Rus, G. and Chiachío, J. and Chiachío, M. (2015) Logical inference for inverse problems. Inverse Problems in Science and Engineering, 24 (3). pp. 448-464. ISSN 1741-5977 , http://dx.doi.org/10.1080/17415977.2015.1047361

Logical inference for inverse problems

Guillermo Rus, Juan Chiachío and Manuel Chiachío

E-mail: grus@ugr.es
Dept. Structural Mechanics and Hydraulic Engineering, University of Granada, Spain

Abstract. Estimating a deterministic single value for model parameters when reconstructing the system response has a limited meaning if one considers that the model used to predict its behavior is just an idealization of reality, and furthermore, the existence of measurements errors. To provide a suitable answer, probabilistic instead of deterministic values should be provided, which carry information about the degree of uncertainty or *plausiblity* of those model parameters providing one or more observations of the system response. This is widely-known as the Bayesian Inverse Problem, which has been covered in the literature from different perspectives, depending on the interpretation or the meaning assigned to the *probability*. In this paper, we revise two main approaches: the one that uses probability as logic, and an alternative one that interprets it as a information content. The contribution of this paper is to highlight their similarities and differences, and eventually provide their links as an unifying formulation. An extension to the problem of model class selection is derived, which is particularly simple under the proposed framework.

PACS numbers: 02.30.Zz, 02.50.Tt, 02.50.Ey

Submitted to: *Inverse Problems*

Keywords: Bayesian updating, Inverse Problem, Model-class selection, Stochastic Inverse Problem, Inference, Probability logic

1. Probability interpretation in physical phenomena

> It is unanimously agreed that statistics depends somehow on probability. But, as to what probability is and how it is connected with statistics, there has seldom been such complete disagreement and breakdown of communication since the Tower of Babel. Doubtless, much of the disagreement is merely terminological and would disappear under sufficiently sharp analysis. However there is a fundamental difference between frequentist and bayesian interpretations that cannot be bridged.
>
> Savage, 1972 [1]

The main statistical frameworks on which inverse problems and inference rely on have rigorously been legitimated after a long history [2]. The following could be an attempt to classify the sequence of physical interpretations of probability:

Classical: if a random experiment can result in a finite number n of mutually exclusive and equally likely outcomes and if n_A of these outcomes result in the occurrence of the event A, the probability of A was defined by Laplace as,

$$P(A) = \frac{n_A}{n} \tag{1}$$

Frequentist: the probability of an event A is its relative frequency of occurrence after repeating a process a large number n of trials under similar conditions,

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n} \tag{2}$$

This definition is commonly used as a physical meaning (R. A. Fisher, J. Neyman and E. Pearson [3, 4, 5, 6]). If the process is repeated a reduced series of times, different relative frequencies will be obtained in different series of trials. If these relative frequencies are to define the probability, the probability of event A will be non-unique. If we acknowledge the fact that we only can estimate a probability we still get into problems as the error of estimation can only be expressed as a probability, the very concept we are trying to define. This renders the frequency definition circular. Hence the relative frequency of a event A informs, but does not define, the parameter representing the probability of the event in a probability model.

Evidential or propensity: the theory of evidential probability studies the impact of evidence on probability. It is motivated by two basic ideas [7]: (i) probability assessments should be based on known relative frequencies, and the assignment of probability to specific individual events should be based on its the available information history, and (ii) Humphreys paradox [8] shows how propensities do not obey Kolmogorov's probability calculus, and reads as follows. Probability calculus implies Bayes' theorem, which allows us to invert a conditional probability $P(A|B) = P(B|A)P(A)/P(B)$, whereas propensities are intented to be interpreted as measures of causal trends, and since the causal relation is not necessarily symmetric, these propensities should not invert. Humphrey's paradox is illustrated by supposing a test for an illness that occasionally gives false positives and false negatives. A given sick patient may have a propensity to give a positive test result, but it apparently makes no sense to say that a given positive test result has a propensity to have come from a sick patient. Thus, propensities, whatever they are, must not obey the usual probability calculus: "if the probability of B, given A exists, then the probability of A, given B exists, however one understands these conditional probabilities". Fetzer and Nute [9] formulated a probabilistic causal calculus different from Kolmogorov's calculus.

Logical: the probability $P[H|E]$ is interpreted as the degree of plausibility of a proposition H (typically a hypothesis) given the information in the proposition E (typically empirical evidence). Logical probabilities are thus objective, logical relations between propositions [10, 11] (states of knowledge), in contrast to the physical propensity of a phenomenon. This views allows to build the Bayesian inference: to compute the posterior probability of a hypothesis, some specified prior probability known about it is updated by new knowledge or data. In contrast to assigning a probability to a hypothesis, in frequentist probability, hypothesis are just formally tested.

Cox [12] postulates enable logical probability interpretation to be applied to any proposition, when supported by new gained information, as a natural extension of Aristotelian logic (by which statements are either true or false) into the realm of reasoning in the presence of uncertainty:

(i) "A double negative is an affirmative" becomes a functional equation $f(f(x)) = x$.

(ii) The plausibility of the conjunction $[A\&B]$ of two propositions A, B, depends only on the plausibility of B and that of A given that B is true, $P(A\&B) = P(A)P(B|A)$.

(iii) Suppose $[A\&B]$ is equivalent to $[C\&D]$. If we acquire new information A and then acquire further new information B, and update all probabilities each time, the updated probabilities will be the same as if we had first acquired new information C and then acquired further new information D, $yf\left(\frac{f(z)}{y}\right) = zf\left(\frac{f(y)}{z}\right)$.

Cox [12] derived the laws of probability from these postulates, which are, assuming that the scale of information measurement ranges from zero to one:

(i) Certainty is represented by $P(A|B) = 1$.

(ii) Negation: $P(A|B) + P(\overline{A}|B) = 1$.

(iii) Conjunction: $P(A, B|C) = P(A|C)P(B|A, C) = P(B|C)P(A|B, C)$.

These laws yield finite additivity of probability, but not countable additivity. Kolmogorov's axioms of probability, which assume that a probability measure is countably additive (necessary for the proof of certain theorems) are,

(i) Non-negativity: $P(A) \geq 0$.

(ii) Finite additivity: $P(A \cup B) = P(A) + P(B)\ \forall\ A, B | A \cap B = \emptyset$.

(iii) Normalization: $P(\Omega) = 1$.

Kolmogorov comments that infinite probability spaces are idealized models of real random processes, and that he limits himself arbitrarily to only those models that satisfy countable additivity. This axiom is the cornerstone of the assimilation of probability theory to measure theory [2]. The conditional probability of A given B is then given by the ratio of unconditional probabilities,

$$P(A|B) = \frac{P(A \cap B)}{P(B)},\ P(B) > 0 \tag{3}$$

Subjective: probabilities are understood as degrees of rational belief [11], rather than logical relations that constrain degrees of rational belief. Ramsey [13] questioned the existence of such objective logical relations and redefined evidential probability as "the logic of partial belief".

Outside physical uses, subjective or personalist probability, and epistemic or inductive probability have recently been developed as an incompatible interpretations to the frequentist one [14].

Predictive inference: stems from Bayesian probability of physical phenomena with errors by assuming De Finetti's [15] idea of exchangeability: that future observations should behave like past observations, and the concept of cross-validation [16].

2. Modeling assumptions

The goal of the inverse problem is to use the observed response of a system to *improve* a single or a set of models that idealize that system, so that they make more accurate predictions of the system response to a prescribed, or uncertain, excitation.

Following the Bayesian formulation of the inverse problem [17], the solution is not a single-valued set of model parameters $\boldsymbol{\theta}$. On the contrary, Bayes' Theorem takes the initial quantification of the plausibility of each model parameterized by $\boldsymbol{\theta}$, which is expressed by the *prior* probability distribution, and updates this plausibility by using the information in the data set $\mathcal{D}$, to obtain the *posterior* probability distribution of model parameters.

The origin of the uncertainties are built into the interpretation of probability as a measure of relative plausibility of the various possibilities conditional to available information. This interpretation is not well known in the engineering community where there is a wide-spread belief that probability only applies to aleatory uncertainty (inherent randomness in nature) and not to epistemic uncertainty (missing information). Jaynes [18] noted that the assumption of inherent randomness is an example of what he called the Mind-Projection Fallacy: our uncertainty is ascribed to an inherent property of nature, or, more generally, our models of reality are confused with reality.

The interpretation of the final inferred model probability can be used either to identify a set of plausible values, or to find the most probable one (expected), or, following Tarantola [17], just to falsify inconsistent models, since according to Popper [19], that is the only thing we can assert.

Furthermore, different model parameterizations or even model hypothesis representing different physics can be formulated and hypothesized to idealize the system, yielding a set of different (Bayesian) *model classes* [20], $\mathbf{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_{N_M}\}$, resulting different values of model hypothesis or classes.

2.1. Notation

From the above description, we highlight three important pieces of information in the Bayesian inverse problem, which are described here:

$\mathcal{D}$: data set containing the system output (or input-output couple, depending on the experimental setup). It can be either the real output $\mathcal{D}^{\text{real}}$, or the ideal output to be predicted $\mathcal{D}^{\text{ideal}}$, or the measured output $\mathcal{D}^{\text{obs}}$. Each of them may belong to different spaces, but need to be comparable in the sense that they can be related.

$\mathcal{M}_j$: j_{th} model class or candidate among alternative model classes hypothesized to idealize the system. A Bayesian model class can be defined by two fundamental probability models: an input-output (I/O) model $\{p(\mathcal{D}^{\text{ideal}}|\mathbf{u}, \boldsymbol{\theta}, \mathcal{M}_j) : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^{N_p}\}$ and a prior probability distribution $p(\boldsymbol{\theta}|\mathcal{M}_j)$, that gives a initial relative plausibility of model parameters defining the I/O model in the class. Here $\mathbf{u}$ denotes the inputs to the system.

$\boldsymbol{\theta}$: set of uncertain model parameters within a specific model class $\mathcal{M}_j$, that calibrate the idealized relationships between input and output of the system.

All the defined variables (output data $\mathcal{D}^{\text{real}}$, $\mathcal{D}^{\text{ideal}}$, $\mathcal{D}^{\text{obs}}$, model parameters $\boldsymbol{\theta}$ or model classes $\mathcal{M}_j$) are defined to lie in manifolds $\mathfrak{D}^{\text{real}}$, $\mathfrak{D}^{\text{ideal}}$, $\mathfrak{D}^{\text{obs}}$, $\mathfrak{M}$ and $\boldsymbol{\Theta}$, respectively.

2.2. Real and ideal system definitions

When observing a real system using prior knowledge about of the physics that governs it, idealized by a model, careful analysis needs to be made about how to combine the elements of these two pieces of information: observations+model.

The first step is to identify which elements of the real system under observation plays a relevant role. Figure 1 schematizes these elements and their relationships. When a physics-based idealization of the system is required, it should follow a parallel scheme to the real one (lower half of the same figure), where all elements are connected by defined relationships. To sum up, the Inverse Problem can be defined as the counterpart of the Forward Problem (aimed at computing the unknown output $\mathcal{D}^{\text{ideal}}$ of a known idealized system $g(\boldsymbol{\theta})$), i.e. computing an unknown part of the system ($\boldsymbol{\theta}$) given some observable part of the output $\mathcal{D}^{\text{obs}}$.

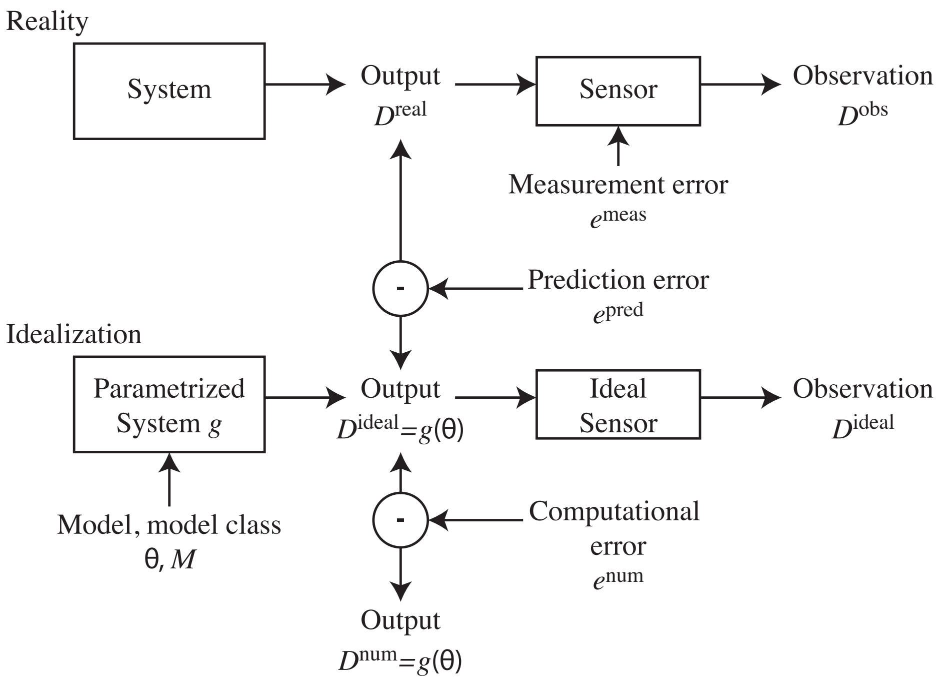

Figure 1. Scheme of real and ideal systems. Note that system input may not necessarily appear explicitly outside the system. In the mathematical idealization, second half, an ideal sensor is conceived with the peculiarity that it is assumed to perfectly interrogate the system output introducing no error or bias.

Note from the figure that the noise in the sensors groups any type of difference between observed and real data, including sensor error (characterized by a probability model) and quantization in the case of digital sensors, yielding the relationship,

$$\mathcal{D}^{\text{obs}} = \mathcal{D}^{\text{real}} + e^{\text{meas}} \tag{4}$$

On the other hand, the assumptions required in the process of idealization of reality are responsible for the differences between real and ideal output,

$$\mathcal{D}^{\text{real}} = \mathcal{D}^{\text{ideal}} + e^{\text{pred}} \tag{5}$$

Then

$$\mathcal{D}^{\text{obs}} = \mathcal{D}^{\text{ideal}} + e^{\text{pred}} + e^{\text{meas}} \tag{6}$$

For some instruments, the measurement errors can be neglected in comparison to modeling errors, thus the last equation can be rewritten as,

$$\mathcal{D}^{\text{obs}} = \mathcal{D}^{\text{ideal}} + e^{\text{pred}} \tag{7}$$

3. IP formulation from the probability logic viewpoint

Following the probability logic formulation of the inverse problem established by Beck [21, 20], the solution is not a single-valued set of optimal model parameters $\boldsymbol{\theta}^*$ but a conditional PDF of the values of the model parameters $\boldsymbol{\theta}$ given a set of data $\mathcal{D}$ and a model class $\mathcal{M}$: $p(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})$. The probability density p is assigned the meaning of relative plausibility of the model values $\boldsymbol{\theta}$ to be true given $\mathcal{D}$ and $\mathcal{M}$.

3.1. Assumptions

Bayesian probabilities in probability logic are always conditioned, i.e. the probability $P[b|c]$ is interpreted as the degree of plausibility of proposition b given the information in proposition c, whose truth we need not know.

The definition is based on logical operators according to Cox [12]. The arbitrary mapping $\phi : [0,1] \to [0,1]$ for defining the conjunction is taken to be the simplest possible definition: the identity. The probability logic axioms based on Boolean logic and Cox's postulate are adopted.

3.2. Formulation in the case of perfect observations

Let's start assuming perfect observations in the sense that the discrepancy due to sensor and idealization is negligible, $\mathcal{D}^{\mathrm{real}} = \mathcal{D}^{\mathrm{ideal}} = \mathcal{D}^{\mathrm{obs}} = \mathcal{D}$. Given observations $\mathcal{D}$ consisting of measured outputs or pairs of outputs response to inputs to the system, their updated relative plausibility can be quantified by $p(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})$ for the uncertain model parameters $\boldsymbol{\theta}$ within the model class $\mathcal{M}$. Using Bayes' Theorem:

$$p(\boldsymbol{\theta}|\mathcal{D},\mathcal{M}) = c^{-1} p(\mathcal{D}|\boldsymbol{\theta},\mathcal{M}) p(\boldsymbol{\theta}|\mathcal{M}) \tag{8}$$

where $c = p(\mathcal{D}|\mathcal{M}) = \int_{\boldsymbol{\Theta}} p(\mathcal{D}|\boldsymbol{\theta},\mathcal{M}) p(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta}$ is a normalizing constant called the *evidence* of data set $\mathcal{D}$ for the model class $\mathcal{M}$; $p(\mathcal{D}|\boldsymbol{\theta},\mathcal{M})$ is the *likelihood function* that quantifies the probability of getting the observations $\mathcal{D}$ by the I/O model specified by $\boldsymbol{\theta}$ in the the model class $\mathcal{M}$; and $p(\boldsymbol{\theta}|\mathcal{M})$ is the *prior* PDF assigned to model parameter values $\boldsymbol{\theta}$ within $\mathcal{M}$ (usually chosen to provide regularization of ill-conditioned inverse problems). ‡

3.3. Formulation for ideal, real and observed output

The case of presence of sensor noise or prediction error can be derived from the relationships in Equations 4 and 5. In the probability logic framework, the relations among ideal, real and observed outputs are derived from conditional probability and a subsequent marginalization, as follows,

‡ Note that, in equation (10) and the sequel, $\mathcal{M}_j$ has been replaced by $\mathcal{M}$ for compactness.

$p(\mathcal{D}^{\text{real}}, \mathcal{D}^{\text{ideal}}) = p(\mathcal{D}^{\text{real}}|\mathcal{D}^{\text{ideal}})p(\mathcal{D}^{\text{ideal}})$ where the conditional probability $p(\mathcal{D}^{\text{real}}|\mathcal{D}^{\text{ideal}})$ incorporates the prediction error. In the case of perfect idealization, this conditional probability is just the identity.

$p(\mathcal{D}^{\text{real}}, \mathcal{D}^{\text{obs}}) = p(\mathcal{D}^{\text{real}}|\mathcal{D}^{\text{obs}})p(\mathcal{D}^{\text{obs}})$ where the conditional probability $p(\mathcal{D}^{\text{real}}|\mathcal{D}^{\text{obs}})$ incorporates the measurement noise (sensor error, bias and quantization). Examples of this conditional probability are given in Figure 2.

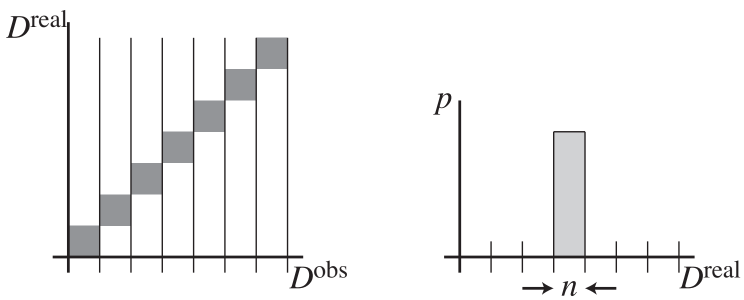

Figure 2. Examples of probability density relating real and ideal output though the error prediction. Left and center: case of perfect measurement with only quantization (center: slice for a single value of $\mathcal{D}^{\text{obs}}$. Right: case of sensor with quantization and uncertainty. Gray tones stand for probability densities, being white null probability, and black maximum probability.

The observed data can be transformed to ideal data, as

$$
\begin{aligned}
p(\mathcal{D}^{\text{real}}) &= \int_{\mathfrak{D}^{\text{obs}}} p(\mathcal{D}^{\text{real}}|\mathcal{D}^{\text{ideal}})p(\mathcal{D}^{\text{ideal}})d\mathcal{D}^{\text{obs}} \quad \Rightarrow \\
p(\mathcal{D}^{\text{ideal}}) &= \int_{\mathfrak{D}^{\text{real}}} \int_{\mathfrak{D}^{\text{obs}}} p(\mathcal{D}^{\text{ideal}}|\mathcal{D}^{\text{real}})p(\mathcal{D}^{\text{real}}|\mathcal{D}^{\text{obs}})p(\mathcal{D}^{\text{obs}})d\mathcal{D}^{\text{obs}}d\mathcal{D}^{\text{real}} \qquad (9)
\end{aligned}
$$

that can subsequently be used to update the ideal model, as

$$
p(\boldsymbol{\theta}|\mathcal{D}^{\text{ideal}}, \mathcal{M}) = c^{-1}p(\mathcal{D}^{\text{ideal}}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M}) \qquad (10)
$$

4. IP formulation from the conjunction of states of information viewpoint

The relationship between the model and the observations provided by a model need not to be an implication due to a cause-effect, which would require to define the conditional probability $p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M})$. Instead, just the joint probability density $f(\boldsymbol{\theta}, \mathcal{D}, \mathcal{M})$ needs to be defined in the following approach, in which the causality between model and observations may be inverted or even not exist.

This formulation does not use conditional probabilities as a elementary notion of information and in turn it uses joint probabilities obtained as a conjunction of states of information [17]. The last two points can be considered as strengths of the formulation.

4.1. Assumptions

The output data (real $\mathcal{D}^{\text{real}}$, ideal $\mathcal{D}^{\text{ideal}}$ and observed $\mathcal{D}^{\text{obs}}$) reside in their own independent manifolds. These manifolds do not need to be intersecting as long as Equations 4 and 5 need not to be written. As defined above, all the variables (output data $\mathcal{D}^{\text{real}}$, $\mathcal{D}^{\text{ideal}}$, $\mathcal{D}^{\text{obs}}$, model parameters $\boldsymbol{\theta}$ or model classes $\mathcal{M}_j$) are defined in their manifolds $\mathfrak{D}^{\text{real}}$, $\mathfrak{D}^{\text{ideal}}$, $\mathfrak{D}^{\text{obs}}$, $\boldsymbol{\Theta}$ and $\mathfrak{M}$, respectively.

An event or realization of them is defined by a region or subset A. The information about them (which is an idealized construct) is defined by a measure ($P(A)$) that satisfies the first two Kolmogorov axioms ($P(A) \geq 0$, $P(A \cup B) = P(A) + P(B) \; \forall \; A, B | A \cap B = \emptyset$). By Radon-Nikodym theorem, a density $f(x)$ can be defined,

$$P(A) = \int_A f(x)dx \tag{11}$$

and the Kolmogorov normality $P(\Omega) = 1$ is not assumed.

The logical inference operations on the information defined above has been defined elsewhere, but can be summarized as follows. Starting from the *and* and *or* operator definition for Boolean logic,

a	b	$P_a \wedge P_b$	$P_a \vee P_b$
0	0	0	0
0	1	0	1
1	0	0	1
1	1	1	1

Without assuming normality, the following relationship are compatible, using De Morgan's law,

$$\begin{array}{lllll} P_a(A) \neq 0 & or & P_b(A) \neq 0 & \Rightarrow & (P_a \vee P_b)A \neq 0 \\ P_a(A) = 0 & or & P_b(A) = 0 & \Rightarrow & (P_a \wedge P_b)A = 0 \end{array} \tag{12}$$

Commutativity is also allowed,

$$P_a \vee P_b = P_b \vee P_a \qquad P_a \wedge P_b = P_b \wedge P_a \tag{13}$$

The simplest solution that fulfills these axioms without normalization is§,

$$\begin{cases} f_1 \vee f_2 = f_1 + f_2 \\ f_1 \wedge f_2 = f_1 f_2 \end{cases} \tag{14}$$

§ This solution is consistent as long as the parameters (observations, model parameters, etc.) are Jeffrey's parameters [17]. If not, the probability densities $f(y)$ just need to be divided by the noninformative probability density $\mu(y)$, i.e. replacing $f(y)$ by $\frac{f(y)}{\mu(y)}$ everytime.

4.2. Case of perfect observations

For presenting the idea behind the formulation in a simpler way, the case when observations are perfect, i.e. discrepancy due to sensor or idealization is negligible, $\mathcal{D}^{\text{real}} = \mathcal{D}^{\text{ideal}} = \mathcal{D}^{\text{obs}} = \mathcal{D}$ is presented without loss of generality.

Assume that the system under test is defined by observations, model parameter and idealized model classes. If we have two sources of information (probabilistic propositions) to infer information about the model parameters $f(\boldsymbol{\theta})$, which are that originated by experimental observations of the system f^o, and that originated from a mathematical model of the system f^m, the probabilistic logic conjunction operator allows to compute the information state that the system parameters fulfill both propositions simultaneously, $f^o \wedge f^m$, as,

$$f(\mathcal{D}, \boldsymbol{\theta}, \mathcal{M}) = f^o(\mathcal{D}, \boldsymbol{\theta}, \mathcal{M}) \wedge f^m(\mathcal{D}, \boldsymbol{\theta}, \mathcal{M}) = f^o(\mathcal{D}, \boldsymbol{\theta}, \mathcal{M}) f^m(\mathcal{D}, \boldsymbol{\theta}, \mathcal{M}) \tag{15}$$

Assuming that the experimental information on observations is carried out with sensors that are independent on techniques to infer experimental information on model parameters, and the same is true for model classes, the joint density can be split as the product $f^o(\mathcal{D}, \boldsymbol{\theta}, \mathcal{M}) = f^o(\mathcal{D}) f^o(\boldsymbol{\theta}) f^o(\mathcal{M})$. This is not true for the model information f^m, since it relates observations and model.

By reusing the mentioned Radon-Nikodym theorem on the density defined in Equation 15, the marginal density for every possible observation $\mathcal{D} \in \mathfrak{D}$ yields the sought information on the model parameters, in a given model class $\mathcal{M} = \mathcal{M}_j$, as ‖

$$f(\boldsymbol{\theta}, \mathcal{M}_j) = \int_{\mathfrak{D}} f(\mathcal{D}, \boldsymbol{\theta}, \mathcal{M}_j) d\mathcal{D} = \int_{\mathfrak{D}} f^o(\mathcal{D}) f^o(\boldsymbol{\theta}) f^o(\mathcal{M}_j) f^m(\mathcal{D}, \boldsymbol{\theta}, \mathcal{M}_j) d\mathcal{D} \tag{16}$$

4.3. Formulation for general ideal, real and observed output

In addition to the a priori information provided by f^o and the information given by the model through f^m, the uncertainty introduced by the idealization of the model and from the sensors can be defined by two new probability densities f^i and f^s respectively. Their treatment is detailed below.

‖ The interpretation of the updated information for identifying the most plausible model parameter just requires to find its maximum, known as the "maximum a posteriori", (MAP)

$$\text{MAP} = \arg\max_{\boldsymbol{\theta} \in \Theta} f(\boldsymbol{\theta}, \mathcal{M}_j)$$

whereas finding plausible model values, or just falsifying inconsistent models, requires comparing information densities, and therefore a normalization. This can be done just by defining a normalized probability density p that satisfies the third Kolmogorov axiom (theorem of total probability),

$$p(\boldsymbol{\theta}) = \frac{f(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} f(y) dy}$$

$f^o(\mathcal{D}^{\text{real}}, \mathcal{D}^{\text{ideal}}, \mathcal{D}^{\text{obs}}, \boldsymbol{\theta}, \mathcal{M}_j) = f^o(\mathcal{D}^{\text{obs}}) f^o(\mathcal{D}^{\text{ideal}}) f^o(\boldsymbol{\theta}) f^o(\mathcal{M}_j) \mu(\mathcal{D}^{\text{real}})$ The prior informations about each magnitude are independent, so they are split as a product. The readings from the sensors are expressed as the prior information on the observations as $f^o(\mathcal{D}^{\text{obs}})$. If some prior information about the system output is available (for example physically impossible values), it can be coded by $f^o(\mathcal{D}^{\text{ideal}})$ and allows, as an example, to reject outliers among the measurements. Since no prior information can be given about the real output, its independent probability is non-informative $\mu(\mathcal{D}^{\text{real}})$. Prior knowledge about the model and the class are given by $f^o(\boldsymbol{\theta})$ and $f^o(\mathcal{M}_j)$.

$f^s(\mathcal{D}^{\text{real}}, \mathcal{D}^{\text{ideal}}, \mathcal{D}^{\text{obs}}, \boldsymbol{\theta}, \mathcal{M}_j) = f^s(\mathcal{D}^{\text{obs}}, \mathcal{D}^{\text{ideal}}) \mu(\boldsymbol{\theta}) \mu(\mathcal{M}_j) \mu(\mathcal{D}^{\text{real}})$. Since the sensor only relates observations to real output by adding noise as described in Equation 4, which is quantified by the joint density $f^s(\mathcal{D}^{\text{obs}}, \mathcal{D}^{\text{ideal}})$, the remaining magnitudes are independent and non-informative, $\mu(\boldsymbol{\theta})$, $\mu(\mathcal{M}_j)$ and $\mu(\mathcal{D}^{\text{real}})$.

$f^i(\mathcal{D}^{\text{real}}, \mathcal{D}^{\text{ideal}}, \mathcal{D}^{\text{obs}}, \boldsymbol{\theta}, \mathcal{M}_j) = f^i(\mathcal{D}^{\text{real}}, \mathcal{D}^{\text{ideal}}) \mu(\boldsymbol{\theta}) \mu(\mathcal{M}_j) \mu(\mathcal{D}^{\text{obs}})$. Since the idealization only relates ideal to real output by adding the prediction error as described in Equation 5, which is quantified by the joint density $f^s(\mathcal{D}^{\text{real}}, \mathcal{D}^{\text{ideal}})$, the remaining magnitudes are independent and non-informative, $\mu(\boldsymbol{\theta})$, $\mu(\mathcal{M}_j)$ and $\mu(\mathcal{D}^{\text{obs}})$.

$f^m(\mathcal{D}^{\text{real}}, \mathcal{D}^{\text{ideal}}, \mathcal{D}^{\text{obs}}, \boldsymbol{\theta}, \mathcal{M}_j) = f^i(\mathcal{D}^{\text{real}}, \mathcal{D}^{\text{ideal}}) \mu(\boldsymbol{\theta}) \mu(\mathcal{M}_j) \mu(\mathcal{D}^{\text{obs}})$. The model only exists in the "ideal world" and therefore only relates ideal output with model parameters given a model class by the density $f^m(\mathcal{D}^{\text{ideal}}, \boldsymbol{\theta}, \mathcal{M}_j)$. The remaining magnitudes $\mu(\mathcal{D}^{\text{obs}})$ and $\mu(\mathcal{D}^{\text{real}})$ are independent and non-informative.

These four pieces of information are simultaneously true yielding a joint probability through the conjunction operator,

$$f(\mathcal{D}^{\text{real}}, \mathcal{D}^{\text{ideal}}, \mathcal{D}^{\text{obs}}, \boldsymbol{\theta}, \mathcal{M}_j) = f^o(\mathcal{D}^{\text{real}}, \mathcal{D}^{\text{ideal}}, \mathcal{D}^{\text{obs}}, \boldsymbol{\theta}, \mathcal{M}_j) f^s(\mathcal{D}^{\text{real}}, \mathcal{D}^{\text{ideal}}, \mathcal{D}^{\text{obs}}, \boldsymbol{\theta}, \mathcal{M}_j) f^i(\mathcal{D}^{\text{real}}, \mathcal{D}^{\text{ideal}}, \mathcal{D}^{\text{obs}}, \boldsymbol{\theta}, \mathcal{M}_j) f^m(\mathcal{D}^{\text{real}}, \mathcal{D}^{\text{ideal}}, \mathcal{D}^{\text{obs}}, \boldsymbol{\theta}, \mathcal{M}_j) \tag{17}$$

In the case of Jeffreys parameters, which have the characteristic of being positive and of being as popular as their inverses [17], all non-informative densities μ are constant and may therefore be dropped from the formulation. By further marginalizing, the sought information is given by,

$$f(\boldsymbol{\theta}, \mathcal{M}_j) = \int_{\mathfrak{D}^{\text{real}}} \int_{\mathfrak{D}^{\text{ideal}}} \int_{\mathfrak{D}^{\text{obs}}} f^o(\mathcal{D}^{\text{obs}}) f^o(\mathcal{D}^{\text{ideal}}) f^o(\boldsymbol{\theta}) f^o(\mathcal{M}_j) f^s(\mathcal{D}^{\text{obs}}, \mathcal{D}^{\text{ideal}}) f^i(\mathcal{D}^{\text{real}}, \mathcal{D}^{\text{ideal}}) f^m(\mathcal{D}^{\text{ideal}}, \boldsymbol{\theta}, \mathcal{M}_j) d\mathcal{D}^{\text{obs}} d\mathcal{D}^{\text{ideal}} d\mathcal{D}^{\text{real}} \tag{18}$$

4.4. Reconstruction of the model parameters

Without loss of generality, and for a simpler notation, we may restrict ourselves to the case when observations are perfect, i.e. discrepancy due to sensor or idealization is negligible, $\mathcal{D}^{\text{real}} = \mathcal{D}^{\text{ideal}} = \mathcal{D}^{\text{obs}} = \mathcal{D}$.

The reconstructed probability for the model parameters $\boldsymbol{\theta}$ providing the model class $\mathcal{M}_j$ is obtained from the joint probability $f(\mathcal{D}, \boldsymbol{\theta}, \mathcal{M})$ by extracting the marginal probability for all possible observations $\mathcal{D} \in \mathfrak{D}$ and provided the model class $\mathcal{M}_j \in \mathfrak{M}$ is assumed to be true ($f^0(\mathcal{M} = \mathcal{M}_j) = 1$) as,

$$f(\boldsymbol{\theta})\big|_{\mathcal{M}=\mathcal{M}_j} = \int_{\mathcal{M}=\mathcal{M}_j} \int_{\mathfrak{D}} f(\mathcal{D}, \boldsymbol{\theta}, \mathcal{M}) d\mathcal{D} d\mathcal{M} = k_1 \int_{\mathfrak{D}} f^0(\mathcal{D}) f^0(\boldsymbol{\theta}) f^m(\mathcal{D}, \boldsymbol{\theta}, \mathcal{M}_j) d\mathcal{D} \quad (19)$$

where k_1 is a normalization constant that replaces the dropped model class probability. The assumption of no prior knowledge about the model parameters is usually made, whereby it is represented by the non-informative distribution, i.e. an arbitrary constant in the assumed case of Jeffrey's parameters $f^0(\boldsymbol{\theta}) = 1$,

$$f(\boldsymbol{\theta})\big|_{\mathcal{M}=\mathcal{M}_j} = k_1 \int_{\mathfrak{D}} f^0(\mathcal{D}) f^m(\mathcal{D}, \boldsymbol{\theta}, \mathcal{M}_j) d\mathcal{D} \quad (20)$$

If we assume the hypothesis of negligible observational uncertainties with respect to modelization uncertainties ($f^0(\mathcal{D}) = f^0(\mathcal{D}^{\text{obs}})$) and that the data manifold $\mathfrak{D}$ is a linear space (whereby the noninformative homogeneous probability density $\mu(\mathcal{D}^{\text{real}})$ is a constant), hence the integral in Equation 20 vanishes yielding the reconstructed model parameters probability density, which is clarified by the example in the next section,

$$f(\boldsymbol{\theta})\big|_{\mathcal{M}=\mathcal{M}_j} = k_2 f^m(\mathcal{D}^{\text{obs}}, \boldsymbol{\theta}, \mathcal{M}_j) \quad (21)$$

The latter formulation is equivalent to the one obtained from the probability logic viewpoint in Equation 10 (after dropping the prior model parameter information for being assumed noninformative), except for a constant since f^m needs not range $[0, 1]$, which proves the correctness and unifies both approaches.

5. Solution for time-domain observations with gaussian uncertainties

Either the final expressions of the probability densities p from the probability logic, or f from the conjunction of states of information can be treated as follows, as both final expressions are equivalent. Assume that the observations are assumed to follow a Gaussian distribution $\mathcal{D} \sim \mathcal{N}(E[\mathcal{D}^{\text{obs}}], C^{\text{obs}})$ whose mean is that of the experimental observations $\mathcal{D}^{\text{obs}}$ and covariance matrix C^{obs} standing for the measurement error noise. Likewise, the numerical errors are also assumed to follow a Gaussian distribution $\mathcal{D} \sim \mathcal{N}(\mathcal{D}^{\text{num}}, C^{\text{num}})$ centered at the numerically computed ones $E[\mathcal{D}^{\text{num}}] = \mathcal{D}(\mathcal{M})$ with covariance matrix C^{num}.

Assume that the observations $\mathcal{D}$ are a vector of functions of time $\mathcal{D} = o_i(t)$ at every measuring time $t \in [0, T]$ and repetition $i \in [1...N_i]$, and that the assumptions made above are valid for every instant t and sensor i. Considering that the compound probability of the information from all sensors and time instants is the productory of that of each one individually, what means information independence, and that this productory is equivalent to a summation within the exponentiation (and an integration along the continuous time, seen as a summation over every infinitesimal dt), the Gaussian distribution allows for an explicit expression of the probability densities,

$$f^0(o_i(t)) = k_3 e^{\left[\begin{array}{l} -\frac{1}{2}\sum_{i,j=1}^{N_i}\int_{t=0}^{t=T}\left(o_i(t) - o_i^{\text{obs}}(t)\right) \\ \qquad\left(c_{ij}^{\text{obs}}\right)^{-1}\left(o_j(t) - o_j^{\text{obs}}(t)\right)dt \end{array}\right]} \tag{22}$$

$$f^m(o_i(t), \boldsymbol{\theta}, \mathcal{M}) = k_4 e^{\left[\begin{array}{l} -\frac{1}{2}\sum_{i,j=1}^{N_i}\int_{t=0}^{t=T}\left(o_i(t) - o_i(t,\boldsymbol{\theta})\right) \\ \qquad\left(c_{ij}^{\text{num}}\right)^{-1}\left(o_j(t) - o_j(t,\boldsymbol{\theta})\right)dt \end{array}\right]} \tag{23}$$

$$\Rightarrow \quad f(\boldsymbol{\theta})\big|_{\mathcal{M}=\mathcal{M}_j} = k_5 e^{\overbrace{\left[\begin{array}{l} -\frac{1}{2}\sum_{i,j=1}^{N_i}\int_{t=0}^{t=T}\left(o_i(t,\boldsymbol{\theta}) - o_i^{\text{obs}}(t)\right) \\ \left(c_{ij}^{\text{obs}} + c_{ij}^{\text{num}}\right)^{-1}\left(o_j(t,\boldsymbol{\theta}) - o_j^{\text{obs}}(t)\right)dt \end{array}\right]}^{J(\boldsymbol{\theta})}} \tag{24}$$

The term $J(\boldsymbol{\theta})$ corresponds to a misfit function between model and observations, then

$$f(\boldsymbol{\theta})\big|_{\mathcal{M}=\mathcal{M}_j} = k_5 e^{-J(\boldsymbol{\theta})} \tag{25}$$

The best-fitting model is found by minimizing $J(\boldsymbol{\theta})$, or equivalently maximizing $f(\boldsymbol{\theta})$, since

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}}\left\{f(\boldsymbol{\theta})\big|_{\mathcal{M}=\mathcal{M}_j} = k_5 e^{-J(\boldsymbol{\theta})}\right\} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}}\left\{J(\boldsymbol{\theta})\right\} \tag{26}$$

Finally, if classical probability densities are desired, the constant k_6 is derived from the theorem of total probability as,

$$I = \int_{\boldsymbol{\Theta}} e^{-J(\boldsymbol{\theta})} d\boldsymbol{\theta} = \int_{\boldsymbol{\Theta}} \frac{f(\boldsymbol{\theta})\big|_{\mathcal{M}=\mathcal{M}_j}}{k_5} d\boldsymbol{\theta} = \frac{1}{k_6} \tag{27}$$

6. Extension to model-class selection

This formulation can be generalized to the case when several model classes $\mathcal{M}$ are candidates to idealize the real excitation-observation. Including this variable into the inverse problem formulation will allow to derive the model-class selection as a particular case of inverse problem.

As introduced in the preceding subsection, the probabilistic nature of the reconstruction is partly motivated by the fact that the model itself may not necessarily reproduce the experimental setup, but is just an approximation. If several models are candidates based on different hypothesis about the system, the former probabilistic formulation of the inverse problem will be shown to be able to provide information to rank them. The bottom idea is the following: if the model-class (based on the candidate hypothesis) is considered as an uncertain discrete variable, its probability can eventually be extracted as a marginal probability from Equation 15. The probability of each model-class will therefore have the sense of degree of certainty of being true in the sense that the probabilistic conjunction of certainty (or information) provided by the experimental measurements and model are coherent.

Let model class $\mathcal{M}$ denote an idealized mathematical model hypothesized to simulate the experimental system, whereas model $\boldsymbol{\theta}$ denotes the set of constants of physical parameters that the model-class depends on. Different model classes can be formulated and hypothesized to idealize the experimental system, and each of them can be used to solve the probabilistic inverse problem in the previous section, yielding different values of model parameters but also physically different sets of parameters. To select among the infinitely many possible model classes that can be defined, user judgement is a criteria, but a probabilistic one can also be defined based on their compatibility between prior information $f^0(\mathcal{D}, \boldsymbol{\theta}, \mathcal{M})$ on observations $\mathcal{D}$, model parameters $\boldsymbol{\theta}$ and model class $\mathcal{M}$, and probabilistic model information given by $f^m(\mathcal{D}, \boldsymbol{\theta}, \mathcal{M})$. The conjunction of probabilities established in Equation 15 will be adopted instead of Bayes' theorem, for its generality [22].

The goal is to find the probability $f(\mathcal{M})$, understood as a measure of plausibility of a model class $\mathcal{M}$ [23]. It can be derived as the marginal probability of the posterior probability $f(\mathcal{D}, \boldsymbol{\theta}, \mathcal{M})$ defined in Equation 15,

$$\begin{aligned} f(\mathcal{M}) &= \int_{\mathfrak{D}} \int_{\boldsymbol{\Theta}} f(\mathcal{D}, \boldsymbol{\theta}, \mathcal{M}) d\boldsymbol{\theta} d\mathcal{D} \\ &= k_1 f^0(\mathcal{M}) \int_{\mathfrak{D}} \int_{\boldsymbol{\Theta}} f^0(\mathcal{D}) f^0(\boldsymbol{\theta}) f^m(\mathcal{D}, \boldsymbol{\theta}, \mathcal{M}) d\boldsymbol{\theta} d\mathcal{D} \end{aligned} \tag{28}$$

If no prior information is provided by the user about the class $f^0(\mathcal{M}) = \mu(\mathcal{M}) \Rightarrow k_1 f^0(\mathcal{M}) = k_6$. Furthermore, this theorem involves exactly the same integral as that for the constant k_5, i.e., allowing to reuse the integral in Equation 27,

$$f(\mathcal{M}) = k_6 \int_{\boldsymbol{\Theta}} \frac{f(\boldsymbol{\theta})\big|_{\mathcal{M}=\mathcal{M}_i}}{k_5} d\boldsymbol{\theta} = k_6 \int_{\boldsymbol{\Theta}} e^{-J(\boldsymbol{\theta})} d\boldsymbol{\theta} = k_6 I \tag{29}$$

where the normalization constant k_6 can be solved from the theorem of total probability over all model classes $\mathfrak{M}$ in order to obtain probabilities in the classical sense,

$$\sum_{\mathfrak{M}} f(\mathcal{M}) = 1 \tag{30}$$

Variations of the probability density at good or bad models may exceed the floating point representation range of a standard operating system. To override this limitation, an alternative computation is proposed in the logarithmic scale. This is carried out redefining all involved PDF in the $-ln$ scale and redefining their relationships as $\tilde{f} = -ln(f)$ or $f = e^{-\tilde{p}}$. Variables expressed in the logarithmic scale are tagged with a tilde (˜).

Once the plausibility $f(\mathcal{M})$ is computed for every class, its value allows to rank the models accordingly to how compatible they are with the observations. This also allows us to find a correct trade-off between model simplicity and fitting to observations [22, 20].

7. Conclusions

The inverse problem of parameter reconstruction from experimental data when a model is available has been derived in a probabilistic way from the theory of conjunction of states of information from observations combined with models. This approach is proposed as an alternative to the logical inference using Bayes theorem, as it relies on different statistical axioms and may be useful. Among them, the input-output relationship needs not to be causal, the axioms that allow the concept conditional probability are not needed, and the incorporation of additional sources of information beyond observation and model become straightforward. As an example of the latter, the extension to model-class selection is derived in a simple way. The validity of the approach is supported by the fact that the final computations are the same for a typical linear gaussian inverse problem.

Acknowledgments

The authors would like to thank the Ministry of Science and Innovation of Spain through Project DPI2010-17065 (MICINN), the Ministry of Education for FPU grants AP2009-4641, AP2009-2390 and the European Union for projects P11-CTS-8089 and GGI3000IDIB. The authors would also like to thank the California Institute of Technology (Caltech, USA) which kindly hosted the two last authors during a part of the course of this work.

References

[1] Leonard J Savage. *The foundations of statistics*. Courier Dover Publications, 1972.
[2] Alan Hajek. Interpretations of probability. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford, 2012.
[3] R. A. Fisher. *Statistical Methods and Scientific Inference*. New York: Hafner Press., 1956.
[4] J. Neyman and E. S. Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, 20A:175–294, 1928.

[5] R. V. Mises. *Probability, statistics and truth, revised English edition.* New York: Macmillan, 1957.

[6] Hans Reichenbach. *The theory of probability.* University of California Press, 1949.

[7] Henry Ely Kyburg. *Probability and inductive logic.* Macmillan London, 1970.

[8] Paul Humphreys. Why propensities cannot be probabilities. *The philosophical review*, 94(4):557–570, 1985.

[9] James H Fetzer. Scientific knowledge. causation, explanation, corroboration. *Boston Studies in the Philosophy of Science New York, NY*, 69, 1981.

[10] John Maynard Keynes. A treatise on probability. *Diamond*, 3(2):12, 1909.

[11] Rudolf Carnap. *Philosophy and logical syntax.* 70. K. Paul, Trench, Trubner & Co., ltd., 1935.

[12] Richard T Cox. Probability, frequency and reasonable expectation. *American journal of physics*, 14:1, 1946.

[13] Frank P Ramsey. Truth and probability (1926). *The foundations of mathematics and other logical essays*, pages 156–198, 1931.

[14] Ian Hacking. *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference.* Cambridge University Press, 1975.

[15] Bruno De Finetti. *Foresight: its logical laws in subjective sources.* Wiley, 1964.

[16] Seymour Geisser. *Predictive inference: an introduction*, volume 55. CRC Press, 1993.

[17] Albert Tarantola. *Inverse Problem Theory and Methods for Model Parameters Estimation.* Siam, 2005.

[18] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[19] Karl R Popper. The logic of scientific discovery. *London: Hutchinson*, 1, 1959.

[20] J.L. Beck. Bayesian system identification based on probability logic. *Structural Control and Health Monitoring*, 17(7):825–847, 2010.

[21] James L. Beck and Siu-Kui Au. Bayesian Updating of Structural Models and Reliability using Markov Chain Monte Carlo Simulation. *Journal of Engineering Mechanics*, 128(4):380–391, 2002.

[22] James Litz. Beck and Ka Veng. Yuen. Model selection using response measurements: Bayesian probabilistic approach. *Journal of Engineering Mechanics*, 130:192, 2004.

[23] R.T. Cox. *The algebra of probable inference.* The Johns Hopkins University Press, 1961.