Geophys. J. Int. (2000) **141,** 713–723

Local Bayesian inversion: theoretical developments

Fernando S. Moraes[1] and John A. Scales[2]

[1] *Laboratório de Engenharia e Exploração de Petróleo, Universidade Estadual do Norte Fluminense, Macaé,* RJ 27973-030, *Brazil. E-mail: fernando@lenep.uenf.br*

[2] *Department of Geophysics and Center for Wave Phenomena, Colorado School of Mines, Golden,* CO 80401, *USA*

Accepted 2000 January 11. Received 1999 December 22; in original form 1997 February 26

SUMMARY

We derive a new Bayesian formulation for the discrete geophysical inverse problem that can significantly reduce the cost of the computations. The Bayesian approach focuses on obtaining a probability distribution (the posterior distribution), assimilating three kinds of information: physical theories (data modelling), observations (data measurements) and prior information on models. Once this goal is achieved, all inferences can be obtained from the posterior by computing statistics relative to individual parameters (e.g. marginal distributions), a daunting computational problem in high dimensions.

Our formulation is developed from the working hypothesis that the local (subsurface) prior information on model parameters supercedes any additional information from other parts of the model. Based on this hypothesis, we propose an approximation that permits a reduction of the dimensionality involved in the calculations via marginalization of the probability distributions. The marginalization facilitates the tasks of incorporating diverse prior information and conducting inferences on individual parameters, because the final result is a collection of 1-D posterior distributions. Parameters are considered individually, one at a time. The approximation involves throwing away, at each step, cross-moment information of order higher than two, while preserving all marginal information about the parameter being estimated. The main advantage of the method is allowing for systematic integration of prior information while maintaining practical feasibility. This is achieved by combining (1) probability density estimation methods to derive marginal prior distributions from available local information, and (2) the use of multidimensional Gaussian distributions, which can be marginalized in closed form.

Using a six-parameter problem, we illustrate how the proposed methodology works. In the example, the marginal prior distributions are derived from the application of the principle of maximum entropy, which allows one to solve the entire problem analytically. Both random and modelling errors are considered. The uncertainty measure for estimated parameters is provided by 95 per cent probability intervals calculated from the marginal posterior distributions.

Key words: Bayesian inversion, local prior probabilities, marginalization.

1 INTRODUCTION

In geophysical inference the goal is to combine information from physical theories and experiments in order to draw conclusions about a given Earth property (for example, density, conductivity, seismic velocity or the shape of a geological body). The first difficulty that has to be addressed comes from the fact that these properties of the Earth are described by functions, whereas the result of any real measurement can only be a finite number of data. In general, it is impossible to infer a function from a finite number of data. In practice, the infinite-dimensional earth models are often approximated by finite-dimensional projections, so that an earth model is described by a finite number of parameters. Further, geophysical data are usually contaminated by noise, both random (not deterministically reproducible) and systematic (unmodelled physics). Whether the noise is truly random or simply information we choose not to fit, the net result is that data fitting—the procedure by which physical theories are linked with experiment—is only performed up to some tolerance. No matter how

many parameters are used, if there is a single model that fits the data, there will be an infinite number of them, since model parameters are continuous variables.

The inference problem is often replaced by an optimization problem; for example, find the model that 'best' fits the data in a least-squares sense. However, as the optimization problem is likely to be ill-posed, these calculations generally require some sort of regularization. For any sufficiently rich parametrization of the subsurface, both reasonable and unreasonable models will fit the data. Therefore, some sort of prior information is essential to narrow the range of inferences, for example, to rule out models with negative densities. This information can be either deterministic (density is positive) or probabilistic (we may have a histogram of previous measurements of a property). In this paper, we will adopt the Bayesian strategy and treat all prior information probabilistically. (However, see Scales & Sneider 1997 for a discussion of both sides of this issue.) The term Bayesian is used here in a broad sense to describe any method that employs model-based probability theory as a method of inference (e.g. Tarantola 1987; Jaynes 1994; Backus 1988a; Gouveia & Scales 1998). This is in contrast to frequentist methods in which probabilities enter via the data alone (e.g. Parker 1975; Backus 1989; Stark 1992a,b).

In principle, the issues of integration of complex prior information and uncertainty analysis can be handled in the Bayesian framework. The Bayesian approach begins with the specification of probability distributions that encapsulate the prior information (the *prior* distribution) and the information from data fit (the *likelihood* function). Once these are available, all inferences can be drawn from their normalized product, which is the *posterior* distribution. Thus, performing integration of prior information and uncertainty analysis depends essentially on one's ability to handle these probability distributions, never a trivial task for high-dimensional problems. The problems with high dimensions are twofold. The first consists of coming up with high-dimensional probabilities that incorporate the available information without overspecifying that information. The second is that even if we manage to conservatively build all required probabilities, extracting information about the parameters from the posterior can be time consuming, since this involves integration or sampling in high dimensions (see e.g. Press *et al.*1992, Section 4.6, and Tierney 1994), and the function we are sampling is often expensive to compute and can only be evaluated pointwise. While the latter problem (sampling) is largely computational, the former (assigning probabilities) involves the fundamental aspects of the probabilistic formulation of the inverse problem.

The very existence of probability distributions over model spaces can be questioned (see e.g. Parker 1994 Section 4.08). Even if one accepts prior probabilities on models, translating available prior information into probability assignments has been the subject of research since the time of Bernoulli (Jeffreys 1939; Jaynes 1957, 1968, 1978; Backus 1988b, 1996; Scott 1992; Scales 1996; Gouveia *et al.* 1996). Jaynes (1968) gives extensive discussions on objective ways of assigning prior probabilities, pointing out that more general principles are still needed. The main rule used by Bayesians for objectively assigning probabilities, which is maximum entropy (Jaynes 1957; Gouveia *et al.* 1996), cannot handle some types of prior information such as non-linear constraints on models. Part of these remarks is related to the old debate involving frequentist and Bayesian interpretations of probabilities. It is not our intention to contribute to this debate, but the interested reader may refer to Jaynes (1994) for an extensive discussion and list of references. This paper is mainly concerned with the development of the Bayesian approach for cases where the main goal is to make inferences about finite-dimensional subsets of earth models and diverse prior information from observations is available. [Good examples of inference in infinite-dimensional spaces can be found in Backus (1989), Fitzpatrick (1991), Stark (1992a) and Parker (1994).]

The developments here are centred on inferences for individual parameters, because results in high dimensions are difficult to interpret. Previous geophysical applications of Bayesian inference have usually taken the multidimensional approach. Most commonly, parametric statistical models, usually Gaussian, are employed. In this way, the work reduces to that of finding the corresponding parameters of the statistical model, exploring connections between inference and optimization. For example, the Gaussian distribution is described entirely by the mean and covariance, both of which can be estimated using standard least-squares methods (Tarantola 1987; Duijndam 1988a,b; Gouveia & Scales 1998). One group of published works takes the Bayesian approach to derive maximum *a posteriori* (MAP) estimators; some examples are given by Richard *et al.* (1984), Yabuki & Matsu'ura (1992) and Sacchi & Ulrych (1995). A more general Bayesian strategy would be to avoid parametric assumptions in the construction of priors. The issue then is how to extract marginal distributions from the posterior; this is discussed by Mosegaard & Tarantola (1995) and Tierney (1994). In the most general case, Monte Carlo sampling methods must be employed, since reliable estimation of statistics requires convergence in probability of the Markov chain used. For a modern comparative review of the convergence properties of Markov chain Monte Carlo methods, see Cowles & Carlin (1996). The application of Monte Carlo integration methods in geophysics is limited to small problems (see e.g. Tarits *et al.* 1994).

This work addresses issues of both building prior probabilities on models and computing posterior marginal distributions in the same way: by reducing the dimensionality involved in Bayesian calculations. To do this, we adopt a Gaussian model for the likelihood function and divide the prior distribution into two parts: the marginal prior distribution for one specific parameter and a normal approximation to the joint distribution for all other parameters. Then, by marginalization of normal distributions, the multidimensional problem can be replaced by a sequence of 1-D problems. Each time, a different parameter is kept in the problem and the rest are eliminated.

Our theoretical study begins in the next section with the derivation of our modified Bayesian formulation. This modification involves making one approximation, which is discussed in detail in Section 3. To conclude, we present a simple analytical gravity example and discuss the results.

2 THE LOCAL BAYESIAN APPROACH

Consider the problem of making inferences about a discrete set of Earth parameters $\mathbf{m} \in \mathcal{M} \subset \mathcal{R}^M$ from experimental data $\mathbf{d} \in \mathcal{D} \subset \mathcal{R}^N$, a physical theory and prior information. The physical theory yields a mathematical operator $\mathbf{g}$ used for predicting observed data according to $\mathbf{d} = \mathbf{g}(\mathbf{m}) + \mathbf{n}$, where $\mathbf{n}$ is the sum of observational and theoretical errors. Prior information, generically represented by $\mathcal{I}$, is all information about $\mathbf{m}$ obtained independently of the data.

The main goal in the Bayesian approach is to obtain the posterior distribution, which is the joint probability distribution for the parameters given all available information: experimental and theoretical data and prior information. The posterior is, by application of Bayes' theorem, the normalized product of the prior distribution and the likelihood function, which carry, respectively, the prior information and information from the data fitting (both observed and modelled data). We write the posterior as

$$p(\mathbf{m} \mid \mathbf{d}, \mathscr{I}) \propto s(\mathbf{m} \mid \mathscr{I})\, l(\mathbf{m} \mid \mathbf{d}), \tag{1}$$

where s and l are, respectively, the prior and the likelihood function (see e.g. Box & Tiao 1973 for a detailed description of the Bayesian formulation).

As all functions in eq. (1) have the same dimensionality of the parameter vector, which is usually high in most geophysical applications, we are interested in alternative Bayesian formulations to avoid solving the full multivariate problem. In particular, we want to avoid having to build the multidimensional prior distribution. Instead, we want the solution to incorporate all prior information $\mathscr{I}$ processed into marginal (local) prior distributions for single parameters, using methods such as non-parametric geostatistics. The problem is how to incorporate the marginals in the full formulation, eq. (1). One way is to seek the solution one parameter at a time. To carry out this approach, the first step is to divide the parameter vector $\mathbf{m}$ into two parts, $\mathbf{m}_1$ and $\mathbf{m}_2$, i.e.

$$\mathbf{m} = \begin{bmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \end{bmatrix}, \tag{2}$$

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

Next, it is necessary to eliminate the parameters $\mathbf{m}_2$ from the problem to have a solution expressed only in terms of $\mathbf{m}_1$. We can then construct an iterative scheme with a different parameter serving as $\mathbf{m}_1$ at each iteration until all parameters have been estimated. This can be done by treating $\mathbf{m}_2$ as nuisance parameters*. If we carry this idea to the general Bayesian formulation, we can rewrite eq. (1) as

$$p(\mathbf{m}_1, \mathbf{m}_2 \mid \mathbf{d}, \mathscr{I}) \propto t(\mathbf{m}_1 \mid \mathscr{I})\, u(\mathbf{m}_2 \mid \mathbf{m}_1, \mathscr{I})\, l(\mathbf{m}_1, \mathbf{m}_2 \mid \mathbf{d}, \mathscr{I}), \tag{3}$$

where $t(\mathbf{m}_1 \mid \mathscr{I})$ is the marginal prior distribution and $t(\mathbf{m}_1 \mid \mathscr{I})\, u(\mathbf{m}_2 \mid \mathbf{m}_1, \mathscr{I}) = s(\mathbf{m}_1, \mathbf{m}_2 \mid \mathscr{I})$. If the marginal prior distribution can be constructed using local regression methods (e.g. parametric and non-parametric geostatistics), the main task is to determine a computing scheme for the posterior distribution for the parameter $\mathbf{m}_1$. In the nuisance parameter approach, the standard procedure is to eliminate parameters $\mathbf{m}_2$ by marginalization, which we discuss next.

2.1 Eliminating parameters $\mathbf{m}_2$

Eliminating parameters $\mathbf{m}_2$ involves finding a marginal distribution for $\mathbf{m}_1$ from the posterior, which can be represented by

$$w(\mathbf{m}_1 \mid \mathbf{d}, \mathscr{I}) = \int_{\mathscr{M}'} p(\mathbf{m}_1, \mathbf{m}_2 \mid \mathbf{d}, \mathscr{I})\, d\mathbf{m}_2 . \tag{4}$$

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

Applying this idea to eq. (3), we obtain

$$w(\mathbf{m}_1 \mid \mathbf{d}, \mathscr{I}) = \kappa\, t(\mathbf{m}_1 \mid \mathscr{I}) \times \int_{\mathscr{M}'} u(\mathbf{m}_2 \mid \mathbf{m}_1, \mathscr{I})\, l(\mathbf{m}_1, \mathbf{m}_2 \mid \mathbf{d}, \mathscr{I})\, d\mathbf{m}_2 , \tag{5}$$

where κ is the normalizing constant providing integration of w to unity. In the integrand of this equation, we have the likelihood function l and the prior distribution u for $\mathbf{m}_2$. The latter is an $(M-1)$-dimensional distribution, which means that we still need to handle integration in a high-dimensional space. The ability to overcome this difficulty will depend on the nature of the data modelling operator and the functions u and l. In dealing with function u, consider that, in each step of this iterative approach, we are only interested in making inferences about $\mathbf{m}_1$. It is, then, intuitive to expect that we may discard some prior information about the parameters $\mathbf{m}_2$, as long as sufficient information about $\mathbf{m}_1$ is introduced through the marginal prior distribution t. Following this idea, the prior information can be divided into two parts: one part that defines a normal distribution (i.e. mean and covariance information) and another part that complements this information (i.e. higher-order moment information) in such a way that

$$\mathscr{I} = \mathscr{I}_{\mathrm{N}} + \mathscr{I}_{\mathrm{C}} . \tag{6}$$

That is, the total information ($\mathscr{I}$) equals the logical sum of normal information ($\mathscr{I}_{\mathrm{N}}$) and its complement ($\mathscr{I}_{\mathrm{C}}$). Furthermore, we may assume that for parameters $\mathbf{m}_2$, only the information $\mathscr{I}_{\mathrm{N}}$ is used in each step. We can then write

$$u(\mathbf{m}_2 \mid \mathbf{m}_1, \mathscr{I}) \rightarrow u'(\mathbf{m}_2 \mid \mathbf{m}_1, \mathscr{I}_{\mathrm{N}}) \tag{7}$$

and

$$u'(\mathbf{m}_2 \mid \mathbf{m}_1, \mathscr{I}_{\mathrm{N}}) = \frac{f(\mathbf{m}_2, \mathbf{m}_1 \mid \mathscr{I}_{\mathrm{N}})}{q(\mathbf{m}_1 \mid \mathscr{I}_{\mathrm{N}})} . \tag{8}$$

Eq. (8) can be substituted into eq. (5) to yield

$$w(\mathbf{m}_1 \mid \mathbf{d}, \mathscr{I}) = \kappa\, \frac{t(\mathbf{m}_1 \mid \mathscr{I})}{q(\mathbf{m}_1 \mid \mathscr{I}_{\mathrm{N}})} \times \int_{\mathscr{M}'} f(\mathbf{m}_2, \mathbf{m}_1 \mid \mathscr{I}_{\mathrm{N}})\, l(\mathbf{m}_1, \mathbf{m}_2 \mid \mathbf{d}, \mathscr{I})\, d\mathbf{m}_2 . \tag{9}$$

The integral in the above equation is still analogous to the likelihood function for the 1-D posterior w because it is a term that carries information from the observed and modelled data. However, it is possible to see that this integral also carries prior information related to moments up to second order (i.e. mean and covariance) that is contained in f. Because of this, we define an *extended likelihood function*, denoted by l_{e} and given by

$$l_{\mathrm{e}}(\mathbf{m}_1 \mid \mathbf{d}, \mathscr{I}) = \int_{\mathscr{M}'} f(\mathbf{m}_2, \mathbf{m}_1 \mid \mathscr{I}_{\mathrm{N}})\, l(\mathbf{m}_1, \mathbf{m}_2 \mid \mathbf{d}, \mathscr{I})\, d\mathbf{m}_2 . \tag{10}$$

With this definition, eq. (9) can be rewritten as

$$w(\mathbf{m}_1 \mid \mathbf{d}, \mathscr{I}) = \kappa\, \frac{t(\mathbf{m}_1 \mid \mathscr{I})}{q(\mathbf{m}_1 \mid \mathscr{I}_{\mathrm{N}})}\, l_{\mathrm{e}}(\mathbf{m}_1 \mid \mathbf{d}, \mathscr{I}) . \tag{11}$$

This is now a 1-D version of Bayes' theorem for $\mathbf{m}_1$, which can be handled in a straightforward way if we have the extended likelihood in closed form. All steps leading to eq. (11) are summarized in Fig. 1. We next consider the case where l is also Gaussian and the forward model is linear (i.e. $\mathbf{d} = \mathbf{G}\mathbf{m} + \mathbf{n}$).

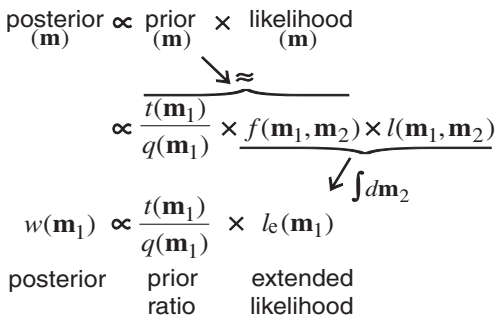

Figure 1. Schematic representation of the local Bayesian inversion. At the top level is the original multidimensional Bayesian problem involving functions of the full vector of parameters **m**. At the intermediate level the prior is approximated by the product of three functions t, $1/q$ and f, where f and q are normal distributions and t is the marginal prior distribution for parameter $\mathbf{m}_1$. Then, for a proper choice of the likelihood function l the parameters $\mathbf{m}_2$ can be integrated out of the problem, leaving a 1-D version of the Bayesian theorem.

This implies that the extended likelihood is also Gaussian. A detailed derivation of the expressions q and l_e can be found in Appendix A. Below we just present the results.

To represent solutions for the whole set of parameters, it is convenient to abandon the vector notation to represent parameters as m_j, $j=1, \ldots, M$, and the marginal posterior distribution for each parameter as

$$w_j(m_j \,|\, \mathbf{d}, \mathscr{I}) = \kappa \, \frac{t_j(m_j \,|\, \mathscr{I})}{q_j(m_j \,|\, \mathscr{I}_N)} \; l_{ej}(m_j \,|\, \mathbf{d}, \mathscr{I}), \tag{12}$$

where

$$q_j(m_j \,|\, \mathscr{I}_N) = \frac{1}{\sqrt{2\pi\sigma^2_{\text{prior}\,j}}} \exp\left[-\frac{1}{2\sigma^2_{\text{prior}\,j}} (m_j - \mu_{\text{prior}\,j})^2 \right] \tag{13}$$

and

$$l_{ej}(m_j \,|\, \mathbf{d}, \mathscr{I}) = \frac{1}{\sqrt{2\pi\sigma^2_{\text{Gauss}\,j}}} \times \exp\left[-\frac{1}{2\sigma^2_{\text{Gauss}\,j}} (m_j - \mu_{\text{Gauss}\,j})^2 \right]. \tag{14}$$

Eq. (14) corresponds to the posterior marginal distribution derived from the familiar Gaussian Bayesian formulation (see e.g. Tarantola 1987). This involves elements $\{\mu_{\text{prior}\,j}\}$ from the prior estimates for the parameter vector and the corresponding prior variances $\{\sigma^2_{\text{prior}\,j}\}$. Means and variances resulting from the Gaussian problem are represented by $\mu_{\text{Gauss}\,j}$ and $\sigma^2_{\text{Gauss}\,j}$, respectively.

3 DISCUSSION

The main goal of this section is to discuss the nature and consequences of using the approximation given by eq. (7) based on three different perspectives.

We first consider an expansion of the general prior distribution u for parameters $\mathbf{m}_2$ (see eq. 3) in terms of normal prior information $\mathscr{I}_N$ and its complementary information $\mathscr{I}_C$ (Appendix B). This expansion shows that the terms that carry information about cross-correlations of order higher than two for the parameters have been eliminated by the approximation (eq. 7). However, if the marginal prior distributions t (see eq. 12) are well constructed (that is, truly represent the state of knowledge about each parameter given by the prior information $\mathscr{I}$), marginal higher-order moment information is still being incorporated through t. Consequently, neglected cross-moment information should not influence the results significantly. The cross-moment information is necessary when lacking sufficient information on individual parameters, since it helps the determination of one parameter upon knowledge of others. In fact, when the true marginal prior distributions are known, it is possible to neglect even second-order cross-correlations, which may be desirable to facilitate the marginalization procedure when using a non-Gaussian likelihood model. The cross-correlations become important in stabilizing the computations for the estimated vector of parameters. Cross-correlations are also important whenever it is necessary to stabilize the computations, as is often the case for Gaussian likelihood functions written in a data-translated form (eq. A5).

Another way to look at the approximation is to use the maximum-entropy distribution and the ratio t/q. For this, consider a parameter m, without any particular subscript. The maximum-entropy distribution for m, for the case of a constant reference and constraints given by moment information up to Kth order (see e.g. Tarantola 1987, prob. 1.15, Jaynes 1994 or Gouveia *et al.* 1996) can be written as

$$t(m \,|\, \mathscr{I}) \propto \exp\left[-\sum_{k=1}^{K} \lambda_k m^k \right], \tag{15}$$

where the λ_k are Lagrange multipliers associated with the maximum-entropy problem. Using eq. (15) for t and eq. (13) for q, the ratio is given by

$$\frac{t(m \,|\, \mathscr{I})}{q(m \,|\, \mathscr{I}_N)} \propto \exp\left[-\left(\lambda_1 + \frac{\mu_{\text{prior}}}{\sigma^2_{\text{prior}}} \right) m - \left(\lambda_2 - \frac{1}{2\sigma^2_{\text{prior}}} \right) m^2 - \sum_{k=3}^{K} \lambda_k m^k \right]. \tag{16}$$

When the total information $\mathscr{I}$ equals the normal part $\mathscr{I}_N$ (i.e. $\mathscr{I}_C$ vanishes), the corresponding maximum-entropy problem is constrained by the first two moments of the unknown distribution. This leads to a normal distribution for the marginal prior t, as discussed in Gouveia *et al.* (1998). More precisely, the Lagrange multipliers $\lambda_k = 0$, for $k = 3, 4, \ldots$, 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 input to the maximum-entropy problem (see Tarantola 1987 problem 1.15). Of course, when constructing the maximum-entropy distribution t, for consistency, we need $\mu = \mu_{\text{prior}}$ and $\sigma^2 = \sigma^2_{\text{prior}}$. For this case, the ratio t/q is unity, which means that q is actually the normal part of t. Thus, all that is left in eq. (12) is the extended likelihood function, which is simply a multinormal Bayesian inversion procedure. This result can be summarized by saying that when all we know are the first- and second-order moments, the proposed methodology reduces to the more traditional Gaussian Bayesian formula (see e.g. Tarantola 1987).

Finally, consider a general prior distribution s in eq. (1), which is also written as

$$s(\mathbf{m}_1, \mathbf{m}_2 \mid \mathscr{I}) = t(\mathbf{m}_1 \mid \mathscr{I})\, u(\mathbf{m}_2 \mid \mathbf{m}_1, \mathscr{I}). \tag{17}$$

By using the approximation given by eq. (7), the prior s becomes

$$s(\mathbf{m}_1, \mathbf{m}_2 \mid \mathscr{I}) \approx t(\mathbf{m}_1 \mid \mathscr{I})\, \frac{f(\mathbf{m}_2, \mathbf{m}_1 \mid \mathscr{I}_N)}{q(\mathbf{m}_1 \mid \mathscr{I}_N)}. \tag{18}$$

If we apply marginalization with respect to parameters $\mathbf{m}_2$ to both sides of eq. (18), we obtain

$$\int s(\mathbf{m}_1, \mathbf{m}_2 \mid \mathscr{I})\, d\mathbf{m}_2 = t(\mathbf{m}_1 \mid \mathscr{I}), \tag{19}$$

because

$$\int \frac{f(\mathbf{m}_2, \mathbf{m}_1 \mid \mathscr{I}_N)}{q(\mathbf{m}_1 \mid \mathscr{I}_N)}\, d\mathbf{m}_2 = 1.$$

This indicates that the approximation (7) preserves the marginal prior for parameter $\mathbf{m}_1$.

These properties of the local Bayesian method make it possible to obtain reliable confidence intervals efficiently for individual parameters. However, there are some limitations. An evident limitation is that the marginalization sacrifices the cross-moment information, which is necessary to construct joint confidence regions for parameters. This means that the local approach is not recommended for such studies. Another important issue is the well-known fact that Gaussian likelihoods are sensitive to outliers. The use of more robust likelihood functions such as Laplace or Cauchy distributions would make it difficult to perform the marginalization step, because the resulting extended likelihood function would be non-Gaussian. The same complication arises when the forward model is non-linear, even in the Gaussian case. For these cases, marginalization would require methods for multidimensional integration that we have been trying to avoid. One attractive alternative is the application of asymptotic methods of approximation (Tierney *et al.* 1989; DiCiccio *et al.* 1993; Shun & McCullagh 1995), which to the authors' knowledge have never been applied to Bayesian geophysical inverse problems. To overcome modest non-linearity of the data modelling operator, one can adopt standard procedures of non-linear optimization and linearize the function around the solution point.

4 ANALYTICAL EXAMPLES

Consider the problem of density inversion from gravity data, where the sources are six rectangular cells of constant density contrasts with respect to some constant background value (Fig. 2). The true density contrasts for the cells are derived by imposing an exponential correlation function on a sequence of uncorrelated Gaussian pseudo-random numbers. From this model, the synthetic gravity field is generated using the formula for the gravity of prismatic bodies (see e.g. Telford *et al.* 1976, p. 74). The synthetic gravity data are corrupted with two different levels of uncorrelated Gaussian noise, respectively 1 and 10 per cent of the maximum synthetic gravity value, modelling errors and a combination of both random and modelling errors (dashed curves of Fig. 3). The modelling errors are generated from an additional source (the grey cell in Fig. 2) not incorporated in the interpretative model (Fig. 4).

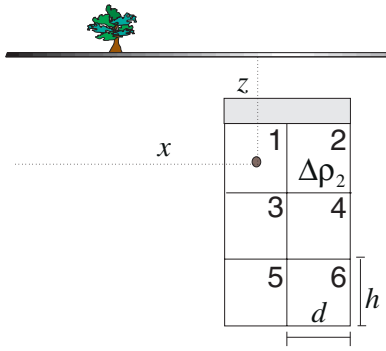

Figure 2. Simple earth model consisting of six rectangular cells, numbered 1–6, with centre coordinates (x, z), width d and height h as indicated in the figure. The problem is to estimate the density contrast in each cell from the gravity field and prior information. The grey box above the cells is used to simulate modelling errors.

For the prior information, we use the true correlation and two well logs measuring density contrasts through the cells (Fig. 4). The well logs are built from pseudo-random numbers from six different probability density distributions, so that each cell density contrast has its own underlying process (solid line plots in Fig. 5). All theoretical probability models used to generate the logs are centred on the true density contrast and truncated to the interval $[-7, 7]$. This introduces an error that causes sample means taken from the logs to deviate from the true contrasts (Table 1).

Table 1. 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.

		Mean values	
	True values	Truncated means	Sample means
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

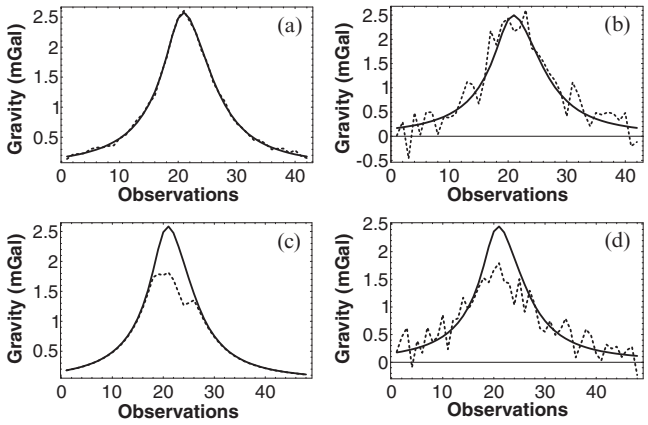

Figure 3. Estimated gravity curve (solid) from the synthetic gravity (dashed) contaminated with two levels of random noise (a and b), modelling errors (c) and modelling and random errors combined (d).

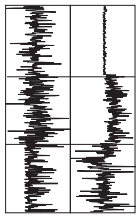

Figure 4. Interpretative model using the inversion and pseudo-random density logs generated from different distributions (Fig. 5) to represent prior information about the density contrast in each cell.

The implementation of the methodology can be summarized in three main steps: analysis of prior information for the determination of the local prior distributions; least-squares inversion (in connection with the determination of the extended likelihood); and Bayesian update, which combines the results from the two previous steps. Each step is discussed in detail below.

4.1 Prior probabilities

The local prior probabilities, t_j, are derived from application of the maximum-entropy principle using the moment information obtained from the logs of Fig. 4. Let α_{ji} be the ith log sample in the jth cell. The prior information is thus defined as $\mathscr{I}_j \equiv \{\mathbf{a}_j : \mathbf{a}_j = (\alpha_{j1}, \ldots, \alpha_{jN})\}$, where $\mathbf{a}_j$ is the vector of log samples corresponding to the jth cell and $N = 1000$ in this example. Therefore, the required moments can be found by using just sample averages given by

$$\langle \alpha \rangle_{kj} = \frac{1}{N} \sum_{i=1}^{N} \alpha_{ji}^{k}, \tag{20}$$

where k is the order of the average value $\langle \alpha \rangle$. The subscript notation for $\mathscr{I}$ reflects the fact that by construction each set of samples corresponding to a cell of the model was generated independently of the others. This greatly simplifies the estimation of the moments, as can be seen from eq. (20). In real applications, methods for conditional moment estimation such as parametric geostatistics have to be employed. This will provide the local uncertainty given the spatial variability of the medium.

The computed sample moments up to fourth order and a constant reference are used to determine the Lagrange multipliers of the maximum-entropy distribution following Mead & Papanicolau (1984). In particular, the implementation for this example uses the Newton method with a line search. The iterations of the algorithm are stopped when the moments of the resulting maximum-entropy distribution agree with the input sample moments to the order of 10^{-6} or better. For this example, it usually takes six or seven iterations for convergence. The final approximations of the theoretical distributions are shown by dashed lines in Fig. 5. The overall agreement between the estimated and true distributions is good, except for the Cauchy distribution (number 6 in Fig. 5).

The maximum-entropy distribution (t_j) is normalized by the corresponding Gaussian marginal prior distribution (q_j) according to eq. (12). The logic behind this procedure is that it avoids incorporating the same information into the problem twice. q_j is given by eq. (13), which is easily constructed by making

$$\mu_{\text{prior}\,j} = \langle \alpha \rangle_{1j} \tag{21}$$

and

$$\sigma^2_{\text{prior}\,j} = \langle \alpha \rangle_{2j} - \langle \alpha \rangle^2_{1j}, \; j = 1, \ldots, 6\,. \tag{22}$$

In real Earth applications, one possible alternative to using maximum-entropy constrained by conditional moment information is the application of non-parametric geostatistics. These methods can take advantage of the much smaller length scales usually found in the subsurface (local) data in comparison with the surface data, thus allowing for inferences about one parameter independently of the others. In this way, we achieve general treatment of the prior information, but the computational cost is reduced by not having to form the full multidimensional prior distribution. The use of both non-parametric geostatistics and maximum-entropy methods is discussed in detail by Moraes (1996).

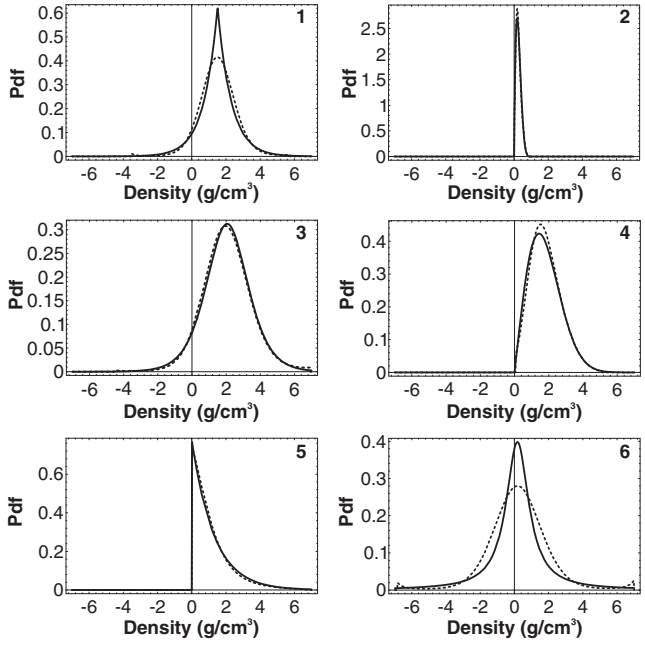

Figure 5. Probability density functions (pdfs) selected from theoretical models to simulate the prior information corresponding to the density contrast in each cell (solid lines). The models in each cell corresponding to the numbers 1 to 6 are 1 Laplace, 2 beta, 3 logistic, 4 Raleigh, 5 exponential and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

4.2 Least squares (extended likelihood)

The extended likelihood corresponds to marginals resulting from the well-known Bayesian inversion using normal variables,

which is extensively discussed in the literature (e.g. Tarantola 1987). Therefore, the parameters of the extended likelihood (eq. 14) can be estimated using conventional stochastic least squares, which requires $\mathbf{m}_{\text{prior}}$, $\mathbf{C}_{\text{m}}$ and $\mathbf{C}_{\text{d}}$, the prior vector of parameters and covariance matrices for parameters and data, respectively.

$\mathbf{m}_{\text{prior}}$ is the vector whose elements are prior means $\{\mu_{\text{prior}\,j}\}$, $j=1, \ldots, 6$. $\mathbf{C}_{\text{m}}$ is built by combining the information from the correlation matrix of the parameters, which is assumed known, and the first two sample moments. Thus, if the normalized correlations for parameters i and j are represented by r_{ij}, the covariance matrix is given by

$$\mathbf{C}_{\text{m}} = \begin{bmatrix} \sigma_1^2 r_{11} & \cdots & \sigma_1\sigma_6 r_{16} \\ \vdots & \ddots & \vdots \\ \sigma_6\sigma_1 r_{61} & \cdots & \sigma_6^2 r_{66} \end{bmatrix}. \tag{23}$$

The data error covariance matrix $\mathbf{C}_{\text{d}}$ can be defined as either the random noise covariance $\mathbf{C}_{\text{o}}$, the covariance for modelling error $\mathbf{C}_{\text{t}}$, or by $\mathbf{C}_{\text{o}}+\mathbf{C}_{\text{t}}$ if both types of errors are present (see e.g. Tarantola 1987, p. 68). Both are computed as the power of the noise vector $\mathbf{n}$, which can be written as

$$\mathbf{C}_{\text{o}} = \text{diag}(n_i^2) \quad \text{and} \quad \mathbf{C}_{\text{t}} = \mathbf{G}_n \boldsymbol{\Delta}\rho_n \boldsymbol{\Delta}\rho_n^{\text{T}} \mathbf{G}_n^{\text{T}}, \tag{24}$$

where $\boldsymbol{\Delta}\rho_n$ and $\mathbf{G}_n$ are, respectively, the density contrast and the Green's function for all the effects not accounted for in the parametrization of the problem, the grey box of Fig. 2 in this case. The above quantities allow for the computation of both the estimated vector of parameters $\mathbf{m}_{\text{Gauss}}$ and the covariance $\mathbf{C}_{\text{p}}$ by least squares (see eq. A6 in Appendix A). It is important to note that the least squares need to be solved only once. To determine the extended likelihood at each iteration, $\mu_{\text{Gauss}\,j}$ and $\sigma^2_{\text{Gauss}\,j}$ are given, respectively, by the jth elements of $\mathbf{m}_{\text{Gauss}}$ and the diagonal of $\mathbf{C}_{\text{p}}$.

4.3 Bayesian update

This last step just involves computing the product between the extended likelihood function and the prior ratio t_j/q_j according to eq. (12). This can be better illustrated by examining Fig. 6, which shows the prior ratio and the extended likelihood for all cells. The extended likelihood (solid line) is Gaussian and the prior ratio (dashed line) has an unusual shape imposed by the Gaussian normalization applied to the maximum-entropy distribution. This multiplication process can be interpreted as a correction to the shape of the extended likelihood to account for the marginal moments of order higher than two. As a result, the final posterior distribution is non-Gaussian. In addition, because all functions involved in this example are known analytically, it is possible to find the posterior distribution in closed form, combining eqs (12), (13), (14) and (15) (after determination of the Lagrange multipliers).

The first series of tests investigates the behaviour of the methodology applied to data contaminated by different levels of random noise (Figs 3a and b). The marginal posteriors,

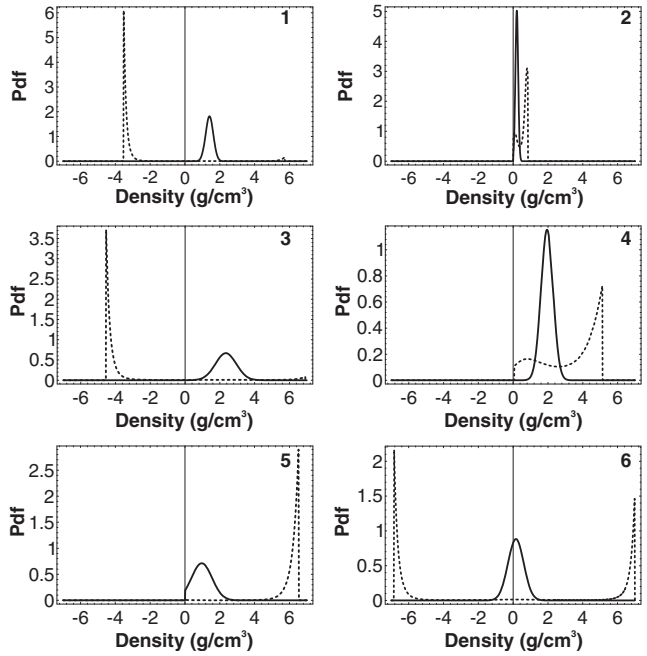

Figure 6. The two probability density functions (pdfs) that comprise the 1-D Bayes theorem of the proposed methodology. They are the prior ratio, which is the maximum-entropy prior normalized by the Gaussian prior, and the extended likelihood, which is equivalent to a Gaussian posterior. The density axis corresponds to values for the density contrast of the corresponding cell indicated by the numbers 1–6.

considering the 10 per cent noise-level problem, are given by

$$w_1(m_1) = 1.5\ 10^{-9} \exp(29.6\, m_1 - 10.3\, m_1^2 - 0.12\, m_1^3 + 0.02\, m_1^4), \tag{25}$$

$$w_2(m_2) = 0.1 \exp(42.5\, m_2 - 144.3\, m_2^2 + 131.8\, m_2^3 - 77.3\, m_2^4), \tag{26}$$

$$w_3(m_3) = 2.0\ 10^{-4} \exp(6.8\, m_3 - 1.4\, m_3^2 - 0.04\, m_3^3 + 0.006\, m_3^4), \tag{27}$$

$$w_4(m_4) = 1.2\ 10^{-7} \exp(17.7\, m_4 - 5.4\, m_4^2 + 0.3\, m_4^3 - 0.03\, m_4^4), \tag{28}$$

$$w_5(m_5) = 0.5 \exp(1.4\, m_5 - 1.3\, m_5^2 + 0.07\, m_5^3 - 0.006\, m_5^4) \tag{29}$$

and

$$w_6(m_6) = 0.8 \exp(0.8\, m_6 - 2.6\, m_6^2 - 0.002\, m_6^3 + 0.005\, m_6^4), \tag{30}$$

where all distributions are conditioned on $\mathbf{d}$ and $\mathscr{I}$. These distributions are shown in Fig. 7. Table 2 presents the resulting means and modes of the posteriors in comparison with the mean of the normal extended likelihood (which is also the mode in this case).

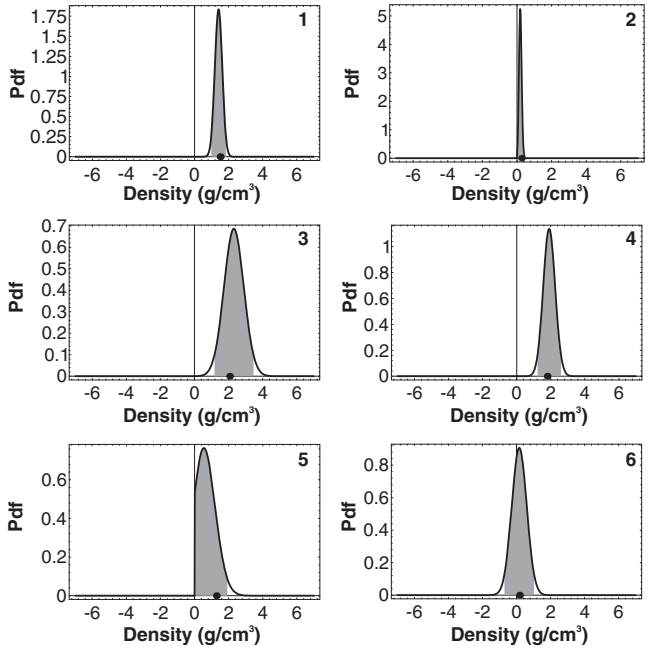

Figure 7. Inversion results depicted by the posterior marginal for each parameter. The 95 per cent interquantile regions are represented by the shaded areas and the true values for the parameter are given by the solid circles. This example uses 10 per cent of the maximum gravity value as the standard deviation for the noise. The density axis corresponds to values for the density contrast of the corresponding cell indicated by the numbers 1–6.

At very low noise levels (1 per cent), there is a general agreement between the three estimates. As we increase the noise level to 10 per cent of the maximum gravity value, the mean and the mode of the posterior distribution start to move apart, but the mean values for the posterior and for the extended likelihood are still very close. Additional examples (not shown) indicate that this behaviour continues for progressively higher noise conditions up to a point where the mode differs significantly from the mean. Overall, the two mean values are equivalent, with only a marginal advantage for the posterior mean in the case of extreme noise (at 80 per cent noise level or higher). Cases where the mean for the posterior has moved away from the true parameter in comparison with the mean for the extended likelihood are indicated by bold numbers in Table 2. Table 3 shows a comparison amongst the variances computed from the prior, the extended likelihood (Gaussian) and the posterior distributions. Overall, there is a marginal reduction in the posterior variances when compared with the Gaussian variances. However, on a few occasions, shown as bold numbers in Table 3, the posterior variances increase. These increases tend to be associated with the asymmetrical prior distributions.

Table 2. Means and modes of the marginal posterior distributions and $\mathbf{m}_{\text{Gauss}}$ (the least-squares solution). Bold numbers indicate that the posterior mean moved away from the true values in comparison with $\mathbf{m}_{\text{Gauss}}$.

	Parameter estimates						
		1% noise			10% noise		
	True	Mean	Mode	$\mathbf{m}_{\text{Gauss}}$	Mean	Mode	$\mathbf{m}_{\text{Gauss}}$
1	1.506	1.491	1.491	1.491	1.410	1.410	1.408
2	0.260	0.291	0.291	0.293	**0.197**	0.189	0.210
3	2.060	2.163	2.163	2.166	2.316	2.310	2.346
4	1.796	**1.659**	1.659	1.663	1.912	1.906	1.947
5	1.284	**1.135**	1.135	1.163	**0.784**	0.549	0.966
6	0.174	**0.303**	0.303	0.304	**0.160**	0.160	0.161

Table 3. Comparison between the variances computed from the prior, the extended likelihood (Gaussian) and the posterior distributions at different noise levels. Bold numbers indicate posterior variances that are greater than their Gaussian counterparts.

	Variances				
		1% noise		10% noise	
	Prior	Gaussian	Posterior	Gaussian	Posterior
1	1.19684	0.00168	0.00168	0.04814	0.04727
2	0.02227	0.00061	**0.00062**	0.00629	0.00567
3	2.04926	0.03321	0.03302	0.35629	0.33958
4	0.81267	0.00947	0.00947	0.11950	**0.12284**
5	1.22102	0.03151	**0.03214**	0.34828	0.25422
6	3.21465	0.01127	0.01124	0.20426	0.19390

Fig. 7 also shows shaded areas representing 95 per cent probability regions. These areas also define on the horizontal axis interquantile intervals, which can be obtained independently of any estimates for the density contrasts. Density estimates, however, are still necessary to compute the estimated gravity field used in the fitting procedure, which is shown in Fig. 3. The mean of each posterior distribution is used 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. Table 4 shows the results for 3 runs using the 10 per cent noise level. The overall behaviour of the solutions is basically the same as discussed above, denoting that the inversion is stable.

The next series of tests considers modelling errors. In the case where only the modelling errors are considered, the data covariance is $\mathbf{C}_t$ and the inversion results are shown in Table 5. The results show that when the correct error information is introduced, the estimated parameters match the true ones almost exactly. When random and modelling errors are combined, the data covariance becomes the sum $\mathbf{C}_t + \mathbf{C}_o$. The results for this case (Table 5) are comparable with the other tests for random noise only. These tests indicate that the right covariance information completely eliminates the effect of unmodelled sources.

Table 4. 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 $\mathbf{m}_{\text{Gauss}}$ (first column).

	Parameter estimates					
	First run		Second run		Third run	
	$\mathbf{m}_{\text{Gauss}}$	Mean	$\mathbf{m}_{\text{Gauss}}$	Mean	$\mathbf{m}_{\text{Gauss}}$	Mean
1	1.301	1.305	1.523	**1.521**	1.464	1.464
2	0.211	**0.198**	0.201	**0.191**	0.242	**0.225**
3	2.408	2.372	2.138	2.115	2.313	2.288
4	1.691	**1.645**	1.366	**1.323**	1.666	**1.617**
5	1.598	1.389	1.293	**1.088**	0.929	**0.758**
6	−0.160	−0.143	0.499	0.473	0.416	0.404

Table 5. Results for inversion considering modelling errors only and a combination of modelling and random errors (10 per cent noise level).

		Modelling errors			
		Modelling		Modelling + Random	
	True	Mean	Variance	Mean	Variance
1	1.506	1.506	10^{-16}	1.527	0.067
2	0.260	0.260	10^{-16}	0.206	0.011
3	2.060	2.060	10^{-16}	1.513	0.363
4	1.796	1.796	10^{-16}	1.643	0.166
5	1.284	1.284	10^{-17}	1.276	0.246
6	0.174	0.174	10^{-19}	0.689	0.372

5 CONCLUSIONS

Probability densities give a synthetic representation of what we know about parameters of earth models, providing an adequate framework to integrate information of diverse origin. In addition, they can serve many different purposes such as the estimation of parameters (e.g. mean or mode of the posterior distribution), uncertainty analysis (e.g. variance and confidence intervals) and simulations (e.g. sampling models from the posterior). Therefore, it is important to focus attention on difficulties that have prevented wide application of Bayesian methods. These difficulties include the specification of the prior probability and the marginalization of the posterior to extract information about specific parameters.

The most important contribution of this research is that it offers an alternative strategy for treating complex multi-dimensional problems by reducing the dimensionality of the problem before the final solution is found. When this is done, the main difficulties in Bayesian inference are automatically addressed. To construct prior distributions, we make available methods of probability density estimation that are awkward in many dimensions such as maximum-entropy and non-parametric density estimation methods. These methods allow one to process subsurface data into marginal distributions, which can be directly incorporated into the calculations via the proposed formulation. Local quantities are a safety device against non-homogeneity in the medium. Of course, the permissible degree of locality is a function of the density of information in a particular region. When the information is sparse, we can expect that local methods will perform like global ones. In this sense, even when all distributions in the proposed methodology are Gaussian, it is important to determine the mean vector and the covariance matrix based on the local information. In addition, the other difficulty of extracting information from the posterior practically disappears because this methodology produces posterior distributions that are already 1-D.

As discussed in Section 3, the proposed methodology is suitable for inferences about individual parameters and is not recommended for high-dimensional problems such as the construction of joint confidence regions. Other limitations arise in the cases of non-Gaussian likelihoods and fully non-linear forward models, where marginalization cannot be performed in closed form. This is one topic where new research can bring significant advances to the methodology.

Finally, a small synthetic problem demonstrates the applicability of the method. In particular, it shows that local probabilities are an efficient way to represent prior knowledge (or uncertainty) about parameters. Using moments of the first four orders in a maximum-entropy probability density, a variety of probability densities are well approximated. The wider applicability of the methodology is currently being investigated with real data examples.

ACKNOWLEDGMENTS

The authors are grateful for the reviews by Philip Stark and an anonymous reviewer. 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, Colorado School of Mines, and the Shell Foundation. In addition, FSM acknowledges the support of the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil).

REFERENCES

Backus, G.E., 1988a. Bayesian inferrence in geomagnetism, *Geophys. J. Int.,* **92,** 125–142.

Backus, G.E., 1988b. Comparing hard and soft prior bounds in geophysical inverse problems, *Geophys. J. Int.,* **94,** 249–261.

Backus, G.E., 1989. Confidence set inference with a prior quadratic bound, *Geophys. J. Int.,* **97,** 119–150.

Backus, G.E., 1996. Trimming and procastination as inverse techniques, *Phys. Earth planet Inter.,* **98,** 101–142.

Box, G.E.P. & Tiao, G.C., 1973. *Bayesian Inference in Statistical Analysis,* Addison-Wesley, Reading, MA.

Cowles, M. & Carlin, B., 1996. Markov chain Monte Carlo convergence diagnostics: a comparative review, *J. Am. stat. Assoc.,* 883–904.

DiCiccio, T.J., Martin, M.A. & Young, G.A., 1993. Analytical approximations to conditional distribution functions, *Biometrika,* **80,** 781–790.

Duijndam, A.J.W., 1988a. Bayesian estimation in seismic inversion. Part i: Principles, *Geophys. Prospect.,* **36,** 878–898.

Duijndam, A.J.W., 1988b. Bayesian estimation in seismic inversion. Part ii: Uncertainty analysis, *Geophys. Prospect.,* **36,** 898–918.

Fitzpatrick, B.G., 1991. Bayesian analysis in inverse problems, *Inverse Problems,* **7,** 675–702.

Gouveia, W. & Scales, J.A., 1998. Bayesian seismic waveform inversion: parameter estimation and uncertainty analysis, *J. geophys. Res.,* **103,** 2759–2779.

Gouveia, W., Moraes, F.S. & Scales, J.A., 1996. Entropy, information and inversion, in *1996 CWP Project Review,* Colorado School of Mines (http://www.cwp.mines.edu/).

Gradshteyn, I.S. & Ryzhik, I.M., 1980. *Table of Integrals, Series, and Products,* corrected and enlarged edn, Academic Press, New York.

Graybill, F.A., 1983. *Matrices With Applications in Statistics,* 2nd edn, Wadsworth.

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

Jaynes, E.T., 1968. Prior probabilities, *IEEE Trans. Systems Cybernetics,* **SSC-4,** 227–241.

Jaynes, E.T., 1978. Where do we stand on maximum entropy?, in *The Maximum Entropy Formalism,* pp. 15–118, eds Levine, R.D. & Tribus, M., MIT Press, Cambridge, MA.

Jaynes, E.T., 1994. *Probability Theory: The Logic of Science,* http://Bayes.Wustl.Edu.

Jeffreys, H., 1939. *Theory of Probability,* Clarendon Press, Oxford.

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

Moraes, F.S., 1996. The application of marginalization and local distributions to multidimensional Bayesian inverse problems, *PhD thesis,* Colorado School of Mines, Golden, CO.

Mosegaard, K. & Tarantola, A., 1995. Monte Carlo sampling of solutions to inverse problems, *J. geophys. Res.,* **100,** 12 431–12 447.

Parker, R.L., 1975. The theory of ideal bodies for gravity interpretation, *Geophys. J. R. astr. Soc.,* **42,** 315–334.

Parker, R.L., 1994. *Geophysical Inverse Theory,* Princeton University Press, Princeton.

Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T., 1992. *Numerical Recipes in C,* 2nd edn, Cambrige University Press, Cambridge.

Richard, V., Bayer, R. & Cuer, M., 1984. An attempt to formulate well-posed questions in gravity: application of linear inverse techniques to mining exploration, *Geophysics,* **49,** 1781–1793.

Sacchi, M.D. & Ulrych, T., 1995. High-resolution velocity gathers and offset space reconstruction, *Geophysics,* **60,** 1169–1177.

Scales, J., 1996. Uncertainties in seismic inverse calculations, in *Inverse Methods,* pp. 79–97, eds Jacobson, B., Mosegaard, K. & Sibani, P., Springer-Verlag, Berlin.

Scales, J.A. & Sneider, R., 1997. To Bayes or not to Bayes?, *Geophysics,* **62,** 1047–1048.

Scott, D.W., 1992. *Multivariate Density Estimation. Theory, Practice and Visualization,* Wiley, New York.

Shun, Z. & McCullagh, P., 1995. Laplace approximation of high dimensional integrals, *J. R. stat. Soc. B,* **57,** 749–760.

Stark, P.B., 1992a. Inference in infinite-dimensional inverse problems: discretization and duality, *J. geophys. Res.,* **97,** 14 055–14 082.

Stark, P.B., 1992b. Minimax confidence intervals in geomagnetism, *Geophys. J. Int.,* **108,** 329–338.

Tarantola, A., 1987. *Inverse Problem Theory—Methods for Data Fitting and Model Parameter Estimation,* Elsevier, Amsterdam.

Tarits, P., Jouanne, V., Menvielle, M. & Roussignol, M., 1994. Bayesian statistics of non-linear inverse problems: example of the magnetotelluric 1-D inverse problem, *Geophys. J. Int.,* **119,** 353–368.

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

Tierney, L., 1994. Markov chains for exploring posterior distributions, *Ann. Stat.,* 1701–1728.

Tierney, L., Kass, R.E. & Kadane, J.B., 1989. Approximate marginal densities of nonlinear functions, *Biometrika,* **76,** 425–433.

Yabuki, T. & Matsu'ura, M., 1992. Geodetic data inversion using a Bayesian information criterion for spatial distribution of fault slip, *Geophys. J. Int.,* **109,** 363–375.

APPENDIX A: MARGINALIZATION OF GAUSSIAN DISTRIBUTIONS

The Gaussian approximation for the prior distribution for parameters $\mathbf{m}_2$ defined in eq. (8) requires that we find q, which is the marginal of f. Let $f \sim N(\mathbf{m}_{\text{prior}}, \mathbf{C}_{\text{m}})$ be given by

$$f(\mathbf{m} \mid \mathscr{I}_{\text{N}}) = (2\pi)^{-M/2} \mid \mathbf{C}_{\text{m}} \mid^{-1/2} \times \exp\left[-\frac{1}{2} (\mathbf{m}-\mathbf{m}_{\text{prior}})^{\text{T}} \mathbf{C}_{\text{m}}^{-1} (\mathbf{m}-\mathbf{m}_{\text{prior}}) \right]. \quad \text{(A1)}$$

The covariance matrix can be partitioned as

$$\mathbf{C}_{\text{m}} = \begin{bmatrix} \mathbf{C}_{\text{m}_{11}} & \mathbf{C}_{\text{m}_{12}} \\ \mathbf{C}_{\text{m}_{21}} & \mathbf{C}_{\text{m}_{22}} \end{bmatrix}, \quad \text{(A2)}$$

where $\mathbf{C}_{\text{m}_{11}}$, $\mathbf{C}_{\text{m}_{12}}^{\text{T}} = \mathbf{C}_{\text{m}_{21}}$ and $\mathbf{C}_{\text{m}_{22}}$ are 1×1, $(M-1) \times 1$ and $(M-1) \times (M-1)$ matrices, respectively. With these definitions, q can be found by integration of all parameters but $\mathbf{m}_1$. For normal distributions the result is also normal (see e.g. Theorem 10.6.1 of Graybill 1983) and can be written as

$$q(\mathbf{m}_1 \mid \mathscr{I}_{\text{N}}) = (2\pi)^{-1/2} \mid \mathbf{C}_{\text{m}11} \mid^{-1/2} \times \exp\left[-\frac{1}{2} (\mathbf{m}_1 - \mathbf{m}_{\text{prior }1})^{\text{T}} \mathbf{C}_{\text{m}11}^{-1} (\mathbf{m}_1 - \mathbf{m}_{\text{prior }1}) \right]. \quad \text{(A3)}$$

Since we are defining $\mathbf{m}_1$ as a 1-D vector, we may drop the vector notation and introduce the prior variances σ^2_{prior} defined as the diagonal elements of $\mathbf{C}_{\text{m}}$. This yields eq. (13).

We now look at the extended likelihood function for the Gaussian case. Taking the definition given in eq. (10) and considering linear forward modelling, we can write

$$l_{\text{e}}(\mathbf{m}_1 \mid \mathbf{d}, \mathscr{I}_{\text{N}}) = (2\pi)^{-N+M/2} \mid \mathbf{C}_{\text{d}} \mid^{-1/2} \mid \mathbf{C}_{\text{m}} \mid^{-1/2} \times \int_{\mathscr{M}'} \exp\left\{ -\frac{1}{2} [(\mathbf{d}-\mathbf{G}\mathbf{m})^{\text{T}} \mathbf{C}_{\text{d}}^{-1} (\mathbf{d}-\mathbf{G}\mathbf{m}) + (\mathbf{m}-\mathbf{m}_{\text{prior}})^{\text{T}} \mathbf{C}_{\text{m}}^{-1} (\mathbf{m}-\mathbf{m}_{\text{prior}})] \right\} d\mathbf{m}_2 . \quad \text{(A4)}$$

Representing eq. (A4) by I, we rewrite it in a data-translated form to obtain

$$I = (2\pi)^{-N+M/2} \mid \mathbf{C}_{\text{d}} \mid^{-1/2} \mid \mathbf{C}_{\text{m}} \mid^{-1/2} \exp[S(\mathbf{m}_{\text{Gauss}})] \times \int_R \exp\left\{ -\frac{1}{2} [(\mathbf{m}-\mathbf{m}_{\text{Gauss}})^{\text{T}} \mathbf{C}_{\text{p}}^{-1} (\mathbf{m}-\mathbf{m}_{\text{Gauss}})] \right\} d\mathbf{m}_2 , \quad \text{(A5)}$$

where $\mathbf{m}_{\text{Gauss}}$ is the estimated vector of parameters given by conventional least squares,

$$\mathbf{m}_{\text{Gauss}} = \mathbf{m}_{\text{prior}} + \mathbf{C}_{\text{p}} \mathbf{G}^{\text{T}} \mathbf{C}_{\text{d}}^{-1} (\mathbf{d} - \mathbf{G}\mathbf{m}_{\text{prior}}) , \quad \text{(A6)}$$

with the posterior covariance matrix $\mathbf{C}_{\text{p}}$ given by

$$\mathbf{C}_{\text{p}} = (\mathbf{G}^{\text{T}} \mathbf{C}_{\text{d}}^{-1} \mathbf{G} + \mathbf{C}_{\text{m}}^{-1})^{-1} . \quad \text{(A7)}$$

$S(\mathbf{m}_{\text{Gauss}})$ is the estimated misfit value given by

$$S(\mathbf{m}_{\text{Gauss}}) = (\mathbf{d}-\mathbf{G}\mathbf{m}_{\text{Gauss}})^{\text{T}} \mathbf{C}_{\text{d}}^{-1} (\mathbf{d}-\mathbf{G}\mathbf{m}_{\text{Gauss}}) \quad \text{(A8)}$$

$$+ (\mathbf{m}_{\text{Gauss}} - \mathbf{m}_{\text{prior}})^{\text{T}} \mathbf{C}_{\text{m}}^{-1} (\mathbf{m}_{\text{Gauss}} - \mathbf{m}_{\text{prior}}) . \quad \text{(A9)}$$

The next main problem is to evaluate the integral I for parameters $\mathbf{m}_2$. This is facilitated if we suppress all the constant terms in eq. (A5) to obtain

$$I_2 = \int_R \exp\left\{ -\frac{1}{2} [(\mathbf{m}-\mathbf{m}_{\text{Gauss}})^{\text{T}} \mathbf{C}_{\text{p}}^{-1} (\mathbf{m}-\mathbf{m}_{\text{Gauss}})] \right\} d\mathbf{m}_2 . \quad \text{(A10)}$$

We now let the inverse of the covariance be given by

$$\mathbf{R} = \mathbf{C}_{\text{p}}^{-1} = \mathbf{G}^{\text{T}} \mathbf{C}_{\text{d}}^{-1} \mathbf{G} + \mathbf{C}_{\text{m}}^{-1} . \quad \text{(A11)}$$

The above matrices can be partitioned as

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{bmatrix} \text{ and } \mathbf{C}_{\text{p}} = \begin{bmatrix} \mathbf{C}_{\text{p}_{11}} & \mathbf{C}_{\text{p}_{12}} \\ \mathbf{C}_{\text{p}_{21}} & \mathbf{C}_{\text{p}_{22}} \end{bmatrix}.$$

According to theorem 10.6.1 of Graybill (1983),

$$I_2 = (2\pi)^{-M-1/2} \mid \mathbf{R}_{22} \mid^{-1/2} \times \exp\left\{ -\frac{1}{2} [(\mathbf{m}_1 - \mathbf{m}_{\text{Gauss }1})^{\text{T}} \mathbf{C}_{\text{p}11}^{-1} (\mathbf{m}_1 - \mathbf{m}_{\text{Gauss }1})] \right\}. \quad \text{(A12)}$$

Substituting this result back into eq. (A5), we finally obtain

$$I = (2\pi)^{-N+1/2} \mid \mathbf{C}_{\text{d}} \mid^{-1/2} \mid \mathbf{C}_{\text{m}} \mid^{-1/2} \times \mid \mathbf{R}_{22} \mid^{-1/2} \exp\left[-\frac{1}{2} S(\mathbf{m}_{\text{Gauss}}) \right] \times \exp\left\{ -\frac{1}{2} [(\mathbf{m}_1 - \mathbf{m}_{\text{Gauss }1})^{\text{T}} \mathbf{C}_{\text{p}11}^{-1} (\mathbf{m}_1 - \mathbf{m}_{\text{Gauss }1})] \right\}. \quad \text{(A13)}$$

Evaluating the constant terms, dropping the vector notation and introducing the Gaussian posterior variances σ^2_{prior} as the diagonal entries of $\mathbf{C}_{\text{p}}$ leads to the definition of the extended likelihood (eq. 14).

The above derivation for the marginal extended likelihood assumes that the inverse matrix appearing in the expression for the posterior covariance matrix exists. This is often not the case in ill-posed problems. For such cases, regularization needs to be applied and, as result, the marginal extended likelihood will be approximated. On the other hand, nothing in Bayesian theory requires that we write the likelihood function in the data-translated form (that is, no explicit inversion is required). We may instead expand the argument of the exponential in eq. (A4) to obtain

$$l_{\text{e}}(\mathbf{m}_1 \mid \mathbf{d}, \mathscr{I}_{\text{N}}) = (2\pi)^{-N+M/2} \mid \mathbf{C}_{\text{d}} \mid^{-1/2} \mid \mathbf{C}_{\text{m}} \mid^{-1/2} \times \exp(\mathbf{d}^{\text{T}} \mathbf{C}_{\text{d}}^{-1} \mathbf{d} + \mathbf{m}_{\text{prior}}^{\text{T}} \mathbf{C}_{\text{m}}^{-1} \mathbf{m}_{\text{prior}}) \times \int_{\mathscr{M}'} \exp\left\{ -\frac{1}{2} [\mathbf{m}^{\text{T}} (\mathbf{G}^{\text{T}} \mathbf{C}_{\text{d}}^{-1} \mathbf{G} + \mathbf{C}_{\text{m}}^{-1}) \mathbf{m} - \mathbf{2}(\mathbf{d}^{\text{T}} \mathbf{C}_{\text{d}}^{-1} \mathbf{G} + \mathbf{m}_{\text{prior}}^{\text{T}} \mathbf{C}_{\text{m}}^{-1}) \mathbf{m}] \right\} d\mathbf{m}_2 , \quad \text{(A14)}$$

or simply

$$l_{\text{e}}(\mathbf{m}_1 \mid \mathbf{d}\ \mathscr{I}_{\text{N}}) = \beta \int_{\mathscr{M}'} \exp\left[-\frac{1}{2} (\mathbf{m}^{\text{T}} \mathbf{A} \mathbf{m} - 2\mathbf{b}^{\text{T}} \mathbf{m}) \right] d\mathbf{m}_2 . \quad \text{(A15)}$$

The resulting marginal extended likelihood is available in closed form (see e.g. Gradshteyn & Ryzhik 1980, p. 307) and can be expressed in the form

$$l_{\text{e}}(\mathbf{m}_1 \mid \mathbf{d}, \mathscr{I}_{\text{N}}) = \gamma \exp\left[-\frac{1}{2} (\mathbf{m}_1^{\text{T}} \mathbf{E} \mathbf{m}_1 - 2\mathbf{f}^{\text{T}} \mathbf{m}_1) \right], \quad \text{(A16)}$$

where $\mathbf{E}$ and $\mathbf{f}$ are constant coefficients. Eq. (14) is obtained by dropping vector notation, completing the squares in the argument of the exponential and evaluating the normalizing constant γ.

APPENDIX B: NATURE OF THE APPROXIMATION

We can better understand the nature of the approximation made in eq. (7), considering the definitions for $\mathscr{I}$, $\mathscr{I}_{\text{N}}$ and $\mathscr{I}_{\text{C}}$ in eq. (6), by fully expanding the conditional probability u according to

$$P(\mathbf{m}_2 \mid \mathbf{m}_1, \mathscr{I}) = P(\mathbf{m}_2 \mid \mathbf{m}_1, \mathscr{I}_{\text{N}} + \mathscr{I}_{\text{C}}) = \frac{S_1}{S_2}, \quad \text{(B1)}$$

where in general

$$S_1 = P(\mathbf{m}_2, \mathbf{m}_1 \mid \mathscr{I}_{\text{N}})\, P(\mathscr{I}_{\text{N}}) + P(\mathbf{m}_2, \mathbf{m}_1 \mid \mathscr{I}_{\text{C}}), \times P(\mathscr{I}_{\text{C}}) - P(\mathbf{m}_2, \mathbf{m}_1 \mid \mathscr{I}_{\text{N}}, \mathscr{I}_{\text{C}})\, P(\mathscr{I}_{\text{N}}, \mathscr{I}_{\text{C}}) \quad \text{(B2)}$$

and

$$S_2 = P(\mathbf{m}_1 \mid \mathscr{I}_{\text{N}})\, P(\mathscr{I}_{\text{N}}) + P(\mathbf{m}_1 \mid \mathscr{I}_{\text{C}})\, P(\mathscr{I}_{\text{C}}) - P(\mathbf{m}_1 \mid \mathscr{I}_{\text{N}}, \mathscr{I}_{\text{C}})\, P(\mathscr{I}_{\text{N}}, \mathscr{I}_{\text{C}}). \quad \text{(B3)}$$

However, according to the definition of $\mathscr{I}_{\text{N}}$ and $\mathscr{I}_{\text{C}}$, they are independent, in which case

$$S_1 = P(\mathbf{m}_2, \mathbf{m}_1 \mid \mathscr{I}_{\text{N}})\, P(\mathscr{I}_{\text{N}}) + P(\mathbf{m}_2, \mathbf{m}_1 \mid \mathscr{I}_{\text{C}})\, P(\mathscr{I}_{\text{C}}) \quad \text{(B4)}$$

and

$$S_2 = P(\mathbf{m}_1 \mid \mathscr{I}_{\text{N}})\, P(\mathscr{I}_{\text{N}}) + P(\mathbf{m}_1 \mid \mathscr{I}_{\text{C}})\, P(\mathscr{I}_{\text{C}}). \quad \text{(B5)}$$

In either case

$$P(\mathscr{I}_{\text{N}}) + P(\mathscr{I}_{\text{C}}) - P(\mathscr{I}_{\text{N}}, \mathscr{I}_{\text{C}}) = 1 \quad \text{(B6)}$$

or

$$P(\mathscr{I}_{\text{N}}) + P(\mathscr{I}_{\text{C}}) = 1. \quad \text{(B7)}$$

Thus the statement made by eq. (7) becomes clear. It says that the weight of the information $\mathscr{I}_{\text{N}}$ for $\mathbf{m}_2$ is such that $P(\mathscr{I}_{\text{N}}) \gg P(\mathscr{I}_{\text{C}}) - P(\mathscr{I}_{\text{N}}, \mathscr{I}_{\text{C}})$ or $P(\mathscr{I}_{\text{N}}) \gg P(\mathscr{I}_{\text{C}})$, depending on the case. This amounts to having practically no information about cross-correlations of order higher than two for the parameters.