Characterization of the interwell space by Bayesian methods

Fernando S. de Moraes and John A. Scales
Center for Wave Phenomena
Colorado School of Mines

ABSTRACT
When facing the task of mapping rock attributes in a reservoir, geophysicists rely on information from the subsurface, such as well logs, and from surface geophysical methods. The most common approach to this problem is to interpret each data set separately and integrate the results manually. Bayesian methods provide a formal way of integration by using probability density functions to represent the information contained in each data set. However, difficulties associated with the construction and manipulation of high-dimensional probabilities have prevented wide application of these methods.

Using first principles of probability theory, one can derive the general Bayesian formulation for the geophysical inverse problem. This formulation can then be specialized to explore the much smaller length scale involved in the subsurface information when compared with the geophysical information. This difference makes it possible to use the subsurface information to independently construct probability densities for specific parameters that account for neighboring information such as well logs and other geological data. The principle of maximum entropy is used to construct these probability densities, thus giving a unified theoretical treatment that generalizes geostatistical methods.

When the specialized Bayesian formulation is used, n-dimensional distributions are replaced by approximations to their marginals, leading to a series of one-dimensional problems involving one parameter at a time. This gives a method that is able to systematically integrate diverse geological and geophysical information while maintaining practical feasibility.

Key words: Bayesian inference, Maximum entropy, Geostatistics.

CONTENTS

1 INTRODUCTION

With the depletion of new exploratory frontiers, there is a growing demand for technology capable of investigating more subtle plays or of giving a more detailed description of reservoirs. Examples of these advances are new discoveries made in the Gulf of Mexico, using 3-D prestack migration methods to image under the salt, and the increasing use of geostatistics to interpolate well data (e.g., porosity), also using surface data (e.g., seismic data) to improve reservoir descriptions. The main goal of this work is to investigate how these demands can be incorporated into geophysical studies aimed at making inferences about the geology of a particular area through the solution of an inverse problem.

Traditionally, inversion methodologies have adopted the perspective of the geophysical data set. That is, we seek to learn about the resolution and the level of noise in the data in order to select a suitable parameterization of the subsurface (Backus and Gilbert, 1968; Parker, 1977). Equivalent results also can be accomplished by regularization, which employs damping and smoothing operators (Constable et al., 1987). This approach faces the fact that the geophysical data can resolve only certain averages of the true earth, thus we must lower demand for resolution in a given model to the appropriate level either by reparameterization or by smoothing. The final results are usually very coarse models that may not serve the inference objectives. For instance, the inversion may generate interpretative models with questionable applicability in describing the geology, which are not useful given the current demands discussed above.

To meet these demands, inversion methodologies cannot rely on only one data set. Instead, additional information from well logs, petrophysics and geological interpretation are necessary to produce satisfactory results, once the parameterization is fixed by the demands for knowledge of a particular area. The order of the day is to explain aspects of the geology, not the geophysical data. To reach an appropriate description of the geology, we usually need parameterizations on a much finer scale than those determined by analysis of the resolution of the geophysical data. Then, according to this framework, the geophysical information becomes relatively less important when a large amount of additional information is available. In such cases, the geophysical data play a role well described as aiding the interpolation of geological attributes. That is, the final picture will be mainly determined by the subsurface information, but will also satisfy the averages that are well resolved by the geophysical data.

Thus, the main challenge is to develop inversion methodologies that can rigorously integrate any additional data sets that carry subsurface information such as well logs, geological sections and core measurements. This task is complicated by the diversity of these data sets. For example, geological information is usually qualitative or semi-quantitative, while all inverse calculations need quantitative data. Furthermore, different types of geophysical data are in different units, they respond to a change in different geological attributes and they have essentially different mathematical formulations.

In addition, to properly solve an inverse problem, it is necessary to address the uncertainties present in the calculations. These include the noise in the geophysical data, the uncertainties in the prior information and in the mathematical model (forward problem). The ability to include all uncertainties in an inverse calculation gives the means of judging proposed models. That is, it provides answer to questions such as: how reliable are the estimates? Or do we need more information (data) to resolve a given objective?

In principle, both issues of integration and uncertainty analysis can be handled by the Bayesian approach. But the burden becomes how to construct the probabilistic models and extract information from a combination of them, which is never a trivial task for high-dimensional problems. This paper attempts to address these difficulties of the Bayesian methodology by replacing multidimensional distributions with approximations to their marginals. Thus, instead of solving the multidimensional problem directly, we can replace it by a series of one-dimensional problems for one parameter at a time. This is possible when we take advantage of the much smaller length scales in the subsurface (local) information in comparison with the geophysical (global) information, thus allowing for inferences on one parameter independently of the others based on subsurface information alone.

In the next section, we derive the general Bayesian formulation of the geophysical inverse problem from the first principles of probability theory. Then, in Section 3, we discuss two methods for deriving distributions for the parameters that account for all the local information: the principle of maximum entropy and kriging. After that, in Section 4, we show how the general Bayesian formulation can be modified to incorporate the probability densities derived from the subsurface information. Finally, we solve a simple computational problem that illustrates the methodology.

2 BAYESIAN METHODOLOGY

The goal of scientific inference is to combine information from physical theories, data and other background knowledge to draw conclusions about a given physical system. Because the total information is always incomplete, that is, insufficient for precise conclusions, the mechanism of inference must involve plausible reasoning rather than deductive reasoning. Of course, it is always possible to modify the original problem so that the information available may provide an exact solution, but then, only a mathematical problem is being solved, not an inference one.

The application of the Bayes' theorem to inference problems addresses the same goal: draw conclusions about a physical system upon incomplete information. But it is not a stand-alone tool for inference, it is just a natural result of using probability theory to conduct plausible (probable) reasoning, which is what we need to

solve problems. In this context, probabilities are used to represent degrees of belief or plausibility about the truth of a proposition.

These ideas were rigorously substantiated by Cox (1946) and Cox (1961). In these works, he started from a desiderata of consistency to show that any method of inference that uses real numbers to represent degrees of plausibility must be based on the basic rules of probability theory (see below); otherwise, it will be inconsistent.

Despite this relatively recent result, the foundations of probability theory as logic, as we know it today, began over two centuries ago. To better understand the context of Bayesian inference, it is useful to make a brief excursion through the sequence of developments in the field. According to the historical accounts given by Jaynes (1978), the Bayes' theorem was derived to solve a problem that it did not solve completely. The problem is that before we can apply the basic rules of probability we need some initial probability assignments. There is nothing within the basic rules that tells how to assign those initial probabilities; thus an extension to these rules is needed. The first formal attempt to do that was the principle of insufficient reason presented by Bernoulli in 1713.

Bernoulli's principle states that if the available evidence give us no reason to consider one proposition neither more nor less favorable than another, the only honest possible probability assignment lends them equally likely. This principle can be extended to consider multiple sets of propositions in which case the probability assignment is the ratio between the number of favorable possibilities (m) and the total number of possibilities (N), that is, $p = \frac{m}{N}$. But Bernoulli himself recognized the enormous limitations with his principle, namely, that it requires the ability to break a given problem into a set of exhaustive equally likely possibilities, which is an impossible task for most problems of inference. He then reasoned as Jaynes tells us: "if a probability p cannot be calculated in the manner $p = \frac{m}{N}$ by direct application of the principle of insufficient reason, then in some cases we may still reason backwards and estimate the ratio $\frac{m}{N}$ approximately by observing frequencies in many trials". Thus Bernoulli started seeking a formal relation between theoretical probabilities and observable frequencies, which yielded the weak law of large numbers for binomially-distributed variables. Later developments by Laplace and de Moivre (Gaussian approximation to the binomial distribution) ended up forming the basis for the central limit theorem. But all these results still assume independence, and they also relate to the direct problem. That is, they all describe the situation where, from known population numbers m and N, we obtain the theoretical probability that particular sample numbers m' and N' can be observed. This indicates that the difficulties observed with the principle of insufficient reason are still present.

The real challenge was to find the solution for the inverse problem: from observable facts to obtain a theoretical probability of a given proposition. Pursuing this problem Bayes found the beta distribution as a solution for the inversion of the binomial distribution (a particular case of the Bayes' theorem), and Laplace found the Bayes' theorem in the form we know it today. The Bayes theorem represents the complete solution for the inversion problem, which was aimed to avoid use of the principle of insufficient reason for direct assignment of probabilities. Ironically, however, the only useful results obtained by Laplace employed a particular case of the Bayes' theorem that relied on priors of the type $\frac{1}{N}$ given by that same principle. According to Jaynes, this was not because Laplace failed to obtain the general form of the Bayes formula, which he wrote down, but rather because he did not have any principle for finding priors in cases where the prior information fails to render the possibilities equally likely. Important breakthroughs were achieved in the works of Shannon (1948) and Jaynes (1957) with the generalization of the principle of maximum entropy to assign prior probabilities.

This clearly illustrates the point that Bayes' theorem is only a consequence of using probability theory as logic, which within itself does not contain rules for assigning probabilities so the calculations can get started. This is why additional principle such as the principles of insufficient reason and maximum entropy are needed.

To introduce the fundamental equations of probability theory, consider the propositions A and B plus some background information C. The sum rule tells that for an exhaustive, mutually exclusive set of probabilities, the sum of their individual probability is the unity. This can be expressed simply in terms of proposition A and its complement $\overline{A}$ by

$$P(A \mid C) + P(\overline{A} \mid C) = 1. \tag{1}$$

The product rule gives the probability for $A\,B$, which is commutative, is given by

$$\begin{aligned} P(A\,B \mid C) &= P(A \mid B\,C)\,P(B \mid C) \qquad \text{or} \\ &= P(B \mid A\,C)\,P(A \mid C). \end{aligned} \tag{2}$$

From (1) and (2), we can find the sum rule for non-exclusive propositions A, B as

$$P(A + B \mid C) = P(A \mid C) + P(B \mid C) - P(A\,B \mid C). \tag{3}$$

Bayes' theorem comes directly from the commutativity of the product rule, which is essentially a logical statement. Thus, if the problem is to draw conclusions about the truth of proposition A, we can equate the two terms in Equation (2) to get

$$P(A \mid B\,C)\,P(B \mid C) = P(B \mid A\,C)\,P(A \mid C) \tag{4}$$

or

$$P(A \mid B\,C) = \frac{P(B \mid A\,C)\,P(A \mid C)}{P(B \mid C)}. \tag{5}$$

If proposition A can be broken down into alternative propositions denoted by $A = \{A_1, A_2, \ldots, A_M\}$ and the problem is to draw conclusions about a particular A_i, we get

$$P(A_i \mid B\,C) = \frac{P(B \mid A_i\,C)\,P(A_i \mid C)}{\sum_{j=1}^{M} P(B \mid A_j\,C)\,P(A_j \mid C)}, \tag{6}$$

where the denominator is still just $P(B \mid C)$. These different forms of the Bayes' theorem are only two out of great number of possibilities, which depend on the specific question being asked. They further illustrate that a single application of these elementary forms of Bayes' theorem gives only a probability, not a probability distribution.

This is an important issue when dealing with continuous probabilities, where the given quantities appearing in the conditional probabilities do not take on simple values. They take on a distribution of values, which makes necessary a marginalization process before a variable can be set to a particular number. To substantiate the understanding of such issues, it is useful to investigate the relationship between marginal and conditional probability distributions. Thus, consider the joint distribution of x and y as given by $p(x, y)$. The marginal distribution for x can be written as

$$\int p(x, y)\,dy = \int p(x \mid y)\,p(y)\,dy = p(x)\,, \tag{7}$$

while the conditional probability for x given y is

$$p(x \mid y = y^*) = \left.\frac{p(x, y)}{\int p(x, y)\,dx}\right|_{y=y^*} = \frac{p(x, y^*)}{p(y^*)}, \tag{8}$$

where the superscript $*$ will always denote a particular value taken by a variable. However, if we let $p(y) = \delta(y - y^*)$, then Equation (7) yields

$$\int p(x \mid y)\,\delta(y - y^*)\,dy = p(x \mid y = y^*)\,, \tag{9}$$

which is the same as the conditional probability. A generalization of this for any $p(y)$ is given by the mean-value theorem (see e.g., Gradshteyn and Ryzhik, 1980, p. 211), which is valid if $p(x \mid y)$ is a bounded, continuous function and $p(y)$ is positive and integrable in the interval $[y_l, y_u]$. In this case, we are guaranteed the existence of a value $\xi \in [y_l, y_u]$, such that

$$\int_{y_l}^{y_u} p(x \mid y)\,p(y)\,dy = p(x \mid y = \xi)\,, \tag{10}$$

using the fact that $p(y)$ integrates to one. From this, it becomes clear that a marginal is always equivalent to a certain conditional distribution, corresponding to an unknown value of the variable being given or integrated out (Figure 1). Only for the particular case of the delta function can the value y^* be immediately substituted for y.

2.1 Basic formulation

Using these ideas, it is possible to derive a Bayesian formulation for geophysical inverse problems. The first step is to define the information that will enter the calculation. This leads to a joint state of knowledge that can be represented by a joint probability distribution, which can be decomposed using the product rule given in Equation (2).

The geophysical inverse problem combines prior information ($\mathcal{I}$), information from theory and from data measurements. Thus, consider the parameterized earth model represented by the vector $\mathbf{m} \in R^M$, and data predicted from theory and obtained by measurements as given by the vectors $\mathbf{d}_t$, $\mathbf{d}_o \in R^N$, respectively.

It also important to establish all sources of uncertainty involved in $\mathbf{d}_t$ and $\mathbf{d}_o$. Those can be introduced as noise terms in the expressions for the data vectors, which we define as given by

$$\mathbf{d}_t^* = \mathbf{d}_t + \mathbf{n}_m + \mathbf{n}_g\,, \tag{11}$$

$$\mathbf{d}_o^* = \mathbf{d}_o + \mathbf{n}_o\,, \tag{12}$$

where $\mathbf{n}_m$ are the errors associated with the model parameterization, $\mathbf{n}_g$ the errors generated by approximations in the forward modeling theory, $\mathbf{n}_o$ are the observational errors, $\mathbf{d}_t^*$ are predicted data values and $\mathbf{d}_o^*$ the observed data values. The predicted data values are the result of computations $\mathbf{d}_t^* = \mathbf{g}(\mathbf{m})$, where $\mathbf{g}$ is the forward modeling operator. The motivation to use the theory to predict data is the underlying assumption that perfect theoretical data ($\mathbf{d}_t$) match the observed data exactly in the absence of observational errors; that is, $\mathbf{d}_o = \mathbf{d}_t$. From this we might as well write that

$$\mathbf{g}(\mathbf{m}) = \mathbf{d}_o + \mathbf{n}_m + \mathbf{n}_g, \tag{13}$$

$$\mathbf{d}_o^* = \mathbf{d}_t + \mathbf{n}_o. \tag{14}$$

With the above definitions, the representation of the joint state of knowledge is given by the joint probability $p(\mathbf{m}, \mathbf{d}_t, \mathbf{d}_o \mid \mathcal{I})$, which can be decomposed by the product rule to yield

$$f(\mathbf{m} \mid \mathbf{d}_t, \mathbf{d}_o, \mathcal{I})\,h(\mathbf{d}_t, \mathbf{d}_o \mid \mathcal{I}) = s(\mathbf{m} \mid \mathcal{I})\,g(\mathbf{d}_o \mid \mathbf{m}, \mathcal{I})\,r(\mathbf{d}_t \mid \mathbf{m}, \mathbf{d}_o, \mathcal{I}) \tag{15}$$

$f(\mathbf{m} \mid \mathbf{d}_t, \mathbf{d}_o, \mathcal{I}) =$

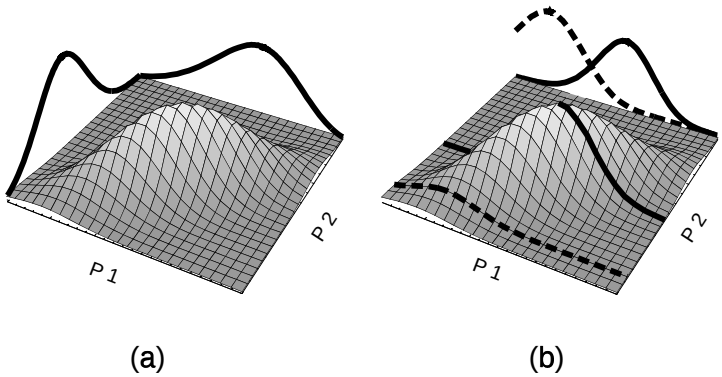

Figure 1. Comparison between marginal (a) and conditional (b) probability distributions, using a two parameter (P 1, P 2) normal distribution. It illustrates that conditional distributions can vary significantly depending on the value of the variable being given, while marginals are uniquely defined.

$$\frac{s(\mathbf{m}\,|\,\mathcal{I})\,g(\mathbf{d}_o\,|\,\mathbf{m},\mathcal{I})\,r(\mathbf{d}_t\,|\,\mathbf{m},\mathbf{d}_o,\mathcal{I})}{h(\mathbf{d}_t,\mathbf{d}_o\,|\,\mathcal{I})}\,. \tag{16}$$

The interpretation of each term in Equation (16) is:

- $r(\mathbf{d}_t\,|\,\mathbf{m},\mathbf{d}_o,\mathcal{I})$ is the probability of the data predicted by the theory. It accounts for all modeling errors, including geological noise and approximations in the forward modeling operator. It also corresponds to the so-called *likelihood function*, which name is used to emphasize the inference role of this distribution when the parameters $\mathbf{m}$ are considered as variables instead of given quantities. We discuss this distribution in more detail in the next section.
- $s(\mathbf{m}\,|\,\mathcal{I})$ is the prior distribution for the model parameters. It accounts for any information at hand before the acquisition of the geophysical data.
- $h(\mathbf{d}_t,\mathbf{d}_o\,|\,\mathcal{I})$ is a term originating from the left side of Equation (15). Thus, for consistency reasons, it cannot be assigned. It must be determined by the terms in the right hand side of the equality, which is why it is only a normalization factor.
- $g(\mathbf{d}_o\,|\,\mathbf{m},\mathcal{I})$ is the data regarded as a probability distribution. It represents the ranking of data vectors according to their plausibility, given our knowledge of the model parameters and prior information. It does not involve the theoretical data, and therefore avoids using the forward modeling operator. If viewed as a likelihood function, it can represent only the qualitative interpretation that can be done using a display of a geophysical data such as seismic sections or gravity maps.

By using this Bayesian formulation and based on the arguments given in last section, it becomes clear that observed data cannot be just substituted into a Bayes' formula, since they are not exactly known numbers, but rather a distribution of values given by g. Observed data must be introduced by a marginalization process, which can be described using Equation (15) by

$$\int_R f(\mathbf{m}\,|\,\mathbf{d}_t,\mathbf{d}_o,\mathcal{I})\,h(\mathbf{d}_t,\mathbf{d}_o\,|\,\mathcal{I})\,d\mathbf{d}_o = s(\mathbf{m}\,|\,\mathcal{I})\int_R g(\mathbf{d}_o\,|\,\mathbf{m},\mathcal{I})\,r(\mathbf{d}_t\,|\,\mathbf{m},\mathbf{d}_o,\mathcal{I})\,d\mathbf{d}_o. \tag{17}$$

Equation (17) is similar to that given by Tarantola (1987), but it has the advantage of having been derived from the product rule of probability theory.

To examine Equation (17) further, it is useful to consider two important particular cases. The first case is when the probability $g(\mathbf{d}_o\,|\,\mathbf{m},\mathcal{I})$ is strongly concentrated; i.e., can be approximated by $\delta(\mathbf{d}_o-\mathbf{d}_o^*)$. Then, the integrals in Equation (17) can be written as

$$\int_R f(\mathbf{m}\,|\,\mathbf{d}_t,\mathbf{d}_o,\mathcal{I})\,h(\mathbf{d}_t,\mathbf{d}_o\,|\,\mathcal{I})\,d\mathbf{d}_o = s(\mathbf{m}\,|\,\mathcal{I})\int_R r(\mathbf{d}_t\,|\,\mathbf{m},\mathbf{d}_o,\mathcal{I})\,\delta(\mathbf{d}_o-\mathbf{d}_o^*)\,d\mathbf{d}_o, \tag{18}$$

which can be evaluated to yield

$$f(\mathbf{m}\,|\,\mathbf{d}_t,\mathcal{I}) = \frac{s(\mathbf{m}\,|\,\mathcal{I})\,r(\mathbf{d}_o^*\,|\,\mathbf{m},\mathcal{I})}{h(\mathbf{d}_t\,|\,\mathcal{I})}, \tag{19}$$

using the assumption that $\mathbf{d}_t=\mathbf{d}_o$.

The other important case is when the predicted data distribution is strongly concentrated; that is, modeling errors are negligible. In this case the probability for $\mathbf{d}_t$ is given by $\delta(\mathbf{d}_t-\mathbf{d}_t^*)$ or equivalently by $\delta(\mathbf{d}_o-\mathbf{d}_t^*)$. Then Equation (17) becomes

$$\int_R f(\mathbf{m}\,|\,\mathbf{d}_t,\mathbf{d}_o,\mathcal{I})\,h(\mathbf{d}_t,\mathbf{d}_o\,|\,\mathcal{I})\,d\mathbf{d}_o = s(\mathbf{m}\,|\,\mathcal{I})\int_R \delta(\mathbf{d}_o-\mathbf{d}_t^*)\,g(\mathbf{d}_o\,|\,\mathbf{m},\mathcal{I})\,d\mathbf{d}_o, \tag{20}$$

which can be evaluated as

$$f(\mathbf{m}\,|\,\mathbf{d}_t,\mathcal{I}) = \frac{s(\mathbf{m}\,|\,\mathcal{I})\,g(\mathbf{d}_t^*\,|\,\mathbf{m},\mathcal{I})}{h(\mathbf{d}_t\,|\,\mathcal{I})}. \tag{21}$$

The similarities between Equations (19) and (21) are striking. We discuss these two extreme cases in more detail in the next two sections.

This formulation also reveals that the usual formulation of a Bayesian inverse problem given in the literature

$$f(\mathbf{m} \,|\, \mathbf{d}_o, \mathcal{I}) = \frac{s(\mathbf{m} \,|\, \mathcal{I})\, r(\mathbf{d}_o \,|\, \mathbf{m}, \mathcal{I})}{h(\mathbf{d}_o \,|\, \mathcal{I})} \tag{22}$$

has some logical difficulties in the way that the observed data values $\mathbf{d}_o^*$ are substituted for $\mathbf{d}_o$. To see that, recall that we can substitute numbers for variables in conditional probabilities, and fully account for the uncertainties (i.e., achieve the correspondence to a marginalization procedure), only when the associated distribution is strongly concentrated. Otherwise, conditional probabilities give only partial account for the uncertainties, since they correspond to looking at the intersection of the joint distribution with a particular hyperplane plane $\mathbf{d}_o = \mathbf{d}_o^*$ (see Figure 1). Thus, if the data are sufficiently precise to justify immediate substitution, what is the probability assignment for r? If the data are not precise and we do make the substitution, how can we guarantee that the particular conditional probability really gives a realistic picture of the uncertainties for $\mathbf{m}$? Questions such as these do not appear when the theoretical data $\mathbf{d}_t$ are introduced.

2.1.1 The likelihood function

To better understand the meaning of the likelihood function, consider the expression for the theoretical data given by Equation (11). It is a sum of the forward modeling equation and some error terms. The forward model $\mathbf{g}(\mathbf{m})$ can be taken as precise numbers, since computer errors can be made arbitrarily small, at least in principle. Therefore, we have that

$$\begin{aligned} r(\mathbf{d}_t \,|\, \mathbf{m}, \mathbf{d}_o, \mathcal{I}) &= r(\mathbf{n}_t \,|\, \mathbf{m}, \mathbf{d}_o, \mathcal{I}) \\ &= r(\mathbf{d}_t - \mathbf{g}(\mathbf{m}) \,|\, \mathbf{m}, \mathbf{d}_o, \mathcal{I}), \end{aligned}$$

where $\mathbf{n}_t \equiv -(\mathbf{n}_m + \mathbf{n}_g)$ (see Equation (11)). The above expression may seem strange at first glance, but all it is saying is that the probability for $\mathbf{d}_t$ remains invariant upon translations, which are indistinguishable by the prior information. However, $\mathbf{d}_t$ can only be assessed through the observed data through the expression given in Equation (14). Then, by substitution, we get

$$\begin{aligned} &r(\mathbf{d}_t - \mathbf{g}(\mathbf{m}) \,|\, \mathbf{m}, \mathbf{d}_o, \mathcal{I}) \\ &\quad = r(\mathbf{d}_o^* - \mathbf{n}_o - \mathbf{g}(\mathbf{m}) \,|\, \mathbf{m}, \mathbf{d}_o, \mathcal{I}), \qquad (23) \\ &\quad = r(\mathbf{d}_o - \mathbf{g}(\mathbf{m}) \,|\, \mathbf{m}, \mathbf{d}_o, \mathcal{I}), \qquad (24) \end{aligned}$$

which gives the same as when the assumption $\mathbf{d}_t = \mathbf{d}_o$ is directly applied to the first equation.

To investigate further, consider again the case where observational errors are negligible, as in Equation (18). The integration over $\mathbf{d}_o$ for the likelihood determined by Equation (24) gives

$$f(\mathbf{m} \,|\, \mathbf{d}_t, \mathcal{I}) = \frac{s(\mathbf{m} \,|\, \mathcal{I})\, r(\mathbf{d}_o^* - \mathbf{g}(\mathbf{m}) \,|\, \mathbf{m}, \mathcal{I})}{h(\mathbf{d}_t \,|\, \mathcal{I})}. \tag{25}$$

This likelihood function has the same form as usually given in the literature, although it involves different arguments.

If a normal distribution with zero mean and covariance matrix $\mathbf{C}_t$ is assigned to the data misfit , we can write

$$\begin{aligned} r(\mathbf{d}_t \,|\, \mathbf{m}, \mathcal{I}) &= (2\pi)^{-\frac{N}{2}} \,|\, \mathbf{C}_t \,|^{-\frac{1}{2}} \\ &\quad \exp\left\{-\frac{1}{2}\left[\mathbf{d}_o^* - \mathbf{g}(\mathbf{m})\right]^T \mathbf{C}_t^{-1} \left[\mathbf{g}(\mathbf{m}) - \mathbf{d}_o^*\right]\right\}. \end{aligned} \tag{26}$$

For the linear problem, the computed data are given by

$$\mathbf{d}_t = \mathbf{G}\mathbf{m} + \mathbf{n}_t, \tag{27}$$

where $\mathbf{G}$ is a $N \times M$ matrix. Then, Equation (26) becomes

$$\begin{aligned} r(\mathbf{d}_t \,|\, \mathbf{m}, \mathcal{I}) &= (2\pi)^{-\frac{N}{2}} \,|\, \mathbf{C}_t \,|^{-\frac{1}{2}} \\ &\quad \exp\left[-\frac{1}{2}(\mathbf{d}_o^* - \mathbf{G}\mathbf{m})^T \mathbf{C}_t^{-1} (\mathbf{d}_o^* - \mathbf{G}\mathbf{m})\right]. \end{aligned} \tag{28}$$

For the linear case, we can rewrite Equation (28) to emphasize its inference role (i.e., $\mathbf{m}$ unknown), using that

$$\begin{aligned} &(\mathbf{d}_o^* - \mathbf{G}\mathbf{m})^T \mathbf{C}_t^{-1} (\mathbf{d}_o^* - \mathbf{G}\mathbf{m}) = \\ &\quad (\mathbf{d}_o^* - \mathbf{G}\hat{\mathbf{m}})^T \mathbf{C}_t^{-1} (\mathbf{d}_o^* - \mathbf{G}\hat{\mathbf{m}}) \\ &\quad + (\mathbf{m} - \hat{\mathbf{m}})^T \mathbf{G}^T \mathbf{C}_t^{-1} \mathbf{G} (\mathbf{m} - \hat{\mathbf{m}}), \end{aligned} \tag{29}$$

where $\hat{\mathbf{m}} = (\mathbf{G}^T \mathbf{C}_t^{-1} \mathbf{G})^{-1} \mathbf{G}^T \mathbf{C}_t^{-1} \mathbf{d}_o^*$ is an estimated parameter vector. This assumes that $\mathbf{G}^T \mathbf{C}_t^{-1} \mathbf{G}$ is invertible. If we substitute this relation into Equation (28), we get

$$\begin{aligned} r(\mathbf{d}_t \,|\, \mathbf{m}, \mathcal{I}) &= (2\pi)^{-\frac{N}{2}} \,|\, \mathbf{C}_t \,|^{-\frac{1}{2}} \exp\left[S(\hat{\mathbf{m}})\right] \\ &\quad \exp\left\{-\frac{1}{2}\left[(\mathbf{m} - \hat{\mathbf{m}})^T \mathbf{G}^T \mathbf{C}_t^{-1} \mathbf{G} (\mathbf{m} - \hat{\mathbf{m}})\right]\right\}, \end{aligned} \tag{30}$$

where $S(\hat{\mathbf{m}}) = (\mathbf{d}_o^* - \mathbf{G}\hat{\mathbf{m}})^T \mathbf{C}_t^{-1} (\mathbf{d}_o^* - \mathbf{G}\hat{\mathbf{m}})$ is the estimated misfit value. This is an important form of likelihood function due to the fact that the location parameters $\hat{\mathbf{m}}$ are completely determined by the data. Because of that it is called in Bayesian literature the *data translated likelihood.*

2.1.2 The observed data distribution

The arguments for the construction of the data distribution (g in Equation 17) are similar to those used for the likelihood function. The exception here is that we are not allowed to introduce the forward modeling equation

since $\mathbf{d}_t$ does not enter the definition of g. However, we show below that for the extreme case where the modeling errors are negligible, the forward modeling equation is introduced into g by marginalization.

Using the definition given in Equation (12), we can write

$$\begin{aligned} g(\mathbf{d}_o \,|\, \mathbf{m}, \mathcal{I}) &= g(\mathbf{n}_o \,|\, \mathbf{m}, \mathcal{I}) \\ &= g(\mathbf{d}_o^* - \mathbf{d}_o \,|\, \mathbf{m}, \mathcal{I}), \end{aligned} \tag{31}$$

which can be justified in the same way as before: the uncertainty about $\mathbf{d}_o$ does not change upon translations, which are not distinguished by the prior information.

To investigate further, we return to the case where modeling errors are negligible, given by Equation (20). For this case, the likelihood function is a delta function $\delta(\mathbf{d}_t - \mathbf{d}_t^*)$, or $\delta(\mathbf{d}_o - \mathbf{d}_t^*)$ when using the assumption that $\mathbf{d}_t = \mathbf{d}_o$. The computed data values $\mathbf{d}_t^*$ are given by the forward modeling equation $\mathbf{g}(\mathbf{m})$. When this and Equation (31) are introduced into Equation (17), the integration over $\mathbf{d}_o$ gives

$$f(\mathbf{m} \,|\, \mathbf{d}_t, \mathcal{I}) = \frac{s(\mathbf{m} \,|\, \mathcal{I})\, g(\mathbf{d}_o^* - \mathbf{g}(\mathbf{m}) \,|\, \mathbf{m}, \mathcal{I})}{h(\mathbf{d}_t \,|\, \mathcal{I})}. \tag{32}$$

Next, consider again the example of a normal distribution with zero mean and covariance matrix $\mathbf{C}_o$ being assigned for g. Then, we can write

$$\begin{aligned} g(\mathbf{d}_o \,|\, \mathbf{m}, \mathcal{I}) &= (2\pi)^{-\frac{N}{2}} \,|\, \mathbf{C}_o \,|\, ^{-\frac{1}{2}} \\ &\exp\left\{ -\frac{1}{2} \left[\mathbf{d}_o^* - \mathbf{g}(\mathbf{m})\right]^T \mathbf{C}_o^{-1} \left[\mathbf{d}_o^* - \mathbf{g}(\mathbf{m})\right] \right\}. \end{aligned} \tag{33}$$

All other equations of the previous section can be derived from this, in the same way. The only notable difference is that, for this case, we are dealing with the covariance matrix for the observational errors ($\mathbf{C}_o$) as opposed to the covariance matrix of the modeling errors ($\mathbf{C}_t$) of the previous section.

3 PRIOR PROBABILITIES

As discussed in the previous sections, inverse problems face the task of producing a detailed picture of the subsurface based on the geophysical data and other prior information. While Bayesian methods seem to be the appropriate framework to combine all the available information, practical difficulties associated with high-dimensionality of the probability distributions limit its application. Thus, further developments are needed to overcome the limitations.

One possible direction to approach the problem is by looking into the characteristics of the geophysical and the subsurface information. When we do this, it is possible to recognize that they are defined over fundamentally different length scales. That is, geophysical information has a global nature, which means that all parameters must be considered simultaneously in order to solve an inverse problem. In contrast, it is possible to explore the subsurface information to draw conclusions about a particular parameter of the model in an isolated way. In fact, this local aspect of the subsurface information has been routinely exploited in other fields such as geostatistics and it is fundamentally an interpolation problem. In this section, two different methods for deriving local probability distributions, the maximum entropy and the indicator kriging, are discussed and their relationships investigated. Then, in the next section, we show how to incorporate local distributions into the full geophysical inverse problem.

A schematic representation of local probabilities computed from a well log is given in Figure 2. At a well location, if the measurement errors are negligible when compared with the inversion errors, the local probability is approximately a delta function. As we go away form the borehole the uncertainty increases, as-reflected in broader and broader local distributions. The intensity of the attenuation of the probabilities, as go away from the borehole, is given by the degree of spatial variability of the medium and by the interpolation algorithm.

The ability to incorporate local distributions into the geophysical inverse corresponds also to the ability to incorporate an unlimited amount of prior information. This becomes clear after going over the methodologies for computing these distributions. Also, when necessary, the same methodologies used to derive the local distributions can be used to construct the full multidimensional distribution. To see this, consider the case where a set of direct measurements of the desired parameter has been made. If we include these measurements in the prior information $\mathcal{I}$ and denote the model parameters by m_j, $j = 1, \ldots, M$, we find that the local probabilities are actually an approximation for the conditional probability given by $f(m_j \,|\, \mathcal{I})$. Because these distributions are estimated one at a time, the current distribution being estimated can also conditioned to the mean value of others previously estimated distributions. Then the multivariate distribution can be obtained by taking the product of all the individual distributions, which is just an application of the product rule:

$$\begin{aligned} f(\mathbf{m} \,|\, \mathcal{I}) &= f(m_1 \,|\, \mathcal{I})\, f(m_2 \,|\, m_1, \mathcal{I}) \cdots \\ &\quad f(m_{M-1} \,|\, m_1, m_2, \ldots, m_{M-2}, \mathcal{I}) \\ &\quad f(m_M \,|\, m_1, m_2, \ldots, m_{M-1}, \mathcal{I}). \end{aligned} \tag{34}$$

But the main goal is to work directly with the local distributions. Thus, we begin the discussion on how to come up with the local probabilities by examining the maximum entropy method. Then, we look into geostatistical

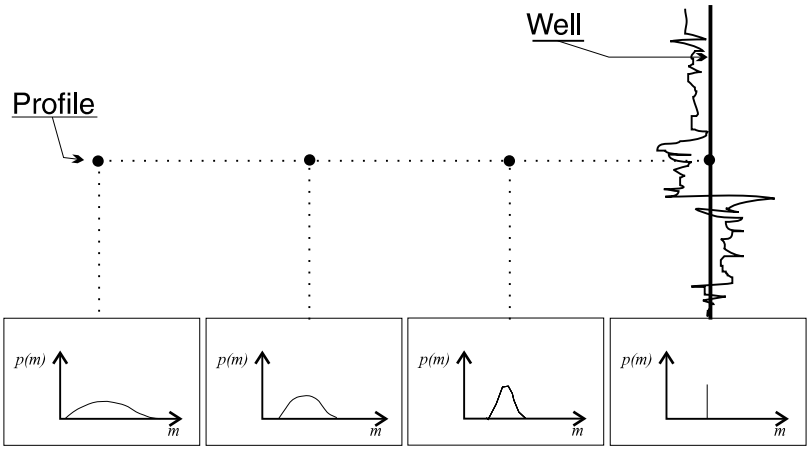

Figure 2. Schematic representation of the nature of the local probabilities. Given a set of well log data, the local uncertainties grow as we go away form the borehole. At the borehole, if the measurements have absolute precision the distribution is a delta function.

methods that have close association to maximum entropy methods

3.1 Maximum entropy

The information entropy was first defined by Shannon (1948) to measure the amount of information (or conversely the uncertainty) in a given distribution. Later, Jaynes (1957) found that the information entropy can be used as formal rule for assigning prior probabilities, as we discuss below. Because entropy is a measure of the uncertainty of a probability distribution, the maximum entropy principle provides the most conservative distribution that agrees with all the given constraints. According to Shannon (1948), the entropy of a discrete probability density function is given by

$$H(p) = -\sum_{i=1}^{n} p_i \log p_i.$$

Many scientific problems involve continuous variables. Thus, it is important to extend the definition of entropy to the continuous case. This extension, which can be found in Jaynes (1963), leads to the relative entropy. However, careful examination of the issues involved suggests that, for inference purposes, relative entropy should be adopted even in the discrete case when the correspondence to continuous variables is important (Gouveia et al., 1996). As the name suggests, relative entropy is a measure of uncertainty in probability p relative to another probability q, which is known *a priori*. Thus, if we consider these probabilities associated with the variable x, the relative entropy is given by

$$H(p;q) = -\sum_{i=1}^{n} p(x_i) \log \frac{p(x_i)}{q(x_i)}, \tag{35}$$

in the case x is a discrete variable, or

$$H(p;q) = -\int_R p(x) \log \frac{p(x)}{q(x)} dx, \tag{36}$$

when x is a continuous variable.

An equivalent measure is given by the *cross-entropy*, which is the negative of the relative entropy as defined above. The cross-entropy was first defined by Kullback (1959) under the name of *directed divergence*. Shore and Johnson (1981) provide an extensive collection of properties of the cross-entropy with proofs.

Maximization of the continuous entropy functional (36) is the variational problem over p, given by max $H(p;q)$, subject to the normalization

$$\int_R p(x)\, dx = 1, \tag{37}$$

and to other constraints given in the form of expectations $\langle f_k(x) \rangle$

$$\int_R f_k(x)\, p(x)\, dx = \mu_k, \qquad k = 1, ..., K, \tag{38}$$

where μ_n is a numerical value that can be computed from the available data. Usually $f_k(x) = x^n$, where $n = 1, 2, \ldots$, which makes $\langle f_k(x) \rangle$ the moments of the distribution. The solution for this problem (Jaynes, 1957, Gouveia et al., 1996) is given by

$$p(x) = q(x) \exp\left[-\lambda_0 - \sum_{k=1}^{K} \lambda_k f_k(x)\right], \tag{39}$$

or

$$p(x) = Z^{-1} q(x) \exp\left[-\sum_{k=1}^{K} \lambda_k f_k(x)\right], \tag{40}$$

with

$$Z \equiv \exp(\lambda_0) = \int_R q(x) \exp\left[-\sum_{k=1}^{K} \lambda_k\, f_k(x)\right] dx\,. \tag{41}$$

The complete solution requires the determination of the Lagrange multipliers λ_k and the reference prior $q(x)$. A well-known result of the maximum entropy problem is that when only the first two moments of the distribution are given as constraints, the maximum entropy distribution relative to a uniform distribution approaches the normal distribution, as the limits of the uniform distribution go to infinity (Gouveia et al., 1996).

3.1.1 Conditional probabilities

To show a simple example of how the principle of maximum entropy can be used to derive local probability densities, consider a set of measurements denoted by the vector $\mathbf{w} = (w_1, \ldots, w_N) \in R^N$, which provide some information about a single parameter m. The measurements need not to correspond to the same attribute. Thus, we can define the vector

$$\mathbf{x} = \begin{bmatrix} m \\ \mathbf{w} \end{bmatrix},$$

and its corresponding covariance matrix

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix}, \tag{42}$$

where

$$\mathbf{C}_{11} = \mathbf{C}_{mm}\,;\ \mathbf{C}_{22} = \mathbf{C}_{ww}\,;\ \mathbf{C}_{12}^T = \mathbf{C}_{21} = \mathbf{C}_{wm},$$

which are 1×1, $N \times N$ and $N \times 1$ matrices, respectively. Also, the corresponding inverse covariance matrix is defined as $\mathbf{V} = \mathbf{C}^{-1}$, which can be partitioned in the same form as Equality (42) and each part is given by

$$\begin{aligned}
\mathbf{V}_{11} &= \left(\mathbf{C}_{11} - \mathbf{C}_{12}\mathbf{C}_{22}^{-1}\mathbf{C}_{21}\right)^{-1}, \\
\mathbf{V}_{12} &= -\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\left(\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\right)^{-1}, \\
\mathbf{V}_{21} &= -\mathbf{C}_{22}^{-1}\mathbf{C}_{21}\left(\mathbf{C}_{11} - \mathbf{C}_{12}\mathbf{C}_{22}^{-1}\mathbf{C}_{21}\right)^{-1}, \\
\mathbf{V}_{22} &= \left(\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\right)^{-1},
\end{aligned} \tag{43}$$

assuming that all matrices are invertible. As mentioned in the last section, the conditional probability $f(m \,|\, \mathbf{w})$, when only the first two moments are being considered, can be written as the ratio of two Gaussians

$$f(m \,|\, \mathbf{w}) = \frac{G(\mathbf{x})}{G(\mathbf{w})} \tag{44}$$

where

$$\begin{aligned}
G(\mathbf{x}) &= (2\pi)^{-\frac{N+1}{2}} \,|\, \mathbf{V} \,|\, ^{-\frac{1}{2}} \\
&\quad \exp\left[-\frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{V} (\mathbf{x} - \bar{\mathbf{x}})\right],
\end{aligned} \tag{45}$$

where

$$\bar{\mathbf{x}} = \begin{bmatrix} m_0 \\ \bar{\mathbf{w}} \end{bmatrix}$$

are the mean values for $\mathbf{x}$, and

$$\begin{aligned}
G(\mathbf{w}) &= (2\pi)^{-\frac{N}{2}} \,|\, \mathbf{C}_{22} \,|\, ^{-\frac{1}{2}} \\
&\quad \exp\left[-\frac{1}{2}(\mathbf{w} - \bar{\mathbf{w}})^T \mathbf{C}_{22}^{-1} (\mathbf{w} - \bar{\mathbf{w}})\right].
\end{aligned} \tag{46}$$

Equation (44) can be expanded as

$$\begin{aligned}
f(m \,|\, \mathbf{w}) &= \alpha \exp\Big\{-\frac{1}{2}\big[\mathbf{V}_{11}(m - m_0)^2 + \\
&\quad 2(m - m_0)\mathbf{V}_{12}(\mathbf{w} - \bar{\mathbf{w}}) + \\
&\quad (\mathbf{w} - \bar{\mathbf{w}})^T(\mathbf{V}_{22} - \mathbf{C}_{22}^{-1})(\mathbf{w} - \bar{\mathbf{w}})\big]\Big\}.
\end{aligned}$$

The mean of a normal distribution also coincides with the mode. Thus we may find the conditional expectation $\langle m \,|\, w_1, \ldots, w_N \rangle$ by maximizing the above conditional distribution, which is equivalent to minimizing the argument of the exponential function. This process represented by

$$\mathbf{V}_{11}(m - m_0) + \mathbf{V}_{12}(\mathbf{w} - \bar{\mathbf{w}}) = 0\,.$$

Note that $\mathbf{V}_{11} = \left(\mathbf{C}_{11} - \mathbf{C}_{12}\mathbf{C}_{22}^{-1}\mathbf{C}_{21}\right)^{-1}$ is just a scalar, and

$$\mathbf{V}_{12} = -\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\left(\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\right)^{-1},$$

a vector in R^N. From this, we can find an estimator for the conditional expectation as given by

$$\hat{m} = m_0 - \mathbf{V}_{11}^{-1}\mathbf{V}_{12}(\mathbf{w} - \bar{\mathbf{w}})\,. \tag{47}$$

To find the conditional variance $\left\langle \sigma_m^2 \,|\, w_1, \ldots, w_N \right\rangle$, we can just rewrite the argument of the exponential function in terms of $\hat{m}$, beginning with the expansion

$$\begin{aligned}
f(m \,|\, \mathbf{w}) = \alpha \exp\Big\{-\frac{1}{2}\big[&\mathbf{V}_{11}(m^2 - 2m_0 m + m_0^2) + \\
&2\mathbf{V}_{12}(\mathbf{w} - \bar{\mathbf{w}})m - 2m_0\mathbf{V}_{12}(\mathbf{w} - \bar{\mathbf{w}}) + \\
&(\mathbf{w} - \bar{\mathbf{w}})^T(\mathbf{V}_{22} - \mathbf{C}_{22}^{-1})(\mathbf{w} - \bar{\mathbf{w}})\big]\Big\}
\end{aligned} \tag{48}$$

or

$$\begin{aligned}
f(m \,|\, \mathbf{w}) = \alpha \exp\Big\{-\frac{1}{2}\big[&\mathbf{V}_{11}(m^2 + m_0^2) - \\
&2\mathbf{V}_{11}[m_0 - \mathbf{V}_{11}^{-1}\mathbf{V}_{12}(\mathbf{w} - \bar{\mathbf{w}})]m - \\
&2m_0\mathbf{V}_{12}(\mathbf{w} - \bar{\mathbf{w}}) + \\
&(\mathbf{w} - \bar{\mathbf{w}})^T(\mathbf{V}_{22} - \mathbf{C}_{22}^{-1})(\mathbf{w} - \bar{\mathbf{w}})\big]\Big\}.
\end{aligned} \tag{49}$$

The second term of the argument can be recognized as the estimator given by Equation (47). By making the substitution and completing the squares we get

$$f(m \mid \mathbf{w}) = \alpha \exp\left\{-\frac{1}{2}\left[\mathbf{V}_{11}(m-\hat{m})^2 - (\mathbf{w}-\bar{\mathbf{w}})^T \mathbf{V}_{12}^T \mathbf{V}_{11}^{-1} \mathbf{V}_{12}(\mathbf{w}-\bar{\mathbf{w}}) + (\mathbf{w}-\bar{\mathbf{w}})^T(\mathbf{V}_{22}-\mathbf{C}_{22}^{-1})(\mathbf{w}-\bar{\mathbf{w}})\right]\right\}$$

or

$$f(m \mid \mathbf{w}) = \alpha \exp\left\{-\frac{1}{2}\left[\mathbf{V}_{11}(m-\hat{m})^2 - (\mathbf{w}-\bar{\mathbf{w}})^T(\mathbf{V}_{22}-\mathbf{C}_{22}^{-1} - \mathbf{V}_{12}^T\mathbf{V}_{11}^{-1}\mathbf{V}_{12})(\mathbf{w}-\bar{\mathbf{w}})\right]\right\}. \quad (50)$$

Since the second term is constant, it can be incorporated into the multiplicative constant α to yield

$$f(m \mid \mathbf{w}) = \alpha \exp\left[-\frac{1}{2}\mathbf{V}_{11}(m-\hat{m})^2\right], \quad (51)$$

where

$$\left\langle \sigma_m^2 \mid w_1, \ldots, w_N \right\rangle = \mathbf{V}_{11}^{-1}. \quad (52)$$

The estimator for the conditional mean, given by Equation (47), can be recognized as the simple kriging equation (see, e.g., Equation (8) in Lesson II of Journel, 1989). Thus, we see that when we have the ability to specify the mean and covariance between variables we can easily obtain the local conditional moments, which lead to the specification of the local normal marginal distribution.

However, in many situations is useful to specify other higher-order moments, such as

$$\langle m \mid w_1, \ldots, w_N \rangle,$$
$$\left\langle m^2 \mid w_1, \ldots, w_N \right\rangle,$$
$$\left\langle m^3 \mid w_1, \ldots, w_N \right\rangle \text{ and}$$
$$\left\langle m^4 \mid w_1, \ldots, w_N \right\rangle.$$

If we did that, we could make use of the maximum entropy principle to find the corresponding probability density function (pdf). In this way, information such skewness and kurtosis would be included in the calculations. McCullagh (1987) discusses methods for calculation of the conditional higher-order moments.

A simple alternative approach is provided by the indicator kriging, which we discuss next.

3.2 Geostatistical approach

Geostatistical methods for estimating local probability distributions are based on the so-called indicator variables, which is defined below. The strengths of the method are twofold. First, is the ability to account for diverse information, which is one of the main difficulties in inversion. Second is that this procedure is specially designed to give approximations to the local distributions we seek. To understand how this is done, let's see how indicators are used to represent probabilities.

3.2.1 Indicator representation of probabilities

We begin this section by recalling that for one particular value y^* taken by the variable y, we have in general that the probability of the event $y \leq y^*$ is given by

$$P(y \leq y^*) = F(y) \text{ and } f(y) = \frac{dF(y)}{dy},$$

where capital letters (F) are used to represent cumulative distribution functions (cdf) and lower case letters (f) are used for probability density functions.

In our problem, we can look at a single model parameter m (e.g., rock density) and its value m^* at a particular point $\mathbf{r}$. The task is to estimate the probability that $m \leq m^*$ given some information in the neighborhood. Thus, if we again consider a data vector $\mathbf{w} = (w_1, \ldots, w_N)$, corresponding to subsurface measurements of the same attribute as m (e.g., density log), we can apply the indicator transformation, which can be defined by

$$i_j(\mu) = \begin{cases} 0, & m = w_j > \mu, \\ 1, & m = w_j \leq \mu, \end{cases} \quad (53)$$

where μ is a selected threshold value for m. Since i is a binary variable, its probabilities are given by the Bernoulli distribution

$$f(i) = P(m \leq \mu)[1 - P(m \leq \mu)].$$

Therefore, the expected value of i is

$$\langle i(\mu) \rangle = 1 \cdot P(m \leq \mu) + 0 \cdot P(m > \mu), \quad (54)$$
$$= P(m \leq \mu) = F(m), \quad (55)$$

where $F(m)$ is the cdf for the model parameter. This means that we can approximate the probability $F(m)$ by averages of indicators. One simple average is just

$$P(m \leq \mu) = \frac{1}{N}\sum_{j=1}^{N} i_j(\mu), \quad (56)$$

which gives the same importance to any value i regardless of its location or the threshold value μ. If we want to infer the uncertainty about m given indicators in the neighborhood, the spatial variability should be accounted for (i.e., the closer the samples, the more likely they are to have similar values). So a weighted version of (56) can be defined as

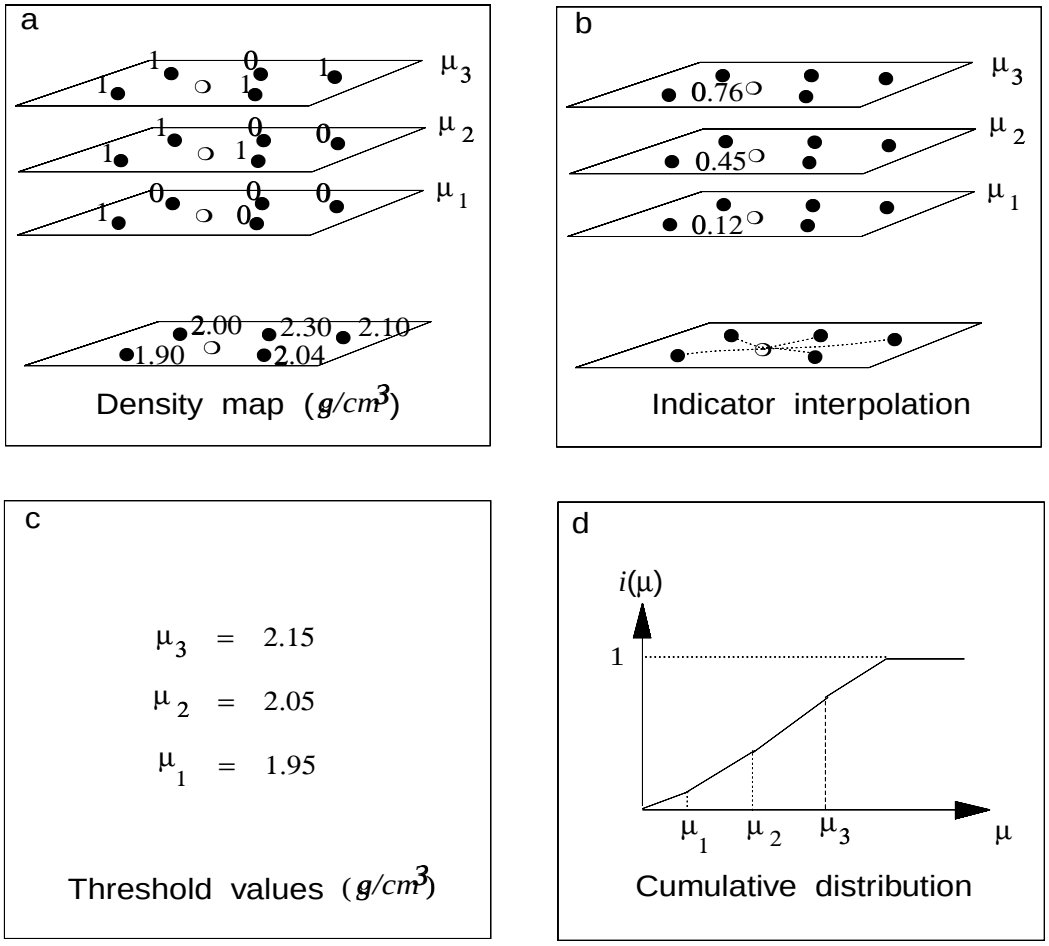

Figure 3. Schematic figure showing: (a) the indicator transformation of rock density measurements (full dot) at three threshold values, (b) indicator interpolation to a location represented by the empty dot for each threshold map, (c) the threshold values corresponding to the transformation in (a) and (d) a sketch of a plot of the cumulative distribution for density obtained by the indicator regression in (b).

$$P(m \leq \mu) = \sum_{j=1}^{N} a_j(\mu) i_j(\mu). \tag{57}$$

This version, which corresponds to an interpolation problem for $i(\mu)$, provides an estimate for $P(m \leq \mu)$ at $\mathbf{r}$ for a given threshold value μ. If this procedure is repeated for a set of appropriately chosen threshold values μ_k, $k = 1, 2, ..., K$, we can build up the discrete approximation for the cdf for the model parameter, namely $F(\mu) \approx F(m)$. For that, the requirements are: the application of the indicator transform to all the N measurements $\mathbf{w}$, for each of the K threshold values, and the determination of the $N \times K$ weights $a_j(\mu)$ (Figure 3). It is important to notice that in order for equation (57) to provide the building block for a legitimate cdf, the following order relations should be observed:

$$\begin{array}{ll} F(\mu_k) \in [0,1] & \text{for all } k, \\ F(\mu_k) \leq F(\mu_l) & \text{for all } \mu_k \leq \mu_l. \end{array} \tag{58}$$

The approximation of $F(m)$ through the indicator formalism can be better understood pictorially, by looking at the indicator transform $i(\mu)$ for different thresholds, at a particular location, as a stack values. This corresponds to examining Figure 3 (a) along a vertical line at a given rock density observation (e.g. 2.04 g/cm^3). In fact, at a location where we have made an observation of the model parameter (e.g., core density

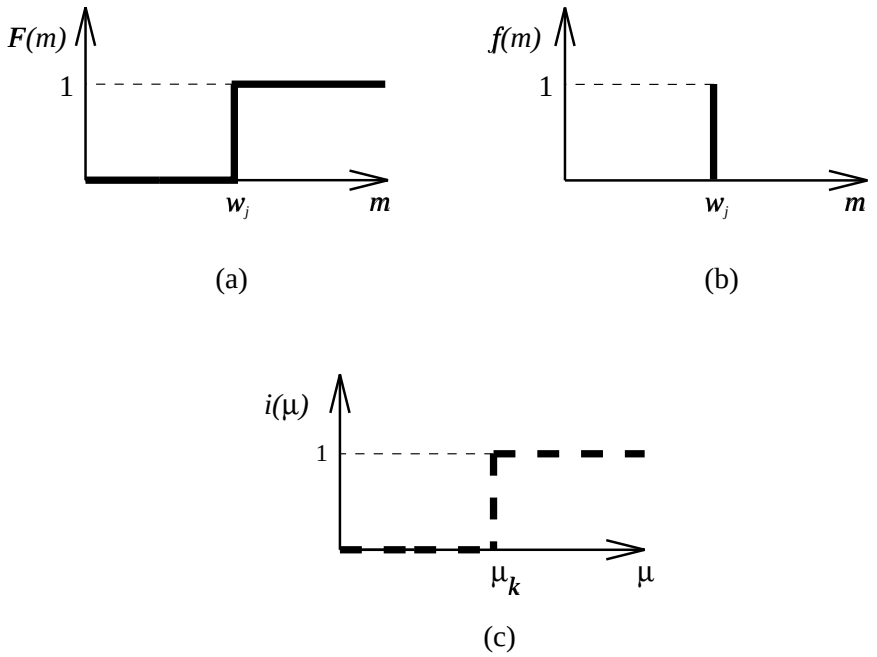

Figure 4. Schematic comparison between the cumulative distribution (a), the corresponding probability density function (b) and the discrete indicator approximation to the cumulative distribution (c) in a location where the attribute corresponding to the parameter m has been observed. In the case where there is a threshold value $\mu_k = w_j$, the approximation will be exact.

measurement), the corresponding cdf is a step function at the observed value w_j (Figure 4). This corresponds to a probability density distribution given by a delta function at w_j, implying an error-free measurement. The indicator variable gives the exact position of the step if there is a threshold value $\mu_k = w_j$. At another location where a sample value is not available, the indicator variable is determined by an interpolation procedure (Figure 3 (b)) such as Equation (57). Thus, we have that

$$P(m \leq \mu) = \langle i(\mu) \rangle \approx i^*(\mu), \tag{59}$$

where i^* is the estimated value for i.

3.2.2 Indicator coding for different types of information

Indicator variables give a great flexibility in representing information. To see that, consider three levels of knowledge in any earth science study: the knowledge prior to any experiment on the area under study, that from secondary information and that from direct observation of the parameter under study.

The knowledge prior to any experiment comes from information of physics or previous experiments in different areas under similar settings. The knowledge from secondary information is gained by performing experiments (geophysical or geological) on parameters that are somehow related to the ones in which we are specifically interested in. For example, in rock density estimation, indirect information on density can be given by the knowledge of seismic velocity. Also, qualitative geological information in the form of lithologic sections derived from drillholes and well logs fall in the same category. Finally, the direct observations are made by in situ measurements, which can be either density logs or core measurements. In summary, if we consider the of rock-density estimation problem, useful information can be provided by

- density measurements from well logs or core analysis,
- velocity values derived from seismic data interpretation,
- physical interval constraints,
- geological information on rock types and
- expertise.

The indicator coding of all these distinct types of information is just a generalization of the indicator transform (53), given by

$$i(\mathbf{r}, \mu) = \begin{cases} 0, & m(\mathbf{r}) > m_l\,, \\ \text{undefined}, & m(\mathbf{r}) \in [m_l, m_u], \\ 1, & m(\mathbf{r}) \leq m_u\,, \end{cases} \tag{60}$$

where m_l and m_u are the lower and upper limits of an interval defining the precision of the corresponding inform-

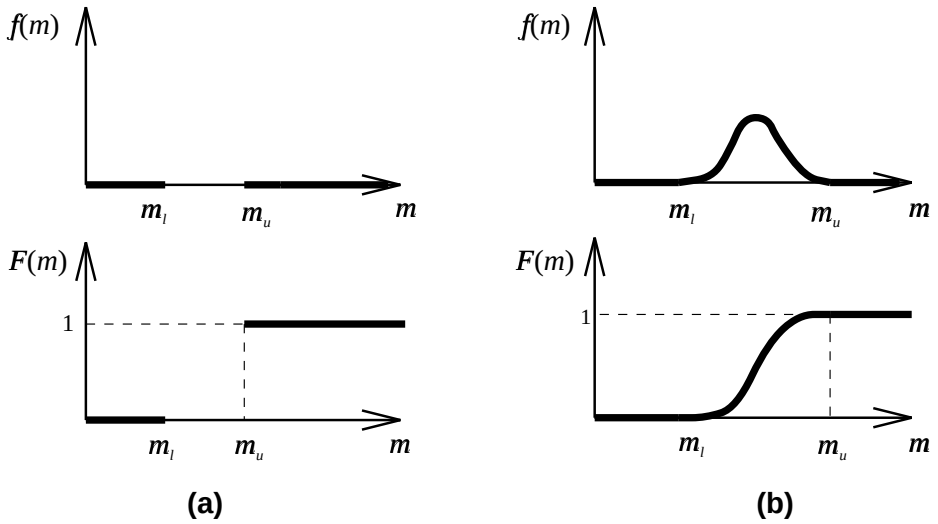

Figure 5. Cumulative distribution and the corresponding probability density function for: (a) interval constraints (type 2) and (b) prior distribution (type 3).

μ_9	1	1	1		1
μ_8	1	1	1		0.9
μ_7	1	1	0.9		0.8
μ_6	1	?	0.6		0.8
μ_5	0	?	0.4	$\longrightarrow$	0.7
μ_4	0	?	0.2		0.4
μ_3	0	0	0.1		0.2
μ_2	0	0	0		0.1
μ_1	0	0	0		0
threshold	type 1	type 2	type 3		posterior cdf

Table 1. Schematic representation for the prior cdfs of the three types described in the text and the corresponding updated posterior cdf.

ation. If the associated errors can be neglected, $m_l = m_u$ and the interval has zero amplitude. For this case Equation (60) reduces to (53), corresponding to a step cdf (Figure 4), which will be referred as a cdf of "type 1." When $m_l < m_u$, the "undefined interval" can be left empty (Figure 5a) to represent an interval constraint limiting the possible occurrence of m, or it can be filled with values between 0 and 1 denoting a prior cdf for m (Figure 5b). The former cdf will be referred as a cdf of "type 2," and the latter will be referred as a cdf of "type 3."

Making reference to the cdfs just defined, it is possible to associate the information, given above, to the proper type. Direct rock density measurements can be coded as a cdf of type 1, when the associated errors are negligible in comparison to those of an inversion procedure. Quantitative secondary information can either be coded as cdfs of type 1 or type 3, while qualitative information, such as geological information, should be represented as type 2 or 3. For quantitative secondary data, such as seismic velocity values, the cdf type depends on how precisely the secondary information can determine the value for m. It usually yields cdfs of type 3. Finally, expertise can be coded either as type 2 or type 3.

To summarize the three types for coding the subsurface information, it is possible to represent each cdf type in the form of indicator columns (i.e., cdfs at various locations), each line corresponding to a threshold value μ_k. An indicator interpolation scheme is the updating procedure leading to the approximation of the cdf for m at a given location $\mathbf{r}$. The interpolation step can be interpreted as a maximum entropy calculation in the same

way as presented in Section 3.1.1, since it usually relies on first and second order-statistics of indicators varying continuously on the interval $[0, 1]$. Table 1 is a schematic representation of this process, considering the case of nine threshold values. The right-hand column is a representation of the updated cdf, which synthesizes all the information available.

Deutsch and Journel (1992) provide an extensive suite of Fortran routines that implement indicator kriging and cokriging, which can be used to estimate the local conditional cdfs.

4 APPROACHES TO THE SOLUTION

As we have seen, the Bayesian approach to inverse problems involves probability distributions in high dimensions both in the data and model spaces. The formal solution is the product of these multivariate probability distributions, and the desired information about a particular parameter can be obtained by marginalization. This requires a sophisticated set of tools, such as Monte Carlo integration and importance sampling, to carry out multidimensional integrations (e.g., Scales and Tarantola, 1994; Mosegaard and Tarantola, 1995).

We seek alternative approaches to the estimation of the parameter vector $\mathbf{m}$, without solving the full multivariate problem just discussed. In addition, the solution must incorporate all the prior information ($\mathcal{I}$), processed into local conditional distributions using the methods discussed in the previous section. To proceed further we have to face a problem arising from the fact that the geostatistical regression is of local nature. Univariate conditional probabilities are estimated for individual parameters, independently of the other, considering only the information in the neighborhood. In contrast, the geophysical inverse problem cannot be done for a single parameter, independently, due to forward modeling relation given by Equation (11). This contrast is basically the origin of the alternative formulations discussed below, which hinge on the idea of reducing the dimensionality of the geophysical problem, using marginalization theory.

The idea is to seek the solution one parameter at a time in a way similar to the geostatistical problem. To carry out this approach, the parameter vector $\mathbf{m}$ must be divided into two parts. One part is just the specific parameter to be estimated in the current iteration, denoted by $\mathbf{m}_1$, and the other part is composed of the remaining components of $\mathbf{m}$, denoted by $\mathbf{m}_2$. That is,

$$\mathbf{m} = \begin{bmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \end{bmatrix}, \tag{61}$$

where $\mathbf{m}_1 \in R$ and $\mathbf{m}_2 \in R^{M-1}$.

Two distinct treatments for $\mathbf{m}_2$ can be considered. To treat $\mathbf{m}_2$ as nuisance parameters*, or as fixed quantities known *a priori*. Now the statement for our problem is: we want a measure of the uncertainty of a particular subset of parameters of the geophysical model $\mathbf{m}_1$ given the synthetic data $\mathbf{d}_t$, the observations $\mathbf{d}_o$ and the prior information $\mathcal{I}$. Then, the general Bayesian formulation given by Equation (17) can be expanded further to give, for the first case

$$\int_R f(\mathbf{m}_1, \mathbf{m}_2 \,|\, \mathbf{d}_t, \mathbf{d}_o, \mathcal{I})\, h(\mathbf{d}_t, \mathbf{d}_o \,|\, \mathcal{I})\, d\mathbf{d}_o = t(\mathbf{m}_1 \,|\, \mathcal{I})\, v(\mathbf{m}_2 \,|\, \mathbf{m}_1, \mathcal{I}) \int_R r(\mathbf{d}_t \,|\, \mathbf{m}_1, \mathbf{m}_2, \mathbf{d}_o, \mathcal{I})\, g(\mathbf{d}_o \,|\, \mathbf{m}_1, \mathbf{m}_2, \mathcal{I})\, d\mathbf{d}_o. \tag{62}$$

and for the second case

$$\int_R f(\mathbf{m}_1 \,|\, \mathbf{m}_2, \mathbf{d}_t, \mathbf{d}_o, \mathcal{I})\, h(\mathbf{d}_t, \mathbf{d}_o, \mathbf{m}_2 \,|\, \mathcal{I})\, d\mathbf{d}_o = t(\mathbf{m}_1 \,|\, \mathcal{I})\, v(\mathbf{m}_2 \,|\, \mathbf{m}_1, \mathcal{I}) \int_R r(\mathbf{d}_t \,|\, \mathbf{m}_1, \mathbf{m}_2, \mathbf{d}_o, \mathcal{I})\, g(\mathbf{d}_o \,|\, \mathbf{m}_1, \mathbf{m}_2, \mathcal{I})\, d\mathbf{d}_o. \tag{63}$$

For simplicity, let us investigate more closely the particular case when the modeling errors are negligible given by Equation (21). When we apply the division of the parameter vector and treat $\mathbf{m}_2$ as nuisance parameters, this equation becomes

$$p(\mathbf{m}_1, \mathbf{m}_2 \,|\, \mathbf{d}_t, \mathcal{I}) = t(\mathbf{m}_1 \,|\, \mathcal{I})\, \frac{u(\mathbf{d}_t^*, \mathbf{m}_2 \,|\, \mathbf{m}_1, \mathcal{I})}{h(\mathbf{d}_t \,|\, \mathcal{I})}, \tag{64}$$

or when $\mathbf{m}_2$ are fixed parameters, it becomes

$$p(\mathbf{m}_1 \,|\, \mathbf{m}_2, \mathbf{d}_t, \mathcal{I}) = t(\mathbf{m}_1 \,|\, \mathcal{I})\, \frac{u(\mathbf{d}_t^*, \mathbf{m}_2 \,|\, \mathbf{m}_1, \mathcal{I})}{v(\mathbf{m}_2 \,|\, \mathcal{I})\, g(\mathbf{d}_t^* \,|\, \mathbf{m}_2, \mathcal{I})}, \tag{65}$$

where

$$u(\mathbf{d}_t^*, \mathbf{m}_2 \,|\, \mathbf{m}_1, \mathcal{I}) = v(\mathbf{m}_2 \,|\, \mathbf{m}_1, \mathcal{I})\, g(\mathbf{d}_t^* \,|\, \mathbf{m}_1, \mathbf{m}_2, \mathcal{I}).$$

Both formulations lead to a product of two terms. One of them is the prior distribution $t(\mathbf{m}_1 \,|\, \mathcal{I})$ that corresponds to the local estimation problem described in Section 3. Thus, the objective here is to study the other term, involving the distributions v and g (h is the normalizing distribution), to determine a computing scheme for the posterior distribution for the parameter $\mathbf{m}_1$.

To put the developments so far into perspective, we have discussed three alternative Bayesian formulations for the geophysical inverse problem that are given by Equations (17), (64) and (65). The first alternative is

* Term usually employed in Bayesian inference to denote parameters one is obligated to infer, but has no immediate interest in.

the full calculation of the posterior distribution; it does not involve the division of the parameter vector (61). To use this approach and the methods for deriving priors (one-dimensional distributions of the type $t(\mathbf{m}_1 \mid \mathcal{I})$) of the previous section, it is necessary to employ a scheme, such as Equation (34), to build the multivariate prior distribution for the parameters from each univariate distribution. But then, the difficulties associated with the high dimensionality, that we have been trying to avoid, need to be addressed. The second approach that treats $\mathbf{m}_2$ as nuisance parameters is the most complete because it reduces the dimensionality of the problem and also takes into account all uncertainties associated with the parameters. Because of that all practical developments follow the nuisance parameter approach, which is discussed in detail in the next section. Finally, the third approach that considers $\mathbf{m}_2$ as numbers fixed *a priori* is limited, since it does not fully account for all the uncertainties (see Figure 1 and associated discussions). This approach looks only at a particular conditional probability for the parameter $\mathbf{m}_1$ given the values for $\mathbf{m}_2$. Because of that, this approach is only briefly discussed below.

4.1 Nuisance parameters

For this case, the $\mathbf{m}_2$ parameters must be eliminated from the problem to produce a final result representing a marginal distribution for $\mathbf{m}_1$. This can be accomplished by integration since

$$w(\mathbf{m}_1 \mid \mathbf{d}_t, \mathcal{I}) = \int_R f(\mathbf{m}_1, \mathbf{m}_2 \mid \mathbf{d}_t, \mathcal{I})\, d\mathbf{m}_2, \tag{66}$$

where R represents the proper domain of integration of $\mathbf{m}_2$. If we apply this idea to Equation (64), we get

$$w(\mathbf{m}_1 \mid \mathbf{d}_t, \mathcal{I}) = \kappa\, t(\mathbf{m}_1 \mid \mathcal{I}) \int_R v(\mathbf{m}_2 \mid \mathbf{m}_1, \mathcal{I})\, g(\mathbf{d}_t^* \mid \mathbf{m}_1, \mathbf{m}_2, \mathcal{I})\, d\mathbf{m}_2, \tag{67}$$

where κ is the normalizing constant that incorporates the distribution h. In the integrand of this equation, we have a pdf on the data space, which is the likelihood function g, and a multivariate prior conditional distribution for $\mathbf{m}_2$ (v). The latter is a $(M-1)$-dimensional distribution, what means that we still need to handle integration in high dimensional space. But the difference here is that this is now an iterative procedure, where at each step a different parameter is considered as $\mathbf{m}_1$ and the others ($\mathbf{m}_2$) will be eliminated from the problem. Thus, it is intuitive to expect that we may discard some information about the parameters $\mathbf{m}_2$ as long as sufficient information about $\mathbf{m}_1$ is introduced through t. Following this idea, the prior information can be divided into two parts: a part that defines only a normal distribution (e.g., mean and covariance information) and another that complements this information (e.g., higher-order moment information) in such way that

$$\mathcal{I}_T = \mathcal{I}_N + \mathcal{I}_C. \tag{68}$$

That is, the total information ($\mathcal{I} \rightarrow \mathcal{I}_T$) equals the logical sum of normal information ($\mathcal{I}_N$) and its complement ($\mathcal{I}_C$). Furthermore, we may assume that for parameters $\mathbf{m}_2$ only the information $\mathcal{I}_N$ is used in each step, then we can write

$$v(\mathbf{m}_2 \mid \mathbf{m}_1, \mathcal{I}_T) \longrightarrow v(\mathbf{m}_2 \mid \mathbf{m}_1, \mathcal{I}_N)$$

and

$$v(\mathbf{m}_2 \mid \mathbf{m}_1, \mathcal{I}_N) = \frac{p(\mathbf{m}_2, \mathbf{m}_1 \mid \mathcal{I}_N)}{q(\mathbf{m}_1 \mid \mathcal{I}_N)}. \tag{69}$$

Equation (69) can be substituted into Equation (67) to yield

$$w(\mathbf{m}_1 \mid \mathbf{d}_t, \mathcal{I}_T) = \kappa \frac{t(\mathbf{m}_1 \mid \mathcal{I}_T)}{q(\mathbf{m}_1 \mid \mathcal{I}_N)} \int_R p(\mathbf{m}_2, \mathbf{m}_1 \mid \mathcal{I}_N)\, g(\mathbf{d}_t^* \mid \mathbf{m}_1, \mathbf{m}_2, \mathcal{I}_N)\, d\mathbf{m}_2. \tag{70}$$

The integrand plays the role of the likelihood function for the one-dimensional posterior w. This can be emphasized by writing Equation (70) after the integration, as given by

$$w(\mathbf{m}_1 \mid \mathbf{d}_t, \mathcal{I}_T) = \kappa \frac{t(\mathbf{m}_1 \mid \mathcal{I}_T)}{q(\mathbf{m}_1 \mid \mathcal{I}_N)}\, s(\mathbf{d}_t^* \mid \mathbf{m}_1, \mathcal{I}_N). \tag{71}$$

Because of that, for easy reference, the function s will be referred by the name of *extended likelihood*.

To illustrate this approach, consider the simple case where the distributions p and g are normal and the forward model is linear (Equation (27)). Thus, the pdf p in Equation (70) is normal with mean $\bar{\mathbf{m}}$ and covariance matrix $\mathbf{C}_m$ ($N(\bar{\mathbf{m}}, \mathbf{C}_m)$), which can be written as

$$f(\mathbf{m} \mid \mathcal{I}) = (2\pi)^{-\frac{M}{2}} \mid \mathbf{C}_m \mid^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{m}-\bar{\mathbf{m}})^T \mathbf{C}_m^{-1} (\mathbf{m}-\bar{\mathbf{m}})\right]. \tag{72}$$

If the covariance matrix of the above equation is partitioned as

$$\mathbf{C}_m = \begin{bmatrix} \mathbf{C}_{m_{11}} \mathbf{C}_{m_{12}} \\ \mathbf{C}_{m_{21}} \mathbf{C}_{m_{22}} \end{bmatrix},$$

where $\mathbf{C}_{m_{11}}$, $\mathbf{C}_{m_{12}}^T = \mathbf{C}_{m_{21}}$ and $\mathbf{C}_{m_{22}}$ are 1×1, $(M-1) \times 1$ and $(M-1) \times (M-1)$ matrices, respectively, then Theorem 10.6.1 of Graybill (1983), which gives the expression for the marginal of a normal distribution, allows us to write

$$q(\mathbf{m}_1 \mid \mathcal{I}_N) = \frac{1}{\sqrt{2\pi \mathbf{C}_{m11}}} \exp\left[-\frac{1}{2}(\mathbf{m}_1 - \bar{\mathbf{m}})^T \mathbf{C}_{m11}^{-1} (\mathbf{m}_1 - \bar{\mathbf{m}}_1)\right]. \tag{73}$$

By adopting these models, it is possible to perform the integral in Equation (70) analytically, what we discuss next.

4.1.1 Normal extended likelihood

In addition to the specification of the model for p, consider the likelihood function given by Equation (33) for g. Using these models, the integral in Equation (70), denoted by I, can be written as

$$I = (2\pi)^{-\frac{N+M}{2}} \mid \mathbf{C}_o \mid^{-\frac{1}{2}} \mid \mathbf{C}_m \mid^{-\frac{1}{2}} \int_R \exp\left\{-\frac{1}{2}\left[(\mathbf{d}_o^* - \mathbf{G}\mathbf{m})^T \mathbf{C}_o^{-1}(\mathbf{d}_o^* - \mathbf{G}\mathbf{m}) + (\mathbf{m} - \bar{\mathbf{m}})^T \mathbf{C}_m^{-1}(\mathbf{m} - \bar{\mathbf{m}})\right]\right\} d\mathbf{m}_2. \quad (74)$$

If we now apply the same manipulation to get the data-translated likelihood in Equation (30) we get

$$I = (2\pi)^{-\frac{N+M}{2}} \mid \mathbf{C}_o \mid^{-\frac{1}{2}} \mid \mathbf{C}_m \mid^{-\frac{1}{2}} \exp[S(\hat{\mathbf{m}})] \int_R \exp\left\{-\frac{1}{2}\left[(\mathbf{m} - \hat{\mathbf{m}})^T \mathbf{C}_p^{-1}(\mathbf{m} - \hat{\mathbf{m}})\right]\right\} d\mathbf{m}_2, \quad (75)$$

where

$$\mathbf{C}_p = (\mathbf{G}^T \mathbf{C}_o^{-1} \mathbf{G} + \mathbf{C}_m^{-1})^{-1}, \quad (76)$$

the estimated parameter vector is given by

$$\hat{\mathbf{m}} = \bar{\mathbf{m}} + \mathbf{C}_p \mathbf{G}^T \mathbf{C}_o^{-1}(\mathbf{d}_o^* - \mathbf{G}\bar{\mathbf{m}}) \quad (77)$$

and the estimated misfit value by

$$S(\hat{\mathbf{m}}) = (\mathbf{d}_o^* - \mathbf{G}\hat{\mathbf{m}})^T \mathbf{C}_o^{-1}(\mathbf{d}_o^* - \mathbf{G}\hat{\mathbf{m}}) + (\hat{\mathbf{m}} - \bar{\mathbf{m}})^T \mathbf{C}_m^{-1}(\hat{\mathbf{m}} - \bar{\mathbf{m}}). \quad (78)$$

Our main problem now is to evaluate the integral for the parameters $\mathbf{m}_2$. This task is facilitated with the above manipulation, since if we suppress all the constant terms in Equation (75) we are left with

$$I_2 = \int_R \exp\left\{-\frac{1}{2}\left[(\mathbf{m} - \hat{\mathbf{m}})^T \mathbf{C}_p^{-1}(\mathbf{m} - \hat{\mathbf{m}})\right]\right\} d\mathbf{m}_2. \quad (79)$$

We now let the inverse of the estimated covariance be given by

$$\mathbf{R} = \mathbf{C}_p^{-1} = \mathbf{G}^T \mathbf{C}_o^{-1} \mathbf{G} + \mathbf{C}_m^{-1}. \quad (80)$$

Using the same matrix partition as before, we can write

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{11}\, \mathbf{R}_{12} \\ \mathbf{R}_{21}\, \mathbf{R}_{22} \end{bmatrix} \text{ and } \mathbf{C}_p = \begin{bmatrix} \mathbf{C}_{p_{11}}\, \mathbf{C}_{p_{12}} \\ \mathbf{C}_{p_{21}}\, \mathbf{C}_{p_{22}} \end{bmatrix}.$$

Now, again according to the theorem of Graybill (1983) we have

$$I_2 = (2\pi)^{-\frac{M-1}{2}} \mid \mathbf{R}_{22} \mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[(\mathbf{m}_1 - \hat{\mathbf{m}}_1)^T \mathbf{C}_{p11}^{-1}(\mathbf{m}_1 - \hat{\mathbf{m}}_1)\right]\right\}. \quad (81)$$

Substituting this result back into Equation (75) we finally get

$$I = (2\pi)^{-\frac{N+1}{2}} \mid \mathbf{C}_o \mid^{-\frac{1}{2}} \mid \mathbf{C}_m \mid^{-\frac{1}{2}} \mid \mathbf{R}_{22} \mid^{-\frac{1}{2}} \exp\left[-\frac{1}{2} S(\hat{\mathbf{m}})\right] \exp\left\{-\frac{1}{2}\left[(\mathbf{m}_1 - \hat{\mathbf{m}}_1)^T \mathbf{C}_{p11}^{-1}(\mathbf{m}_1 - \hat{\mathbf{m}}_1)\right]\right\}. \quad (82)$$

4.1.2 Nature of the approximation

We can better understand the nature of the approximation made in Equation (69), considering the definitions for $\mathcal{I}_T$, $\mathcal{I}_N$ and $\mathcal{I}_C$ in Equation (68), by fully expanding the conditional probability v according to

$$P(\mathbf{m}_2 \mid \mathbf{m}_1, \mathcal{I}_T) = P(\mathbf{m}_2 \mid \mathbf{m}_1, \mathcal{I}_N + \mathcal{I}_C) = \frac{S_1}{S_2}, \quad (83)$$

where in general

$$S_1 = P(\mathbf{m}_2, \mathbf{m}_1 \mid \mathcal{I}_N)\, P(\mathcal{I}_N) + P(\mathbf{m}_2, \mathbf{m}_1 \mid \mathcal{I}_C)\, P(\mathcal{I}_C) - P(\mathbf{m}_2, \mathbf{m}_1 \mid \mathcal{I}_N, \mathcal{I}_C)\, P(\mathcal{I}_N, \mathcal{I}_C),$$

and

$$S_2 = P(\mathbf{m}_1 \mid \mathcal{I}_N)\, P(\mathcal{I}_N) + P(\mathbf{m}_1 \mid \mathcal{I}_C)\, P(\mathcal{I}_C) - P(\mathbf{m}_1 \mid \mathcal{I}_N, \mathcal{I}_C)\, P(\mathcal{I}_N, \mathcal{I}_C).$$

However, according to the definition of $\mathcal{I}_N$ and $\mathcal{I}_C$, they are independent, in which case

$$S_1 = P(\mathbf{m}_2, \mathbf{m}_1 \mid \mathcal{I}_N)\, P(\mathcal{I}_N) + P(\mathbf{m}_2, \mathbf{m}_1 \mid \mathcal{I}_C)\, P(\mathcal{I}_C). \quad (84)$$

and

$$S_2 = P(\mathbf{m}_1 \mid \mathcal{I}_N)\, P(\mathcal{I}_N) + P(\mathbf{m}_1 \mid \mathcal{I}_C)\, P(\mathcal{I}_C). \quad (85)$$

In either case

$$P(\mathcal{I}_N) + P(\mathcal{I}_C) - P(\mathcal{I}_N, \mathcal{I}_C) = 1 \quad (86)$$

or

$$P(\mathcal{I}_N) + P(\mathcal{I}_C) = 1. \quad (87)$$

Thus the statement made by Equation (69) becomes clear. It says that the weight of the information $\mathcal{I}_N$ for $\mathbf{m}_2$ is such that $P(\mathcal{I}_N) \ll P(\mathcal{I}_C) - P(\mathcal{I}_N, \mathcal{I}_C)$ or $P(\mathcal{I}_N) \ll P(\mathcal{I}_C)$, depending on the case. This amounts to having practically no information about cross-correlations of order higher than two for the parameters. But we still can make use of the higher-order information on individual parameter through the probability density t in Equation (70).

Analytically, this can be illustrated using the maximum entropy distribution, Equation (40) of Section 3. Substituting that equation for t in Equation (70), using

Equation (73) for q, gives

$$\frac{t(\mathbf{m}_1 \mid \mathcal{I}_T)}{q(\mathbf{m}_1 \mid \mathcal{I}_N)} \propto \exp\left[-\left(\lambda_1 + \frac{\bar{m}_1}{C_{m_{11}}}\right) m_1 - \left(\lambda_2 - \frac{1}{2C_{m_{11}}}\right) m_1^2 - \sum_{n=3}^{N} \lambda_n m_1^n\right]. \tag{88}$$

A very important particular case to consider is when all we have is actually $\mathcal{I}_N$ (i.e., $\mathcal{I}_C$ vanishes). This corresponds to a maximum entropy problem constrained by the first two moments of the unknown distribution, which leads to a normal distribution. More precisely, the Lagrange multipliers $\lambda_n = 0$, for $n = 3, 4, \ldots$ and

$$\lambda_1 = -\frac{\mu}{\sigma^2} \qquad \text{and} \qquad \lambda_2 = \frac{1}{2\,\sigma^2},$$

where μ and σ^2 are the mean and the variance supplied to the maximum entropy problem. Of course to be consistent we need $\mu = \bar{m}_1$ and $\sigma^2 = C_{m_{11}}$, which yields a ratio of unity for $\frac{t}{q}$. Thus, all that is left is the extended likelihood function, which is simply a multinormal Bayesian inversion procedure. To summarize, when all we know is really only the first and second-order moments, Equation (70) reduces to the more traditional Gaussian Bayesian formula (see, e.g., Tarantola, 1987).

4.2 Fixed parameters

To analyze this case, we can use the same simplifications of the previous section to rewrite (65) as

$$p(\mathbf{m}_1 \mid \mathbf{m}_2, \mathbf{d}_t, \mathcal{I}) = t(\mathbf{m}_1 \mid \mathcal{I}) \frac{v(\mathbf{m}_2 \mid \mathbf{m}_1)\, g(\mathbf{d}_t^* \mid \mathbf{m}_1, \mathbf{m}_2)}{v(\mathbf{m}_2)\, g(\mathbf{d}_t^* \mid \mathbf{m}_2)}.$$

or if we rewrite it in terms of the extended likelihood function

$$s(\mathbf{m}_1 \mid \mathbf{d}_t, \mathbf{m}_2, \mathcal{I}) = \frac{t(\mathbf{m}_1 \mid \mathcal{I})}{q(\mathbf{m}_1)} \frac{v(\mathbf{m}_1, \mathbf{m}_2)\, g(\mathbf{d}_t^* \mid \mathbf{m}_1, \mathbf{m}_2)}{v(\mathbf{m}_2)\, g(\mathbf{d}_t^* \mid \mathbf{m}_2)}. \tag{89}$$

To analyze this expression, the ratio in the second factor can be represented by

$$\frac{v(\mathbf{m}_1, \mathbf{m}_2)\, g(\mathbf{d}_t^* \mid \mathbf{m}_1, \mathbf{m}_2)}{v(\mathbf{m}_2)\, g(\mathbf{d}_t^* \mid \mathbf{m}_2)} = \frac{v(H_0)\, g(\mathbf{d}_t^* \mid H_0)}{v(H_1)\, g(\mathbf{d}_t^* \mid H_1)}, \tag{90}$$

which is similar to the ratios used for simple hypothesis testing with H_0 being the *simple hypothesis*† and H_1 the *simple alternative*‡. This term provides a measure of how significantly the parameters $\mathbf{m}_1$ influences the fit to the data. This may suggest a connection between this ratio and resolution. For instance, if the parameter $\mathbf{m}_1$ has no significance in explaining the observed data it can be removed, causing the ratio to become just unity. Conversely, if the ratio is close to one, it is an indication that parameter $\mathbf{m}_1$ plays no significant role in explaining the data. To proceed further in the analysis, let us consider again the same normal distributions of the previous section. Then, the second term of Equation (89), denoted by f_1, can be written as

$$f_1 = \frac{\exp\left[-\frac{1}{2}\left(\delta\mathbf{d}^T \mathbf{C}_o^{-1} \delta\mathbf{d} + \delta\mathbf{m}^T \mathbf{C}_m^{-1} \delta\mathbf{m}\right)\right]}{\exp\left[-\frac{1}{2}\left(\delta\mathbf{d}'^T \mathbf{C}_o^{-1} \delta\mathbf{d}' + \delta\mathbf{m}_2^T \mathbf{C}_{m22}^{-1} \delta\mathbf{m}_2\right)\right]}, \tag{91}$$

with

$$\begin{aligned} \delta\mathbf{d} &= \mathbf{d}_o^* - \mathbf{G}\mathbf{m}, \\ \delta\mathbf{d}' &= \mathbf{d}_o^* - \mathbf{G}'\mathbf{m}_2, \end{aligned}$$

and

$$\begin{aligned} \delta\mathbf{m} &= \mathbf{m} - \bar{\mathbf{m}}, \\ \delta\mathbf{m}_2 &= \mathbf{m}_2 - \bar{\mathbf{m}}_2. \end{aligned}$$

$\mathbf{G}'$ is the forward operator $\mathbf{G}$ without the columns corresponding to the parameter vector $\mathbf{m}_1$. If we rewrite the previous expression in terms of the data translated likelihoods we have

$$f_1 \propto \frac{\exp\left\{-\frac{1}{2}\left[(\mathbf{m} - \hat{\mathbf{m}})^T \mathbf{C}_p^{-1} (\mathbf{m} - \hat{\mathbf{m}})\right]\right\}}{\exp\left\{-\frac{1}{2}\left[(\mathbf{m}_2 - \hat{\mathbf{m}}_2)^T \mathbf{C}_p'^{-1} (\mathbf{m}_2 - \hat{\mathbf{m}}_2)\right]\right\}}, \tag{92}$$

with

$$\mathbf{C}_p = (\mathbf{G}^T \mathbf{C}_o^{-1} \mathbf{G} + \mathbf{C}_m^{-1})^{-1}$$

and

$$\mathbf{C}_p'^{-1} = (\mathbf{G}'^T \mathbf{C}_o^{-1} \mathbf{G}' + \mathbf{C}_{m22}^{-1})^{-1}.$$

At this point the usefulness of this approach is not completely clear. For inference purposes, it is not very attractive in comparison with the approach of the previous section, because it only makes partial use of the uncertainty. This is because we are considering only a particular cross-section of the joint distribution corresponding to values for the parameters $\mathbf{m}_2$ fixed *a priori* (see Figure 1).

5 EXAMPLE

To develop a better understanding of the proposed methodology, the results of some calculations are presented, beginning from a simple problem. Consider the problem of density inversion from gravity data, where the sources are six rectangular cells of constant density (Figure 6).

† This is the term used in the statistical literature to denote a statistical hypothesis that completely specifies the distribution, as opposed to a *composite hypothesis* which does not.

‡ This is just the alternative hypothesis to be considered against the simple hypothesis.

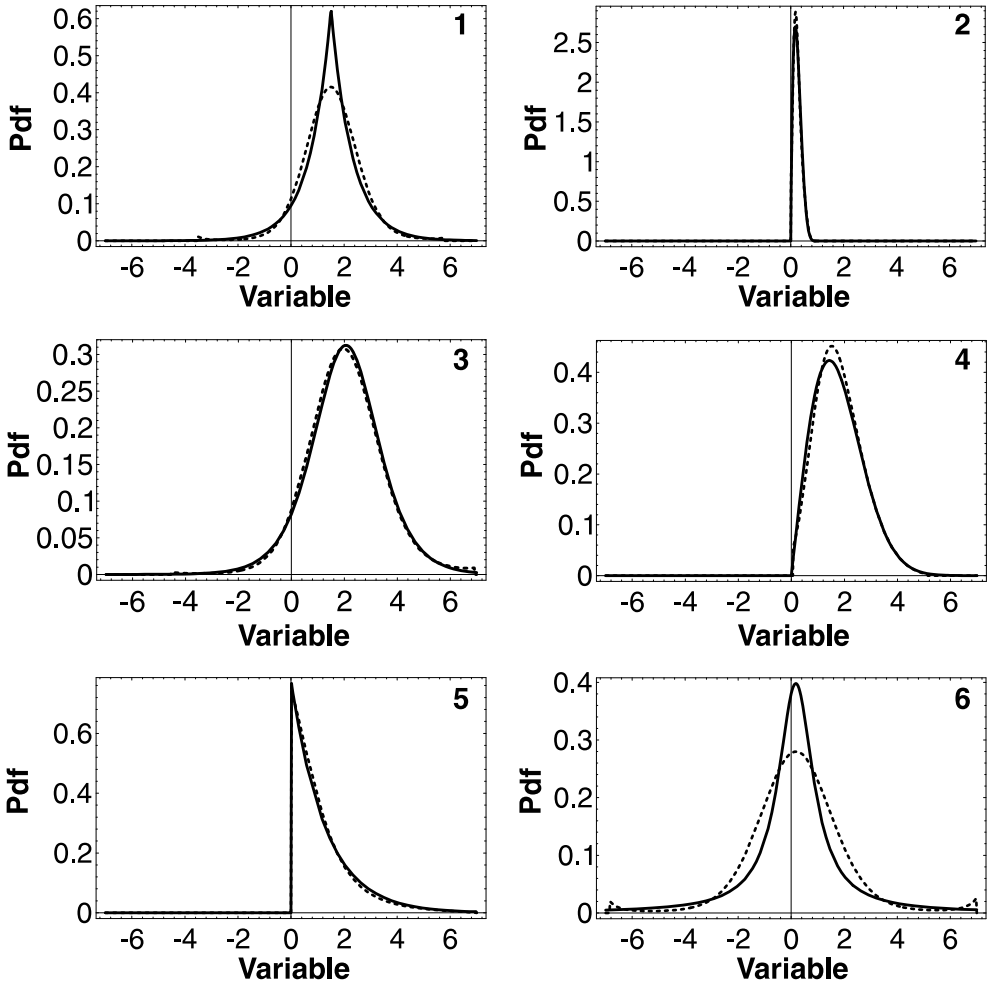

Figure 7. Probability density functions selected from theoretical models to simulate the prior information corresponding to the density contrast in each cell (solid lines). The models in each cell numbered from 1 to 6 are: 1- Laplace, 2- beta, 3- logistic, 4- Raleigh, 5- Exponential, 6- Cauchy. The distributions shown in dashed lines are the approximations to the theoretical distributions computed using maximum entropy and sample moments up to fourth order.

The true density contrast is derived by imposing an exponential correlation function on a sequence of uncorrelated Gaussian pseudo-random numbers. The formula for the gravity of prismatic bodies, which is available from many different sources (see, e.g., Telford et al., 1976, p. 74), is used to compute the synthetic gravity field.

In addition to the earth model and observed field, it is also necessary to generate the prior information. Besides assuming that the exponential correlation function is known, we simulate the uncertainties for the density in each cell based on selected probability density functions (solid line plots in Figure 7). This is done by setting the mean of the selected pdfs to the true density contrasts, what makes it possible to generate samples whose averages approximate these true values and whose higher-order moments give a synthetic degree of uncertainty (e.g., variance, skewness and kurtosis). The range of possible values taken by the random samples is limited to a certain interval, what introduces an error between two means computed from each pdf using its natural domain and the truncated domain, respectively. This error can be made small by choosing the truncation interval large enough. The numbers for this example were computed from one thousand samples generated from each distribution truncated at $[-7, 7]$. If we consider the physical property density, this range of density contrasts makes sense only if we think in terms of chemical elements and not in terms of rocks. Table 2 shows the true value for the density contrast and the values obtained considering the particular choice of truncation interval.

The averages computed from the samples generated using the selected pdfs can be used as prior information

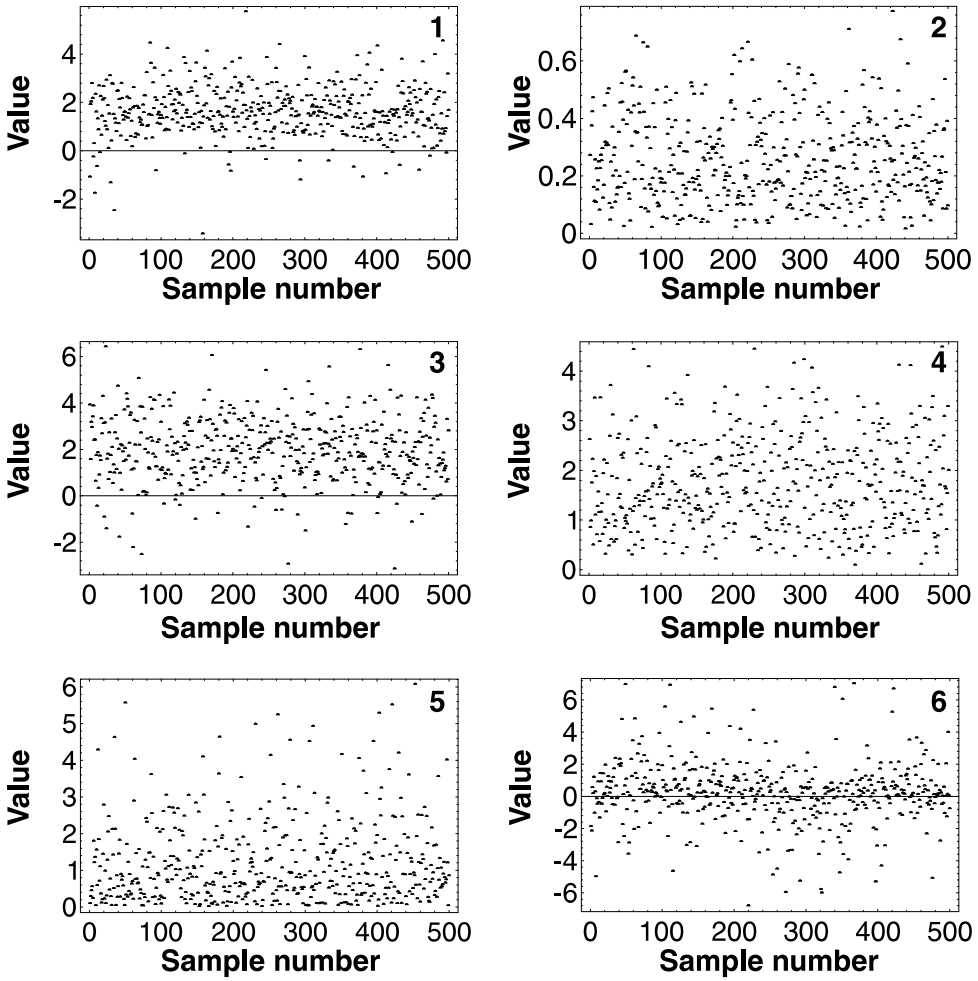

Figure 8. First five hundred samples generated from the selected theoretical distributions (Figure 6) to represent the uncertainty about the density contrast in each cell.

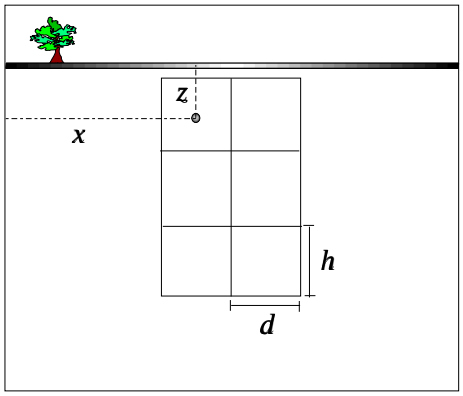

Figure 6. Simple earth model consisting of six rectangular cells with the parameters center coordinates (x, z), width (d) and height (h) indicated in the figure. The problem is to estimate the density contrast in each cell from a synthetic gravity field and synthetic prior information.

	True values	Truncated mean	Sample mean
1	1.50634	1.50237	1.51462
2	0.26004	0.26004	0.25899
3	2.05991	2.04381	2.03918
4	1.79565	1.79561	1.82460
5	1.28441	1.24882	1.16753
6	0.17437	0.14920	0.21767

Table 2. Comparison between the true value set for the density contrast in each cell and values derived from the truncated distributions and from the samples drawn from these distributions.

for a maximum entropy distribution calculation. The results from this calculation provide the local uncertainty for each parameter. Thus within the proposed methodology,

the next task, according to Equation (70), is to derive the global part of the problem using the normal approximation, where marginals can be computed analytically.

The global part of the problem is just the full Gaussian Bayesian inversion, which is composed of the likelihood function and a prior, both taken to be normal distributions. The likelihood function is constructed in the usual way, according to Equation (33), with the data error covariance matrix assumed to be known. The prior distribution is built by combining the information from the exponential correlation function between the parameters and the first two sample moments. This information is used to derive the covariance matrix, which is just correlations (ρ) scaled by the sample variances (σ), i.e.,

$$\mathbf{C}_m = \begin{bmatrix} \sigma_1^2\rho_{11} & \cdots & \sigma_1\sigma_6\rho_{16} \\ \vdots & \ddots & \vdots \\ \sigma_6\sigma_1\rho_{61} & \cdots & \sigma_6^2\rho_{61} \end{bmatrix},$$

and the vector of initial guesses, which are given by the first sample moments. These two functions are combined in analytical expression arising from the integration of the parameters $\mathbf{m}_2$, which is given in Equation (82). Next, to get the final answer (i.e., the posterior distribution) for a specific parameter, we need to take the product of this equation and the maximum entropy distribution, normalized by the marginal of the Gaussian prior q. To get all the parameters, it is necessary to iterate as many times as the number of parameters.

The maximum entropy distribution calculations follows the work of Mead and Papanicolau (1984) modified to include a uniform reference prior. The particular implementation for this example uses the Newton method with line search. Sample moments up to the fourth order are used as input to the optimization routine. The iterations of the algorithm are stopped when the moments of the computed maximum entropy distribution agree with the input sample moments to the order 10^{-6} or better. For this run, it usually took 6 to 7 iterations for convergence. The final approximations of the theoretical distributions are shown by dashed lines in Figure 7. The overall agreement between the estimated and true distributions is good, except for the Cauchy distribution (6). However, this is not very important since the estimated distribution carries the right moment information, according to the convergence criterion. It only tells that moments up to fourth order fall short of being able to correctly represent the Cauchy distribution, but this is not part of the goals. The only goal is to make inferences about the density contrasts. The decision as to what are the relevant moments for this goal have already been made *a priori*. In fact, later results show that the gain for going up to fourth order is marginal for this problem; the ability to do so is important nevertheless, since it permits the introduction of diverse prior information.

To understand the behavior of the proposed estimation methodology, three levels of uncorrelated Gaussian random numbers are used to corrupt the synthetic gravity data. The standard deviations for each noise level are respectively 1, 10 and 100% of the maximum synthetic gravity value. Then, several measures of central tendency are computed as estimates for the density contrast in each cell. They are the mean and the mode of the posterior distribution and the mean of the Gaussian Bayesian problem alone, which are given in Table 3. At very low noise level (1%), there is a general agreement between the three estimates. As we increase the noise level to 10% of the maximum gravity value, the mean and the mode of the posterior distribution start to pull apart, but the posterior means and the Gaussian means are still essentially equivalent. At extreme noise condition, the 100% level, the mean of the posterior distributions are overall slightly closer to the true values than are the Gaussian means, and the modes become extremely biased. This reflects the degradation of the gravity information, which allows for more features contained in the prior information to show up. Overall the posterior means and the Gaussian means are equivalent, with only a marginal advantage for the posterior means for the case of extreme noise. When the posterior means have pulled away from the true parameter in comparison with the Gaussian means, the numbers are written in bold face in Table 3. In some of those occasions the effect seems to be systematic.

It is important to point out that the fact that the posterior contains more features of the distributions used to generate the samples as noise level increases is just a natural consequence of the lack of resolution in the data. This is a feature desired only in the sense that it is supplying information not contained in the gravity data, since our goal is not to reproduce those theoretical distributions of Figure 7. What we want is to draw inferences about the density contrast of the cells. Thus, to see how the methodology is performing, we must look for the errors associated with the density estimates. Table 4 shows a comparison between the prior, the Gaussian and the posterior variances for the density contrast in each cell. Overall, there is a marginal reduction in the posterior variances when compared with the Gaussian variances. However, for few occasions shown as bold numbers in Table 4, the posterior variances increase. These increases tend to be associated with the asymmetrical distributions.

As observed above, at low noise level the prior information is made irrelevant by the gravity data and we

Parameter estimates										
		1% noise			10% noise			100% noise		
	true	mean	mode	$\hat{\mathbf{m}}$	mean	mode	$\hat{\mathbf{m}}$	mean	mode	$\hat{\mathbf{m}}$
1	1.506	1.491	1.491	1.491	1.410	1.410	1.408	1.097	1.104	1.057
2	0.260	0.291	0.291	0.293	**0.197**	0.189	0.210	0.262	0.223	0.276
3	2.060	2.163	2.163	2.166	2.316	2.310	2.346	**1.990**	1.980	2.010
4	1.796	**1.659**	1.659	1.663	1.912	1.906	1.947	1.999	1.961	2.057
5	1.284	**1.135**	1.135	1.163	**0.784**	0.549	0.966	**1.081**	0.612	1.336
6	0.174	**0.303**	0.303	0.304	**0.160**	0.160	0.161	0.396	0.382	0.457

Table 3. Values of central tendency taken from the posterior distribution (means and mode) and from the Gaussian Bayesian solution ($\hat{\mathbf{m}}$). Bold numbers indicate that the posterior mean moved away from the true values in comparison with $\hat{\mathbf{m}}$.

Variances							
		1% noise		10% noise		100% noise	
	Prior	Gaussian	Posterior	Gaussian	Posterior	Gaussian	Posterior
1	1.19684	0.00168	0.00168	0.04814	0.04727	0.33775	0.30732
2	0.02227	0.00061	**0.00062**	0.00629	0.00567	0.01450	**0.01535**
3	2.04926	0.03321	0.03302	0.35629	0.33958	0.73136	0.66580
4	0.81267	0.00947	0.00947	0.11950	**0.12284**	0.29358	**0.31679**
5	1.22102	0.03151	**0.03214**	0.34828	0.25422	0.61830	0.55005
6	3.21465	0.01127	0.01124	0.20426	0.19390	1.14011	0.91929

Table 4. Comparison between the prior, the Gaussian and the posterior variances for the parameters at different noise levels. Bold numbers indicate posterior variances that are greater than their Gaussian counterpart.

are basically free to choose any one of the estimates. At high noise level, however, this is not the case, the choice for estimates can make a difference. If we use only the Gaussian approach, the decision is still simple. Take the mean and the variance for the estimates and the error bars, respectively. However, for the posterior distribution, the decision is not straightforward. What should we pick for the estimate and the associated error for the density contrast? Here that we have the true values available, it is clear that the mean is the best approximation. We can understand this intuitively by thinking that for asymmetrical distributions the mode can be extremely biased towards high or low values, depending on the nature of the asymmetry. This happens, for example, for a monotonic distribution such as the exponential, where the mode will always be zero regardless of its mean (see Figure 7). For the error bars, the issues are basically the same. The asymmetry of the posterior distribution makes it inappropriate to have centered error bar of the type mean $\pm n\sigma$, where n is an integer and σ is the standard deviation. Instead, as illustrated in Figures 10 – 12, we can take 95% interquantile intervals, which can are computed independently of any estimates for the density contrasts. This is done by finding the parameter values corresponding to probability $1 - 0.025$ and 0.025, which can be computed by inversion of the posterior cumulative distribution (Figure 13).

However, density estimates are necessary to compute the synthetic gravity field used in the fitting procedure, which is shown in Figure 9 with the synthetic, the noisy and the estimated gravity fields overlapped. We use the mean of each posterior distribution to compute the synthetic gravity field.

Another useful type of analysis is to perform several runs of the inversion scheme for different noise values with the same standard deviation. A possible choice from the previous section is the 10% noise level. The results are shown in Tables 5 and 6. The overall behavior of

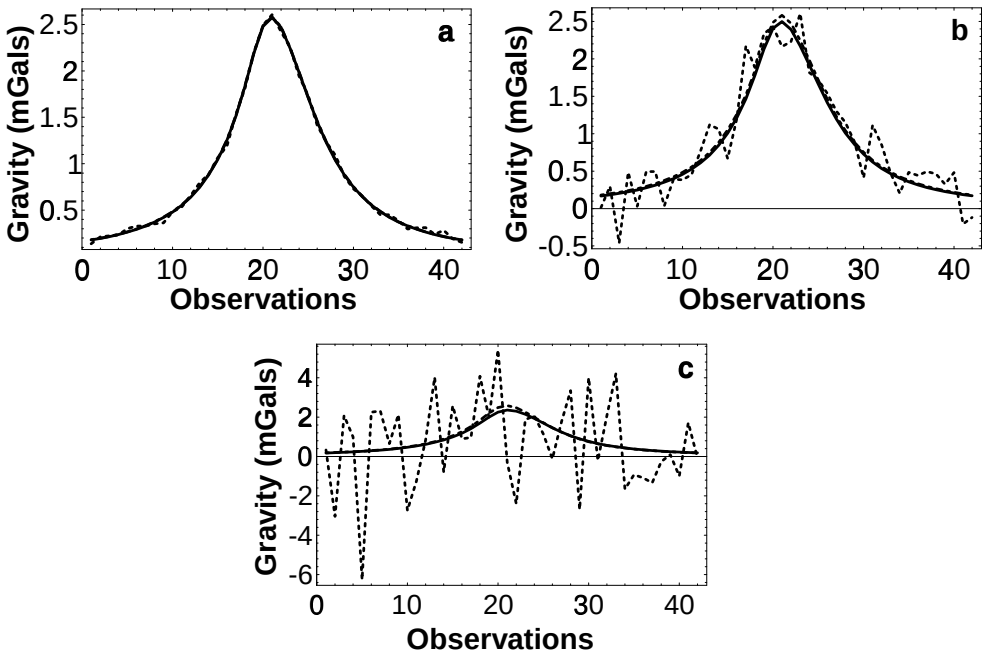

Figure 9. The synthetic (dot-dashed line), the noisy (dashed line) and the estimated (solid line) gravity fields for 1 (a), 10 (b) and 100% (c) of the maximum synthetic gravity values. The estimated field uses the posterior mean for the parameters.

Parameter estimates										
	First run		Second run		Third run		Fourth run		Fifth run	
1	1.301	1.305	1.523	**1.521**	1.464	1.464	1.465	1.465	1.274	1.279
2	0.211	**0.198**	0.201	**0.191**	0.242	**0.225**	0.217	**0.202**	0.233	**0.217**
3	2.408	2.372	2.138	2.115	2.313	2.288	2.162	2.141	1.886	1.878
4	1.691	**1.645**	1.366	**1.323**	1.666	**1.617**	1.825	1.789	1.591	**1.538**
5	1.598	1.389	1.293	**1.088**	0.929	**0.758**	1.212	**1.032**	1.600	1.426
6	-0.160	-0.143	0.499	0.473	0.416	0.404	0.252	0.247	0.392	0.363

Table 5. Comparison between the Gaussian means and the posterior means in several inversion runs considering different noise realizations with the same variance. Bold numbers indicate that the posterior mean (second column) moved away from the true values in comparison with $\hat{\mathbf{m}}$ (first column).

the solution is basically the same as discussed above. It is interesting to note that the differences between the Gaussian and posterior means tend to be systematic.

6 CONCLUSIONS

Bayesian methods provide an adequate framework to integrate information of diverse origin and to carry uncertainty analysis associated with an inverse problem. The general Bayesian formulation for the geophysical inverse problem can be obtained by the application of the product rule of probability theory which carries logical mechanisms of commutativity. Furthermore, when we introduce a new variable for predicted data, all particular cases, such as noiseless observed data or perfect forward modeling, comes out in a coherent way. The greatest difficulty with Bayesian methods comes from the high dimensionality of the parameter and data vectors, which leads to probability distributions with the same dimensions. When we explore the possibility of using local prior information to derive priors for each parameter independently of the others, we can minimize the dimensionality problem. This is done by using part of the local information to construct a high-dimensional normal distribution that can be mar-

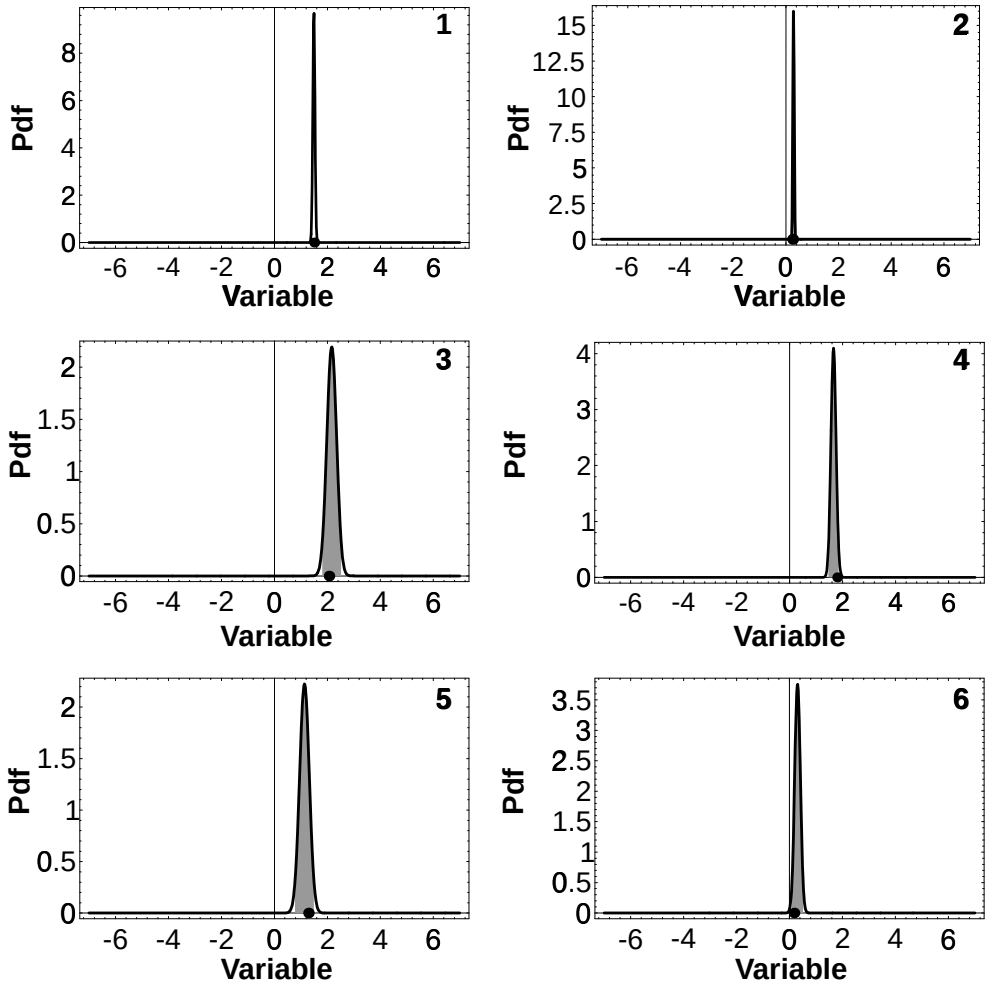

Figure 10. Inversion results depicted by the posterior marginal for each parameter. The 95% interquantile regions are represented by the shaded areas and the true value for the parameter are given by the solid circles. This example uses the 1% of the maximum gravity value as standard deviation for the noise.

	Variances									
	First run		Second run		Third run		Fourth run		Fifth run	
1	0.06899	0.06725	0.08705	0.08433	0.05096	0.04999	0.03201	0.03162	0.07586	0.07378
2	0.01051	0.00866	0.01196	0.00922	0.01006	0.00943	0.00764	0.00680	0.01150	0.01022
3	0.39925	0.37971	0.42777	0.40204	0.31180	0.29838	0.37332	0.35329	0.51726	0.47890
4	0.14969	0.149331	0.20168	0.18589	0.16488	0.16360	0.11519	**0.11688**	0.19007	0.18522
5	0.35299	**0.41494**	0.27640	**0.28377**	0.29010	0.21888	0.22422	**0.22884**	0.28123	**0.33119**
6	0.23107	0.21816	0.30497	0.28345	0.16861	0.16156	0.17777	0.16985	0.54031	0.47809

Table 6. Comparison between the Gaussian variances and the posterior variances for several inversion runs using different noise values with the same variance. Bold numbers indicate the cases where the posterior variance is greater than the Gaussian variance.

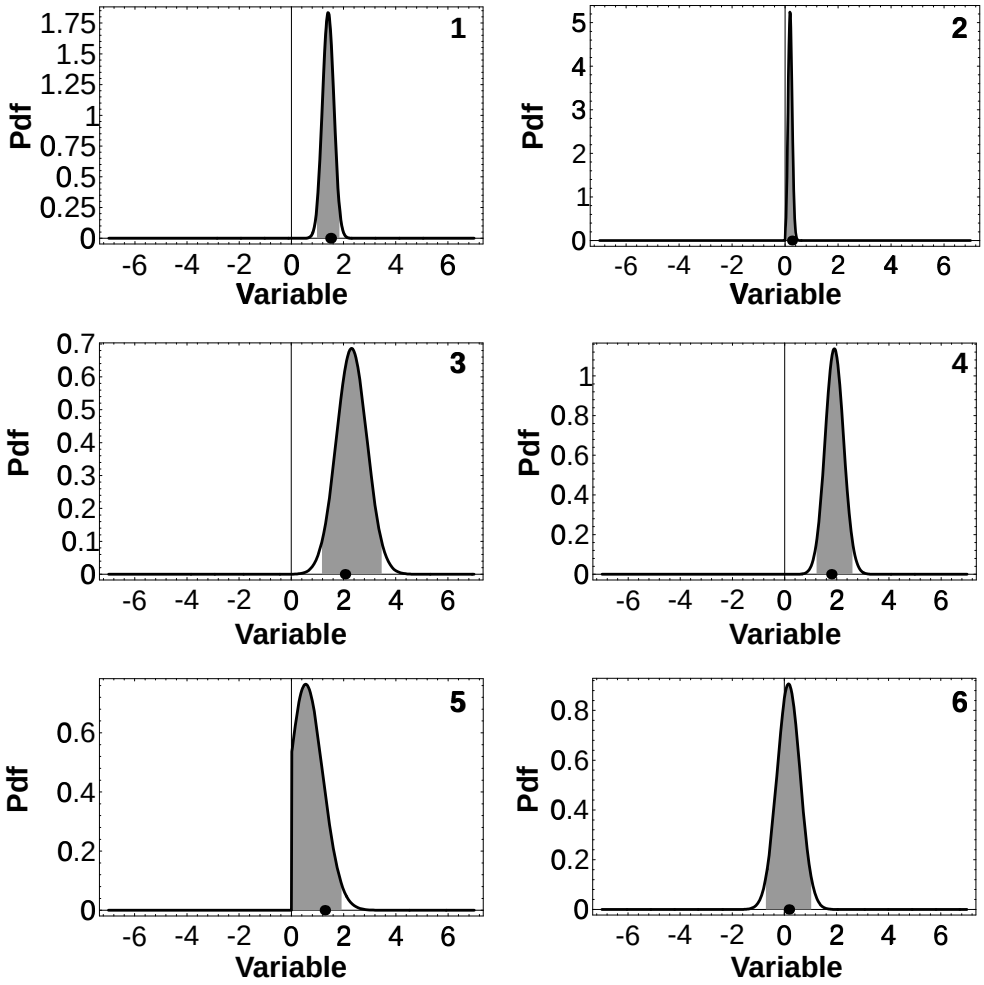

Figure 11. Inversion results depicted by the posterior marginal for each parameter. The 95% interquantile regions are represented by the shaded areas and the true value for the parameter are given by the solid circles. This example uses the 10% of the maximum gravity value as standard deviation for the noise.

ginalized analytically. After marginalization we are left with only a one-dimensional distribution for a specific parameter that carries the geophysical information. This distribution can be combined with the rest of the local information by simple multiplication. In this way, we replace a high-dimensional inverse problem by series of estimation steps, involving one parameter at a time.

7 ACKNOWLEDGMENTS

The authors are grateful for the discussions with Richard O. Hansen, Kadri Dagdelen, Wences Gouveia and Ken Larner. This work was partially supported by the sponsors of the Consortium Project on Seismic Inverse Methods for Complex Structures at the Center for Wave Phenomena, the sponsors of the Gravity and Magnetics Project, both at the Colorado School of Mines, the Shell Foundation and the Army Research Office. In addition, the first author acknowledges the support of the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil).

References

Backus, G. E. and J. F. Gilbert (1968). The resolving power of gross earth data. *Geophys. J. R. astr. Soc.*, **16**, 169–205.

Constable, S. C., R. L. Parker, and C. G. Constable (1987). Occam's inversion: a practical algorithm for generating smooth models from eletromagnetic sounding data. *Geophysics*, **52**, 289–300.

Cox, R. T. (1946). Probability, frequency and reasonable expectations. *Am. J. Phys.*, **14**, 1–13.

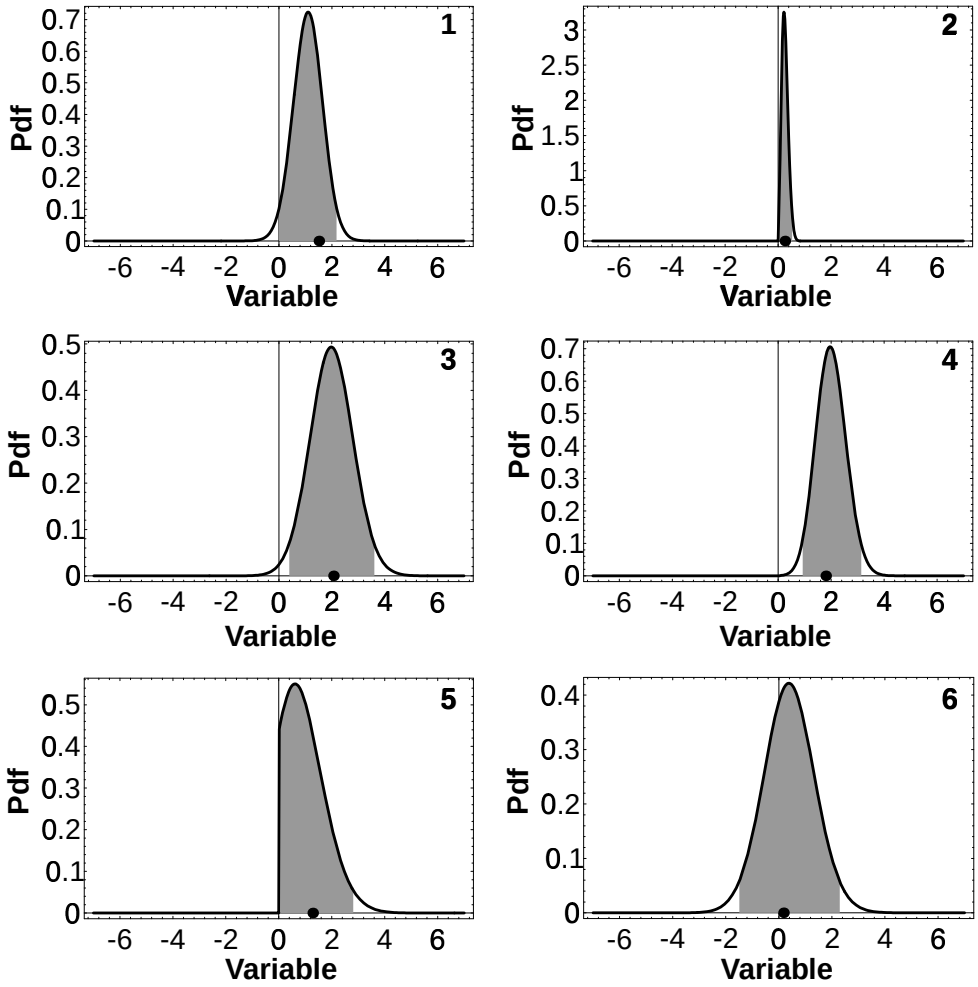

Figure 12. Inversion results depicted by the posterior marginal for each parameter. The 95% interquantile regions are represented by the shaded areas and the true value for the parameter are given by the solid circles. This example uses the 100% of the maximum gravity value as standard deviation for the noise.

Cox, R. T. (1961). *Algebra of the Probable Inference*. Johns Hopkins Press, Baltimore, MD.

Deutsch, C. V. and A. G. Journel (1992). *GSLIB: Geostatistical Software Library and User's Guide*. Oxford University Press, Oxford, New York.

Gouveia, W., F. S. Moraes, and J. A. Scales (1996). Entropy, information and inversion. In *CWP project review*, CWP-203, Golden, CO, Center for Wave Phenomena, Colorado School of Mines.

Gradshteyn, I. S. and I. M. Ryzhik (1980). *Table of integrals, series, and products* (Corrected and enlarged ed.). Academic Press.

Graybill, F. A. (1983). *Matrices With Applications in Statistics* (Second ed.). Wadsworth Publ. Co.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Phys. Rev.*, **106**, 171–190.

Jaynes, E. T. (1963). Information theory and statistical mechanics. In K. W. Ford (Ed.), *Statistical Physics*, pp. 181–218. W. A. Benjamin, Inc. Reprinted in Jaynes (1983).

Jaynes, E. T. (1978). Where do stand on maximum entropy ? In R. D. Levine and M. Tribus (Eds.), *The maximum entropy formalism*, Cambridge, Mass. M.I.T. Press. Reprinted in Jaynes (1983).

Jaynes, E. T. (1983). *Papers on probability, statistics and statistical physics. A reprint collection*. D. Reidel, Dordrecht, Holland.

Journel, A. G. (1989). Fundamental of geostatistics in five lessons. In *Short Course in Geology*, Volume 8, Washington, D. C. American Geophysical Union.

Kullback, S. (1959). *Information Theory and Statistics*. Wiley, New York, N. Y. Plublished by Dover in 1968.

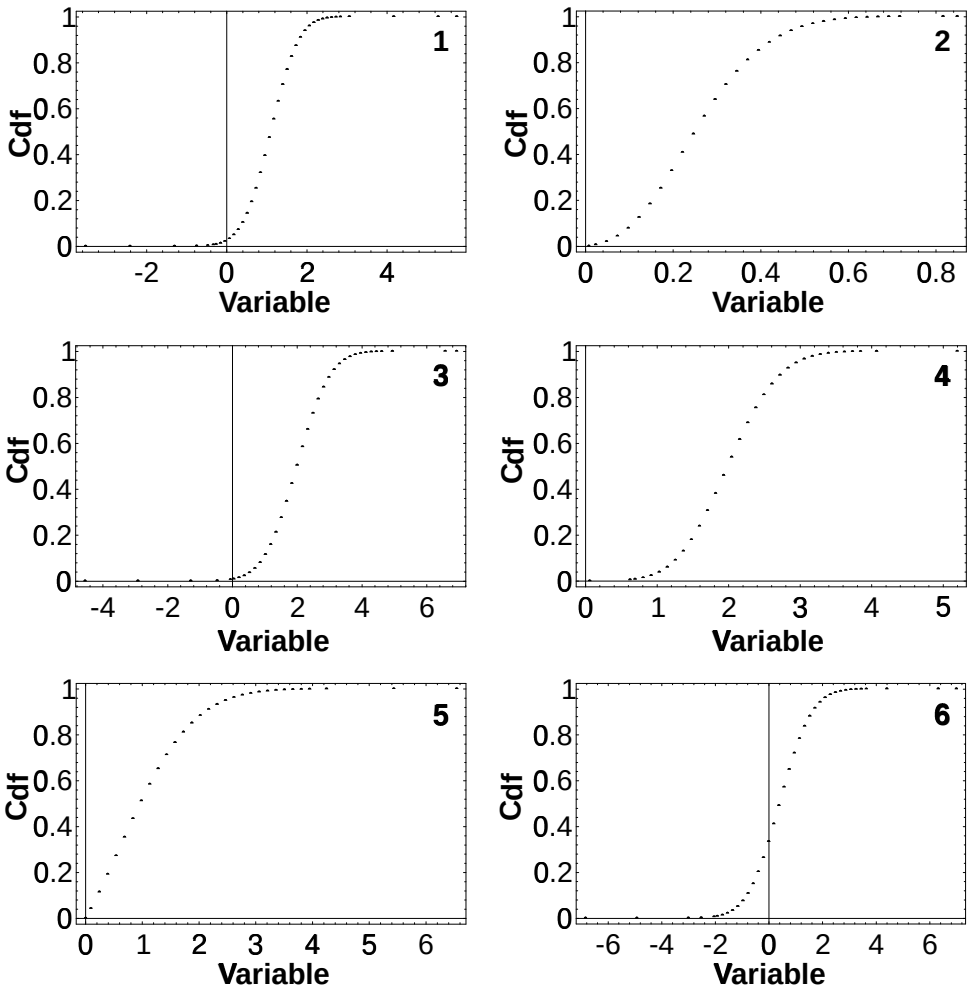

Figure 13. Points of the cumulative distribution computed from the posterior distribution for each parameter, using an adaptive method.

McCullagh, P. (1987). *Tensor Methods in Statistics.* Chapman and Hall, New York, N. Y.

Mead, L. R. and N. Papanicolau (1984). Maximum entropy in the problem of moments. *J. Math. Phys.*, **25**, 2404–2417.

Mosegaard, K. and A. Tarantola (1995). Monte carlo sampling of solutions to inverse problems. *J. Geophys. Res.*, **100**, 12,431–12,447.

Parker, R. L. (1977). Understanding inverse theory. *Ann. Rev. Earth Panet. Sci.*, **5**, 35–64.

Scales, J. A. and A. Tarantola (1994). Bayesian inversion with realistic *a priori* information. Technical Report 159, Center for Wave Phenomena, Colorado School of Mines, Golden, CO.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Jour.*, **27**, 379–423,623–656.

Shore, J. E. and R. W. Johnson (1981). Properties of cross–entropy minimization. *IEEE Trans. on Information Theory*, **IT-27**, 472–482.

Tarantola, A. (1987). *Inverse Problem Theory - Methods for Data Fitting and Model Parameter Estimation.* Elsevier.

Telford, W. M., L. P. Geldard, R. E. Sheriff, and D. A. Keys (1976). *Applied Geophysics.* Cambridge University Press.