I. INTRODUCTION

Many recent Bayesian applications in acoustics hint at ‘Bayesian revolution’ in Science and Engineering (Ballard et al., 2020; Landschoot and Xiang, 2019; Lee et al., 2020; Martiartu et al., 2019; Nannuru et al., 2018). However, logical foundations of the Bayesian methods cannot be put in fully satisfactory form until the classical problem of arbitrariness (sometimes called ‘subjectivity’) in assigning prior probability is resolved (Jaynes, 2003). These recent Bayesian applications have already reached a level where the problem of prior probabilities can no longer be ignored.

This paper describes model-based Bayesian analysis at an introductory level to the readers of the journal, leaning to sacrifice some rigorous complexity for increasing clarity in a way to keep mathematical handling of Bayesian probability methods as minimum as feasible. Model-based approaches rely crucially on parametric models which can be derived from physical/acoustical principles, can also be phenomenological, and numerical. Another class of Bayesian data analysis which does not necessarily rely on parametric models, so-called nonparametric Bayesian analysis (Müller et al., 2015) is outside the scope of this paper.

This tutorial paper is organized as follows: Sec. II introduces logical concept of probability in Bayesian view. Sec. III discusses two simple examples within Bayesian framework. Sec. IV applies the principle of maximum entropy to assign prior probabilities. Sec. V explains two levels of Bayesian inference. Sec. VI presents a number of recent applications of model-based Bayesian analysis, and Sec. VII summarizes the paper.

II. BAYESIAN PROBABILITY

When considering the interpretation of probability, debate is now in its third century between the two main schools of statistical inference, the Bayesian school and the frequentist school. Frequentists consider probability to be a proportion in a large ensemble of repeatable observations (Cox, 2006). This interpretation has been dominant in statistics until recent decades (McGrayne, 2011). But in considering, for instance, the probability of the mass of the universe falling between certain bounds, the probability cannot be interpreted in terms of frequencies of repeatable observations, because there is only one universe. The Bayesian school [henceforth just ‘Bayesian’ (Fienberg, 2006)] views probability instead as a degree or strength of implication: how strongly one thing implies another (Garrett, 1998; Keynes, 1921). Carnap (1950) taking a similar view, considered it as the degree of confirmation (or conclusion) of a hypothesis on basis of some given evidence (or promises). In detail, \( p(B|A) \) represents how strongly the assumed truth of one binary proposition \( A \) implies the truth of another, \( B \), according to the relations known between the things that the propositions refer to; in this expression the proposition in question \( (B) \) appears first, to the left of the vertical line or ‘conditioning solidus’, and the proposition held to be true \( (A) \) appears to its right. Use of implication or confirmation, rather than belief, does away with common criticisms of Bayesianism relating to psychology,
and also demonstrates clearly that all probabilities are conditional: they depend on at least two propositions, one of which is the conditioning information, so there is no such thing as an unconditional probability (Carnap, 1950; de Finetti, 2017). Furthermore, since propositions obey an algebra – Boolean algebra – and since propositions are the arguments of probabilities, an algebra for the probabilities follows from the algebra of propositions. This algebra turns out to be the sum and product rules (Cox, 1946), and this is their deepest rationale. Herein is the justification for calling degree of implication ‘probability’ – it obeys the two ‘laws of probability’, and it is what is actually needed in all problems involving uncertainty.

If there were objections to this meaning of ‘probability’, Bayesians do not waste time on semantics but get on with calculating the degree of implication, because it is what is needed in solving any real problem, regardless of the name attached to it. Below, ‘probability’ should be understood as a shorthand for ‘degree of implication’. Note that state of knowledge or state of information can be quantitatively encoded in probabilities, namely in ‘degree of implication’.

Probability is a real-valued quantity ranging between 0 and 1, with \( p(B|A) = 0 \) meaning that truth of \( A \) implies falsehood of \( B \) with certainty, and \( p(B|A) = 1 \) meaning that truth of \( A \) implies truth of \( B \) with certainty. Bayesian probability theory is not restricted to applications involving a large number of repeatable events or so-called random variables. Using Bayesian probability it is possible to reason in a consistent and rational manner about single events (such as the mass of the universe) when the information needed for certainty is lacking, so-called inductive reasoning (Jaynes, 2003; Keynes, 1921). At the same time, Bayesian probability can also be applied in the ‘frequentist’ case of repeated trials with an uncontrolled variable. The extent of control is included in the conditioning proposition. In the case of repeated trials, the value of the probability of an outcome is often equal numerically to the relative frequency (proportion) of that outcome, but the concepts are distinct.

The frequentist view is associated with ‘randomness’ about what member of an ensemble is chosen by nature, in a ‘random process’. But when people speak of a random process (perhaps yielding a ‘random number’) they really mean a process which they believe nobody can work out how to predict. That statement is as much about human ingenuity as about the actual process, for a smarter person might work out how to analyze the process better or gather further information about it. Randomness is not intrinsic to a system, which is why mathematicians have not been able to come up with any agreed definition of it. Accordingly, the Bayesian view downplays the notion (Jaynes, 2003).

**A. Product and Sum Rules**

Given the probabilities of various propositions, how can the probability of any compound proposition assem-

bled from them, using the Boolean operations of logical product, logical sum and negation, be calculated? Two relations turn out to be enough to decompose the probability of any compound proposition. These are known as the **product rule** and the **sum rule**. They follow principally from the associativity property of the Boolean logical product, by decomposing the probability of the logical product of three propositions in differing ways and equating the results (Cox, 1946; 1961). For their interesting history and a full tutorial derivation, see Jaynes (2003).

**Product rule**: given two propositions \( A \) and \( B \), the probability of the logical product \( AB \) (i.e., the probability of them both being true) is given by

\[
p(AB|Z) = p(B|AZ) p(A|Z),
\]

where \( Z \) is a proposition specifying the ‘background information’. This rule is named after the product of probabilities on its right-hand side, although it is best understood as specifying how to move a proposition (here, \( A \)) across the conditioning solidus. An insertion of a comma between the propositions in a logical product, for instance, \( p(A, B|Z) \), often makes the probability expression clearer, where they appear as an argument of a probability, and in particular where continuous variables are involved (as discussed later in Sec. II B), so as to maintain consistency with conventional mathematical notation for functions of multiple arguments. If knowledge of whether or not \( A \) is true has no bearing on one’s knowledge of whether \( B \) is true, so that

\[
p(B|A, Z) = p(B|Z),
\]

then the product rule reduces to

\[
p(A, B|Z) = p(A|Z) p(B|Z).
\]

In this case, \( A \) and \( B \) are said to be **logically independent** of each other, given \( Z \).

The product rule of two propositions can be generalized to more propositions. In particular, if multiple propositions \( A, B, C \ldots \) are logically independent of each other, given \( Z \), then

\[
p(A, B, C \ldots |Z) = p(A|Z) p(B|Z) p(C|Z) \cdots.
\]

**Sum rule**: if the probability \( p(A|Z) \) that a proposition \( A \) is true is known, given \( Z \), then the probability that it is false, \( p(\overline{A}|Z) \) must be calculable from it: \( p(\overline{A}|Z) \) is a unique function of \( p(A|Z) \), where \( \overline{A} \) represents the logical negation of \( A \). The relation between these two probabilities is

\[
p(A|Z) + p(\overline{A}|Z) = 1
\]

and is known, from its form, as the sum rule. For any two propositions \( A, B \), it can be shown using the sum and product rules, together with de Morgan’s laws (Patrick,
2015) relating the Boolean operations of logical sum, logical product and negation, that
\[ p(A + B|Z) + p(A, B|Z) = p(A|Z) + p(B|Z). \] (6)

Suppose now that \( A \) and \( B \) are exclusive, given \( Z \), so that at most one of them is true. Then \( p(A|B, Z) \) and \( p(B|A, Z) \) are zero, from which it follows that \( p(A, B|Z) \) is zero upon decomposing it using the product rule. Hence, in that case,
\[ p(A + B|Z) = p(A|Z) + p(B|Z). \] (7)

Upon writing \( B \) as the logical sum of two further exclusive propositions and re-applying this relation repeatedly (the ‘method of induction’), it follows that, for a set of propositions \( A, B, C \ldots \) that are mutually exclusive (one at most is true), given \( Z \), then
\[ p(A|Z) + p(B|Z) + p(C|Z) + \ldots = p(A + B + C \ldots |Z). \] (8)

If \( A, B, C \ldots \) are also exhaustive given \( Z \), meaning that one of the set is always true and any one member is the negation of the logical sum of the rest, then
\[ p(A|Z) + p(B|Z) + p(C|Z) + \ldots = 1. \] (9)

When a discrete variable is considered, this relation normalizes the probabilities for its values. For instance, if \( D_k \) is the proposition that ‘the \( k \)th face of a dice shows’, and \( Z \) specifies that ‘the dice has \( K \) faces’, then
\[ \sum_{k=1}^{K} p(D_k|Z) = 1. \] (10)

B. Marginalization and Probability Density

Marginalization is a means to reduce dimensions of a multivariate distribution/density, in general. In Bayesian data analysis particularly, it enables the removal of nuisance parameters. These are variables, along with the variable of interest, which contain relevant information, but which are not themselves of interest. The rule for marginalizing follows from the sum and product rules. Consider the expression
\[ p(A, B|Z) + p(A, \bar{B}|Z). \] (11)

Using the product rule to move \( A \) across the conditioning solidus in these two joint probabilities, this expression becomes
\[ [p(B|A, Z) + p(\bar{B}|A, Z)]p(A|Z). \] (12)

Since the term in square brackets takes value one by the sum rule, it follows that
\[ p(A|Z) = p(A, B|Z) + p(A, \bar{B}|Z). \] (13)

Generalization is routine to the case of a discrete variable \( \theta \) taking one of the set of \( K \) possible values \( \theta_k \), \( k = 1, 2 \ldots K \), when the value of \( \theta \) has a bearing on a proposition of interest, \( A \). Denote by \( \Theta_k \) the proposition that \( \theta \) takes the value \( \theta_k \), so that the set of propositions \( \Theta_k, k = 1, 2 \ldots K \), is exclusive and exhaustive. Then
\[ p(A|Z) = \sum_{k=1}^{K} p(A, \Theta_k|Z). \] (14)

Now suppose that the value of \( \theta \) may have a bearing on another variable, \( x \), taking one of the set of \( J \) possible values \( x_j, j = 1, 2 \ldots J \). Denote by \( X_j \) the proposition that \( x \) takes the value \( x_j \), so that the set of propositions \( X_j, j = 1, 2 \ldots J \), is exclusive and exhaustive. The preceding relation is true when \( A \) is replaced by any \( X_j \):
\[ p(X_j|Z) = \sum_{k=1}^{K} p(X_j, \Theta_k|Z) \]
\[ = \sum_{k=1}^{K} p(\Theta_k|X_j, Z)p(X_j|Z). \] (15)

This is a relation between probabilities of propositions. One can now replace the probabilities by functional forms, and think of \( p(X_j|Z) \) as a function of the discrete variable \( j \), denoted \( p(x_j|Z) \), and think of \( p(\Theta_k|X_j, Z) \) as a function of \( j \) and \( k \), denoted \( p(\theta_k|x_j,Z) \).

In the case of continuous variables it is necessary to introduce the idea of probability densities, by applying the sum and product rules to propositions of type ”the continuous variable takes a value between \( x \) and \( x + dx \), and defining the probability of this proposition to be a probability density multiplied by \( dx \)” which is specified as given background information, \( z \). It follows now that
\[ p(x, \theta|z) + p(x, \bar{\theta}|z) = p(x|z). \] (16)

This can be derived by applying the product rule
\[ p(x, \theta|z) = p(\theta|x, z)p(x|z), \] (17)

Upon adding the product rule for \( p(x, \bar{\theta}|z) \) to both sides of Eq. (17),
\[ p(x, \theta|z) + p(x, \bar{\theta}|z) = [p(\theta|x, z) + p(\bar{\theta}|x, z)]p(x|z). \] (18)

This result for marginalization can be generalized to multiple alternative propositions as a consequence of both the product and the sum rule: given a joint probability \( p(x_j, \theta_k) \) of \( K \) alternative propositions:
\[ p(x_j|y) = \sum_{k=1}^{K} p(x_j, \theta_k|y) = \sum_{k=1}^{K} p(\theta_k|x_j, y)p(x_j|y), \] (19)

where \( y \) specifies that ”\( x_j, \theta_k \) are discrete variables, and \( 1 \leq j \leq J, 1 \leq k \leq K \) with \( J, K \) alternative propositions, respectively”. In similar fashion, applying the product and sum rule for a continuous variable yields
\[ p(x|I) = \int_{\Theta} p(x, \theta|I) d\theta = \int_{I} p(\theta|x, I)p(x|I) d\theta, \] (20)
with background information \( I \) for now specifying that ”\( x, \theta \) are continuous variables.”
C. Bayes’ Theorem

Bayes’ theorem can be straightforwardly derived from the product rule. Consider again the joint probability \( p(A, B|Z) \) in Eq.(1); since the logical product of two propositions is commutative, it is equal to \( p(B, A|Z) \), and so by the product rule

\[
\]

Bayes’ theorem

A special case of Bayes’ theorem was published posthumously in 1763 in Philosophical Transactions of the Royal Society (Bayes, 1763) through the effort of Richard Price (Hooper, 2013), an amateur mathematician and a close friend of Reverend Thomas Bayes (1702-1761), two years after Bayes had died. While sorting through Bayes’ unpublished mathematical papers, Price recognized the importance of an essay by Bayes giving a solution to an inverse probability problem in moving mathematically from observations of the natural world inversely back to its ultimate cause (McGrayne, 2011). The general mathematical form of the theorem is attributable to Laplace (1812), who was also the first to apply Bayes’ theorem to astronomy, earth science and social sciences.

III. BAYESIAN INVERSION EXAMPLES

An example in seismic/acoustics research is the investigation of earthquakes in California, based on a limited number of globally deployed seismic sensors. Must an ensemble of repeatable, independent devastating earthquakes at the same location occur in order to infer the location of their epicenter?

To introduce model-based Bayesian analysis in acoustic studies, consider a data analysis task common not only in acoustic investigations but in many scientific and engineering fields. This example begins with an acoustic measurement in a room, which records a discrete dataset expressed as the sequence \( D = [d_1, d_2, \ldots, d_K] \). Figure 1 illustrates the data, which consist of a finite number (K) of observation points. These data represent a sound energy decay process in an enclosed space. Based on visual inspection of Fig. 1, an experienced architectural acoustician will formulate a hypothesis that this set of data points probably represents an exponential decay. This hypothesis can be formulated into an analytical function of time \( t \), specifically a parametric model,

\[
H(\theta) = \theta_0 + \theta_1 e^{-\theta_2 t}.
\]

This model contains a set of parameters \( \theta = [\theta_0, \theta_1, \theta_2] \), in which \( \theta_0 \) represents background noise, \( \theta_1 \) represents the initial amplitude, and \( \theta_2 \) is the decay constant. The aim is to estimate this set of parameters \( \theta \) so that the modeled curve (solid line) is consistent with the data points (black dots). The data analysis task is to estimate the relevant parameters \( \theta \) contained in the model \( H(\theta) \), particularly the decay coefficient \( \theta_2 \), from the experimental observations \( D \). This is known as an inverse problem.

![FIG. 1. (Color online) Comparison between experimental data for a sound energy decay process and data predicted by a parametric model. The sequence \( D = [d_1, d_2, \ldots, d_K] \) expresses experimental data values collected in an enclosure, plotted on a logarithmic scale. A parametric model function can be specified by a single exponential decay function \( H(\theta) \) having three parameters \( \theta = [\theta_0, \theta_1, \theta_2] \). When the three parameters \( \theta \) are estimated well, the model prediction yields the model curve shown as a solid line, whereas poorly estimated parameters yield the mismatched model curves shown by dotted lines.](image)

To highlight the practical importance of Bayes’ theorem in the context of data analysis and model-based inference, the propositions in Eq.(21) can relate to concrete experimental data and the parameters of interest. Again, the sound energy decay example above is helpful in relating more general problems to concrete data analysis tasks often encountered in acoustics. Proposition \( A \) can be stated as ‘the experimental data \( D \) took the values stated’, namely \( A = D \), and Proposition \( B \) represents ‘the set of parameters takes certain values \( \theta’ \), so that \( B = \theta’ \); the experimenter wishes to estimate the actual values of these parameters \( \theta \) from the experimental data \( D \). In addition, there is an important proposition, reflecting the relevant background information \( I \) known to the experimenter, of the form

‘an appropriate set of parameters \( \theta \) will generate a hypothesized dataset via a well-established prediction model \( H(\text{hypothesis}) \) which approximates the experimental data well.’

Given this background information \( I \), substitution of \( A = D, B = \theta \), and \( Z = I \) into Eq (21) yields Bayes’ theorem for this data analysis task as

\[
p(\theta|D, I) = \frac{p(D|\theta, I) p(\theta|I)}{p(D|I)}
\]
supposing that \( p(\mathbf{D}|I) \neq 0 \). Bayes’ theorem represents the principle of inverse probability (Jeffreys 1965). The significance of Bayes’ theorem, as recognized by R. Price, is that an initial implication of \( \theta \), based on the background information \( I \) expressed by \( p(\theta|I) \), will be updated by new information, \( p(\mathbf{D}|\theta, I) \), about the probable cause of those items of data. The new information comprising the data \( \mathbf{D} \) comes from experiments.

In the room-acoustic example mentioned above, the quantity \( p(\theta|I) \) is a probability representing the initial implication of the parameter values from the background information before taking into account the experimental data \( \mathbf{D} \). It is consequently referred to as the prior probability for \( \theta \), or \( \text{(prior for short)} \).

The term \( p(\mathbf{D}|\theta, I) \) reads “the probability of the data \( \mathbf{D} \) given the parameter \( \theta \) and the background information \( I \), and represents the strength of implication that the measured data \( \mathbf{D} \) would have been generated for a given value of \( \theta \). It represents the probability of observing the dataset \( \mathbf{D} \) if the parameters take any particular set of values \( \theta \). It is often called the likelihood function (‘likelihood for short’). This likelihood represents the probability of getting the measured data \( \mathbf{D} \) supposing that the model \( H(\theta) \) holds with values of its defining parameters \( \theta \).

The term \( p(\theta|\mathbf{D}, I) \) represents the probability of the parameter values \( \theta \) after taking the data values \( \mathbf{D} \) into account. It is therefore referred to as the posterior probability, also known as the posterior predictive distribution. The quantity \( p(\mathbf{D}|I) \) in the denominator on the right-hand side of Eq. (23) represents the probability that the observed data occur no matter what the values of the parameters. It can be interpreted as a normalization factor ensuring that the posterior probability for the parameters integrates to unity.

Acousticians who face data analysis tasks from experimental observations are typically challenged to estimate a set of parameters encapsulated in the model \( H(\theta) \), also called the hypothesis. Figure 2 illustrates this task schematically. A model \( H(\theta) \) is specified by a functional form containing a set of parameters \( \theta \) whose values are not known in advance. The task is to go backwards: given the experimental data \( \mathbf{D} \) and the well-established model form containing the parameters \( \theta \), estimate the unknown (hopefully optimal) set of parameters.

Figure 3 illustrates Bayes’ theorem as shown in Eq. (23). The posterior probability, up to a normalization constant \( p(\mathbf{D}|I) \), arises by updating the prior probability \( p(\theta|I) \). Once the experimental data \( \mathbf{D} \) become available, the likelihood function \( p(\mathbf{D}|\theta, I) \) represents a multiplicative factor that updates the prior probability so as to transform it into the posterior probability (up to a normalization constant). Bayes’ theorem represents how one’s initial assignment is updated in the light of the data. This corresponds exactly to the process of scientific exploration: acoustical scientists seek to gain new knowledge from acoustic experiments. The prior knowledge in many acoustics fields is the fruit of long development, leading to well-understood hypotheses—models. These models, as part of the prior knowledge (Candy, 2016) are typically based on generations of learning/education.

In Eq. (23), Bayes’ theorem requires two probabilities in the calculation of the posterior probability up to a normalization constant. These two are the prior probability, \( p(\theta|I) \), and the likelihood function, \( p(\mathbf{D}|\theta, I) \). Use of the prior probability in the data analysis was an element of controversy between the Bayesian and frequentist schools. Frequentists have criticized the ‘subjective’
The assignment of prior probability should proceed in a noninformative manner, since different prior assignments lead to different posterior distributions according to Bayes’ theorem [see Fig. 3 and Cowan (2007)]. Figure 3 (a) shows that, if the prior probability is sharply peaked in parameter space, the likelihood function multiplied by the prior probability may give rise to a posterior probability peaked at a significantly different position than either the prior or the likelihood. Sharply peaked prior probability density functions encode a strong implication that the parameter values fall in certain (narrow) ranges in parameter space. Assignment of the prior probability in this way implies injection of information that may differ from person to person into the data analysis. In the case of Fig. 3 (a), the parameters are already known accurately, in which case there is almost no need for experiment, unless the data can sharpen the peak significantly.

The use of prior probability is actually a strength of Bayesian analysis, not a weakness, for in the extreme case that the experimenter knows the parameters exactly, the prior probability is all heaped at a single value and the prior is zero elsewhere as shown in an example in Fig. 4. Bayes’ theorem shows immediately that this feature carries through to the posterior, in accord with intuition but not with frequentist methods of data analysis.

The assignment of prior probability should proceed from the prior information. If it is not known how to assign a prior distribution from the prior information, as is often the case when limited prior knowledge about the parameter values is available, or no preference is intended to be incorporated into the data analysis, then a broad prior probability density can safely be assigned, as in Fig. 3(b). Bayesian analysis often involves a broad or flat prior, representing maximal non-commitment to any particular value. Below, the model-based Bayesian analysis proceeds by assigning such a noninformative prior probability.

IV. MAXIMUM ENTROPY PRIOR PROBABILITY

Bayes’ theorem takes account of both the prior information about the process (parameters) under investigation, \( p(\theta | I) \), and the experimental data observed in the experiment, through the likelihood function, \( p(D | \theta, I) \). The prior probability encodes the experimenter’s initial implication of the possible values of parameters, and the likelihood function encodes how likely are the data given particular values of the parameters. The parameters are coefficients in a functional form which specifies the model. The prior probability \( p(\theta | I) \) must be assigned based on available information prior to the data analysis in order to apply Bayes’ theorem.

Berger (2006) rebutted a common criticism of the Bayesian school arising from the supposedly subjective use of prior probabilities. The first Bayesians, including Bayes (1763) and Laplace (1812), conducted probabilistic analysis using constant prior probability for unknown parameters. A fundamental technique that is often used in Bayesian analysis relies on whatever is already known (prior knowledge) about a probability distribution in order to assign it. This technique developed in recent decades (Jaynes, 1968) encodes whatever is already known about the probability distribution in mathematical form. This is the maximum entropy method, and it generates a so-called maximum entropy prior probability. Jaynes (1968) applied a continuum version of the Shannon (1948) information-theoretic entropy, which is a measure of uncertainty, to encode the available information into a probability assignment. To assign a probability distribution \( p(x) \), the information entropy of this probability \( S[p(x)] \), given by:

\[
S[p(x)] = -\int p(x) \ln \left[ \frac{p(x)}{m(x)} \right] dx \quad (24)
\]

needs to be examined, with

\[
\int m(x) dx = 1, \quad (25)
\]

where \( m(x) \) is a Lebesgue measure which ensures that the entropy remains invariant under change of variables (Gregory, 2005). The probability \( p(x) \) is assigned by maximization of the entropy in Eq. (24), subject to whatever is known about it as constraints on the maximization process. Knowledge directly about a probability distribution is in a different category from knowledge about samples drawn from it; moments of the distribution provide a good example. In the Bayesian literature (Gregory, 2005; Jaynes, 1968) this technique is termed the principle of maximum entropy, and it provides a consistent and rigorous way to encode such ‘testable’ information into a unique probability distribution. The principle of maximum entropy assigns the probability distribution as non-committally as possible, while satisfying all known constraints on the distribution. The resulting
distribution is also guaranteed to be non-negative. The following sections demonstrate how to arrive at two common probability assignments using the principle of maximum entropy. The method of Lagrange multipliers is well suited to solve such a constrained maximization problem. Detailed derivations are given elsewhere (Gregory, 2005; Jaynes, 1968; Sivia and Skilling, 2006; Woodward, 1953).

A. Prior probability assignment

Assignment of the prior probability [in Eq. (23)] requires that no possible value of a parameter should be preferred to any other, except to the extent necessary to conform to any known constraints on the probability distribution. The following illustration involves a one-dimensional distribution, for simplicity. Normalization is a universal constraint such that the prior probability density \( p(x) \) integrates to unity:

\[
\int p(x) \, dx = 1. \tag{26}
\]

In the absence of further constraints, incorporating this constraint in Eq. (26) into the maximization of the entropy in Eq. (24) with respect to \( p(x) \) yields

\[
\frac{\partial}{\partial p} \left\{ -\int p(x) \ln \left( \frac{p(x)}{m(x)} \right) \, dx \right\} - \lambda \left( \int p(x) \, dx - 1 \right) = 0, \tag{27}
\]

where \( \lambda \) is the undetermined Lagrange multiplier. Solution of Eq. (27) leads to

\[
-\ln \left[ \frac{p(x)}{m(x)} \right] - 1 - \lambda = 0, \quad \text{or} \quad p(x) = m(x) e^{-(1+\lambda)}. \tag{28}
\]

Upon using the only constraint in Eq. (26),

\[
\int m(x) e^{-(1+\lambda)} \, dx = 1 = e^{-(1+\lambda)} \int m(x) \, dx, \tag{29}
\]

and, since the measure is normalized as in Eq. (25), the (undetermined) Lagrange multiplier becomes \( \lambda = -1 \), and

\[
p(x) = m(x). \tag{30}
\]

In other words, subject only to the constraint of normalization, the probability assignment is equal to the measure.

Jaynes (1968) used group theory to show how the measure may be assigned in the case that nothing known to the experimenter distinguishes one value of \( x \) from another; for example, if \( x \) describes a location, and nothing is known about where. In that case, if the coordinate system describing \( x \) were to be shifted by an amount \( x_0 \), the state of knowledge about the location should not change, so that

\[
p(x) \, dx \approx p(x + x_0) \, d(x + x_0), \tag{31}
\]

For this to be true, the measure \( m(x) \) must be constant (since \( m(x) = m(x + x_0) \) for all \( x_0 \)), and for a quantity located on an interval \( (a, b) \), the probability assignment encoding this state of knowledge is then the (bounded) uniform distribution

\[
p(x) = \begin{cases} 1/(b - a) & \text{for } a \leq x \leq b \\ 0 & \text{otherwise}, \end{cases} \tag{32}
\]

so as to fulfill the normalization constraint in Eq. (26) within the finite-valued range \( (a, b) \)

\[
\int_{a}^{b} p(x) \, dx = 1. \tag{33}
\]

If the parameter in question represents not a location but a magnitude or size, then relative changes in the scale of the quantity, rather than in its position, should be invariant; it is a so-called scale parameter (Jeffreys, 1946). Mathematically, scaling the quantity by an arbitrary factor \( c \) should not change the state of knowledge on the parameter, leading to (Sivia and Skilling, 2006)

\[
p(x) \, dx \approx p(c \, x) \, d(c \, x). \tag{34}
\]

This leads to a functional equation having solution

\[
p(x) = \frac{1}{x}, \tag{35}
\]

also known as the Jeffreys’ prior, and equivalent to a uniform distribution on a logarithmic domain. In this form the Jeffreys’ prior represents an ‘improper’ probability, since it is not normalizable over the entire positive domain. Since the posterior probability is given, via Bayes’ theorem, by multiplying the prior probability by the likelihood and then normalizing the result, any factor by which the prior is multiplied cancels out. As a result, the use of an improper prior is harmless provided that the resulting posterior is normalizable. (If it is not, the experiment was not very well designed!)

In summary, encoding the state of ‘no knowledge’ about the value of a parameter, but with use of its physical meaning, gives the measure for it, via invariance arguments; then the prior probability for it, in the absence of testable information, is equal to the (normalized) measure according to the principle of maximum entropy. The idea of maximum entropy is to distribute the prior probability as non-committally as possible (Gregory, 2005; Jaynes, 1968). The result for a location parameter is a uniform prior, and for a scale parameter it is the logarithmic prior.

This maximum entropy assignment does not involve any consideration that the prior probability distribution shall be assigned according to the results of any prior
‘random’ experiment. Nevertheless, if testable information about the distribution is declared to be known based on such results, it can be incorporated (Jaynes, 1968). As mentioned, this enables wider application of Bayesian analysis to problems where the prior probabilities have no reasonable frequency interpretation. Bayesian probabilistic analysis, equipped with the principle of the maximum entropy, represents a quantitative tool capable of dealing with ignorance in the sense of complete lack of knowledge. The least informative prior probability is consistent with the principle of indifference (Keynes, 1921), or the principle of insufficient reasoning, as historically applied to Bayesian inference by its pioneers Bayes (1763), Richard Price (Hooper, 2013), and Laplace (1812), among others. The principle of indifference was eventually given a deeper logical justification by Jaynes (1968), based on the principle of transformation groups followed by the principle of maximum entropy. The methods of probability theory constitute consistent reasoning in scenarios where insufficient information for certainty is available, namely, inductive reasoning. Thus, probability is always the appropriate tool for dealing with uncertainty, lack of scientific data, and ignorance.

B. Likelihood function assignment

The likelihood function represents the probability of getting the measured data assuming the values of the parameters. Often the mean and standard deviation of the data measurements are of physical relevance to the inference problem for the parameters. For instance, the mean of the distribution might actually be the value of a parameter which is measured directly, and the standard deviation might relate to a physically significant noise process in the measuring apparatus. In that case the likelihood for a single datum is given by maximizing the entropy subject to (unknown) values of the mean and standard deviation. The experimenters know in advance that the model [such as Eq. (22) in the example mentioned above] is capable of representing the data well, so that the variance of the residual errors should feature finite values, expressed mathematically as

$$\int (x - \mu)^2 p(x) dx = \sigma^2 < \infty,$$

where \(\mu\) is the mean and \(\sigma^2\) is the variance.

To assign the likelihood function, the data are supposed to be defined on a space having uniform measure, for simplicity. (They might be a location measurement, for instance.) The likelihood function \(p(x)\) is given by maximizing the entropy (Gregory, 2005; Woodward, 1953)

$$- \int p(x) \ln p(x) \, dx,$$

subject to the values of the mean and standard deviation (and normalization). With appropriate Lagrange multipliers, the maximization condition is

$$\frac{\partial}{\partial p} \left\{ - \int p(x) \ln p(x) \, dx - \lambda \left[ \int p(x) \, dx - 1 \right] \right\} = 0,$$

subject to two constraints:

$$\lambda_1 \left[ \int (x - \mu)^2 p(x) \, dx - \sigma^2 \right] = 0, \quad \lambda_2 \left[ \int p(x) \, dx - 1 \right] = 0,$$

which yields

$$- \ln p(x) - 1 - \lambda - \lambda_1 (x - \mu) = 0$$

or

$$p(x) = e^{-\lambda_0} e^{-\lambda_1 (x - \mu)}.$$

with \(\lambda_0 = 1 + \lambda\). The Lagrange multipliers \(\lambda_0\) and \(\lambda_1\) can be determined and eliminated in favor of \(\mu\) and \(\sigma\) by substituting this expression into the constraints, Eq. (26) and Eq. (36), giving (see Appendix A)

$$p(x) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(x - \mu)^2}{2\sigma^2}}.$$  

It is straightforward to generalize this for discrete variables:

$$p_k = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(x_k - \mu)^2}{2\sigma^2}}.$$

where \(x_k\) can take only one of a number (1 \(\leq k \leq K\)) of specific discrete values.

Upon taking into account the finite variance of the errors about the mean in the maximum entropy procedure on a continuous space of uniform measure, the result is the Gaussian or normal distribution (Gregory, 2005; Jaynes, 1968).

The following discussion specifies this assignment in terms of the data and the model parameters. The residual errors are taken to have a zero mean by a transformation of variables provided that the mean is tied to a model parameter of physical relevance. Then, for residual errors \(\epsilon_k = d_k - \hat{h}_k\) for the experimental data \(\mathcal{D} = [d_1, d_2, \ldots, d_K]\) of \(K\) points, the principle of maximum-entropy assigns, as in Eq. (42), the probability of individual residual error \(\epsilon_k\) given the model \(H\), the model parameters \(\theta\), an undetermined error standard deviation \(\sigma_k\) and the background information \(I\), as

$$p(\epsilon_k | \theta, H, \sigma_k, I, \mu_k = 0) = \frac{1}{\sqrt{2\pi} \sigma_k} e^{-\frac{\epsilon_k^2}{2\sigma_k^2}}.$$  

In formulating the likelihood function, which is the probability that a particular dataset would be observed given the model parameters, the probability of each datum is equal to the probability of the residual error at each datum \(p(d_k | \theta, H, \sigma_k) = p(\epsilon_k | \theta, H, \sigma_k)\). The overall
The likelihood function is then equal to the joint probability of all data points,

\[
p(D | \sigma, \theta, H, I) = p(\epsilon_1, \epsilon_2, \ldots, \epsilon_K | \sigma, \theta, H, I, I) \]

(44)

The individual measurements \(d_k\) are logically independent of each other: given \(\sigma, \theta, H, I\), the value of one observation indicates nothing about any other. This independence is justified by the principle of maximum entropy as well, since any dependence of one value from any other would lower the entropy (Jaynes, 1968). Consequently, substituting Eq. (43) into Eq. (44) and ‘chaining’ via repeated application of the product rule (see Eq. (4) in Sec. II A), lead to

\[
p(D | \sigma_k, \theta, H, I) = \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi} \sigma_k} \exp \left( -\frac{\epsilon_k^2}{2\sigma_k^2} \right),
\]

(45)

where different values \(\sigma_k\) of the standard deviation for the \(k\)th measurement are allowed for. On the other hand, in most experiments the noise is the same for each measurement, and

\[
p(D | \sigma, \theta, H, I) = \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi} \sigma} \exp \left( -\frac{\epsilon_k^2}{2\sigma^2} \right),
\]

(46)

so that

\[
p(D | \sigma, \theta, H, I) = \left( \sigma \sqrt{2\pi} \right)^{-K} \exp \left( -\frac{Q}{\sigma^2} \right),
\]

(47)

where

\[
Q = \frac{1}{2} \sum_{k=1}^{K} \epsilon_k^2.
\]

(48)

In summary, this assignment of the likelihood function incorporates no more than the available information; the experimenter knows a priori that the data model \(H(\theta)\) describes the experimental data \(D\) well such that the residual errors between the model and the data are expressed in a finite error variance \(\sigma^2\). The resulting Gaussian likelihood distribution in Eq. (42) is then the consequence of the principle of the maximum entropy which is fundamentally different from assuming probability of the residual errors being Gaussian. The end result in Eq. (47) is due also to the residual errors being logically independent of each other, given the conditioning information which is also according to the principle of the maximum entropy. The assigned distribution ensures no bias for which there is no prior evidence.

**C. Student’s t-distribution**

In the above formulation, the likelihood function encodes the extent of implication of the experimental data \(D\), by the values \(\theta\) of the parameters in the model \(H(\theta)\). However, the likelihood also contains a hyperparameter, the error standard deviation \(\sigma\), stemming from the maximum entropy assignment discussed in the previous section. For some applications, this is often a nuisance parameter of no interest to the analysis. Marginalization (as introduced in Sec. II B) may be applied to the joint probability of the \(D\) and the standard deviation \(\sigma\) (Bretthorst, 1988-1990). Recent acoustic applications of Bayesian analysis also utilized a similar marginalization process to remove nuisance parameters (Jasa and Xiang, 2009; Xiang et al., 2011). The following derivation relies heavily on these references, particularly (Bretthorst, 1988-1990).

To perform the marginalization [as in Eq. (20)], the joint probability \(p(D, \sigma | \theta, H, I)\) of data \(D\) and the standard deviation \(\sigma\), namely the product of the conditional likelihood in Eq. (47) and the marginal distribution \(p(\sigma | \theta, H, I)\) of \(\sigma\), is integrated over all possible values of \(\sigma\), resulting in a likelihood free from the error standard deviation:

\[
p(D | \theta, H, I) = \int p(D, \sigma | \theta, H, I) \, d\sigma
= \int p(D | \sigma, \theta, H, I) \, p(\sigma | \theta, H, I) \, d\sigma.
\]

(49)

The marginal distribution \(p(\sigma | \theta, H, I)\) is regarded as a prior probability for the standard deviation \(\sigma\), which is a scale parameter discussed in Section IV A above, so that the principle of maximum entropy assigns the Jeffreys prior:

\[
p(\sigma | \theta, H, I) = \frac{1}{\sigma},
\]

(50)

Substituting Eq. (47) and Eq. (50) in Eq. (49) and integrating over all possible values for \(\sigma\), result

\[
p(D | \theta, H, I) = \int_0^\infty \frac{\sigma^{-K}}{\sigma} (2\pi)^{-K/2} \exp \left( -\frac{Q}{\sigma^2} \right) \, d\sigma.
\]

(51)

This integral can be performed (see Appendix B) to give a marginalized likelihood function taking the form of Student’s t-distribution:

\[
p(D | \theta, H, I) = \frac{\Gamma(K/2)}{\Gamma(M/2)} \left( \frac{\sigma^2}{\sum_{k=1}^{K} \epsilon_k^2} \right)^{-(K+M)/2},
\]

(52)

where \(\Gamma(\cdot)\) is the standard Gamma function (Abramowitz and Stegun, 1964), and the residuals \(\epsilon_k = d_k - h_k(\theta)\).

An extended Bayesian parameter estimation scheme is a sequential Bayesian process (Candy, 2015; Carrierie and Hermand, 2012; Özdamar et al., 1990), such as the work recently reported on underwater applications (Yardim et al., 2011), in which the parameters encapsulated in the model evolve in time and space, with the data arriving consecutively.

**V. TWO LEVELS OF BAYESIAN INFERENCE**

In many acoustic experiments there are a finite number of models (hypotheses) \(H_1, H_2, \ldots, H_M\) that compete against one other to explain the data. In the room-acoustic example above, \(H_1\) is specified in Eq. (22) as containing one exponential decay term and one noise term, but the same data in Fig. 1 may alternatively be
Bayesian data analysis applied to solving parameter estimation problems, as in the example above, is referred to as the first level of inference, while solving model selection problems, as in the example above, is referred to as the second level of inference. Bayesian data analysis is capable of performing either task. In Bayesian model selection, the likelihood of models, is referred to as the marginal likelihood of the data or the Bayesian evidence [as further explained by Eq. (59) in Sec. V C], while \( p(H_S | \mathcal{D}, I) \) is the posterior probability of the model \( H_S \), given the data.

Using Eq. (53), model comparison between two different models \( H_i \) and \( H_j \) evaluates the so-called Bayes’ factor, \( K_{i,j} \) (Kass and Raftery, 1995),

\[
K_{i,j} = \frac{p(H_i | \mathcal{D}, I)}{p(H_j | \mathcal{D}, I)} = \frac{p(\mathcal{D} | H_i, I)}{p(H_i | I)} \frac{p(H_j | I)}{p(\mathcal{D} | H_j, I)},
\]

where \( 1 \leq i, j \leq M; i \neq j \). In the right-hand side of the Bayes’ factor, the second fraction, termed the prior ratio, represents one’s prior knowledge as to how strongly model \( H_i \) is preferred over \( H_j \) before considering the data \( \mathcal{D} \). Often one is unable to incorporate any prior preference to either of the models, and the principle of maximum entropy as discussed previously assigns equal prior probability

\[
p(H_j | I) = \frac{1}{M}, \quad 1 \leq j \leq M
\]

to each of \( M \) models. In this case, the Bayes’ factor for model comparison between two different models \( H_i \) and \( H_j \) relies solely on the posterior ratio between models,

\[
K_{i,j} = \frac{p(\mathcal{D} | H_i, I)}{p(\mathcal{D} | H_j, I)}, \quad 1 \leq i, j \leq M; \ i \neq j,
\]

which is consequently equal to the marginal likelihood ratio when the model prior probabilities are uniform. The marginal likelihood \( p(\mathcal{D} | H_i, I) \) therefore plays a crucial role in Bayesian model selection. For computational convenience, the Bayes factor is determined using a logarithmic scale with the unit ‘decibans’ (Jeffreys, 1961),

\[
L_{ij} = 10 \log_{10}(B_{ij}) = 10 \log_{10}(Z_i) - 10 \log_{10}(Z_j), \quad [\text{decibans}],
\]

with simplified notations for the Bayesian evidence, \( Z_i = p(\mathcal{D} | H_i, I) \), and \( Z_j = p(\mathcal{D} | H_j, I) \). This enables the evidence values (marginal likelihoods) for two models to be compared quantitatively against one another. Among a
finite set of competing models, the highest positive Bayes factor, $L_{ij}$, implies that the data prefer model $H_i$ over $H_j$ the most. Therefore, the Bayes factor is also applied to select a finite number of models under consideration.

Although more complex models may fit the data better, they pay a penalty by strewing some of the prior probability for their parameters where the data subsequently indicate that those parameters are extremely unlikely to be. There is therefore a trade-off between goodness of fit and simplicity of model. This can be seen as a quantitative generalization of the qualitative principle known as Occam’s Razor, which is to prefer the simpler theory that fits the facts well; there is a trade-off between simplicity of model and closeness of fit, and Occam’s Razor makes this trade-off quantitative. (Garrett, 1991; Jefferys and Berger, 1992; MacKay, 2003).

**B. Parameter estimation: the first level of inference**

Once a model has been chosen, based on the experimental data, denote it as model $H_S$; then the Bayesian framework is used with this selected model to estimate its parameters $\theta_S$. For this purpose the model is specified with its parameters on view as $H_S(\theta_S)$; for instance, model $H_S(\theta_1)$ in Eq. (22) contains one exponential decay term with its parameters collectively denoted as $\theta_1 = [\theta_0, \theta_1, \theta_2]$. Bayes’ theorem is now applied as before in order to estimate the parameters $\theta_S$, given the data $D$ and the model $H_S$. The quantity $p(D|I)$ in Eq. (23), now becomes $p(D|H_S)$, the probability of the data $D$ given the model $H_S$. In this context, the background information $I$ now includes the fact that a specific model $H_S$ is selected or given. Bayes’ theorem for the parameter estimation problem is written as:

$$p(\theta | D, H) = \frac{p(D | \theta, H) p(\theta | H)}{p(D | H)} ,$$

where the subscript $S$ and background information $I$ have been dropped for simplicity. Bayes’ theorem represents here how one’s prior knowledge about parameters $\theta$, given the specific model $H$ encoded in $p(\theta | H)$, is updated by incorporating the data $D$ through the likelihood, $p(D | \theta, H)$.

The prior $p(\theta | H)$ encodes all of the knowledge about the parameters before the data are incorporated, and it is denoted by $\Pi(\theta) \equiv p(\theta | H)$ in the following discussion for simplicity. Once the data have been observed or measured, the likelihood $p(D | \theta, H)$ incorporates the data for updating the prior probability for the parameters. To emphasize that the data are settled, once observed, and that the likelihood is a function of the parameter values, it is denoted by $\mathcal{L}(\theta) \equiv p(D | \theta, H)$. The posterior probability for the parameters, $p(\theta | D, H)$ encodes the updated knowledge of the parameters in the light of the data.

**C. Two levels of inference in one unified framework**

The posterior $p(\theta | D, H)$ must be normalized and integrate to unity over the entire parameter space [see Eq. (26)]. With the notational changes of the previous paragraph, this normalization constraint involves integrating both sides of Eq. (58) over the entire parameter space:

$$p(D | H) = \int_{\theta} p(D | \theta, H) p(\theta | H) d\theta$$

$$\equiv Z = \int_{\theta} \mathcal{L}(\theta) \Pi(\theta) d\theta ,$$

where the integral marginalizes out $\theta$ which appear in the likelihood $p(D | \theta, H)$, by assigning the prior $p(\theta | H)$ for the parameters $\theta$ [compare Eq. (20)].

At the first level of inference, namely parameter estimation, the quantity $p(D | H)$ plays the role of a normalization constant in Eq. (58). Furthermore, it is identical to the Bayesian evidence in Eq. (53), where the simplified notation $Z \equiv p(D | H)$, is used. This quantity is central to model selection – the second level of inference. Rearrangement of the terms of Bayes’ theorem in Eq. (58), with the notational changes, gives

$$p(\theta | D, H) \times Z = \mathcal{L}(\theta) \times \Pi(\theta) ,$$

$$\text{posterior} \times \text{evidence} = \text{likelihood} \times \text{prior} ,$$

which shows the logical relationship among the quantities of Bayesian inference (Skilling, 2006): the likelihood function $\mathcal{L}(\theta)$ and the prior probability $\Pi(\theta)$ are the inputs, while the posterior probability $p(\theta | D, H)$ and the evidence $Z$ are the outputs of Bayesian inference. The evidence $Z$ is then needed in the second level of inference, model selection, while the posterior probability is the output for the first level of inference, parameter estimation. Bayesian evidence automatically encapsulates the principle of parsimony, and quantitatively embodies Occam’s razor (Garrett, 1991; Jefferys and Berger, 1992; Knuth et al., 2015; MacKay, 2003); when two competing theories explain the data equally accurately, the simpler one is preferred.

Figure 6 displays the quantitative embodiment of Occam’s Razor, in which a sharply peaked posterior distribution at position $\theta_{\text{MAP}}$ in parameter space is assumed relative to a broad prior distribution; $\theta_{\text{MAP}}$ stands for the parameter set which maximizes the posterior probability, so-called maximum a posterior (MAP). Equation (59) can be simplified (Bishop, 2006; Knuth et al., 2015) as

$$p(D) = \int_{\theta} \mathcal{L}(\theta) \Pi(\theta) d\theta \approx \mathcal{L}(\theta_{\text{MAP}}) \frac{\Delta w_{\text{post}}}{\Delta w_{\text{prior}}},$$

where the prior distribution is taken as flat with width $\Delta w_{\text{prior}}$, so that $\Pi(\theta) = 1/\Delta w_{\text{prior}}$. The integral, which is essentially the area under the likelihood curve, is approximated by the peak-value $\mathcal{L}(\theta_{\text{MAP}})$ multiplied by the width $\Delta w_{\text{post}}$ of the posterior, since the prior is flat.
so that the posterior and the likelihood peak at the same location in parameter space; the conditioning ‘given the model $H$’ is dropped for notational simplicity. The logarithm of Eq. (61) yields

$$\ln p(\mathcal{D}) \approx \ln \hat{L}(\theta_{\text{MAP}}) - \ln \left( \frac{\Delta w_{\text{prior}}}{\Delta w_{\text{post}}} \right).$$  \hspace{1cm} (62)

The first term on the right-hand side, the logarithm of the peak-value of the likelihood, represents a goodness of model fit to the data. The second term represents the penalty on over-parameterized models, since over-parameterized models give a larger value of the first term, but more parameters cause the probability to be distributed over a larger space. Equation (62) represents a quantitative trade-off between the goodness of model fit and simplicity of the model.

A similar approximation used for model selection that has been used recently in some acoustical applications (Dettmer et al., 2011; Steininger et al., 2014; Xiang et al., 2011) is the Bayesian information criteria (BIC) (Schwarz, 1978) and its variation, the deviance information criteria (DIC) (Spiegelhalter et al., 2002). The BIC asymptotically approximates the Bayesian evidence (Schwarz, 1978) under the assumption that a (multi-dimensional) Gaussian distribution can approximate the posterior probability distribution around the global extremum of the likelihood (Stoica and Selen, 2004). Supposing that the dataset involved in the analysis is large (large $K$), the (base ten logarithmic) inverse BIC for ranking a set of decay models $H_1, H_2, H_3, \ldots$ is given (Xiang et al., 2011) by

$$\text{IBIC} \approx 20 \cdot \log_{10}[\hat{L}(\theta_{\text{MAP}})] - 10 \cdot N_{\theta} \cdot \log_{10}(K), \text{ [decibans]},$$ \hspace{1cm} (63)

where $N_{\theta}$ is the number of parameters involved in model $H$, and $K$ is total number of data points involved. The quantity $\hat{L}(\theta_{\text{MAP}})$ is the peak value of the likelihood, at a location in the parameter space specified as $\theta_{\text{MAP}}$.

As in Eq. (62), the first term in Eq. (63) represents the goodness of the model fit to the data, and the second term represents a penalty on over-parameterized models, the DIC (Spiegelhalter et al., 2002; Steininger et al., 2014) ranks the competing models similarly, yet with a slightly different penalty term. In accordance with the (natural logarithmic) simplified evidence in Eq. (62), the present paper prefers the inverse BIC definition [Eq. (63)] which differs from that in Schwarz (1978) in sign and in the base of logarithms. The IBIC as defined in Eq. (63) is then in units of decibans, as advocated by Jeffreys (1961).

To calculate the evidence in Eq. (39), the product of the likelihood and the prior probability must be integrated over the entire parameter space. Estimation of the evidence therefore requires substantial computational effort. Often it is necessary to use numerical sampling methods based on Markov chain Monte Carlo (MCMC) approaches such as a trans-dimensional approach using reversible jump MCMC (Dettmer et al., 2010; Green, 1995), and nested sampling (Jasa and Xiang, 2005-2012; Skilling, 2004). In fact, Eq. (60) indicates implicitly that model selection and parameter estimation can both be accomplished within a unified Bayesian framework. Once sufficient exploration has been performed to estimate the evidence, the explored likelihood function multiplied by the prior probability can contribute to the normalized posterior probability, since the evidence is also estimated. From estimates of the posterior distributions it is straightforward to estimate the mean values of the relevant parameters, and also their uncertainties in terms of associated individual variances, and inter-relationships between the parameters of interest (Xiang et al., 2011).

VI. APPLICATIONS

Acoustics has increasingly been seen both levels of model-based Bayesian inference applications in porous media analysis, (Fackler et al., 2018; Jeong et al., 2017; Roncen et al., 2018), in sound absorber design (Xiang et al., 2018), in acoustic surface impedance estimation (Bockman et al., 2015), in sound source localization, (Bush and Xiang, 2018; Xiang and Landschoot, 2019), community annoyance (Lee et al., 2020), in room-acoustic analysis (Balint et al., 2019; Beaton and Xiang, 2017; Jasa and Xiang, 2005; Sü Gül et al., 2016), in musical acoustics (Davy et al., 2006), in ultrasound spectroscopy (Bales et al., 2018) and imaging (Martiartu et al., 2019), in photoacoustic tomography (Tick et al., 2018), in geoacoustic inversion (Bonnel et al., 2013; Michalopoulou and Amsri, 2018), in ecosystem monitoring (Ballard et al., 2020), in audiogram analysis, (Özdamar et al., 1990; Schlittenlacher et al., 2018), acoustic compressive and sparse sensing (Gerstoft et al., 2018; Namuru et al., 2018), and many other acoustical applications. This paper does not aim to cover all of the tools and applications, rather briefly sets out some recent examples, including applications in room acoustics, communication acoustics, noise control and physical acoustics (Xiang and Fackler, 2015). These applications all have in common the existence of applications...
well established, or at least well understood models for predicting the processes under investigation, yet only a set of experimental data is available to study the process.

A. Modal, Decay, and Direction of Arrival Analysis

This section discusses room-acoustic modal and decay analysis, and the direction of arrival estimation using a microphone array. They are in common that the prediction models used are in form of generalized linear models (Brethorst, 1988; Ó Ruanaidh and Fitzgerald, 2016), consisting in essence of a sum of simple functions.

1. Room-Acoustic Modal Analysis

This application explores a method to identify multiple decaying modes in measured room impulse responses from existing spaces. Beaton and Xiang (2017) reported a method employing Bayesian framework working in the time domain to identify numerous decaying modes in a room impulse response. Experimental measurement is carried out at one strategic location in the room under investigation using one monophonic microphone. A model describing the room impulse response \( h(t_k) \), employed in this application is so-called Prony model (d. Prony, 1795)

\[
    h(\Theta_S, t_k) = \sum_{s=1}^{S} A_s \exp \left( -\frac{6.9t_k}{T_s} \right) \cos(2\pi f_s t_k + \phi_s), \quad (64)
\]

where \( A_s, T_s, f_s \) and \( \phi_s \) are modal parameters, amplitude, decay time, modal frequency, and phase, respectively. The model contains altogether \( S \) room modes, namely \( S \) sets of parameters \( \{A_s, T_s, f_s, \phi_s\} \) with \( s = 1, \ldots, S \). Figure 7 compares an experimentally measured room-impulse response with predicted data based on the model in eq. (64) when the number of modes, \( S \), and modal parameters were known or well estimated.

Two levels of Bayesian inference are suitable to this application, the model selection estimates the number of modes, \( S \), being present in the room impulse response, while the parameter estimation determines the relevant parameters upon the selected model (e.g. decay times, modal frequency and amplitude) of each mode.

2. Sound Energy Decay Analysis

Due to recent developments in concert hall design (Jaffe, 2005) and the high variations in chamber-based sound absorption measurement (Balint et al., 2019), there is an increasing interest in the analysis of sound energy decays consisting of multiple exponential decay rates. To meet the need of characterizing energy decays of potentially multiple decay processes, Balint et al. (2019); Xiang et al. (2011) applied Bayesian analysis to experimentally measured sound energy decay data. A fully parameterized Schroeder’s decay model (Xiang, 2017a; Xiang et al., 2011)

\[
    H_S(\Theta_S, t_k) = \theta_0 (t_k - t_K) + \sum_{s=1}^{S} \theta_{2s-1} \left[ \exp \left( -\frac{13.8}{T_s} t_k \right) - \exp \left( -\frac{13.8}{T_s} t_K \right) \right], \quad 0 \leq k \leq K \quad (65)
\]

has been found to be capable of characterization of multiple-slope decays beyond the single-slope and double-slope energy decays, with \( T_s = \theta_{2s} \) being the \( s \)th decay time (parameter), \( \Theta_S = \{\theta_0, \theta_1, \ldots, \theta_{2S+1}\} \) then includes all \( 2S + 1 \) parameters with \( S \) being the number of exponential decaying terms (slopes). Variable \( t_k \) represents a discrete time variable, value \( t_K \) is the upper limit of Schroeder’s integration, and term \( \theta_0 (t_k - t_K) \) is associated with background noise in the experimentally measured room impulse responses (Xiang, 2017a). Within the Bayesian framework, sound energy decays more complicated than single-slope and double-slope nature, such as triple-slope decays have been identified and characterized (Sü Güll et al., 2016; Xiang et al., 2011). Figure 8 illustrates the experimentally measured, sound energy decay functions, so-called Schroeder decay curve in a monumental mosque in comparison with the predicted curve based on the model in eq. (65) where \( S = 3 \) decay slopes are identified and the decay parameters are well estimated. (Sü Güll et al., 2016). Using the estimated decay parameters three exponentially decaying and the noise terms can be decomposed as depicted in Fig. 8.

3. Estimation of Multiple Direction of Arrivals

Estimating direction of arrival (DoA) of sound sources is an important acoustical problem often tackled using microphone arrays. Coprime linear microphone arrays represent an innovative sparse sensing technique which extends the frequency range of a given number of array elements by exceeding the spatial Nyquist limit. Whereas initial coprime array theory was derived based
on an operating frequency dictated by the specific co-prime spacing (Vaidyanathan and Pal, 2011; Xiang et al., 2015), recent investigations demonstrate advantages of broadband beamforming (Bush and Xiang, 2015). Parametric models describing this broadband behavior enable the use of model-based Bayesian inference for estimating not only source directions, but also number of sources present in the sound field, which is often unknown prior to estimation. The following section discusses the DoA of recently published by Bush and Xiang (2018) to demonstrate this line of Bayesian applications.

\[ H_{S}(\Theta_{S}, \theta) = A_0 + \sum_{s=1}^{S} A_s e^{-|\phi_s - \theta|/\delta_s} \]  

sufficiently predicts the beamforming data with \( \theta \) being azimuth angular variable. \( A_0 \) is a constant parameter for accounting noise floor and the three parameters per sound source include amplitude, \( A_s \), angle of arrival, \( \phi_s \), and beam width, \( \delta_s \) of each sound sources. The model parameter for \( S \) number of sound sources, \( \Theta_{S} = \{A_0, A_1, \ldots, A_S, \phi_1, \ldots, \phi_S, \delta_1, \ldots, \delta_S\} \), includes all the amplitude and angular parameters. Figure 9 illustrates directional response experimentally measured using a comprime linear microphone array of 16 elements, compared with the predicted one based on the model in eq. (66), where two simultaneous sound sources are in the data. Figure 9 (b) also illustrates the residual error function when the predictive model in eq. (66) fits the experimental data well. The residual errors fulfill the condition as expressed in eq. (36) when assigning the likelihood function in Sec. IV B.

The DoA model using a coprime microphone array expressed in Eq.(66), including those in Eqs.(64,65), represents a class of models; generalized linear models, consisting of a linear superposition of \( S \) number of nonlinear functions. Escolano et al. (2014)’s DoA estimation using two microphones and the room-acoustic modal analysis (Beaton and Xiang, 2017) and decay analysis (Xiang et al., 2011) all employ the generalized linear models, where the Bayesian model selection is applied to estimate the number \( S \) of the nonlinear functions in their models.

\[ H_{S}(\Phi_{S}, \Phi) = \sum_{s=1}^{S} A_s g_s(\Phi_{s}, \Phi), \]  

with \( A_s \) representing strength associated with \( s \)th sound source. \( \Phi_{S} = \{\theta_1, \ldots, \theta_S; \phi_1, \ldots, \phi_S\} \) are \( S \) number of sound source directions. Instead of basic nonlinear functions, a normalized sound energy function, \( g_s(\Phi_{s}, \Phi) \),

\[ g_s(\Phi_{s}, \Phi) = \frac{|g(\Phi_{s}, \Phi)|^2}{\max \left[ |g(\Phi_{s}, \Phi)|^2 \right]} \]  

(a) Broadband directional pattern in response to two noise sources (solid line). The two-sound source Laplace distribution function model with reasonable values of the model parameters is superimposed onto the experimental data (dashed line). (b) Errors between model and data are finite with mean zero.
exploits the completeness property of the spherical harmonics with a finite spherical harmonic order is expressed by the truncated completeness (Williams, 1999) as

\[ g(\Phi_s, \Phi) = 2\pi \sum_{n=1}^{N} \sum_{m=-n}^{n} Y_n^m(\Phi_s)^* Y_n^m(\Phi), \]  

(69)

where \( \Phi_s = \{\theta_s, \phi_s\} \) denotes the specific filtering direction, and \( g(\Phi_s, \Phi) \) represents specific beamforming function oriented towards direction, \( \Phi_s \), over angular range specified by \( \Phi \). The maximum order, \( N \), is dictated by the number of microphone channels, determines the sharpness of the beam patterns (Landschoot and Xiang, 2019).

This section, Sec. VI A has dealt with three different applications, yet they are in common that the prediction models are nested analytical expressions in form of generalized linear models (Brethorst, 1988; O Ruanaidh and Fitzgerald, 1996), consisting in essence of a sum of simple nonlinear functions. The sum of total number, \( S \) leads to different competing models for \( s = 1, \ldots, S \) for these problems, respectively. The Bayesian model selection, the higher level of inference is applied to selection of the suitable model given the experimental data, before the parameters encapsulated in the selected model are inferred. Three above discussed applications employed the nested sampling (Skilling, 2004);(Jasa and Xiang, 2012) to estimate the evidence in eqs. (59-60).

B. Multi-Layered Porous Media

In physical acoustics and many noise control applications, porous materials are of practical interest. Recently, Chazot et al. (2012); Roncen et al. (2018) apply Bayesian parameter estimation (the first level of inference) in rigid frame porous media analysis. When depth-dependent anisotropy of the porous media occurs, multilayered porous absorbers of finite-thickness layers approximate the depth anisotropy with each layer being considered as isotropic. Fackler et al. (2018) reported an application employing the two levels of inference within Bayesian framework to analyze multilayer porous materials, developing a method to determine simultaneously the number of constituent layers as well as the macroscopic physical properties of each layer.

1. Multi-Layered Model

In the work by Fackler et al. (2018), the rigid-frame porous materials are modeled based on a Miki-model (Miki, 1990), this model contains three physical parameters; the flow resistivity \( \sigma_f \), porosity \( \phi \), and tortuosity \( \alpha_\infty \) of a porous material. The Miki-model predicts the propagation coefficient \( \gamma(\omega) \) and the characteristic impedance \( Z_c(\omega) \) of such a material as (Miki, 1990)

\[ \gamma(\omega) = \frac{\omega\sqrt{\alpha_\infty}}{c_0} \left[ 0.160 \kappa^{-0.618} + j (1 + 0.109 \kappa^{-0.618}) \right], \]

(70)

\[ Z_c(\omega) = \rho_0 c_0 \sqrt{\alpha_\infty} \left( 1 + 0.070 \kappa^{-0.632} - j 0.107 \kappa^{-0.632} \right), \]

(71)

where \( \kappa = \omega/2\pi/\sigma_e \), with \( \sigma_e = \phi/\alpha_\infty \sigma_f \),

(72)

with \( \sigma_e \) being the effective flow resistivity of the porous material, \( j = \sqrt{-1} \), and \( \omega \) is the angular frequency.

When combining multiple distinct layers into multilayered media, the transfer matrix method is used to model the overall response of the media (Allard and Atalla, 2009; Blauert and Xiang, 2009). A two-by-two transfer matrix relates the acoustic pressure and normal component of particle velocity between the two sides of the \( m \)th layer with a thickness \( d_m \)

\[ T_{eq}^{(m)} = \begin{bmatrix} \cosh(\gamma_m d_m) & \sinh(\gamma_m d_m) \cdot Z_c^{(m)} \\ \sinh(\gamma_m d_m)/Z_c^{(m)} & \cosh(\gamma_m d_m) \end{bmatrix}, \]

(73)

and

\[ T_{rigid} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \]

(74)

where \( \gamma_m \) and \( Z_c^{(m)} \) of the \( m \)th layer are given in Eq. (70) and Eq. (71). \( T_{rigid} \) is a rigid termination matrix appended to the end of the transfer matrix chain of the composite material (Allard and Atalla, 2009) to model the

FIG. 10. Logarithm of the Bayesian evidence among competing models for three simultaneous sound sources. Twenty trials of evidence estimations are run per model. In this case, the three-source model correctly shows significant increase over

the one- and two-source models. The four- and five-source model correctly shows significant increase over}

tals of evidence estimations are run per model. In this case, the

three-source model correctly shows significant increase over

the one- and two-source models. The four- and five-source

models show slight increases in evidence, though not enough
to justify their higher complexity (Bush and Xiang, 2018).
mounting and termination of the material against a rigid backing. Matrix multiplication of this chain of distinct, multiple $T_{eq}^{(m)}$ in Eq. (73) along with the rigid termination results in the complex-valued surface impedance

$$Z_s = \frac{T_{11}}{T_{21}}, \quad \text{with} \quad \left[ \begin{array}{c} T_{11} \\ T_{21} \end{array} \right] = T_{eq}^{(1)} \cdots T_{eq}^{(M)} T_{\text{rigid}},$$ (75)

with $T_{11}, T_{21}$ associated with the acoustic pressure and normal component of the particle velocity at the front surface of potentially ($M$) multilayer media with a rigid backing (Fackler et al., 2018).

2. Likelihood function

The squared error between measured ($Z_{s,\text{meas}}$) and modeled ($Z_{s,\text{mod}}$) complex-valued surface impedance data as function of frequency is given as

$$e_k^2 = \text{Re} \left[ Z_{s,\text{meas}}(f_k) - Z_{s,\text{mod}}(f_k) \right]^2 + \text{Im} \left[ Z_{s,\text{meas}}(f_k) - Z_{s,\text{mod}}(f_k) \right]^2,$$ (76)

at each measured frequency point $f_k$, the real and imaginary parts of the complex surface impedance contribute separately to the error function. The likelihood function is assigned as a Student’s $t$-distribution as in Eq. (52),

$$L(\theta) = \frac{\Gamma(K/2)}{2} \left( \pi \sum_{k=1}^{K} e_k^2 \right)^{-K/2},$$ (77)

where the squared errors $e_k^2$ have been summed across all $K$ measured data points over the frequency range of interest. Note that the model predicted surface impedance $Z_{s,\text{mod}}(f_k)$ is dictated by parameter vector $\theta$ containing 3-4 parameters for each layer with potentially $M$ layers, as discussed below.

3. Parameter priors

Before involving any data, limited knowledge is available about the parameters under study. The prior probability distributions for each parameter encode this state of knowledge into a Bayesian analysis. For realistic porous materials, the physical parameters describing the pore structure fall into broad ranges of physically realistic values. Following the principle of maximum entropy introduced in Sec. IV A, a uniform prior distribution in form of Eq. (32) is assigned to each of the physical porous material parameters, encoding a lack of specific prior knowledge:

$$\Pr(\text{flow resist.} \sigma_f) = \text{Uniform}(0.1, 1000 \text{ kNs/m}^4),$$ (78)

$$\Pr(\text{porosity} \phi) = \text{Uniform}(0.1, 1),$$ (79)

$$\Pr(\text{tortuosity} \alpha_\infty) = \text{Uniform}(1, 7).$$ (80)

The material layers experimentally tested are on the order of a few centimeters thick. To remain impartial when considering the layer thickness as an unknown parameter to be estimated, it is assigned the following prior:

$$\Pr(\text{layer thickness} d) = \text{Uniform}(0.1 \text{ mm}, 10 \text{ cm});$$ (81)

otherwise, it is fixed at the physically-measured value.

C. Concise Representation of Head-Related Transfer Functions

Head-related transfer functions (HRTFs) represent directionally-dependent responses between an external source and the eardrums of a binaural listener. They are crucial part of databank for auditory virtual reality (Blauert, 2001; Xie, 2013) in the communication-acoustics applications. Experimental measurement results of the HRTFs are often represented on several hundred data points in frequency domain, approximation by a recursive filter can result in substantially less computation and data storage. Botts et al. (2013) represent them by a pole-zero model of digital filters with infinite-impulse responses (IIR filters). In order to represent measured HRTFs in a parsimonious form, Botts et al. (2013) applied Bayesian model-based analysis to filter design in form of infinite impulse response (IIR) filter.

1. Parametric representation

The head-related transfer functions, $H(z)$, can be presented in frequency domain (Botts et al., 2013) through a zeros-poles model

$$H(z) = G \prod_{l=1}^{L} \frac{z - q_l}{z - p_l} \times \prod_{m=1}^{M} \frac{z^2 - 2r_{q,m} \cos \phi_{q,m} z + r_{q,m}^2}{z^2 - 2r_{p,m} \cos \phi_{p,m} z + r_{p,m}^2},$$ (82)

where $z = e^{-j\omega}$, $p_l, q_l \in \mathbb{R}$ are real-valued zeros and poles,

$$q_m = r_{q,m} e^{j\phi_{q,m}}, \quad p_m = r_{p,m} e^{j\phi_{p,m}},$$ (83)

and $G$ is a real-valued gain factor. In the second line of above equation, the complex-valued zeros, $(z - q_m)(z - \overline{q_m})$, and poles, $(z - p_m)(z - \overline{p_m})$, are presented as magnitudes, $r_{p,m}, r_{q,m}$ and phases $\phi_{p,m}, \phi_{q,m}$ with $-1 \leq \cos \phi_{p,m}, \cos \phi_{q,m} \leq 1$. $N = L + M$ is the filter order, collectively.

2. Likelihood function and parameter prior

Botts et al. (2013) apply this model to derive filter coefficients of concise IIR filters for the HRTFs. The squared error, $\epsilon_k$, between complex-valued HRTFs, and the model $D(\omega_k) = HRTF(\omega_k)$ at discrete angular frequency, $\omega_k$,

$$\epsilon_k = \left[ |D(\omega_k)| - |H(\omega_k)| \right]^2 - \left[ \angle D(\omega_k) - \angle H(\omega_k) \right]^2,$$ (84)

with symbol $\angle$ representing phase angle. The likelihood function is expressed in Student’s $t$-distribution as in Eq. (52) and Eq. (77). The parameter priors also incorporate the stability constraints of the HRTF filters, assigned as $r_{p,m} < 1, p_l < 1$, as maximum-entropy priors, the assignment takes a form as

$$\Pr(\text{Gain}, G) = \text{Uniform}(0.01, 5.0),$$ (85)

$$\Pr(\text{poles}, q_{l,m}) = \text{Uniform}(0.01, 0.99),$$ (86)

$$\Pr(\text{zeros}, p_{l,m}) = \text{Uniform}(0, 0.99).$$ (87)
FIG. 11. Bayes factors and magnitude spectra of different filter orders (Botts et al., 2013). (a) Bayes factors for filter order 5 - 10 both in linear and logarithmic form. (b) Measured and estimated head-related transfer functions of filter order 5, 7, and 9.

Figure 11 exemplifies one HRTF data set from an experimentally measured data bank (Gardner and Martin, 1995). The two levels of Bayesian inference is formulated so as to estimate concise infinite impulse response filters with the parsimonious filter order \( N = L + M \), also using the nested sampling (Jasa and Xiang, 2012; Skilling, 2004) for the filter order selection followed by filter coefficient estimation as detailed in Sec. V. The effort benefits virtual auditory reality due to the computational demand in dynamic changes of auditory scenes in real-time (Xiang et al., 2019; Xie, 2013). The magnitude spectra as shown in Figure 11(b) indicates that an IIR filter of order 7 approximates the experimentally measured data concisely, yet reasonably well, while order 9 slightly improves the filter response. When considering a trade-off between simplicity of model and closeness of fit, the model selection narrows a number of alternatives down to filter order 7 and 9. As in Sec. VA among the two models, the closeness of fit favors filter order 9 (Botts et al., 2013), while the simplicity of the model favors filter order 7.

VII. SUMMARY

Experimental investigations are vital for scientists and engineers to better understand underlying theories in acoustics. Scientific experiments are often subject to uncertainties and randomness. Probability theory is the appropriate mathematical language for quantification of uncertainties and randomness. This tutorial paper emphasizes logical interpretation of probability, that the probability itself quantitatively represents strength of implication, or equivalently, state of knowledge. The arguments of the probabilities within Bayesian school represent propositions expressed by statements, not necessarily limited to large ensembles of repeatable events or random variables. Bayesian probability theory, centered around the Bayes’ theorem includes all the necessary calculus to manipulate probabilities which are essentially relying on product and sum rules of the probability which are the same rules within the propositional calculus, as detailed in Sec. II C.

Bayes’ theorem includes the prior probability and the likelihood function as its input quantities. Bayes’ theorem represents how one’s prior knowledge, encoded in the initial assignment, is updated once experimental data are incorporated into the analysis through the likelihood function. The model-based Bayesian inference critically relies on well understood parametric models which can be physical, phenomenological, or numerical. The parametric model is considered as a part of prior information when incorporating the observation data into the likelihood function, it ensures that the output of Bayes theorem is a posterior probability distribution based precisely on the available information put into the prior probability assignment, the information encoded in the prior probabilities is utilized rigorously based on principle of the maximum entropy. Sec. IV introduces, in details, how the information-theoretic entropy is maximized to assign the prior probability and the likelihood. The application of maximum entropy principle also constitutes consistent reasoning in scenarios where insufficient information for certainty is available. Therefore, Bayesian probability has been considered as the logic tool of inductive reasoning, for dealing with uncertainty, lack of scientific data, and ignorance.

Bayesian analysis has recently been applied to an increasing extent in acoustic science and engineering. Many data analysis tasks in acoustics and other fields require two levels of inference, namely, model selection, and within the chosen model, parameter estimation. The model selection quantitatively embodies Occam’s razor in favor of the concise and simpler model, Sec. V C has given this point of view a detailed explanation. Bayesian probabilistic methods provide solutions of both levels using Bayes theorem within a unified framework. One of two Bayes’ theorem outputs, Bayesian evidence, represents the key quantity for the higher level of inference, the model selection. The posterior probability as another output is responsible for the parameter estimation. In demonstrating the model-based data analysis within the unified Bayesian framework, the last section
(Sec. VI) briefly discusses a number of recent applications in room-acoustics, noise control, and communication acoustics. These examples all have in common the existence of well established models for predicting the processes under consideration and the two levels of inference are required for solving these acoustic problems.

Implementation of the model-based Bayesian analysis for both levels of inference often requires numerical exploration technique, so-called random walk or Markov chain Monte Carlo sampling methods, they are beyond the scope of the current tutorial paper.

VIII. APPENDICES

For completeness, this Chapter derives in certain details of maximum entropy assignment of the likelihood function (Appendix A) and marginalization of the hyperparameter (Appendix B). The derivations heavily rely on prior work (Brethorst, 1988; Gregory, 2005; Jaynes, 1968; Xiang and Goggans, 2001· 2003).

A. Appendix A

When assigning the likelihood function, the solution of Lagrange multipliers \( \lambda_0 \) and \( \lambda_1 \) in Eq. (40) relies on nothing else, but the two known constraints expressed in Eq. (36). The minimal constraint in Eq. (33) applied to Eq. (40) yields

\[
e^{-\lambda_0} \int_a^b e^{-\lambda_1 (x-\mu)^2} \, dx = 1, \quad (88)
\]

leading to

\[
e^{\lambda_0} = \int_a^b e^{-\lambda_1 (x-\mu)^2} \, dx. \quad (89)
\]

The integral on the right-hand side can be converted into the error function \( \text{erf}(z) \) (Abramowitz and Stegun, 1964; Gregory, 2005)

\[
\text{erf}(\tau) = \frac{2}{\sqrt{\pi}} \int_0^\tau \exp(-u^2) \, du, \quad (90)
\]

with a variable replacement of \( u = \sqrt{\lambda_1} (x-\mu) \). The error function has the properties: \( \text{erf}(-\tau) = -\text{erf}(\tau) \), \( \text{erf}(2) = 0.995 \), \( \text{erf}(3) = 0.999978 \), and \( \text{erf}(\infty) = 1 \). Rearranging eq. (89) in terms of eq. (90) yields

\[
e^{\lambda_0} = \sqrt{\frac{\pi}{2 \lambda_1}} \left[ \int_0^b e^{-u^2} \, du - \int_a^0 e^{-u^2} \, du \right]. \quad (91)
\]

Solving Eq. (91) for \( \lambda_0 \) with asymptotic limits \( a \to -\infty \) and \( b \to \infty \) yields

\[
\lambda_0 = \ln \left( \sqrt{\frac{\pi}{2 \lambda_1}} \right) = \ln \left( \sqrt{\frac{\pi}{2 \lambda_1}} \right) + \ln(2). \quad (92)
\]

The second constraint is used along with Eq. (92) for solving \( \lambda_1 \) as what follows. Differentiating, then multiplying \( -e^{-\lambda_0} \) on both sides of Eq. (89) yields

\[
-e^{-\lambda_0} \partial (e^{\lambda_0}) = -e^{-\lambda_0} \partial \left[ \int_a^b e^{-\lambda_1 (x-\mu)^2} \, dx \right]. \quad (93)
\]

\[
-\frac{\partial \lambda_0}{\partial \lambda_1} = \int_a^b (x-\mu)^2 p(x) \, dx = \sigma^2. \quad (95)
\]

Finally, substituting Eq. (97) into Eq. (40) yields

\[
p(x) = \frac{1}{\sqrt{2\pi \sigma}} e^{-(x-\mu)^2/2\sigma^2}, \quad (98)
\]

which is a Gaussian assignment as in Eq. (41). It is straightforwardly generalized for discrete variables as in Eq. (42). Note that this Gaussian assignment is the result of applying the principle of the maximum entropy which is distinctly different from assuming the statistics of the residual errors being Gaussian.

B. Appendix B

Expressing the likelihood function as function of parameter \( \theta \) for the marginalization integral in Eq. (51) with notation change \( L(\theta) = p(\mathbf{d} \mid \theta, H, I) \)

\[
L(\theta) = \int_0^\infty \left( \sigma \sqrt{2\pi} \right)^{-K} \exp \left( \frac{-Q}{\sigma^2} \right) \frac{1}{\sigma} \, d\sigma. \quad (99)
\]

Regrouping terms,

\[
L(\theta) = (2\pi)^{-K/2} \int_0^\infty \sigma^{-(K+1)} \exp(-Q \sigma^{-2}) \, d\sigma. \quad (100)
\]

Substituting \( u = \sigma^{-2} \), \( du = -2\sigma^{-3} \, d\sigma \) yields

\[
L(\theta) = \frac{1}{2} (2\pi)^{-K/2} \int_0^\infty u^{(K/2)-1} \exp(-Qu) \, du. \quad (101)
\]

Using the gamma function identity (Abramowitz and Stegun, 1964)

\[
\Gamma(n) = x^n \int_0^\infty u^{n-1} \exp(-xu) \, du, \quad (102)
\]
letting $x = Q$, and $n = K/2$, the marginalization integral is completed as

$$L(\theta) = \frac{1}{2} (2\pi)^{-K/2} K/2 \Gamma(K/2),$$

(103)

which is a Student’s $t$-distribution with $K - 1$ degrees of freedom. The derivations here in Appendix B lead to Eq. (52).

IX. ACKNOWLEDGMENT

The author is grateful to Drs. John Skilling, Anthony Garrett, Paul Goggans for their insightful discussion, to Drs. Jose Escolano, Tomislav Jaza, Zuhre Sü-Gül, Jonathan Botts, Cameron Fackler, Dane Bush, Mr. Douglas Beaton, and Mr. Christopher Landschoot for their collaborative research applications in Bayesian analysis.

1 Extended from the Acoustics Today’s article, vol 11 55-61 (2015). The portions of this work have also been presented at the 175th Troy Lectures, second edition (Springer-Verlag, Heidelberg, New York), pp. 161–166.


References


