \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{10}
$$

The conditional distribution $p(\varpi \mid \boldsymbol{\theta})$ in (8) must be derived based on the structural prediction model that gives at least a probabilistic prediction of ϖ for a given $\boldsymbol{\theta}$. As a remark, (10) shows that the prior distribution of ϖ is completely determined by $p(\boldsymbol{\theta})$ and $p(\varpi \mid \boldsymbol{\theta})$, and so unlike $p(\boldsymbol{\theta})$ it is not open to choice by the analyst. This is an important issue in the two-stage problem that shall be revisited later.

3.1. Basic form

We now derive an expression for $p(\boldsymbol{\theta} \mid D)$ in terms of $p(\varpi \mid D)$ in Stage I and other probabilistic information available in the problem. We start with the relationship between joint and marginal distribution:

$$
p(\boldsymbol{\theta} \mid D) = \int \int p(\varpi, \upsilon, \boldsymbol{\theta} \mid D) d\varpi d\upsilon \tag{11}
$$

Using Bayes' Theorem,

$$
p(\varpi, \upsilon, \boldsymbol{\theta} \mid D) = p(D)^{-1} p(D \mid \varpi, \upsilon, \boldsymbol{\theta}) p(\varpi, \upsilon, \boldsymbol{\theta}) \tag{12}
$$

Substituting (12) into (11),

$$
p(\boldsymbol{\theta} \mid D) = p(D)^{-1} \int \int p(D \mid \varpi, \upsilon, \boldsymbol{\theta}) p(\varpi, \upsilon, \boldsymbol{\theta}) d\varpi d\upsilon \tag{13}
$$

For a given (ϖ, υ) the probability distribution of D is completely defined and so the additional information from $\boldsymbol{\theta}$ is redundant, i.e., $p(D \mid \varpi, \upsilon, \boldsymbol{\theta}) = p(D \mid \varpi, \upsilon)$ from (3). A further application of Bayes' Theorem on $p(D \mid \varpi, \upsilon)$ on the RHS gives

$$
p(D \mid \varpi, \upsilon, \boldsymbol{\theta}) = \frac{p(\varpi, \upsilon \mid D)}{p(\varpi, \upsilon)} p(D) \tag{14}
$$

Substituting (14) into (13),

$$p(\boldsymbol{\theta} \mid D) = \int \int p(\varpi, \upsilon \mid D) \frac{p(\varpi, \upsilon, \boldsymbol{\theta})}{p(\varpi, \upsilon)} d\varpi d\upsilon \tag{15}$$

This equation expresses the posterior distribution of $\boldsymbol{\theta}$ in terms of the posterior distribution $p(\varpi, \upsilon \mid D)$ in Stage I and other associated prior distributions, $p(\varpi, \upsilon, \boldsymbol{\theta})$ and $p(\varpi, \upsilon)$. It is generally applicable regardless of the nature of ϖ and υ as we have not made use of the assumption in (8).

Computationally, (15) is not useful because $p(\varpi, \upsilon)$ and $p(\varpi, \upsilon, \boldsymbol{\theta})$ are non-trivial functions which must be derived based on the prior information about $\boldsymbol{\theta}$ and the relationship between $\boldsymbol{\theta}$ and (ϖ, υ). We next simplify the expression by taking into consideration the different nature of ϖ and υ in the problem. Using (8) and the implied form (9), the ratio of prior distributions in (15) becomes

$$\begin{aligned} \frac{p(\varpi, \upsilon, \boldsymbol{\theta})}{p(\varpi, \upsilon)} &= \frac{p(\varpi, \upsilon \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\varpi, \upsilon)} \\ &= \frac{p(\varpi \mid \boldsymbol{\theta}) p(\upsilon) p(\boldsymbol{\theta})}{p(\varpi) p(\upsilon)} \\ &= \frac{p(\varpi \mid \boldsymbol{\theta})}{p(\varpi)} p(\boldsymbol{\theta}) \end{aligned} \tag{16}$$

which does not depend on υ. Substituting (16) into (15),

$$\begin{aligned} p(\boldsymbol{\theta} \mid D) &= \int \int p(\varpi, \upsilon \mid D) \frac{p(\varpi \mid \boldsymbol{\theta})}{p(\varpi)} p(\boldsymbol{\theta}) d\varpi d\upsilon \\ &= p(\boldsymbol{\theta}) \int \int p(\varpi, \upsilon \mid D) d\upsilon \frac{p(\varpi \mid \boldsymbol{\theta})}{p(\varpi)} d\varpi \end{aligned} \tag{17}$$

Note that the inner integral is simply the posterior marginal distribution of ϖ in Stage I, i.e.,

$$\int p(\varpi, \upsilon \mid D) d\upsilon = p(\varpi \mid D) \tag{18}$$

We thus obtain the basic result:

$$p(\boldsymbol{\theta} \mid D) = p(\boldsymbol{\theta}) \int \frac{p(\varpi \mid D)}{p(\varpi)} p(\varpi \mid \boldsymbol{\theta}) d\varpi \tag{19}$$

As a simple check, if even ϖ were not related to $\boldsymbol{\theta}$ then $p(\varpi \mid \boldsymbol{\theta}) \equiv p(\varpi)$ and (19) gives

$$p(\boldsymbol{\theta} \mid D) = p(\boldsymbol{\theta}) \qquad (\varpi \text{ not related to } \boldsymbol{\theta}) \tag{20}$$

That is, there is no updating effect, which is a consistent result.

Equation (19) is a simpler form than (15) as the effect of $\boldsymbol{\upsilon}$ has been eliminated. However, it involves the prior distribution $p(\varpi)$, which is non-trivial and must be determined according to (10). In the next section we derive a better form that can further eliminate this inconvenience. Not only does it provide a convenient formulation for computation, it also reveals the processing of information in the two-stage problem in a fundamental manner.

4. Standard formulation

Equation (19) expresses the posterior distribution of $\boldsymbol{\theta}$ given D in terms of the posterior distribution of ϖ (Stage I), the relationship between $\boldsymbol{\theta}$ and ϖ in terms of $p(\varpi|\boldsymbol{\theta})$, and the prior distributions $p(\boldsymbol{\theta})$ and $p(\varpi)$. It cannot be directly used for computations because $p(\varpi)$ has to be determined from (10). In this section we investigate this issue of prior information, which leads to an important principle in the two-stage approach. It also leads to an alternative but equivalent form of (19) that is conducive to computations and is recommended as the standard form.

4.1. Consistent prior information

In (19) the role of $p(\boldsymbol{\theta})$ is intuitive and appears in a conventional manner. The non-trivial part lies in the terms $p(\varpi|\boldsymbol{\theta})$ and $p(\varpi)$, which are respectively the prior distribution of ϖ with and without the knowledge of $\boldsymbol{\theta}$. The conditional distribution $p(\varpi|\boldsymbol{\theta})$ reflects the uncertainty about ϖ when $\boldsymbol{\theta}$ is known. It must be formulated based on the structural prediction model. On the other hand, $p(\varpi)$ reflects the information about ϖ when $\boldsymbol{\theta}$ is not known. It is not arbitrary, however, because ϖ and $\boldsymbol{\theta}$ are related and there is prior information about $\boldsymbol{\theta}$. The information contained in $p(\varpi)$ is induced by $p(\boldsymbol{\theta})$ and $p(\varpi|\boldsymbol{\theta})$, through the theorem of total probability in (10) recalled here:

$$p(\varpi) = \int p(\varpi|\boldsymbol{\theta})\,p(\boldsymbol{\theta})\,d\boldsymbol{\theta} \tag{21}$$

Note that a uniform prior distribution for $\boldsymbol{\theta}$, i.e., $p(\boldsymbol{\theta}) = \text{constant}$, does not necessarily induce a uniform prior distribution for ϖ.

4.2. Double-counting principle

The posterior distribution $p(\varpi \mid D)$ in Stage I has incorporated the prior information induced on ϖ. The ratio $p(\varpi \mid D)/p(\varpi)$ in (19) effectively removes this prior information. The derivation shows that as far as the posterior distribution of $\boldsymbol{\theta}$ is concerned this is the correct way to process the prior information on ϖ. Intuitively, the prior information of $\boldsymbol{\theta}$ has been directly used in (19) through the term $p(\boldsymbol{\theta})$ and so it should not be 'double-counted'. Whatever information induced by $p(\boldsymbol{\theta})$ and used in Stage I must be removed in Stage II. We refer this as the 'double-counting principle'. Intuitive as it sounds, the mathematical form is by no means trivial.

4.3. Standard form

Equation (19) is not conducive to computations because the analytical expression for the induced prior distribution $p(\varpi)$ may not be available and it must also be used in obtaining the posterior distribution $p(\varpi \mid D)$. Philosophically, this complication may be unnecessary because the prior information induced on ϖ should not play any role in the identification of $\boldsymbol{\theta}$, by virtue of the double-counting principle as just discussed. We next derive an equivalent form of (19) that removes its pseudo-dependence on the prior information induced on ϖ, thereby allowing more straightforward insights and computations.

The idea stems from the double-counting principle. For the purpose of identifying $\boldsymbol{\theta}$ in Stage II, if we *artificially* ignore the induced prior information on ϖ in Stage I, then there is no need to remove it in Stage II. Of course under this hypothetical situation the posterior distribution of ϖ in Stage I will not reflect correctly its updated information but as far as the posterior distribution of $\boldsymbol{\theta}$ in Stage II is concerned it can still lead to the correct answer. In the following we provide a mathematical proof for this argument.

The key result is that the factor $p(\varpi \mid D)/\, p(\varpi)$ in (19) can be expressed as the *hypothetical* posterior distribution of ϖ in Stage I where its prior distribution is artificially taken to be uniform, i.e., constant. From Bayes' Theorem,

$$\frac{p(\varpi \mid D)}{p(\varpi)} = \frac{p(D \mid \varpi)}{p(D)} \tag{22}$$

Using the theorem of total probability,

$$\begin{aligned} p(D \mid \varpi) &= \int p(D \mid \varpi, \boldsymbol{\upsilon})\, p(\boldsymbol{\upsilon} \mid \varpi)\, d\boldsymbol{\upsilon} \\ &= \int p(D \mid \varpi, \boldsymbol{\upsilon})\, p(\boldsymbol{\upsilon})\, d\boldsymbol{\upsilon} \end{aligned} \tag{23}$$

since from (9) ϖ and $\boldsymbol{\upsilon}$ are independent based on prior information. Combining (22) and (23),

$$\frac{p(\varpi \mid D)}{p(\varpi)} = \int p(D)^{-1}\, p(D \mid \varpi, \boldsymbol{\upsilon})\, p(\boldsymbol{\upsilon})\, d\boldsymbol{\upsilon} \tag{24}$$

Consider now the *hypothetical* case where a uniform prior distribution is used for ϖ in Stage I, i.e.,

$$p(\varpi) \equiv c_{\varpi} = \text{constant} \qquad \text{(hypothetical case)} \tag{25}$$

For this case we specifically denote the resulting posterior distribution of $[\varpi, \boldsymbol{\upsilon}]$ by $p_0(\varpi, \boldsymbol{\upsilon} \mid D)$, where the subscript '0' reminds that it is not the actual posterior distribution but is only an instrumental one under the hypothetical case. Using Bayes' Theorem, this hypothetical posterior distribution is given by

$$\begin{aligned} p_0(\varpi, \boldsymbol{\upsilon} \mid D) &= p(D)^{-1}\, p(D \mid \varpi, \boldsymbol{\upsilon})\, p(\varpi, \boldsymbol{\upsilon}) \\ &= p(D)^{-1}\, p(D \mid \varpi, \boldsymbol{\upsilon})\, p(\varpi)\, p(\boldsymbol{\upsilon}) \\ &= c_{\varpi}\, p(D)^{-1}\, p(D \mid \varpi, \boldsymbol{\upsilon})\, p(\boldsymbol{\upsilon}) \end{aligned} \tag{26}$$

where the second equality follows from (9) and the third equality from (25). Except for the constant c_{ϖ}, the RHS of this equation coincides with the integrand of (24). Substituting into (24) gives

$$\begin{aligned} \frac{p(\varpi \mid D)}{p(\varpi)} &= c_{\varpi}^{-1} \int p_0(\varpi, \boldsymbol{\upsilon} \mid D)\, d\boldsymbol{\upsilon} \\ &= c_{\varpi}^{-1}\, p_0(\varpi \mid D) \end{aligned} \tag{27}$$

where

$$p_0(\varpi \mid D) = \int p_0(\varpi, \boldsymbol{\upsilon} \mid D) d\boldsymbol{\upsilon} \tag{28}$$

is the hypothetical marginal posterior distribution of ϖ in Stage I. Substituting (27) into (19),

$$\begin{aligned} p(\boldsymbol{\theta} \mid D) &= c_{\varpi}^{-1} p(\boldsymbol{\theta}) \int p_0(\varpi \mid D) p(\varpi \mid \boldsymbol{\theta}) d\varpi \\ &\propto p(\boldsymbol{\theta}) \int p_0(\varpi \mid D) p(\varpi \mid \boldsymbol{\theta}) d\varpi \end{aligned} \tag{29}$$

This equation expresses the posterior distribution of $\boldsymbol{\theta}$ in terms of the prior distribution of $\boldsymbol{\theta}$, the hypothetical posterior distribution of ϖ and the conditional distribution of ϖ given $\boldsymbol{\theta}$. The integral over the product $p_0(\varpi \mid D) p(\varpi \mid \boldsymbol{\theta})$ fundamentally accounts for the identification uncertainty of $\boldsymbol{\theta}$ through $p_0(\varpi \mid D)$ (Stage I) and structural prediction uncertainty through $p(\varpi \mid \boldsymbol{\theta})$.

Equation (29) has important philosophical significance. It says that for the purpose of identifying $\boldsymbol{\theta}$, we can simply postpone using our prior information until Stage II, where in Stage I we artificially use a uniform prior distribution for ϖ. This still leads to the same identification result as in (19) and it avoids the unnecessary complications arising from double-counting the prior information from $\boldsymbol{\theta}$ induced on ϖ. Note that $p_0(\varpi \mid D)$ is not the actual posterior distribution of ϖ because it has not incorporated the prior information in $p(\varpi)$. For the purpose of obtaining the posterior distribution of $\boldsymbol{\theta}$, however, it is the one that facilitates computations and it still leads to the correct answer. Equation (29) is recommended as the standard form to be used for computations.

The foregoing formulation is fundamental and unique as it follows strictly from Bayes' Theorem. The result is simple and intuitive, but non-trivial to derive. The posterior distribution of $\boldsymbol{\theta}$ results consistently from the data prediction model, the structural prediction model and prior information. No heuristics are involved.

5. The role of structural prediction error

By the very nature of prediction error it should play a role, and needs to be modeled, only when it is not possible to match reality with the theoretical model. Otherwise a consistent theory should simply indicate that the prediction error is zero, up to the evidence provided by the data. A mathematical result analogous to this statement can be derived based on the standard form (29).

When there is no structural prediction error, in the sense that the knowledge of $\boldsymbol{\theta}$ completely determines the value of ϖ through the model prediction $\tilde{\varpi}(\boldsymbol{\theta})$ (say), then

$$p(\varpi \mid \boldsymbol{\theta}) = \delta(\varpi - \tilde{\varpi}(\boldsymbol{\theta})) \tag{30}$$

where $\delta(\cdot)$ is the Dirac-Delta function. Substituting (30) into (29) gives

$$p(\boldsymbol{\theta} \mid D) = c_{\varpi}^{-1} p(\boldsymbol{\theta})\, p_0(\tilde{\varpi}(\boldsymbol{\theta}) \mid D) \tag{31}$$

That is, the posterior distribution of $\boldsymbol{\theta}$ is proportional to the product of the prior distribution and the posterior distribution of ϖ in Stage I evaluated at the structural model prediction $\tilde{\varpi}(\boldsymbol{\theta})$.

In the general case when structural prediction error is modeled, suppose we partition the set of model parameters as $\boldsymbol{\theta} = [\boldsymbol{\theta}_S, \boldsymbol{\theta}_P]$, where $\boldsymbol{\theta}_S$ is related to the structural model $\tilde{\varpi}(\cdot)$ and $\boldsymbol{\theta}_P$ is related to the prediction error statistics (e.g., variance) of $\tilde{\varpi}$. With little loss of generality assume that the prior distribution of $\boldsymbol{\theta}_S$ is uniform, i.e., $p(\boldsymbol{\theta}_S) \equiv$ constant. Let $\boldsymbol{\theta}_P = \mathbf{0}$ correspond to the case when the structural prediction error is not modeled. Equation (30) is then written more specifically as

$$p(\varpi \mid \boldsymbol{\theta}_S, \mathbf{0}) = \delta(\varpi - \tilde{\varpi}(\boldsymbol{\theta}_S)) \tag{32}$$

Let $\hat{\varpi}$ be the most probable value (MPV) of ϖ in Stage I, i.e.,

$$p_0(\varpi \mid D) \le p_0(\hat{\varpi} \mid D) \qquad \text{for any } \varpi \tag{33}$$

Suppose there is a $\hat{\boldsymbol{\theta}}_S$ that gives a structural prediction matching exactly the MPV $\hat{\varpi}$ in Stage I, i.e.,

$$\tilde{\varpi}(\hat{\boldsymbol{\theta}}_S) = \hat{\varpi} \tag{34}$$

Clearly, when the structural prediction error is not modeled, $\hat{\boldsymbol{\theta}}_S$ is an MPV of $\boldsymbol{\theta}_S$ because in this case according to (31) with $p(\boldsymbol{\theta}_S) \equiv \text{constant}$, $p(\boldsymbol{\theta}_S \mid D) \propto p_0(\tilde{\varpi}(\boldsymbol{\theta}_S) \mid D)$ is maximized at $\hat{\boldsymbol{\theta}}_S$. Intuitively, it can be expected that when the structural prediction error is modeled, the MPV of $\boldsymbol{\theta}_S$ is still $\hat{\boldsymbol{\theta}}_S$ and the MPV of $\boldsymbol{\theta}_P$ is always zero, unless the prior distribution suggests otherwise. This is expressed in Proposition 1 below.

Proposition 1 (Invariance of MPV when perfect match is possible)

Let $\hat{\varpi}$ be the posterior MPV (most probable value) that maximizes the likelihood $p(D \mid \varpi)$, or equivalently the hypothetical posterior distribution $p_0(\varpi \mid D)$ in Stage I, i.e.,

$$p_0(\varpi \mid D) \le p_0(\hat{\varpi} \mid D) \qquad \text{for any } \varpi \tag{35}$$

Let $\hat{\boldsymbol{\theta}}_S$ be such that it can match the theoretical prediction with $\hat{\varpi}$, i.e.,

$$\tilde{\varpi}(\hat{\boldsymbol{\theta}}_S) = \hat{\varpi} \tag{36}$$

and $(\hat{\boldsymbol{\theta}}_S, \mathbf{0})$ maximizes the prior distribution

$$p(\boldsymbol{\theta}_S, \boldsymbol{\theta}_P) \le p(\hat{\boldsymbol{\theta}}_S, \mathbf{0}) \qquad \text{for any } (\boldsymbol{\theta}_S, \boldsymbol{\theta}_P) \tag{37}$$

Then

$$p(\boldsymbol{\theta}_S, \boldsymbol{\theta}_P \mid D) \le p(\hat{\boldsymbol{\theta}}_S, \mathbf{0} \mid D) \qquad \text{for any } (\boldsymbol{\theta}_S, \boldsymbol{\theta}_P) \tag{38}$$

That is, $(\hat{\boldsymbol{\theta}}_S, \mathbf{0})$ is an MPV of the posterior distribution $p(\boldsymbol{\theta}_S, \boldsymbol{\theta}_P \mid D)$ in Stage II when structural prediction error is modeled. Moreover, if $\hat{\boldsymbol{\theta}}_S$ is unique, the MPV $(\hat{\boldsymbol{\theta}}_S, \mathbf{0})$ is also unique.

■

Proof of Proposition 1:

First of all, maximizing $p(D \mid \varpi)$ and $p_0(\varpi \mid D)$ are equivalent because by definition the latter is the posterior distribution of ϖ with (hypothetically) a uniform prior and from Bayes' Theorem

$$p_0(\varpi \mid D) = p(D)^{-1} p(D \mid \varpi) p(\varpi) \propto p(D \mid \varpi) \tag{39}$$

when $p(\varpi)$ is taken as a constant. When there is no prediction error ($\boldsymbol{\theta}_P = \mathbf{0}$) the conditional distribution $p(\varpi \mid \boldsymbol{\theta}) \equiv p(\varpi \mid \boldsymbol{\theta}_S, \mathbf{0})$ is simply a Dirac-Delta function at $\tilde{\varpi}(\boldsymbol{\theta}_S)$:

$$p(\varpi \mid \boldsymbol{\theta}_S, \mathbf{0}) = \delta(\varpi - \tilde{\varpi}(\boldsymbol{\theta}_S)) \tag{40}$$

Correspondingly, the posterior distribution of $\boldsymbol{\theta}_S$ is given by, according to (29),

$$\begin{aligned} p(\boldsymbol{\theta}_S, \mathbf{0} \mid D) &= c_{\varpi}^{-1} p(\boldsymbol{\theta}_S, \mathbf{0}) \int p_0(\varpi \mid D) \delta(\varpi - \tilde{\varpi}(\boldsymbol{\theta}_S)) d\varpi \\ &= c_{\varpi}^{-1} p(\boldsymbol{\theta}_S, \mathbf{0}) p_0(\tilde{\varpi}(\boldsymbol{\theta}_S) \mid D) \end{aligned} \tag{41}$$

Evaluating at $\boldsymbol{\theta}_S = \hat{\boldsymbol{\theta}}_S$ and using $\tilde{\varpi}(\hat{\boldsymbol{\theta}}_S) = \hat{\varpi}$ from (36) gives

$$\begin{aligned} p(\hat{\boldsymbol{\theta}}_S, \mathbf{0} \mid D) &= c_{\varpi}^{-1} p(\hat{\boldsymbol{\theta}}_S, \mathbf{0}) p_0(\tilde{\varpi}(\hat{\boldsymbol{\theta}}_S) \mid D) \\ &= c_{\varpi}^{-1} p(\hat{\boldsymbol{\theta}}_S, \mathbf{0}) p_0(\hat{\varpi} \mid D) \end{aligned} \tag{42}$$

On the other hand, for any $(\boldsymbol{\theta}_S, \boldsymbol{\theta}_P)$, using the standard form in (29) and $p_0(\varpi \mid D) \le p_0(\hat{\varpi} \mid D)$ from (35),

$$\begin{aligned} p(\boldsymbol{\theta}_S, \boldsymbol{\theta}_P \mid D) &= c_{\varpi}^{-1} p(\boldsymbol{\theta}_S, \boldsymbol{\theta}_P) \int p_0(\varpi \mid D) p(\varpi \mid \boldsymbol{\theta}_S, \boldsymbol{\theta}_P) d\varpi \\ &\le c_{\varpi}^{-1} p(\boldsymbol{\theta}_S, \boldsymbol{\theta}_P) \int p_0(\hat{\varpi} \mid D) p(\varpi \mid \boldsymbol{\theta}_S, \boldsymbol{\theta}_P) d\varpi \\ &= c_{\varpi}^{-1} p(\boldsymbol{\theta}_S, \boldsymbol{\theta}_P) p_0(\hat{\varpi} \mid D) \int p(\varpi \mid \boldsymbol{\theta}_S, \boldsymbol{\theta}_P) d\varpi \\ &= c_{\varpi}^{-1} p(\boldsymbol{\theta}_S, \boldsymbol{\theta}_P) p_0(\hat{\varpi} \mid D) \end{aligned} \tag{43}$$

since $\int p(\varpi \mid \boldsymbol{\theta}_S, \boldsymbol{\theta}_P) d\varpi = 1$. From (37), $p(\boldsymbol{\theta}_S, \boldsymbol{\theta}_P) \le p(\hat{\boldsymbol{\theta}}_S, \mathbf{0})$ for any $(\boldsymbol{\theta}_S, \boldsymbol{\theta}_P)$, and so (43) implies

$$p(\boldsymbol{\theta}_S, \boldsymbol{\theta}_P \mid D) \le c_{\varpi}^{-1} p(\hat{\boldsymbol{\theta}}_S, \mathbf{0}) p_0(\hat{\varpi} \mid D) = p(\hat{\boldsymbol{\theta}}_S, \mathbf{0} \mid D) \tag{44}$$

and hence $(\hat{\boldsymbol{\theta}}_S, \mathbf{0})$ is an MPV of the posterior distribution $p(\boldsymbol{\theta}_S, \boldsymbol{\theta}_P \mid D)$. If $\hat{\boldsymbol{\theta}}_S$ is unique, then the MPV $(\hat{\boldsymbol{\theta}}_S, \mathbf{0})$ is also unique because then it is the only value that satisfies the above inequality.

■

6. Illustrative example

In this section we illustrate the developed theory using a simple example. In the companion paper we apply the theory to the case of structural model identification where the modal properties are identified using ambient data in Stage I and the structural model parameters (stiffness, mass) are identified in Stage II. The example here is designed to be simple so that it can be easily understood and it can illustrate the main theory explicitly. In reality the data prediction model (Stage I) and structural prediction model (Stage II) can be much more complicated to formulate. The number of parameters can also be much larger.

Consider the following situation. The output response quantity y of a system is thought to depend on the input x. The theoretical relationship between y and x is described by

$$\tilde{y}(x;\mathbf{a}) = \frac{a_1}{2}(x - a_2)^2 + a_3 \tag{45}$$

where $\mathbf{a} = [a_1, a_2, a_3]$ contains the parameters associated with the model. Among these parameters, a_1 and a_2 are of primary interest because they in turn depend on another parameter θ_S that is directly associated with the property of the system. Theoretically this relationship may be described as $a_1 = f_1(\theta_S)$ and $a_2 = f_2(\theta_S)$ where $f_1(\theta_S)$ and $f_2(\theta_S)$ are known but complicated functions of θ_S. In reality, modeling error exists and it is addressed by the following probabilistic description

$$\begin{aligned} a_1 &= f_1(\theta_S) + \xi_1 \\ a_2 &= f_2(\theta_S) + \xi_2 \end{aligned} \tag{46}$$

where ξ_1 and ξ_2 are assumed to be i.i.d. (independent and identically distributed) Gaussian with zero mean and (unknown) variance v_2.

Output data of y when the system is subjected to inputs $\{x_k : k = 1,..., N\}$ ($N =$ number of samples) has been measured:

$$D = \{\hat{y}_k : k = 1,..., N\} \tag{47}$$

Because of measurement error, $\hat{y}_k$ and x_k do not obey exactly the theoretical relationship in (45). To reflect the measurement error their relationship is modeled by augmenting (45) as

$$\hat{y}_k = \tilde{y}(x_k; \mathbf{a}) + \varepsilon_k \tag{48}$$

where $\{\varepsilon_k : k = 1,..., N\}$ are modeled as i.i.d. Gaussian random variables with zero mean and (unknown) variance v_1.

Using the data D, it is of interest to identify (a_1, a_2) and, if possible, θ_S. From (45) and (46), it is clear that y depends on (a_1, a_2) in a much more intuitive manner than on θ_S. According to the theoretical model, a_1 is the curvature, a_2 is the location of x when y is minimum and a_3 is the minimum value of y. Provided that the inputs $\{x_k : k = 1,..., N\}$ cover a broad enough region around the minimum value of y, $\mathbf{a} = [a_1, a_2, a_3]$ should be globally identifiable and is therefore much easier to identify than θ_S. This suggests the potential advantage of a two-stage approach where in Stage I we first identify $\mathbf{a}$ from the data and then use the identification result of a_1 and a_2 to identify θ_S in Stage II. The identification result in Stage I immediately yields useful information and understanding towards the problem, regardless of whether θ_S can be identified. Stage II attempts to further identify θ_S. By this time there is already some insights gained from Stage I and the problem has been condensed in the sense that only the mappings $f_1(\theta_S)$ and $f_2(\theta_S)$ are involved, rather than their composite action through the quadratic relationship (45) in Stage I. Issues regarding identifiability and identification precision of θ_S are then directly related to the mappings $f_1(\theta_S)$ and $f_2(\theta_S)$, and the identification precision of a_1 and a_2 in Stage I. In this manner the problem is better structured.

6.1. Likelihood function in Stage I

In the context of our theory, the data prediction model is given by (48) where ε_k is the data prediction error. The inputs $\{x_k : k = 1,..., N\}$ are assumed to be given and they shall be omitted in the conditioning statement to maintain consistency in notation with the main theory. The structural prediction model is given by (46), where (ξ_1, ξ_2) is the structural prediction error. Although the primary interest in Stage I is to identify (a_1, a_2), the likelihood function

$p(D \mid a_1, a_2)$ cannot be derived directly based on (48) because the full probabilistic description of the data requires the knowledge of a_3 and v_1 as well. The full set of parameters in Stage I should therefore be

$$\boldsymbol{\alpha} = [a_1, a_2, a_3, v_1] \tag{49}$$

Among these parameters, only a_1 and a_2 depend on θ_S. Thus,

$$\boldsymbol{\varpi} = [a_1, a_2] \tag{50}$$

$$\boldsymbol{\upsilon} = [a_3, v_1] \tag{51}$$

We shall assume a uniform prior distribution for $\boldsymbol{\upsilon} = [a_3, v_1]$. We cannot assign the prior distribution for $\boldsymbol{\varpi} = [a_1, a_2]$, however, as it is already determined by $p(\theta_S, v_2)$ and $p(\boldsymbol{\varpi} \mid \theta_S, v_2)$; see (59) later.

From (48), given (a_1, a_2, a_3, v_1) and inputs $\{x_k : k = 1,..., N\}$, the data $\{\hat{y}_k : k = 1,..., N\}$ are i.i.d. Gaussian with mean $\{\tilde{y}(x_k; \mathbf{a}) : k = 1,..., N\}$ and variance v_1. The likelihood function in Stage I is then given by

$$\begin{aligned} p(D \mid \boldsymbol{\varpi}, \boldsymbol{\upsilon}) &\equiv p(\{\hat{y}_k\}_{k=1}^N \mid a_1, a_2, a_3, v_1) \\ &= (2\pi)^{-N/2} v_1^{-N/2} \exp\left\{ -\frac{1}{2v_1} \sum_{k=1}^{N} [\hat{y}_k - \tilde{y}(x_k; \mathbf{a})]^2 \right\} \end{aligned} \tag{52}$$

where the conditioning on $\{x_k : k = 1,..., N\}$ has been omitted for consistency with the main theory.

6.2. Conditional distribution $p(\boldsymbol{\varpi} \mid \boldsymbol{\theta})$

Although the primary interest in Stage II is to identify θ_S, the conditional distribution $p(\boldsymbol{\varpi} \mid \theta_S) \equiv p(a_1, a_2 \mid \theta_S)$ cannot be derived directly because according to (46) the full probabilistic description of $\boldsymbol{\varpi} = [a_1, a_2]$ requires the knowledge of v_2 as well. The full set of parameters to be identified in Stage II should therefore be defined as

$$\boldsymbol{\theta} = [\theta_S, v_2] \tag{53}$$

In the context of Proposition 1,

$$\begin{aligned}\boldsymbol{\theta}_S &= \theta_S \\ \boldsymbol{\theta}_P &= v_2\end{aligned} \tag{54}$$

The independence prior assumption $p(\varpi, \boldsymbol{\upsilon} \mid \boldsymbol{\theta}) = p(\varpi \mid \boldsymbol{\theta}) p(\boldsymbol{\upsilon})$ in (8) manifests into

$$p(a_1, a_2, a_3, v_1 \mid \theta_S, v_2) = p(a_1, a_2 \mid \theta_S, v_2) p(a_3, v_1) \tag{55}$$

which is justified in this example.

According to (46), given $\boldsymbol{\theta} = [\theta_S, v_2]$, a_1 and a_2 are i.i.d. Gaussian with mean $f_1(\theta_S)$ and $f_2(\theta_S)$, respectively, and a common variance v_2. The conditional distribution $p(\varpi \mid \boldsymbol{\theta})$ is then given by

$$\begin{aligned} p(\varpi \mid \boldsymbol{\theta}) &\equiv p(a_1, a_2 \mid \theta_S, v_2) \\ &= (2\pi)^{-1} v_2^{-1} \exp\left\{ -\frac{1}{2v_2}[a_1 - f_1(\theta_S)]^2 - \frac{1}{2v_2}[a_2 - f_2(\theta_S)]^2 \right\} \end{aligned} \tag{56}$$

6.3. Posterior distribution in Stage I

Given the data D, the posterior distribution of $[\varpi, \boldsymbol{\upsilon}] \equiv [a_1, a_2, a_3, v_1]$ is given by, using Bayes' Theorem,

$$\begin{aligned} p(\varpi, \boldsymbol{\upsilon} \mid D) &\equiv p(a_1, a_2, a_3, v_1 \mid \{\hat{y}_k\}_{k=1}^N) \\ &\propto p(\{\hat{y}_k\}_{k=1}^N \mid a_1, a_2, a_2, v_1) p(a_1, a_2, a_3, v_1) \\ &= (2\pi)^{-N/2} v_1^{-N/2} \exp\left\{ -\frac{1}{2v_1} \sum_{k=1}^N [\hat{y}_k - \tilde{y}(x_k; \mathbf{a})]^2 \right\} p(a_1, a_2, a_3, v_1) \end{aligned} \tag{57}$$

where $p(a_1, a_2, a_3, v_1)$ is the prior distribution. Using (55) it can be shown that (see also (9))

$$p(a_1, a_2, a_3, v_1) = p(a_1, a_2) p(a_3, v_1) \tag{58}$$

Note that $p(a_3, v_1)$ has been assigned to be uniform in Section 6.1. On the other hand, $p(a_1, a_2)$ is not open to assignment by the analyst because it is completely determined by $p(a_1, a_2 \mid \theta_S, v_2)$ (from (56)) and $p(\theta_S, v_2)$ (assigned by analyst) through (see also (21)):

$$p(a_1, a_2) = \int \int p(a_1, a_2 \mid \theta_S, v_2) p(\theta_S, v_2) d\theta_S dv_2 \tag{59}$$

6.4. Hypothetical posterior distribution in Stage I

The posterior distribution $p(\varpi, \mathbf{\upsilon} \mid D)$ in (57) is the one that reflects the updated knowledge on $(\varpi, \mathbf{\upsilon})$ after incorporating the information from both the prior distribution and the data. As far as the knowledge of $(\varpi, \mathbf{\upsilon})$ is concerned, it is the one that is relevant. However, it is not the posterior distribution to be used in Stage II in the standard formulation (Section 4) to avoid double-counting the prior information. Instead, the following hypothetical posterior distribution (which has ignored the prior information induced on ϖ in Stage I) should be used:

$$\begin{aligned} p_0(\varpi, \mathbf{\upsilon} \mid D) &\equiv p_0(a_1, a_2, a_3, v_1 \mid \{\hat{y}_k\}_{k=1}^N) \\ &\propto p(\{\hat{y}_k\}_{k=1}^N \mid a_1, a_2, a_3, v_1)\, p(a_3, v_1) \\ &\propto (2\pi)^{-N/2} v_1^{-N/2} \exp\left\{ -\frac{1}{2v_1} \sum_{k=1}^N [\hat{y}_k - \tilde{y}(x_k; \mathbf{a})]^2 \right\} \end{aligned} \tag{60}$$

since $p(a_3, v_1)$ has been assigned to be uniform in Section 6.1. Based on this, the hypothetical posterior marginal distribution $p_0(\varpi \mid D)$ can be obtained by integrating out $\mathbf{\upsilon} = [a_3, v_1]$:

$$\begin{aligned} p_0(\varpi \mid D) &\equiv p_0(a_1, a_2 \mid \{\hat{y}_k\}_{k=1}^N) \\ &= \int_0^\infty \int_{-\infty}^\infty p_0(a_1, a_2, a_3, v_1 \mid \{\hat{y}_k\}_{k=1}^N) da_3 dv_1 \\ &\propto \int_0^\infty \int_{-\infty}^\infty (2\pi)^{-N/2} v_1^{-N/2} \exp\left\{ -\frac{1}{2v_1} \sum_{k=1}^N [\hat{y}_k - \tilde{y}(x_k; \mathbf{a})]^2 \right\} da_3 dv_1 \end{aligned} \tag{61}$$

Substituting $\tilde{y}(x_k; \mathbf{a})$ from (45) shows that the exponent in the integrand is a quadratic function of a_3 and so the integration with respect to a_3 can be performed analytically, giving (details omitted)

$$p_0(\varpi \mid D) \propto \int_0^\infty v_1^{-(N-1)/2} \exp\left[-\frac{N}{2v_1} (\mu_2 - \mu_1^2) \right] dv_1 \tag{62}$$

where

$$\mu_1(a_1, a_2) = \frac{1}{N} \sum_{k=1}^N [\hat{y}_k - \frac{a_1}{2} (x_k - a_2)^2] \tag{63}$$

$$\mu_2(a_1, a_2) = \frac{1}{N} \sum_{k=1}^N [\hat{y}_k - \frac{a_1}{2} (x_k - a_2)^2]^2 \tag{64}$$

The integral with respect to v_1 in (62) be evaluated analytically, giving (omitting constants)

$$p_0(\varpi \mid D) \propto (\mu_2 - \mu_1^2)^{-(N-3)/2} \tag{65}$$

Note that μ_1 and μ_2 depend on $\varpi = [a_1, a_2]$ through (63) and (64), respectively.

It should be noted that in general applications the integrals can be complicated and they may not be obtained analytically. When the integral is of Laplace type its asymptotic behavior for large N can be expressed analytically [18][19] in terms of the Hessian matrix of the log of the integrand. More generally, as long as $p_0(\varpi, \upsilon \mid D)$ has a single peak, it may be (heuristically) approximated by a Gaussian distribution. Correspondingly the marginal distribution $p_0(\varpi \mid D)$ is also Gaussian. This, however, is only an approximation and need not be asymptotically correct.

6.5. Posterior distribution in Stage II

Based on (29), the posterior distribution $p(\boldsymbol{\theta} \mid D) \equiv p(\theta_S, v_2 \mid \{\hat{y}_k\}_{k=1}^N)$ in Stage II is given by

$$\begin{aligned} p(\boldsymbol{\theta} \mid D) &\propto p(\boldsymbol{\theta}) \int p_0(\varpi \mid D)\, p(\varpi \mid \boldsymbol{\theta})\, d\varpi \\ &\equiv p(\theta_S, v_2) \int\int p_0(a_1, a_2 \mid \{\hat{y}_k\}_{k=1}^N)\, p(a_1, a_2 \mid \theta_S, v_2)\, da_1 da_2 \end{aligned} \tag{66}$$

where $p(\theta_S, v_2)$ is the prior distribution assigned by the analyst; $p_0(a_1, a_2 \mid \{\hat{y}_k\}_{k=1}^N)$ is given by (65); and $p(a_1, a_2 \mid \theta_S, v_2)$ is given by (56). Evaluating this distribution for each value of $\boldsymbol{\theta}$ requires efficient evaluation of the integral with respect to $\varpi = [a_1, a_2]$. The difficulty of evaluating the integral depends on the complexity of the theoretical relationship between θ_S and $\varpi = [a_1, a_2]$, i.e., $f_1(\theta)$ and $f_2(\theta)$. It is generally problem-dependent and will not be further discussed here.

6.6. Remarks on basic form

For instructional purpose, if one insists on using the posterior distribution $p(\varpi, \upsilon \mid D)$ in (57) for making inference on $\boldsymbol{\theta}$ in Stage II, then one must use (19):

$$p(\boldsymbol{\theta}\,|\,D) = p(\boldsymbol{\theta})\int \frac{p(\varpi\,|\,D)}{p(\varpi)}\, p(\varpi\,|\,\boldsymbol{\theta})\,d\varpi$$
$$= p(\theta_S, v_2)\int\int \frac{p(a_1, a_2\,|\,\{\hat{y}_k\}_{k=1}^{N})}{p(a_1, a_2)}\, p(a_1, a_2\,|\,\theta_S, v_2)\,da_1 da_2 \tag{67}$$

so that the prior information already used in Stage I is correctly removed in Stage II. To use this form, however, the prior distribution $p(\varpi) \equiv p(a_1, a_2)$ must be obtained first from (59). Evaluating $p(\varpi)$ for each ϖ involves evaluating an integral over $\boldsymbol{\theta} = [\theta_S, v_2]$. Unlike the integral in (61), this integral is more difficult to evaluate or approximate because the relationship between $\varpi = [a_1, a_2]$ and $\boldsymbol{\theta}$ implied by $f_1(a_1, a_2)$ and $f_2(a_1, a_2)$ can be very complicated. Through the standard formulation, our theory shows that (66) and (67) gives the same distribution, which is non-trivial from the appearance of these expressions.

7. Conclusions

This paper has made several contributions that lay down the foundation of the Bayesian two-stage system identification framework:

1) general formulation of the posterior distribution in Stage II in terms of the identification results in Stage I and the structural prediction model in Stage II, as in (19);
2) formulation of the double-counting principle that leads to the standard form of the posterior distribution in (29); see also Theorem 1;
3) theoretical result regarding the role of the structural prediction error, see Proposition 1 and Section 5;

These are fundamental discoveries that resolve the heuristic nature of the existing two-stage formulations in structural system identification. The following critical comments are in order:

1) The form of the posterior distribution of the structural model parameters in Stage II is not subjected to heuristic choice. It is given by the product integral of the posterior distribution of the parameters in Stage I and the conditional distribution characterizing the relationship between the parameters in Stage I and II, as in (19) or (29).
2) It has been argued previously [20][21] that in modal identification (Stage I) the prior distribution of the modal parameters (natural frequency, damping, etc.) can be assumed to be uniform because for sufficient data encountered in practice it is slowly varying compared to

the likelihood function. That is, even if a uniform prior distribution is used the identification results will be practically the same. In this paper the double-counting principle and the standard form in Section 4 now shows that a uniform prior distribution MUST be used in Stage I when the modal identification results are to be used conveniently for further identifying the structural parameters in Stage II.

8. Acknowledgements

The work in this paper is partially supported by Grant EGG10034 from the University of Liverpool, the National Basic Research Program of China (973 Program) (Project No. 2014CB049100) and Fundamental Research Funds for the Central Universities, China (Grant No. 2014KJ040).

References

[1] D. Malakoff, Bayes Offers a 'New' Way to Make Sense of Numbers, Science 286 (1999) 460-1464.

[2] R. T. Cox, The Algebra of Probable Inference, Johns Hopkins, Baltimore,1961.

[3] E. T. Jaynes, Probability Theory: The Logic of Science, Cambridge University Press, UK, 2003.

[4] L. S. Katafygiotis, H.F. Lam, Tangential-Projection Algorithm for Manifold Representation in Unidentifiable Model Updating Problems, Earthquake Engineering and Structural Dynamics 31(4) (2002) 791-812.

[5] D. Ewins, Modal Testing, 2nd Edition, Research Studies Press, Baldock, 2000.

[6] J. L. Beck, L.S. Katafygiotis, Updating Models and Their Uncertainties. I: Bayesian Statistical Framework, Journal of Engineering Mechanics, ASCE 124(4) (1998) 455-461.

[7] J.L. Beck, Bayesian System Identification Based on Probability Logic. Structural Control and Health Monitoring 17(7) (2010) 825-847.

[8] M.W. Vanik, J. L. Beck, S.K. Au, Bayesian Probabilistic Approach to Structural Health Monitoring, Journal of Engineering Mechanics, ASCE 126(7) (2000) 738-745.

[9] J. L. Beck, S. K. Au, M.W. Vanik, Monitoring Structural Health Using a Probabilistic Measure, Computer-Aided Civil and Infrastructure Engineering 16(1)(2001) 1-11.

[10] K.V. Yuen, J. L. Beck, L.S. Katafygiotis, Efficient Model Updating and Health Monitoring Methodology Using Incomplete Modal Data without Mode Matching, Structural Control and Health Monitoring 13(2006) 91-107.

[11] K. Christodoulou, C. Papadimitriou, Structural Identification Based on Optimally Weighted Modal Residuals, Mechanical Systems and Signal Processing 21(2007) 4-23.

[12] K. Christodoulou, E. Ntotsios, C. Papadimitriou, C. Panetsos, Structural Model Updating and Prediction Variability Using Pareto Optimal Models, Computational Methods and Applications in Mechanical Engineering 198(2008) 138-149.

[13] K.V. Yuen, S.K. Au, J. L. Beck, Two-Stage Structural Health Monitoring Approach for Phase I Benchmark Studies, Journal of Engineering Mechanics, ASCE 130(1) (2004) 16-33.

[14] B. Goller, M. Broggi, A. Calvi, G.I. Schueller, A Stochastic Model Updating Technique for Complex Aerospace Structures, Finite Elements in Analysis and Design, 47(2011) 739-752.

[15] E. Simoen, G. De Roeck, G. Lombaert, Resolution and Uncertainty Analysis of FE Model Updating Results, In Proceedings of the 8th International Conference on Structural Dynamics, pp.2318-2325; Leuven, Belgium, 4-6 July 2011.

[16] C. Papadimitriou, C. Argyris, D.C. Papadioti, P. Panetsos, Uncertinty Calibration of Large-Order Models of Bridges Using Ambient Vibration Measurements, In Proceedings of 7th European Workshop on Structural Health Monitoring, La Cité, Nantes, France, 8-11 July 2014.

[17] S.K. Au, F.L. Zhang, A Fundamental Two-Stage Formulation for Bayesian System Identification, Part II: Application to Ambient Vibration Data, Mechanical Systems and Signal Processing, under review.

[18] N. Bleistein, R.A. Handelsman, Asymptotic Expansions of Integrals, Dover, New York, 1986.

[19] C. Papadimitriou, J.L. Beck, L.S. Katafygiotis, Asymptotic Expansions for Reliabilities and Moments of Uncertain Dynamic Systems, Journal of Engineering Mechanics, 123(12) (1997) 1219-1229.

[20] S.K. Au, Fast Bayesian ambient modal identification in the frequency domain, Part I: posterior most probable value, Mechanical Systems and Signal Processing, 26(1) (2012) 60-75.

[21] S.K. Au, F.L. Zhang, Y.C. Ni, Bayesian operational modal analysis: theory, computation, practice, Computers and Structures, 126 (2013) 3-14.

List of tables

List of figures

Table 1. Nomenclature of two-stage Bayesian identification problem

Parameter	Descriptions
D	Available data for Bayesian inference
$\boldsymbol{\theta} = [\boldsymbol{\theta}_S, \boldsymbol{\theta}_P]$	Parameters identified in Stage II, partitioned according to the role in the structural prediction model; $\boldsymbol{\theta}_S$ related to structural model prediction; $\boldsymbol{\theta}_P$ related to structural prediction error
$\boldsymbol{\alpha} = [\varpi, \boldsymbol{\upsilon}]$	Parameters identified in Stage I, ϖ related to $\boldsymbol{\theta}$, $\boldsymbol{\upsilon}$ not related to $\boldsymbol{\theta}$
$\tilde{\varpi}(\boldsymbol{\theta})$	Structural model prediction of ϖ for given $\boldsymbol{\theta}$
$p(\boldsymbol{\theta})$	Prior distribution of $\boldsymbol{\theta}$
$p(\boldsymbol{\upsilon})$	Prior distribution of $\boldsymbol{\upsilon}$
$p(\varpi \mid \boldsymbol{\theta})$	Conditional distribution of ϖ given $\boldsymbol{\theta}$, formulated based on structural prediction model
$p(D \mid \varpi, \boldsymbol{\upsilon})$	Likelihood function used in Stage I, formulated based on data prediction model

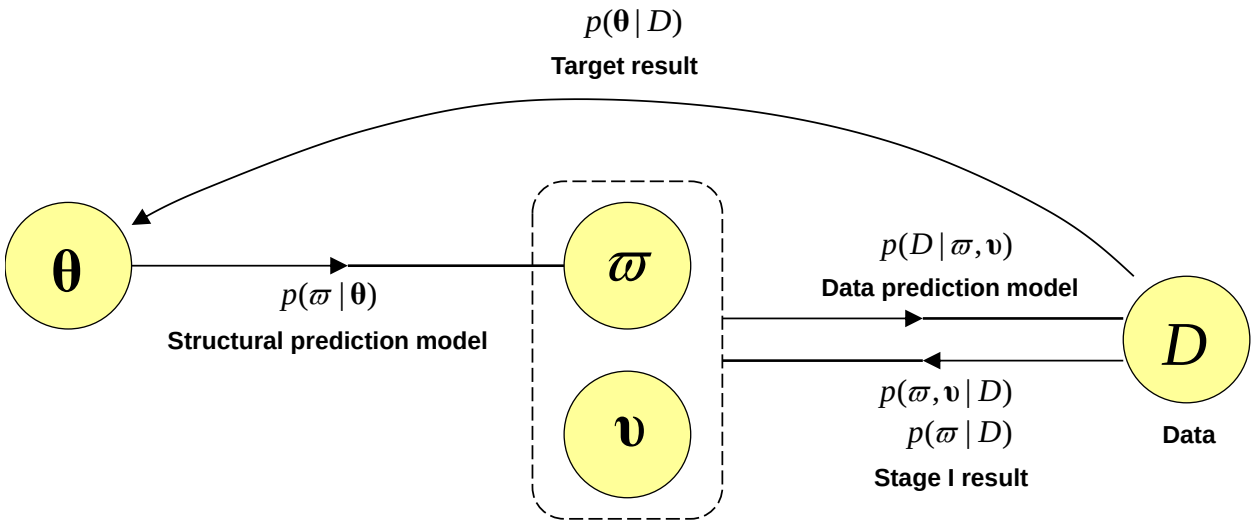

Figure 1. Schematic diagram of two-stage Bayesian identification problem