16

Probabilistic Approach to Inverse Problems

Klaus Mosegaard
Niels Bohr Institute, Copenhagen, Denmark
Albert Tarantola
Institut de Physique du Globe, Paris, France

1. Introduction

In 'inverse problems' data from indirect measurements are used to estimate unknown parameters of physical systems. Uncertain data (possibly vague) prior information on model parameters, and a physical theory relating the model parameters to the observations are the fundamental elements of any inverse problem. Using concepts from probability theory, a consistent formulation of inverse problems can be made, and, while the most general solution of the inverse problem requires extensive use of Monte Carlo methods, special hypotheses (e.g., Gaussian uncertainties) allow, in some cases, an analytical solution to part of the problem (e.g., using the method of least squares).

1.1 General Comments

Given a physical system, the 'forward' or 'direct' problem consists, by definition, in using a physical theory to predict the outcome of possible experiments. In classical physics this problem has a unique solution. For instance, given a seismic model of the whole Earth (elastic constants, attenuation, etc. at every point inside the Earth) and given a model of a seismic source, we can use current seismological theories to predict which seismograms should be observed at given locations at the Earth's surface.

The 'inverse problem' arises when we do not have a good model of the Earth, or a good model of the seismic source, but we have a set of seismograms, and we wish to use these observations to infer the internal Earth structure or a model of the source (typically we try to infer both).

There are many reasons that make the inverse problem underdetermined (nonunique). In the seismic example, two different Earth models may predict the same seismograms,[1] the finite bandwidth of our data will never allow us to resolve very small features of the Earth model, and there are always experimental uncertainties that allow different models to be 'acceptable.'

The name 'inverse problem' is widely used. The authors of this chapter only like this name moderately, as we see the problem more as a problem of 'conjunction of states of information' (theoretical, experimental, and prior information). In fact, the equations used below have a range of applicability well beyond 'inverse problems': they can be used, for instance, to predict the values of observations in a realistic situation where the parameters describing the Earth model are not 'given' but only known approximately.

We take here a probabilistic point of view. The axioms of probability theory apply to different situations. One is the traditional statistical analysis of random phenomena, another one is the description of (more or less) subjective states of information on a system. For instance, estimation of the uncertainties attached to any measurement usually involves both uses of probability theory: Some uncertainties contributing to the total uncertainty are estimated using statistics, while some other uncertainties are estimated using informed scientific judgment about the quality of an instrument, about effects not explicitly taken into account, etc. The International Organization for Standardization (ISO) in *Guide to the Expression of Uncertainty in Measurement* (1993), recommends that the uncertainties evaluated by statistical methods are named 'type A' uncertainties, and those evaluated by other means (for instance, using Bayesian arguments) be named 'type B' uncertainties. It also recommends that former classifications, for instance into 'random' and 'systematic uncertainties,' should be avoided. In the present text, we accept ISO's basic point of view, and extend it by downplaying the role assigned by ISO to the particular Gaussian model for uncertainties (see Section 4.3) and by not assuming that the uncertainties are 'small.'

INTERNATIONAL HANDBOOK OF EARTHQUAKE AND ENGINEERING SEISMOLOGY, VOLUME 81A ISBN: 0-12-440652-1

In fact, we like to think of an 'inverse' problem as merely a 'measurement.' A measurement that can be quite complex, but the basic principles and the basic equations to be used are the same for a relatively complex 'inverse problem' as for a relatively simple 'measurement.'

We do not normally use, in this text, the term 'random variable,' as we assume that we have probability distributions over 'physical quantities.' This is a small shift in terminology that we hope will not disorient the reader.

An important theme of this paper is *invariant formulation* of inverse problems, in the sense that solutions obtained using different, equivalent, sets of parameters should be consistent, i.e., probability densities obtained as the solution of an inverse problem, using two different set of parameters, should be related through the well-known rule of multiplication by the Jacobian of the transformation.

This chapter is organized as follows. After a brief historical review of inverse problem theory, with special emphasis on seismology, we give a short introduction to probability theory. In addition to being a tutorial, this introduction also aims at fixing a serious problem of classical probability, namely the noninvariant definition of conditional probability. This problem, which materializes in the so-called Borel paradox, has profound consequences for inverse problem theory.

A probabilistic formulation of inverse theory for general inverse problems (usually called 'nonlinear inverse problems') is not complete without the use of Monte Carlo methods. Section 3 is an introduction to the most versatile of these methods, the Metropolis sampler. Apart from being versatile, it also turns out to be the most natural method for implementing our probabilistic approach.

In Sections 4, 5, and 6 time has come for applying probability theory and Monte Carlo methods to inverse problems. All the steps of a careful probabilistic formulations are described, including parametrization, prior information over the parameters, and experimental uncertainties. The hitherto overlooked problem of uncertain physical laws ('forward relations') is given special attention in this text, and it is shown how this problem is profoundly linked to the resolution of the Borel paradox.

Section 7 treats the special case of the mildly nonlinear inverse problems, where deterministic (non-Monte Carlo) methods can be employed. In this section, invariant forms of classical inversion formulas are given.

1.2 Brief Historical Review

For a long time scientists have estimated parameters using optimization techniques. Laplace explicitly stated the least absolute values criterion. This, and the least-squares criterion were later popularized by Gauss (1809). While Laplace and Gauss were mainly interested in overdetermined problems, Hadamard (1902, 1932) introduced the notion of an 'ill-posed problem,' which can be viewed in many cases as an underdetermined problem.

The late 1960s and early 1970s were a golden age for the theory of inverse problems. In this period the first uses of Monte Carlo theory to obtain Earth models were made by Keilis-Borok and Yanovskaya (1967) and by Press (1968). At about the same time, Backus and Gilbert, and Backus alone, in the years 1967–1970, made original contributions to the theory of inverse problems, focusing on the problem of obtaining an unknown *function* from discrete data. Although the resulting mathematical theory is elegant, its initial predominance over the more 'brute force' (but more powerful) Monte Carlo theory was only possible due to the quite limited capacities of the computers at that time. It is our feeling that Monte Carlo methods will play a more important role in the future (and this is the reason why we put emphasis on these methods in this chapter). An investigation of the connection between analog models, discrete models, and Monte Carlo models can be found in a paper by Kennett and Nolet (1978).

Important developments of inverse theory in the fertile period around 1970 were also made by Wiggins (1969), with his method of suppressing 'small eigenvalues,' and by Franklin (1970) by introducing the right mathematical setting for the Gaussian, functional (i.e., infinite dimensional) inverse problem (see also Lehtinen *et al.*, 1989). Other important papers from the period are those of Gilbert (1971) and Wiggins (1972).

A reference that may interest some readers is Parzen *et al.* (1998), where the probabilistic approach of Akaike is described.

To the 'regularizing techniques' of Tikhonov (1963), Levenberg (1944), and Marquardt (1970), we prefer, in this chapter, the approach where the a priori information is used explicitly.

For seismologists, the first bona fide solution of an inverse problem was the estimation of the hypocenter coordinates of an earthquake using the 'Geiger method' (Geiger, 1910), which present-day computers have made practical. In fact, seismologists have been the originators of the theory of inverse problems (for data interpretation), and this is because the problem of understanding the structure of the Earth's interior using only surface data is a difficult one.

3-D tomography of the Earth, using travel times of seismic waves, was developed by Keiiti Aki and his coworkers in a couple of well known papers (Aki and Lee, 1976; Aki, Christofferson and Husebye 1977). Minster and Jordan (1978) applied the theory of inverse problems to the reconstruction of the tectonic plate motions, introducing the concept of 'data importance.' Later, tomographic studies have provided spectacular images of the Earth's interior. Interesting papers on these inversions are by van der Hilst *et al.* (1997) and Su *et al.* (1992).

One of the major current challenges in seismic inversion is the nonlinearity of wave field inversions. This is accentuated by the fact that major experiments in the future most likely will allow us to sample the whole seismic wave field. For low frequencies, wave field inversion is linear. Dahlen (1976) investigated the influence of lateral heterogeneity on the free oscillations. He showed that the inverse problem of estimating lateral heterogeneity of even degree from multiplet variance and skewance is linear. At the time this was published, data accuracy and unknown ellipticity splitting parameters hindered its application to real data, but later developments, including the works of Woodhouse and Dahlen (1978) on discontinuous Earth models, led to present-day successful inversions of low-frequency seismograms. In this connection the works of Woodhouse, Dziewonski, and others spring to mind.[2] Later, the first attempts to go to higher frequencies and nonlinear inversion were made by Nolet *et al.* (1986), and Nolet (1990).

Purely probabilistic formulations of inverse theory saw the light around 1970 (see, for instance, Kimeldorf and Wahba, 1970). In an interesting paper, Rietsch (1977) made nontrivial use of the notion of a 'noninformative' prior distribution for positive parameters. Jackson (1979) explicitly introduced prior information in the context of linear inverse problems, an approach that was generalized by Tarantola and Valette (1982a,b) to nonlinear problems.

There are three monographs in the area of inverse problems (from the viewpoint of data interpretation). In Tarantola (1987), the general, probabilistic formulation for nonlinear inverse problems is proposed. The small book by Menke (1984) covers several viewpoints on discrete, linear, and nonlinear inverse problems, and is easy to read. Finally, Parker (1994) exposes his view of the general theory of linear problems.

Recently, the interest in Monte Carlo methods, for the solution of inverse problems, has been increasing. Mosegaard and Tarantola (1995) proposed a generalization of the Metropolis algorithm (Metropolis *et al.*, 1953) for analysis of general inverse problems, introducing explicitly prior probability distributions, and they applied the theory to a synthetic numerical example. Monte Carlo analysis was recently applied to real data inverse problems by Mosegaard *et al.* (1997), Dahl-Jensen *et al.* (1998), Mosegaard and Rygaard-Hjalsted (1999), and Khan *et al.* (2000).

2. Elements of Probability

Probability theory is essential to our formulation of inverse theory. This chapter therefore contains a review of important elements of probability theory, with special emphasis on results that are important for the analysis of inverse problems. Of particular importance is our explicit introduction of *distance* and *volume* in data and model spaces. This has profound consequences for the notion of *conditional probability density*, which plays an important role in probabilistic inverse theory.

Also, we replace the concept of conditional probability by the more general notion of 'conjunction' of probabilities, this allowing us to address the more general problem where not only the data, but also the physical laws, are uncertain.

2.1 Volume

Let us consider an abstract space $\mathcal{S}$, where a point $\mathbf{x}$ is represented by some coordinates $\{x^1, x^2, \ldots\}$, and let $\mathcal{A}$ be some region (subspace) of $\mathcal{S}$. The measure associating a volume $V(\mathcal{A})$ to any region $\mathcal{A}$ of $\mathcal{S}$ will be denoted the *volume measure*

$$V(\mathcal{A}) = \int_{\mathcal{A}} d\mathbf{x}\, v(\mathbf{x}) \ , \tag{1}$$

where the function $v(\mathbf{x})$ is the *volume density*, and where we write $d\mathbf{x} = dx^1\, dx^2 \ldots$ The *volume element* is then[3]

$$dV(\mathbf{x}) = v(\mathbf{x})\, d\mathbf{x} \ , \tag{2}$$

and we may write $V(\mathcal{A}) = \int_{\mathcal{A}} dV(\mathbf{x})$. A manifold is called a *metric manifold* if there is a definition of distance between points, such that the distance ds between the point of coordinates $\{x^i\}$ and the point of coordinates $\{x^i + dx^i\}$ can be expressed as[4]

$$ds^2 = g_{ij}(\mathbf{x})\, dx^i\, dx^j \ , \tag{3}$$

i.e., if the notion of distance is 'of the L_2 type.'[5] The matrix whose entries are g_{ij} is the *metric matrix*, and an important result of differential geometry and integration theory is that the volume density of the space, $v(\mathbf{x})$, equals the square root of the determinant of the metric:

$$v(\mathbf{x}) = \sqrt{\det \mathbf{g}(\mathbf{x})} \ . \tag{4}$$

Example 1. *In the Euclidean 3D space, using spherical coordinates, the distance element is* $ds^2 = dr^2 + r^2\, d\theta^2 + r^2 \sin^2\theta\, d\varphi^2$, *from which it follows that the metric matrix is*

$$\begin{pmatrix} g_{rr} & g_{r\theta} & g_{r\varphi} \\ g_{\theta r} & g_{\theta\theta} & g_{\theta\varphi} \\ g_{\varphi r} & g_{\varphi\theta} & g_{\varphi\varphi} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2 \sin^2\theta \end{pmatrix} . \tag{5}$$

The volume density equals the metric determinant $v(r, \theta, \varphi) = \sqrt{\det \mathbf{g}(r, \theta, \varphi)} = r^2 \sin\theta$ *and therefore the volume element is* $dV(r, \vartheta, \varphi) = v(r, \vartheta, \varphi)\, dr\, d\vartheta\, d\varphi = r^2 \sin\theta\, dr\, d\vartheta\, d\varphi$.

2.2 Probability

Assume that we have defined over the space, not only the volume $V(\mathcal{A})$ of a region $\mathcal{A}$ of the space, but also its *probability* $P(\mathcal{A})$, which is assumed to satisfy the Kolmogorov axioms (Kolmogorov, 1933). This probability is assumed to be descriptible in terms of a probability density $f(x)$ through the expression

$$P(\mathcal{A}) = \int_{\mathcal{A}} d\mathbf{x}\ f(\mathbf{x}) \ . \tag{6}$$

It is well known that, in a change of coordinates over the space, a probability density changes its value: it is multiplied by the Jacobian of the transformation (this is the *Jacobian rule*). Normally, the probability of the whole space is normalized to one. If it is not normalizable, we do not say that we have a probability, but a 'measure.' We can state here the following postulate.

Postulate 1. *Given a space $\mathcal{X}$ over which a volume measure $V(\cdot)$ is defined. Any other measure (normalizable or not) $M(\cdot)$ considered over $\mathcal{X}$ is* absolutely continuous *with respect to $V(\cdot)$, i.e., the measure $M(\mathcal{A})$ of any region $\mathcal{A} \subset \mathcal{X}$ with vanishing volume must be zero: $V(\mathcal{A}) = 0 \Rightarrow M(\mathcal{A}) = 0$.*

2.3 Homogeneous Probability Distributions

In some parameter spaces, there is an obvious definition of distance between points, and therefore of volume. For instance, in the 3D Euclidean space the distance between two points is just the Euclidean distance (which is invariant under translations and rotations). Should we choose to parametrize the position of a point by its Cartesian coordinates $\{x, y, z\}$, the volume element in the space would be $dV(x, y, z) = dx\, dy\, dz$, while if we choose to use geographical coordinates, the volume element would be $dV(r, \theta, \varphi) = r^2 \sin\theta\, dr\, d\vartheta\, d\varphi$.

Definition. *The* homogeneous probability distribution *is the probability distribution that assigns to each region of the space a probability proportional to the volume of the region.*

Then, which probability density represents such a homogeneous probability distribution? Let us give the answer in three steps.

- If we use Cartesian coordinates $\{x, y, z\}$, as we have $dV(x, y, z) = dx\, dy\, dz$, the probability density representing the homogeneous probability distribution is constant: $f(x, y, z) = k$.
- If we use geographical coordinates $\{r, \theta, \varphi\}$, as we have $dV(r, \theta, \varphi) = r^2 \sin\theta\, dr\, d\theta\, d\varphi$, the probability density representing the homogeneous probability distribution is $g(r, \theta, \varphi) = kr^2 \sin\theta$.
- Finally, if we use an arbitrary system of coordinates $\{u, v, w\}$, in which the volume element of the space is $dV(u, v, w) = v(u, v, w)\, du\, dv\, dw$, the homogeneous probability distribution is represented by the probability density $h(u, v, w) = kv(u, v, w)$.

This is obviously true, since if we calculate the probability of a region $\mathcal{A}$ of the space, with volume $V(\mathcal{A})$, we get a number proportional to $V(\mathcal{A})$.

From these observations we can arrive at conclusions that are of general validity. First, the homogeneous probability distribution over some space is represented by a constant probability density **only** if the space is flat (in which case rectilinear systems of coordinates exist) and if we use Cartesian (or rectilinear) coordinates. The other conclusions can be stated as rules:

Rule 1. *The probability density representing the homogeneous probability distribution is easily obtained if the expression of the volume element $dV(u_1, u_2, \ldots) = v(u_1, u_2, \ldots)\, du_1\, du_2 \ldots$ of the space is known, as it is then given by $h(u_1, u_2, \ldots) = kv(u_1, u_2, \ldots)$, where k is a proportionality constant (that may have physical dimensions).*

Rule 2. *If there is a metric $g_{ij}(u_1, u_2, \ldots)$ in the space, then the volume element is given by $dV(u_1, u_2, \ldots) = \sqrt{\det \mathbf{g}(u_1, u_2, \ldots)}\, du_1\, du_2 \cdots$, i.e., we have $v(u_1, u_2, \ldots) = \sqrt{\det \mathbf{g}(u_1, u_2, \ldots)}$. The probability density representing the homogeneous probability distribution is, then, $h(u_1, u_2, \ldots) = k\sqrt{\det \mathbf{g}(u_1, u_2, \ldots)}$.*

Rule 3. *If the expression of the probability density representing the homogeneous probability distribution is known in one system of coordinates, then it is known in any other system of coordinates, through the Jacobian rule.*

Indeed, in the expression above, $g(r, \theta, \varphi) = kr^2 \sin\theta$, we recognize the Jacobian between the geographical and the Cartesian coordinates (where the probability density is constant).

For short, when we say *the homogeneous probability density* we mean *the probability density representing the homogeneous probability distribution.* **One should remember that, in general, the homogeneous probability density is *not* constant.**

Let us now examine 'positive parameters,' like a temperature, a period, or a seismic wave propagation velocity. One of the properties of the parameters we have in mind is that they occur in pairs of mutually reciprocal parameters:

Period	$T = 1/\nu$	;	*Frequency*	$\nu = 1/T$
Resistivity	$\rho = 1/\sigma$	;	Conductivity	$\sigma = 1/\rho$
Temperature	$T = 1/(k\beta)$	;	Thermodynamic parameter	$\beta = 1/(kT)$
Mass density	$\rho = 1/\ell$	;	Lightness	$\ell = 1/\rho$
Compressibility	$\gamma = 1/\kappa$	;	Bulk modulus (uncompressibility)	$\kappa = 1/\gamma$
Wave velocity	$c = 1/n$	;	Wave slowness	$n = 1/c$.

When working with physical theories, one may freely choose one of these parameters or its reciprocal.

Sometimes these pairs of equivalent parameters come from a definition, like when we define frequency ν as a function of the period T, by $\nu = 1/T$. Sometimes these parameters arise when analyzing an idealized physical system. For instance, Hooke's law, relating stress $\sigma_{\ell j}$ to strain $\varepsilon_{\ell j}$ can be expressed as $\sigma_{ij} = c_{ij}{}^{k\ell}\varepsilon_{k\ell}$, thus introducing the stiffness tensor $c_{ijk\ell}$, or as $\varepsilon_{ij} = d_{ij}{}^{k\ell}\sigma_{k\ell}$, thus introducing the compliance tensor

$d_{ijk\ell}$, the inverse of the stiffness tensor. Then the respective eigenvalues of these two tensors belong to the class of scalars analyzed here.

Let us take, as an example, the pair conductivity–resistivity (which may be thermal, electric, etc.). Assume we have two samples in the laboratory S_1 and S_2 whose resistivities are respectively ρ_1 and ρ_2. Correspondingly, their conductivities are $\sigma_1 = 1/\rho_1$ and $\sigma_2 = 1/\rho_2$. How should we define the 'distance' between the 'electrical properties' of the two samples? As we have $|\rho_2 - \rho_1| \neq |\sigma_2 - \sigma_1|$, choosing one of the two expressions as the 'distance' would be arbitrary. Consider the following definition of 'distance' between the two samples:

$$D(S_1, S_2) = \left| \log \frac{\rho_2}{\rho_1} \right| = \left| \log \frac{\sigma_2}{\sigma_1} \right| . \tag{7}$$

This definition (i) treats symmetrically the two equivalent parameters ρ and σ and, more importantly, (ii) has an *invariance of scale* (what matters is how many 'octaves' we have between the two values, not the plain difference between the values). In fact, it is the only definition of distance between the two samples S_1 and S_2 that has an invariance of scale and is additive (i.e., $D(S_1, S_2) + D(S_2, S_3) = D(S_1, S_3)$).

Associated to the distance $D(x_1, x_2) = |\log(x_2/x_1)|$ is the distance element (differential form of the distance)

$$dL(x) = \frac{dx}{x} . \tag{8}$$

This being a 'one-dimensional volume,' we can now apply Rule 1 above to get the expression of the homogeneous probability density for such a positive parameter:

$$f(x) = \frac{k}{x} . \tag{9}$$

Defining the reciprocal parameter $y = 1/x$ and using the Jacobian rule, we arrive at the homogeneous probability density for y:

$$g(y) = \frac{k}{y} . \tag{10}$$

These two probability densities have the same form: the two reciprocal parameters are treated symmetrically. Introducing the logarithmic parameters

$$x^* = \log \frac{x}{x_0} ; \qquad y^* = \log \frac{y}{y_0} , \tag{11}$$

where x_0 and y_0 are arbitrary positive constants, and using the Jacobian rule, we arrive at the homogeneous probability densities:

$$f'(x^*) = k ; \qquad g'(y^*) = k . \tag{12}$$

This shows that the logarithm of a positive parameter (of the type considered above) is a 'Cartesian' parameter. In fact, it is the consideration of Eqs. (12), together with the Jacobian rule, that allows full understanding of the (homogeneous) probability densities (9) and (10).

The association of the probability density $f(u) = k/u$ with positive parameters was first made by Jeffreys (1939). To honor him, we propose to use the term *Jeffreys parameters* for all the parameters of the type considered above. The $1/u$ probability density was advocated by Jaynes (1968), and a nontrivial use of it was made by Rietsch (1977) in the context of inverse problems.

Rule 4. *The homogeneous probability density for a Jeffreys quantity u is $f(u) = k/u$.*

Rule 5. *The homogeneous probability density for a 'Cartesian parameter' u (like the logarithm of a Jeffreys parameter, an actual Cartesian coordinate in an Euclidean space, or the Newtonian time coordinate) is $f(u) = k$. The homogeneous probability density for an angle describing the position of a point in a circle is also constant.*

If a parameter u is a Jeffreys parameter with the homogeneous probability density $f(u) = k/u$, then its inverse, its square, and, in general, any power of the parameter is also a Jeffreys parameter, as it can easily be seen using the Jacobian rule.

Rule 6. *Any power of a Jeffreys quantity (including its inverse) is a Jeffreys quantity.*

It is important to recognize when we do *not* face a Jeffreys parameter. Among the many parameters used in the literature to describe an isotropic linear elastic medium we find parameters like the Lamé's coefficients λ and μ, the bulk modulus κ, the Poisson ratio σ, etc. A simple inspection of the theoretical range of variation of these parameters shows that the first Lamé parameter λ and the Poisson ratio σ may take negative values, so they are certainly not Jeffreys parameters. In contrast, Hooke's law $\sigma_{ij} = c_{ijk\ell}\, \varepsilon^{k\ell}$, defining a linearity between stress σ_{ij} and strain ε_{ij}, defines the positive definite stiffness tensor $c_{ijk\ell}$ or, if we write $\varepsilon_{ij} = d_{ijk\ell}\, \sigma^{k\ell}$, defines its inverse, the compliance tensor $d_{ijk\ell}$. The two reciprocal tensors $c_{ijk\ell}$ and $d_{ijk\ell}$ are 'Jeffreys tensors.' This is a notion whose development is beyond the scope of this paper, but we can give the following rule.

Rule 7. *The eigenvalues of a Jeffreys tensor are Jeffreys quantities.*[6]

As the two (different) eigenvalues of the stiffness tensor $c_{ijk\ell}$ are $\lambda_\kappa = 3\kappa$ (with multiplicity 1) and $\lambda_\mu = 2\mu$ (with multiplicity 5), we see that the incompressibility modulus κ and the shear modulus μ are Jeffreys parameters[7] (as are any parameters proportional to them, or any powers of them, including the inverses). If, for some reason, instead of working with κ and μ, we wish to work with other elastic parameters, for instance, the Young modulus Y and the

Poisson ratio σ, or the two elastic wave velocities, then the homogeneous probability distribution must be found using the Jacobian of the transformation (see Appendix H).

Some probability densities have conspicuous 'dispersion parameters,' like the σ's in the normal probability density $f(x) = k \exp\left(-\frac{(x-x_0)^2}{2\sigma^2}\right)$, in the log-normal probability $g(X) = \frac{k}{X} \exp\left(-\frac{(\log X/X_0)^2}{2\sigma^2}\right)$ or in the Fisher probability density (Fisher, 1953) $h(\vartheta, \varphi) = k \sin\theta \exp(\cos\theta/\sigma^2)$. A consistent probability model requires that when the dispersion parameter σ tends to infinity, the probability density tends to the homogeneous probability distribution. For instance, in the three examples just given, $f(x) \to k$, $g(X) \to k/X$, and $h(\theta, \varphi) \to k \sin\theta$, which are the respective homogeneous probability densities for a Cartesian quantity, a Jeffreys quantity, and the geographical coordinates on the surface of the sphere. We can state the following rule.

Rule 8. *If a probability density has some 'dispersion parameters,' then, in the limit where the dispersion parameters tend to infinity, the probability density must tend to the homogeneous one.*

As an example, using the normal probability density $f(x) = k \exp\left(-\frac{(x-x_0)^2}{2\sigma^2}\right)$, for a Jeffreys parameter is not consistent. Note that it would assign a finite probability to negative values of a positive parameter that, by definition, is positive. More technically, this would violate our Postulate 1. Using the log-normal probability density for a Jeffreys parameter is consistent.

There is a problem of terminology in the Bayesian literature. The homogeneous probability distribution is a very special distribution. When the problem of selecting a 'prior' probability distribution arises in the absence of any information, except the fundamental symmetries of the problem, one may select as prior probability distribution the homogeneous distribution. But enthusiastic Bayesians do not call it 'homogeneous,' but 'noninformative.' We cannot recommend using this terminology. The homogeneous probability distribution is as informative as any other distribution, it is just the homogeneous one (see Appendix D).

In general, each time we consider an abstract parameter space, each point being represented by some parameters $\mathbf{x} = \{x^1, x^2 \dots x^n\}$, we will start by solving the (sometimes nontrivial) problem of defining a distance between points that respects the necessary symmetries of the problem. Only exceptionally this distance will be a quadratic expression of the parameters (coordinates) being used (i.e., only exceptionally our parameters will correspond to 'Cartesian coordinates' in the space). From this distance, a volume element $dV(\mathbf{x}) = v(\mathbf{x})\, d\mathbf{x}$ will be deduced, from where the expression $f(\mathbf{x}) = k\, v(\mathbf{x})$ of the homogeneous probability density will follow. Sometimes, we can directly define the volume element, without the need of a distance. We emphasize the need of defining a distance—or a volume element—in the parameter space, from which the notion of homogeneity will follow. With this point of view, we slightly depart from the original work by Jeffreys and Jaynes.

2.4 Conjunction of Probabilities

We shall here consider two probability distributions P and Q . We say that a probability R is a product of the two given probabilities, and is denoted $(P \wedge Q)$ if

- $P \wedge Q = Q \wedge P$;
- for any subset $\mathcal{A}$, $(P \wedge Q)(\mathcal{A}) \neq 0 \Rightarrow P(\mathcal{A}) \neq 0$ and $Q(\mathcal{A}) \neq 0$;
- if M denotes the homogeneous probability distribution, then $P \wedge M = P$.

The realization of these conditions leading to the simplest results can easily be expressed using probability densities (see Appendix G for details). If the two probabilities P and Q are represented by the two probability densities $p(\mathbf{x})$ and $q(\mathbf{x})$, respectively, and if the homogeneous probability density is represented by $\mu(\mathbf{x})$, then the probability $P \wedge Q$ is represented by a probability density, denoted $(p \wedge q)(\mathbf{x})$, that is given by

$$(p \wedge q)(\mathbf{x}) = k \frac{p(\mathbf{x})\; q(\mathbf{x})}{\mu(\mathbf{x})} \ , \qquad (13)$$

where k is a normalization constant.[8]

The two left columns of Figure 1 represent these probability densities.

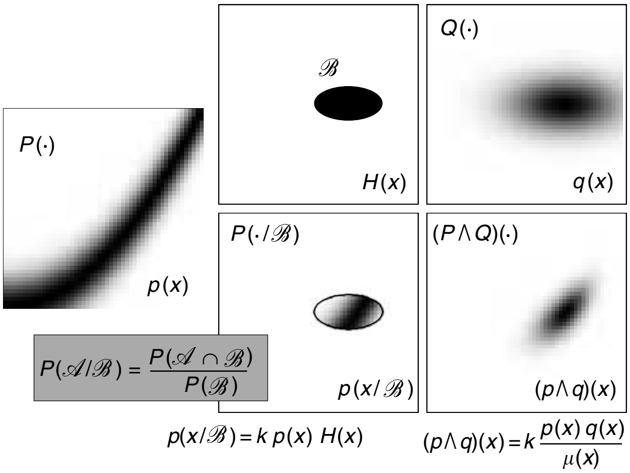
$Q(\cdot)$
$\mathscr{B}$
$P(\cdot)$
$H(x)$
$q(x)$
$P(\cdot/\mathscr{B})$
$(P \wedge Q)(\cdot)$
$p(x)$
$P(\mathscr{A}/\mathscr{B}) = \frac{P(\mathscr{A} \cap \mathscr{B})}{P(\mathscr{B})}$
$p(x/\mathscr{B})$
$(p \wedge q)(x)$
$p(x/\mathscr{B}) = k\, p(x)\, H(x)$
$(p \wedge q)(x) = k \frac{p(x)\, q(x)}{\mu(x)}$

FIGURE 1 The two left columns of the figure illustrate the definition of conditional probability (see text for details). The right of the figure explains that the definition of the AND operation is a generalization of the notion of conditional probability. While a conditional probability combines a probability distribution $P(\cdot)$ with an 'event' $\mathscr{B}$, the AND operation combines two probability distributions $P(\cdot)$ and $Q(\cdot)$ defined over the same space. See text for a detailed explanation.

Example 2. *On the surface of the Earth, using geographical coordinates (latitude ϑ and longitude φ), the homogeneous probability distribution is represented by the probability density $\mu(\vartheta,\varphi)=\frac{1}{4\pi}\cos\vartheta$. An estimation of the position of a floating object at the surface of the sea by an airplane navigator gives a probability distribution for the position of the object corresponding to the probability density $p(\vartheta,\varphi)$, and an independent, simultaneous estimation of the position by another airplane navigator gives a probability distribution corresponding to the probability density $q(\vartheta,\varphi)$. How do we 'combine' the two probability densities $p(\vartheta,\varphi)$ and $q(\vartheta,\varphi)$ to obtain a 'resulting' probability density? The answer is given by the conjunction of the two probability densities:*

$$(p \wedge q)(\vartheta,\varphi) = k\,\frac{p(\vartheta,\varphi)\;q(\vartheta,\varphi)}{\mu(\vartheta,\varphi)}\ . \tag{14}$$

We emphasize here the following:

Example 2 is at the basis of the paradigm that we use below to solve inverse problems.

More generally, the conjunction of the probability densities $f_1(\mathbf{x}),\ f_2(\mathbf{x})\ldots$ is

$$\begin{aligned} h(\mathbf{x}) &= (f_1 \wedge f_2 \wedge f_3 \cdots)(\mathbf{x}) \cdots \\ &= k\,\mu(\mathbf{x})\,\frac{f_1(\mathbf{x})}{\mu(\mathbf{x})}\,\frac{f_2(\mathbf{x})}{\mu(\mathbf{x})}\,\frac{f_3(\mathbf{x})}{\mu(\mathbf{x})}\cdots\ . \end{aligned} \tag{15}$$

For a formalization of the notion of conjunction of probabilities, the reader is invited to read Appendix G.

2.5 Conditional Probability Density

Given a probability distribution over a space $\mathcal{X}$, represented by the probability density $f(\mathbf{x})$, and given a subspace $\mathcal{B}$ of $\mathcal{X}$ of lower dimension, can we, in a consistent way, infer a probability distribution over $\mathcal{B}$, represented by a probability density $f(\mathbf{x}|\mathcal{B})$ (to be named the conditional probability density 'given $\mathcal{B}$')? The answer is: Using only the elements given, NO, THIS IS NOT POSSIBLE.

The usual way to induce a probability distribution on a subspace of lower dimension is to assign a 'thickness' to the subspace $\mathcal{B}$, to apply the general definition of conditional probability (this time to a region of $\mathcal{X}$, not to a subspace of it) and to take the limit when the 'thickness' tends to zero. But, as suggested in Figure 2, there are infinitely many ways to take this limit, each defining a different 'conditional probability density' on $\mathcal{B}$. Among the infinitely many ways to define a conditional probability density, there is one that is based on the notion of distance between points in the space, and therefore corresponds to an intrinsic definition (see Fig. 2).

Assume that the space $\mathcal{U}$ has p dimensions, the space $\mathcal{V}$ has q dimensions, and define in the $(p+q)$-dimensional space $\mathcal{X}=(\mathcal{U},\mathcal{V})$ a p-dimensional subspace by the p relations

$$\begin{aligned} v_1 &= v_1(u_1,u_2,\ldots,u_p) \\ v_2 &= v_2(u_1,u_2,\cdots,u_p) \\ \cdots &= \cdots \\ v_q &= v_q(u_1,u_2,\ldots,u_p)\ . \end{aligned} \tag{16}$$

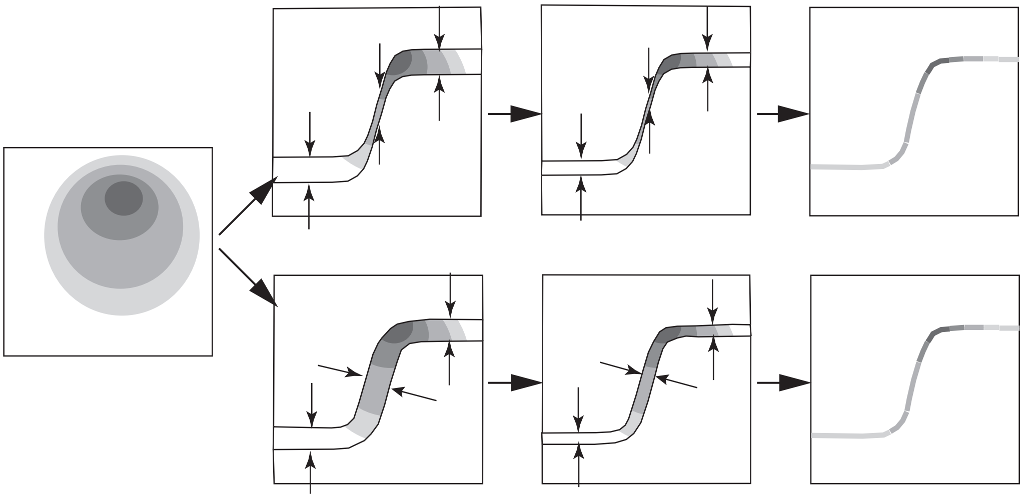

FIGURE 2 An original 2D probability density, and two possible ways (among many) of defining a region of the space whose limit is a given curve. At the top is the 'vertical' limit, while at the bottom is the normal (or orthogonal) limit. Each possible limit defines a different 'induced' or 'conditional' probability density. Only the orthogonal limit gives an intrinsic definition (i.e., a definition invariant under any change of variables). It is, therefore, the only one examined in this work.

The restriction of a probability distribution represented by the probability density $f(\mathbf{x}) = f(\mathbf{u},\mathbf{v})$ into the subspace defined by the constraint $\mathbf{v} = \mathbf{v}(\mathbf{u})$, can be defined with all generality when it is assumed that we have a metric defined over the $(p+q)$-dimensional space $\mathcal{X} = (\mathcal{U}, \mathcal{V})$. Let us limit here to the special circumstance (useful for a vast majority of inverse problems[9]) where there the $(p+q)$-dimensional space $\mathcal{X}$ is built as the Cartesian product of $\mathcal{U}$ and $\mathcal{V}$ (then we write, as usual, $\mathcal{X} = \mathcal{U} \times \mathcal{V}$) . In this case, there is a metric $\mathbf{g}_u$ over $\mathcal{U}$, with associated volume element $dV_u(\mathbf{u}) = \sqrt{\det \mathbf{g}_u}\, d\mathbf{u}$, there is a metric $\mathbf{g}_v$ over $\mathcal{V}$, with associated volume element $dV_v(\mathbf{v}) = \sqrt{\det \mathbf{g}_v}\, d\mathbf{v}$, and the global volume element is simply $dV(\mathbf{u},\mathbf{v}) = dV_u(\mathbf{u})\, dV_v(\mathbf{v})$.

The restriction of the probability distribution represented by the probability density $f(\mathbf{u},\mathbf{v})$ on the subspace $\mathbf{v} = \mathbf{v}(\mathbf{u})$ (i.e., the conditional probability density given $\mathbf{v} = \mathbf{v}(\mathbf{u})$) is a probability distribution *on* the submanifold $\mathbf{v} = \mathbf{v}(\mathbf{u})$. We could choose ad-hoc coordinates over this manifold, but as there is a one-to-one correspondence between the coordinates $\mathbf{u}$ and the points on the manifold, the conditional probability density can be expressed using the coordinates $\mathbf{u}$. The restriction of $f(\mathbf{u},\mathbf{v})$ over the submanifold $\mathbf{v} = \mathbf{v}(\mathbf{u})$ defines the probability density (see Appendix B for the more general case)

$$
\begin{aligned}
&f_{u|v(u)}(\mathbf{u}|\mathbf{v} = \mathbf{v}(\mathbf{u})) \\
&\quad = k\, f(\mathbf{u},\mathbf{v}(\mathbf{u}))\, \left. \frac{\sqrt{\det(\mathbf{g}_u + \mathbf{V}^T \mathbf{g}_v \mathbf{V})}}{\sqrt{\det \mathbf{g}_u}\, \sqrt{\det \mathbf{g}_v}} \right|_{\mathbf{v}=\mathbf{v}(\mathbf{u})} ,
\end{aligned} \tag{17}
$$

where k is a normalizing constant, and where $\mathbf{V} = \mathbf{V}(\mathbf{u})$ is the matrix of partial derivatives (see Appendix M for a simple explicit calculation of such partial derivatives)

$$
\begin{pmatrix} V_{11} & V_{12} & \cdots & V_{1p} \\ V_{21} & V_{22} & \cdots & V_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ V_{q1} & V_{q2} & \cdots & V_{qp} \end{pmatrix} = \begin{pmatrix} \frac{\partial v_1}{\partial u_1} & \frac{\partial v_1}{\partial u_2} & \cdots & \frac{\partial v_1}{\partial u_p} \\ \frac{\partial v_2}{\partial u_1} & \frac{\partial v_2}{\partial u_2} & \cdots & \frac{\partial v_2}{\partial u_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial v_q}{\partial u_1} & \frac{\partial v_q}{\partial u_2} & \cdots & \frac{\partial v_q}{\partial u_p} \end{pmatrix} . \tag{18}
$$

Example 3. *If the hypersurface* $\mathbf{v} = \mathbf{v}(\mathbf{u})$ *is defined by a constant value of* $\mathbf{v}$, *say* $\mathbf{v} = \mathbf{v}_0$, *then Eq. (17) reduces to*

$$
f_{u|v}(\mathbf{u}|\mathbf{v} = \mathbf{v}_0) = k\, f(\mathbf{u}, \mathbf{v}_0) = \frac{f(\mathbf{u}, \mathbf{v}_0)}{\int_{\mathcal{U}} d\mathbf{u}\ f(\mathbf{u}, \mathbf{v}_0)} . \tag{19}
$$

Elementary definitions of conditional probability density are not based on this notion of distance-based uniform convergence, but use other, ill-defined limits. This is a mistake that, unfortunately, pollutes many scientific works. See Appendix P, in particular, for a discussion on the 'Borel paradox.'

Equation (17) defines the conditional $f_{u|v(u)}(\mathbf{u}\,|\,\mathbf{v} = \mathbf{v}(\mathbf{u}))$. Should the relation $\mathbf{v} = \mathbf{v}(\mathbf{u})$ be invertible, it would correspond to a change of variables. It is then possible to show that the alternative conditional $f_{v|u(v)}(\mathbf{v}\,|\,\mathbf{u} = \mathbf{u}(\mathbf{v}))$ is related to $f_{u|v(u)}(\mathbf{u}\,|\,\mathbf{v} = \mathbf{v}(\mathbf{u}))$ through the Jacobian rule. This is a property that elementary definitions of conditional probability do not share.

2.6 Marginal Probability Density

In the special circumstance described above, where we have a Cartesian product of two spaces, $\mathcal{X} = \mathcal{U} \times \mathcal{V}$, given a 'joint' probability density $f(\mathbf{u},\mathbf{v})$, it is possible to give an intrinsic sense to the definitions

$$
f_u(\mathbf{u}) = \int_{\mathcal{V}} d\mathbf{v}\ f(\mathbf{u},\mathbf{v}) \ ; \qquad f_v(\mathbf{v}) = \int_{\mathcal{U}} d\mathbf{u}\ f(\mathbf{u},\mathbf{v}) \ . \tag{20}
$$

These two densities are called *marginal probability densities*. Their intuitive interpretation is clear, as the 'projection' of the joint probability density respectively over $\mathcal{U}$ and over $\mathcal{V}$.

2.7 Independence and Bayes Theorem

Dropping the index 0 in Eq. (19) and using the second of Eqs. (20) gives

$$
f_{u|v}(\mathbf{u}|\mathbf{v}) = \frac{f(\mathbf{u},\mathbf{v})}{f_v(\mathbf{v})} , \tag{21}
$$

or, equivalently, $f(\mathbf{u},\mathbf{v}) = f_{u\,|\,v}(\mathbf{u}\,|\,\mathbf{v})\, f_v(\mathbf{v})$. As we can also define $f_{v}\,|\,u(\mathbf{v}\,|\,\mathbf{u})$, we have the two equations

$$
\begin{aligned}
f(\mathbf{u},\mathbf{v}) &= f_{u|v}(\mathbf{u}|\mathbf{v})\ f_v(\mathbf{v}) \\
f(\mathbf{u},\mathbf{v}) &= f_{v|u}(\mathbf{v}|\mathbf{u})\ f_u(\mathbf{u}) ,
\end{aligned} \tag{22}
$$

that can be read as follows: 'When we work in a space that is the Cartesian product $\mathcal{U} \times \mathcal{V}$ of two subspaces, a joint probability density can always be expressed as the product of a conditional times a marginal.'

From these last equations there follows the expression

$$
f_{u|v}(\mathbf{u}|\mathbf{v}) = \frac{f_{v|u}(\mathbf{v}|\mathbf{u})\ f_u(\mathbf{u})}{f_v(\mathbf{v})} , \tag{23}
$$

known as the *Bayes theorem*, and generally used as the starting point for solving inverse problems. We do not think this is a useful setting, and we prefer here *not* to use the Bayes theorem (or, more precisely, not to use the intuitive paradigm usually associated with it).

It also follows from Eqs. (22) that the two conditions

$$
f_{u|v}(\mathbf{u}|\mathbf{v}) = f_u(\mathbf{u}) \ ; \qquad f_{v|u}(\mathbf{v}|\mathbf{u}) = f_v(\mathbf{v}) \tag{24}
$$

are equivalent. It is then said that $\mathbf{u}$ and $\mathbf{v}$ are *independent parameters* (with respect to the probability density $f(\mathbf{u}, \mathbf{v})$). The term 'independent' is easy to understand, as the conditional of any of the two (vector) variables, given the other variable equals the (unconditional) marginal of the variable. Then, one clearly has

$$f(\mathbf{u}, \mathbf{v}) = f_u(\mathbf{u})\ f_v(\mathbf{v}) \tag{25}$$

i.e., for independent variables, the joint probability density can be simply expressed as the product of the two marginals.

3. Monte Carlo Methods

When a probability distribution has been defined, we face the problem of how to 'use' it. The definition of 'central estimators' (such as the mean or the median) and 'estimators of dispersion' (such as the covariance matrix) lacks generality as it is quite easy to find examples (such as multimodal distributions in highly-dimensional spaces) where these estimators fail to have any interesting meaning.

When a probability distribution has been defined over a space of low dimension (say, from one to four dimensions) we can directly represent the associated probability density. This is trivial in one or two dimensions. It is easy in three dimensions, and some tricks may allow us to represent a four-dimensional probability distribution, but clearly this approach cannot be generalized to the high dimensional case.

Let us explain the only approach that seems practical, with help of Figure 3. At the left of the figure, there is an explicit representation of a 2D probability distribution (by means of the associated probability density or the associated (2D) volumetric probability). In the middle, some random points have been generated (using the Monte Carlo method about to be described). It is clear that, if we make a histogram with these points, in the limit of a sufficiently large number of points we recover the representation at the left. Disregarding the histogram possibility, we can concentrate on the individual points. In the 2D example of the figure we have actual points in a plane. If the problem is multidimensional, each 'point' may correspond to some abstract notion. For instance, for a geophysicist a 'point' may be a given model of the Earth. This model may be represented in some way, for instance, by a color plot. Then a collection of 'points' is a collection of such pictures. Our experience shows that, given a collection of randomly generated 'models,' the human eye–brain system is extremely good at apprehending the basic characteristics of the underlying probability distribution, including possible multimodalities, correlations, etc.

When such a (hopefully large) collection of random models is available, we can also answer quite interesting questions. For instance, a geologist might ask: *at which depth is that subsurface structure?* To answer this, we can make a histogram of the depth of the given geological structure over the collection of random models, and the histogram *is* the answer to the question. *What is the probability of having a low-velocity zone around a given depth?* The ratio of the number of models presenting such a low-velocity zone over the total number of models in the collection gives the answer (if the collection of models is large enough).

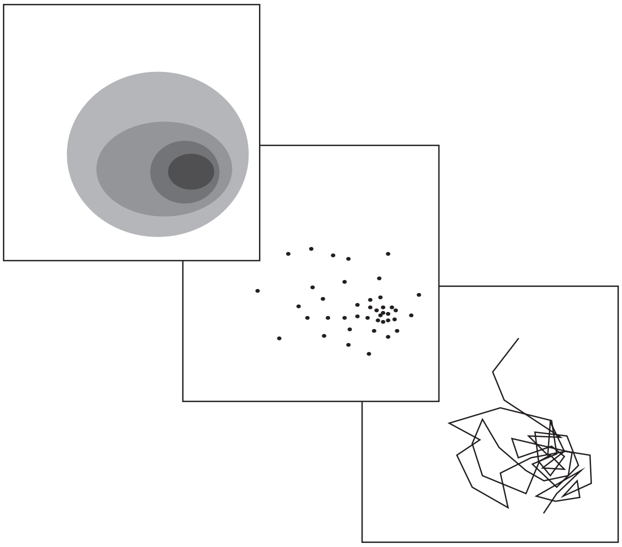

FIGURE 3 An explicit representation of a 2D probability distribution and the sampling of it, using Monte Carlo methods. While the representation at the top left cannot be generalized to high dimensions, the examination of a collection of points can be done in arbitrary dimensions. Practically, Monte Carlo generation of points is done through a 'random walk' where a 'new point' is generated in the vicinity of the previous point.

This is essentially what we propose: looking at a large number of randomly generated models in order to intuitively apprehend the basic properties of the probability distribution, followed by calculation of the probabilities of all interesting 'events.'

Practically, as we will see, the random sampling is not made by generating points independently of each other. Rather, as suggested in the last image of Figure 3, it is done through a 'random walk' where a 'new point' is generated in the vicinity of the previous point.

Monte Carlo methods have a random generator at their core. At present, Monte Carlo methods are typically implemented on digital computers, and are based on pseudo random generation of numbers.[10] As we shall see, any conceivable operation on probability densities (e.g., computing marginals and conditionals, integration, conjunction (the AND operation), etc.) has its counterpart in an operation on/by their corresponding Monte Carlo algorithms.

Inverse problems are often formulated in high-dimensional spaces. In this case a certain class of Monte Carlo algorithms, the so-called *importance sampling algorithms*, come to the rescue, allowing us to sample the space with a sampling density proportional to the given probability density. In this case excessive (and useless) sampling of low-probability areas of the space is avoided. This is not only important, but in fact vital in high-dimensional spaces.

Another advantage of the importance sampling Monte Carlo algorithms is that we need not have a closed-form mathematical expression for the probability density that we want to sample. Only an algorithm that allows us to evaluate it at a given point in the space is needed. This has considerable practical advantage in analysis of inverse problems where computer-intensive evaluation of, for example, misfit functions plays an important role in calculation of certain probability densities.

Given a probability density that we wish to sample, and a class of Monte Carlo algorithms that samples this density, which one of the algorithms should we choose? Practically, the problem here is to find the most efficient of these algorithms. This is an interesting and difficult problem for which we will not go into detail here. We will, later in this chapter, limit ourselves to only two general methods that are recommendable in many practical situations.

3.1 Random Walks

To escape the dimensionality problem, *any* sampling of a probability density for which point values are available only upon request has to be based on a *random walk*, i.e., in a generation of successive points with the constraint that point $\mathbf{x}_{i+1}$ sampled in iteration $(i+1)$ is in the vicinity of the point $\mathbf{x}_i$ sampled in iteration i . The simplest of the random walks are generated by the so-called Markov Chain Monte Carlo (MCMC) algorithms, where the point $\mathbf{x}_{i+1}$ depends on the point $\mathbf{x}_i$, but not on previous points. We will concentrate on these algorithms here.

If random rules have been defined to select points such that the probability of selecting a point in the infinitesimal 'box' $dx_1 \cdots dx_N$ is $p(\mathbf{x})\, dx_1 \cdots dx_N$, then the points selected in this way are called *samples* of the probability density $p(\mathbf{x})$. Depending on the rules defined, successive samples $i, j, k, \ldots$ may be dependent or independent.

3.2 The Metropolis Rule

The most common Monte Carlo sampling methods are the Metropolis sampler (described below) and the Gibbs sampler (Geman and Geman, 1984). As we believe that the Gibbs sampler is only superior to the Metropolis sampler in low-dimensional problems, we restrict ourselves here to the presentation of the latter.

Consider the following situation. Some random rules define a random walk that samples the probability density $f(\mathbf{x})$. At a given step, the random walker is at point $\mathbf{x}_j$, and the application of the rules would lead to a transition to point $\mathbf{x}_i$. By construction, when all such 'proposed transitions' $\mathbf{x}_i \leftarrow \mathbf{x}_j$ are always accepted, the random walker will sample the probability density $f(\mathbf{x})$. Instead of always accepting the proposed transition $\mathbf{x}_i \leftarrow \mathbf{x}_j$, we reject it sometimes by using the following rule to decide if it is allowed to move to $\mathbf{x}_i$ or if it must stay at $\mathbf{x}_j$:

- If $g(\mathbf{x}_i)/\mu(\mathbf{x}_i) \geq g(\mathbf{x}_j)/\mu(\mathbf{x}_j)$, then accept the proposed transition to $\mathbf{x}_i$.
- If $g(\mathbf{x}_i)/\mu(\mathbf{x}_i) < g(\mathbf{x}_j)/\mu(\mathbf{x}_j)$, then decide randomly to move to $\mathbf{x}_i$, or to stay at $\mathbf{x}_j$, with the following probability of accepting the move to $\mathbf{x}_i$:

$$P = \frac{g(\mathbf{x}_i)/\mu(\mathbf{x}_i)}{g(\mathbf{x}_j)/\mu(\mathbf{x}_j)} \ . \tag{26}$$

Then we have the following theorem.

Theorem 1. *The random walker samples the conjunction $h(\mathbf{x})$ of the probability densities $f(\mathbf{x})$ and $g(\mathbf{x})$*

$$h(\mathbf{x}) = k\, f(\mathbf{x}) \frac{g(\mathbf{x})}{\mu(\mathbf{x})} = k \frac{f(\mathbf{x})\, g(\mathbf{x})}{\mu(\mathbf{x})} \tag{27}$$

(*see Appendix O for a demonstration*).

It should be noted here that this algorithm nowhere requires the probability densities to be normalized. This is of vital importance in practice, since it allows sampling of probability densities whose values are known only in points already sampled by the algorithm. Obviously, such probability densities cannot be normalized. Also , the fact that our theory also allows unnormalizable probability densities will not cause any trouble in the application of the above algorithm.

The algorithm above is reminiscent (see Appendix O) of the Metropolis algorithm (Metropolis *et al.*, 1953), originally designed to sample the Gibbs–Boltzmann distribution.[11] Accordingly, we will refer to the above acceptance rule as the *Metropolis rule*.

3.3 The Cascaded Metropolis Rule

As above, assume that some random rules define a random walk that samples the probability density $f_1(\mathbf{x})$. At a given step, the random walker is at point $\mathbf{x}_j$.

(1) Apply the rules that unthwarted would generate samples distributed according to $f_1(\mathbf{x})$, to propose a new point $\mathbf{x}_i$.

(2) If $f_2(\mathbf{x}_i)/\mu(\mathbf{x}_i) \geq f_2(\mathbf{x}_j)/\mu(\mathbf{x}_j)$, go to point 3; if $f_2(\mathbf{x}_i)/\mu(\mathbf{x}_i) < f_2(\mathbf{x}_j)/\mu(\mathbf{x}_j)$, then decide randomly to go to point 3 or to go back to point 1, with the following probability of going to point 3: $P = (f_2(\mathbf{x}_i)/\mu(\mathbf{x}_i))/(f_2(\mathbf{x}_j)/\mu(\mathbf{x}_j))$.

(3) If $f_3(\mathbf{x}_i)/\mu(\mathbf{x}_i) \geq f_3(\mathbf{x}_j)/\mu(\mathbf{x}_j)$, go to point 4; if $f_3(\mathbf{x}_i)/\mu(\mathbf{x}_i) < f_3(\mathbf{x}_j)/\mu(\mathbf{x}_j)$, then decide randomly to go to point 4 or to go back to point 1, with the following probability of going to point 4: $P = (f_3(\mathbf{x}_i)/\mu(\mathbf{x}_i))/(f_3(\mathbf{x}_j)/\mu(\mathbf{x}_j))$.

... ...

(n) If $f_n(\mathbf{x}_i)/\mu(\mathbf{x}_i) \geq f_n(\mathbf{x}_j)/\mu(\mathbf{x}_j)$, then accept the proposed transition to $\mathbf{x}_i$; if $f_n(\mathbf{x}_i)/\mu(\mathbf{x}_i) < f_n(\mathbf{x}_j)/\mu(\mathbf{x}_j)$, then decide randomly to move to $\mathbf{x}_i$, or to stay at $\mathbf{x}_j$, with the following probability of accepting the move to $\mathbf{x}_i$: $P = (f_n(\mathbf{x}_i)/\mu(\mathbf{x}_i))/(f_n(\mathbf{x}_j)/\mu(\mathbf{x}_j))$.

Then we have the following theorem.

Theorem 2. *The random walker samples the conjunction $h(\mathbf{x})$ of the probability densities $f_1(\mathbf{x})$, $f_2(\mathbf{x}), \ldots, f_n(\mathbf{x})$:*

$$h(\mathbf{x}) = k\, f_1(\mathbf{x})\, \frac{f_2(\mathbf{x})}{\mu(\mathbf{x})} \cdots \frac{f_n(\mathbf{x})}{\mu(\mathbf{x})} \ . \tag{28}$$

(*see the supplementary materials to this chapter on the attached Handbook CD for a demonstration*).

3.4 Initiating a Random Walk

Consider the problem of obtaining samples of a probability density $h(\mathbf{x})$ defined as the conjunction of some probability densitites $f_1(\mathbf{x})$, $f_2(\mathbf{x})$, $f_3(\mathbf{x}) \ldots$,

$$h(\mathbf{x}) = k\, f_1(\mathbf{x})\, \frac{f_2(\mathbf{x})}{\mu(\mathbf{x})}\, \frac{f_3(\mathbf{x})}{\mu(\mathbf{x})} \cdots \ , \tag{29}$$

and let us examine three common situations.

We start with a random walk that samples $f_1(\mathbf{x})$ (optimal situation): This corresponds to the basic algorithm where we know how to produce a random walk that samples $f_1(\mathbf{x})$, and we only need to modify it, taking into account the values $f_2(\mathbf{x})/\mu(\mathbf{x})$, $f_3(\mathbf{x})/\mu(\mathbf{x}) \ldots$, using the cascaded Metropolis rule, to obtain a random walk that samples $h(\mathbf{x})$.

We start with a random walk that samples the homogeneous probability density $\mu(\mathbf{x})$: We can write Eq. (29) as

$$h(\mathbf{x}) = k \left(\left(\left(\mu(\mathbf{x})\, \frac{f_1(\mathbf{x})}{\mu(\mathbf{x})} \right) \frac{f_2(\mathbf{x})}{\mu(\mathbf{x})} \right) \cdots \right) \ . \tag{30}$$

The expression corresponds to the case where we are not able to start with a random walk that samples $f_1(\mathbf{x})$, but we have a random walk that samples the homogeneous probability density $\mu(\mathbf{x})$. Then, with respect to the example just mentioned, there is one extra step to be added, taking into account the values of $f_1(\mathbf{x})/\mu(\mathbf{x})$.

We start with an arbitrary random walk (worst situation): In the situation where we are not able to directly define a random walk that samples the homogeneous probability distribution, but only one that samples some arbitrary (but known) probability distribution $\psi(\mathbf{x})$, we can write Eq. (29) in the form

$$h(\mathbf{x}) = k \left(\left(\left(\left(\psi(\mathbf{x})\, \frac{\mu(\mathbf{x})}{\psi(\mathbf{x})} \right) \frac{f_1(\mathbf{x})}{\mu(\mathbf{x})} \right) \frac{f_2(\mathbf{x})}{\mu(\mathbf{x})} \right) \cdots \right) . \tag{31}$$

Then, with respect to the example just mentioned, there is one more extra step to be added, taking into account the values of $\mu(\mathbf{x})/\psi(\mathbf{x})$. Note that the closer $\psi(\mathbf{x})$ is be to $\mu(\mathbf{x})$, the more efficient will be the first modification of the random walk.

3.5 Convergence Issues

When has a random walk visited enough points in the space so that a probability density has been sufficiently sampled? This is a complex issue, and it is easy to overlook its importance. There is no general rule: Each problem has its own 'physics,' and the experience of the 'implementer' is crucial here.

Many methods that work for low dimension completely fail when the number of dimensions is high. Typically, a random walk select a random direction and, then, a random step along that direction. The notion of 'direction' in a high-dimensional space is far from the intuitive one we get in the familar three-dimensional space. Any serious discussion on this issue must be problem-dependent, so we do not even attempt one here.

Obviously, a necessary condition for adequate sampling is that any 'output' from the algorithm must 'look stationary.'

4. Probabilistic Formulation of Inverse Problems

A so-called 'inverse problem' arises when a usually complex measurement is made, and information on unknown parameters of the physical system is sought. Any measurement is indirect (we may weigh a mass by observing the displacement of the cursor of a balance), and therefore a possibly nontrivial analysis of uncertainties must be done. Any guide describing good experimental practice (see, for instance the ISO's *Guide to the Expression of Uncertainty in Measurement* (ISO, 1993) or the shorter description by Taylor and Kuyatt, 1994) acknowledges that a measurement involves, at least, two different sources of uncertainties: those estimated using statistical methods, and those estimated using subjective, common-sense estimations. Both are described using the axioms of probability theory, and this chapter clearly takes the probabilistic point of view for developing inverse theory.

4.1 Model Parameters and Observable Parameters

Although the separation of all the variables of a problem into two groups, 'directly observable parameters' (or 'data') and 'model parameters', may sometimes be artificial, we take this point of view here, since it allows us to propose a simple setting for a wide class of problems.

We may have in mind a given physical system, like the whole Earth or a small crystal under our microscope. The system (or a given state of the system) may be described by assigning values to a given set of parameters $\mathbf{m} = \{ m^1, m^2, \ldots, m^{NM} \}$, which we will name the *model parameters*.

Let us assume that we make observations on this system. Although we are interested in the parameters **m**, they may not be directly observable, so we make indirect measurements such as obtaining seismograms at the Earth's surface for analyzing the Earth's interior, or making spectroscopic measurements for analyzing the chemical properties of a crystal. The set of (*directly*) *observable parameters* (or, by abuse of language, the set of *data parameters*) will be represented by $\mathbf{d} = \{ d^1, d^2, \ldots, d^{ND} \}$.

We assume that we have a physical theory that can be used to solve the *forward problem*, i.e., that given an arbitrary model **m**, it allows us to predict the theoretical data values **d** that an ideal measurement should produce (if **m** were the actual system). The generally nonlinear function that associates with any model **m** the theoretical data values **d** may be represented by a notation such as

$$d^i = f^i(m^1, m^2, \ldots, m^{NM}) \; ; \qquad i = 1, 2, \ldots, ND \; , \quad (32)$$

or, for short,

$$\mathbf{d} = \mathbf{f}(\mathbf{m}) \; . \quad (33)$$

It is in fact this expression that separates the whole set of our parameters into the subsets **d** and **m** , although sometimes there is no difference in nature between the parameters in **d** and the parameters in **m** . For instance, in the classical inverse problem of estimating the hypocenter coordinates of an earthquake, we may put in **d** the arrival times of the seismic waves at seismic observatories, and we need to put in **m**, besides the hypocentral coordinates, the coordinates defining the location of the seismometers—as these are parameters that are needed to compute the travel times—although we estimate arrival times of waves and coordinates of the seismic observatories using similar types of measurements.

4.2 Prior Information on Model Parameters

In a typical geophysical problem, the model parameters contain geometrical parameters (positions and sizes of geological bodies) and physical parameters (values of the mass density, of the elastic parameters, the temperature, the porosity, etc.).

The *prior information* on these parameters is all the information we possess independently of the particular measurements that will be considered as 'data' (to be described below). This prior probability distribution is generally quite complex, as the model space may be high-dimensional, and the parameters may have nonstandard probability densities.

To this generally complex probability distribution over the model space corresponds a probability density that we denote $\rho_m(\mathbf{m})$.

If an explicit expression for the probability density $\rho_m(\mathbf{m})$ is known, it can be used in analytical developments. But such an explicit expression is, by no means, necessary. Using Monte Carlo methods, all that is needed is a set of probabilistic rules that allows us to generate samples distributed according to $\rho_m(\mathbf{m})$ in the model space (Mosegaard and Tarantola, 1995).

Example 4. *Appendix E presents an example of prior information for the case of an Earth model consisting of a stack of horizontal layers with variable thickness and uniform mass density.*

4.3 Measurements and Experimental Uncertainties

Observation of geophysical phenomena is represented by a set of parameters **d** that we usually call data. These parameters result from prior measurement operations, and they are typically seismic vibrations on the instrument site, arrival times of seismic phases, gravity or electromagnetic fields. As in any measurement, the data are determined with an associated uncertainty, described by a probability density over the data parameter space, that we denote here $\rho_d(\mathbf{d})$. This density describes not only marginals on individual datum values, but also possible cross-relations in data uncertainties.

Although the instrumental errors are an important source of data uncertainties, in geophysical measurements there are other sources of uncertainty. The errors associated with the positioning of the instruments, the environmental noise, and the human factor (like for picking arrival times) are also relevant sources of uncertainty.

Example 5. Nonanalytic Probability Density. *Assume that we wish to measure the time t of occurrence of some physical event. It is often assumed that the result of a measurement corresponds to something like*

$$t = t_0 \pm \sigma \; . \quad (34)$$

An obvious question is the exact meaning of the $\pm\sigma$. Has the experimenter in mind that she or he is absolutely certain that the actual arrival time satisfies the strict conditions

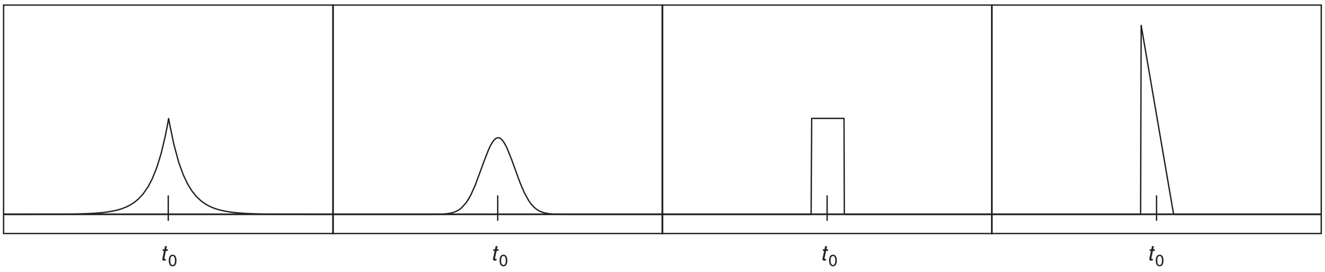

FIGURE 4 What has an experimenter in mind when she or he describes the result of a measurement by something like $t = t_0 \pm \sigma$?

$t_0 - \sigma \leq t \leq t_0 + \sigma$, or has she or he in mind something like a Gaussian probability, or some other probability distribution (see Fig. 4)? We accept, following ISO's recommendations (1993) that the result of any measurement has a probabilistic interpretation, with some sources of uncertainty being analyzed using statistical methods ('type A' uncertainties), and other sources of uncertainty being evaluated by other means (for instance, using Bayesian arguments) ('type B' uncertainties). But, contrary to ISO suggestions, we do not assume that the Gaussian model of uncertainties should play any central role. In an extreme example, we may well have measurements whose probabilistic description may correspond to a multimodal probability density. Figure 5 shows a typical example for a seismologist: the measurement on a seismogram of the arrival time of a certain seismic wave, in the case one hesitates in the phase identification or in the identification of noise and signal. In this case the probability density for the arrival of the seismic phase does not have an explicit expression like $f(t) = k \exp(-(t-t_0)^2/(2\sigma^2))$, but is a numerically defined function.

Example 6. *The Gaussian model for uncertainties. The simplest probabilistic model that can be used to describe experimental uncertainties is the Gaussian model*

$$\rho_d(\mathbf{d}) = k \exp\left(-\frac{1}{2}(\mathbf{d} - \mathbf{d}_{\text{obs}})^T \mathbf{C}_D^{-1} (\mathbf{d} - \mathbf{d}_{\text{obs}})\right) . \qquad (35)$$

It is here assumed that we have some 'observed data values' $\mathbf{d}_{\text{obs}}$ with uncertainties described by the covariance matrix $\mathbf{C}_D$. If the uncertainties are uncorrelated,

$$\rho_d(\mathbf{d}) = k \exp\left(-\frac{1}{2}\sum_i \left(\frac{d^i - d^i_{\text{obs}}}{\sigma^i}\right)^2\right) , \qquad (36)$$

where the σ^i are the 'standard deviations.'

Example 7. *The generalized Gaussian model for uncertainties. An alternative to the Gaussian model is to use the*

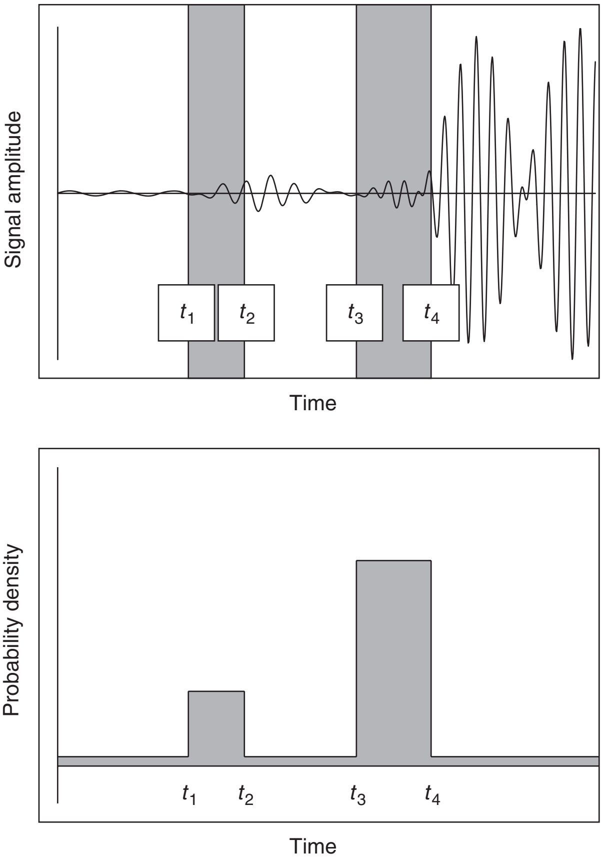

FIGURE 5 A seismologist tries to measure the arrival time of a seismic wave at a seismic station, by 'reading' the seismogram at the top of the figure. The seismologist may find quite likely that the arrival time of the wave is between times t_3 and t_4, and believe that what is before t_3 is just noise. But if there is a significant probability that the signal between t_1 and t_2 is not noise but the actual arrival of the wave, then the seismologist should define a bimodal probability density, as the one suggested at the bottom of the figure. Typically, the actual form of each peak of the probability density is not crucial (here, box-car functions are chosen), but the position of the peaks is important. Rather than assigning a zero probability density to the zones outside the two intervals, it is safer (more 'robust') to attribute some small 'background' value, as we may never exclude some unexpected source of error.

Laplacian (double exponential) model for uncertainties,

$$\rho_d(\mathbf{d}) = k \exp\left(- \sum_i \frac{|d^i - d^i_{\text{obs}}|}{\sigma^i} \right) . \tag{37}$$

While the Gaussian model leads to least-squares-related methods, this Laplacian model leads to absolute-values methods (see Section 4.5.2), well known for producing robust[12] *results. More generally, there is the* L_p *model of uncertainties*

$$\rho_p(\mathbf{d}) = k \exp\left(- \frac{1}{p} \sum_i \frac{|d^i - d^i_{\text{obs}}|^p}{(\sigma^i)^p} \right) \tag{38}$$

(see Fig. 6).

4.4 Joint "Prior" Probability Distribution in the $(\mathcal{M}, \mathcal{D})$ Space

We have just seen that the prior information on model parameters can be described by a probability density in the model space, $\rho_m(\mathbf{m})$, and that the result of measurements can be described by a probability density in the data space $\rho_d(\mathbf{d})$. As by 'prior' information on model parameters we mean information obtained *independently* from the measurements (it often represents information we had before the measurements were made), we can use the notion of independency of variables of Section 2.6 to define a joint probability density in the $\mathcal{X} = (\mathcal{M}, \mathcal{D})$ space as the product of the two 'marginals'

$$\rho(\mathbf{x}) = \rho(\mathbf{m}, \mathbf{d}) = \rho_m(\mathbf{m})\, \rho_d(\mathbf{d}) \ . \tag{39}$$

Although we have introduced $\rho_m(\mathbf{m})$ and $\rho_d(\mathbf{d})$ separately, and we have suggested building a probability distribution in the $(\mathcal{M}, \mathcal{D})$ space by the multiplication (39), we may have a more general situation where the information we have on $\mathbf{m}$ and on $\mathbf{d}$ is not independent. So, in what follows, let us assume that we have some information in the $\mathcal{X} = (\mathcal{M}, \mathcal{D})$ space, represented by the 'joint' probability density

$$\rho(\mathbf{x}) = \rho(\mathbf{m}, \mathbf{d}) \ , \tag{40}$$

and let us consider Eq. (39) as just a special case.

Let us in the rest of this chapter denote by $\mu(\mathbf{x})$ the probability density representing the homogeneous probability distribution, as introduced in Section 2.2. We may remember here the Rule 8, stating that the limit of a consistent probability density must be the homogeneous one, so we may formally write

$$\mu(\mathbf{x}) = \lim_{\text{infinite dispersions}} \rho(\mathbf{x}) \ . \tag{41}$$

When the partition (39) holds, then, typically (see Rule 8),

$$\mu(\mathbf{x}) = \mu(\mathbf{m}, \mathbf{d}) = \mu_m(\mathbf{m})\, \mu_d(\mathbf{d}) \ . \tag{42}$$

4.5 Physical Laws as Mathematical Functions

4.5.1 Physical Laws

Physics analyzes the correlations existing between physical parameters. In standard mathematical physics, these correlations are represented by 'equalities' between physical parameters (as when we write $\mathbf{F} = m\,\mathbf{a}$ to relate the force

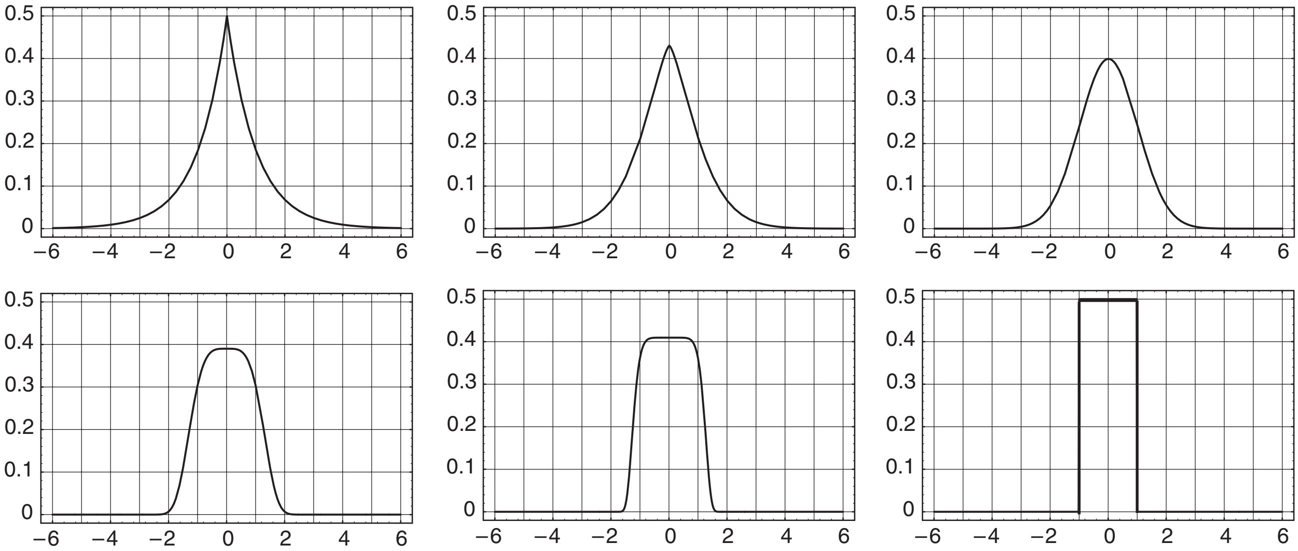

FIGURE 6 Generalized Gaussian for values of the parameter $p = 1$, $\sqrt{2}$, 2, 4, 8, and ∞ .

$\mathbf{F}$ applied to a particle, the mass m of the particle and the acceleration $\mathbf{a}$). In the context of inverse problems, this corresponds to assuming that we have a function from the 'parameter space' to the 'data space' that we may represent as

$$\mathbf{d} = \mathbf{f}(\mathbf{m}) \ . \tag{43}$$

We do not mean that the relation is necessarily explicit. Given $\mathbf{m}$ we may need to solve a complex system of equations in order to get $\mathbf{d}$, but this nevertheless defines a function $\mathbf{m} \to \mathbf{d} = \mathbf{f}(\mathbf{m})$.

At this point, given the probability density $\rho(\mathbf{m}, \mathbf{d})$ and given the relation $\mathbf{d} = \mathbf{f}(\mathbf{m})$, we can define the associated conditional probability density $\rho_{m\,|\,d\,(m)}(\mathbf{m} \mid \mathbf{d} = \mathbf{f}(\mathbf{m}))$. We could here use the more general definition of conditional probability density of Appendix B, but let us simplify the text by using a simplifying assumption: that the total parameter space $(\mathcal{M}, \mathcal{D})$ is just the Cartesian product $\mathcal{M} \times \mathcal{D}$ of the model parameter space $\mathcal{M}$ times the space of directly observable parameters (or 'data space') $\mathcal{D}$. Then, rather than a general metric in the total space, we have a metric $\mathbf{g}_m$ over the model parameter space $\mathcal{M}$ and a metric $\mathbf{g}_d$ over the data space, and the total metric is just the Cartesian product of the two metrics. In particular, then, the total volume element in the space, $dV(\mathbf{m}, \mathbf{d})$ is just the product of the two volume elements in the model parameter space and the data space: $dV(\mathbf{m}, \mathbf{d}) = dV_m(\mathbf{m})\, dV_d(\mathbf{D})$. Most inverse problems satisfy this assumption.[13] In this setting, the formulas of Section 2.4 are valid.

4.5.2 Inverse Problems

In the $(\mathcal{M}, \mathcal{D}) = \mathcal{M} \times \mathcal{D}$ space, we have the probability density $\rho(\mathbf{m}, \mathbf{d})$ and we have the hypersurface defined by the relation $\mathbf{d} = \mathbf{f}(\mathbf{m})$. The natural way to 'compose' these two kinds of information is by defining the conditional probability density induced by $\rho(\mathbf{m}, \mathbf{d})$ on the hypersurface $\mathbf{d} = \mathbf{f}(\mathbf{m})$,

$$\sigma_m(\mathbf{m}) \equiv \rho_{m|d(m)}(\mathbf{m}|\mathbf{d} = \mathbf{f}(\mathbf{m})) \ , \tag{44}$$

this gives (see Eq. (17))

$$\sigma_m(\mathbf{m}) = k\, \rho(\mathbf{m}, \mathbf{f}(\mathbf{m}))\, \frac{\sqrt{\det\left(\mathbf{g}_m + \mathbf{F}^T \mathbf{g}_d \mathbf{F}\right)}}{\sqrt{\det \mathbf{g}_m}\sqrt{\det \mathbf{g}_d}} \Bigg|_{\mathbf{d}=\mathbf{f}(\mathbf{m})} , \tag{45}$$

where $\mathbf{F} = \mathbf{F}(\mathbf{m})$ is the matrix of partial derivatives, with components $\mathbf{F}_{i\alpha} = \partial f_i / \partial m_\alpha$, where $\mathbf{g}_m$ is the metric in the model parameter space $\mathcal{M}$ and where $\mathbf{g}_d$ is the metric in the data space $\mathcal{D}$.

Example 8. *Quite often, $\rho(\mathbf{m}, \mathbf{d}) = \rho_m(\mathbf{m})\, \rho_d(\mathbf{d})$. Then, Eq. (45) can be written*

$$\sigma_m(\mathbf{m}) = k\, \rho_m(\mathbf{m}) \left(\frac{\rho_d(\mathbf{d})}{\sqrt{\det \mathbf{g}_m}} \frac{\sqrt{\det(\mathbf{g}_m + \mathbf{F}^T \mathbf{g}_d \mathbf{F})}}{\sqrt{\det \mathbf{g}_m}} \right) \Bigg|_{\mathbf{d}=\mathbf{f}(\mathbf{m})} . \tag{46}$$

Example 9. *If $g_{\mathbf{m}}(\mathbf{m}) = constant$ and $g_{\mathbf{d}}(\mathbf{d}) = constant$, and the nonlinearities are weak ($\mathbf{F}(\mathbf{m}) = constant$), then Eq. (46) reduces to*

$$\sigma_m(\mathbf{m}) = k\, \rho_m(\mathbf{m}) \frac{\rho_d(\mathbf{d})}{\mu_d(\mathbf{d})} \Bigg|_{\mathbf{d}=\mathbf{f}(\mathbf{m})} , \tag{47}$$

where we have used $\mu_d(\mathbf{d}) = k\, \sqrt{\det \mathbf{g}_d(\mathbf{d})}$ (see Rule 2).

Example 10. *We examine here the simplification that we arrive at when assuming that the 'input' probability densities are Gaussian*:

$$\rho_m(\mathbf{m}) = k\ \exp\left(-\frac{1}{2} \left(\mathbf{m} - \mathbf{m}_{\text{prior}}\right)^t\, \mathbf{C}_M^{-1} \left(\mathbf{m} - \mathbf{m}_{\text{prior}}\right) \right) \tag{48}$$

$$\rho_d(\mathbf{d}) = k\ \exp\left(-\frac{1}{2} (\mathbf{d} - \mathbf{d}_{\text{obs}})^t\, \mathbf{C}_D^{-1} (\mathbf{d} - \mathbf{d}_{\text{obs}}) \right) . \tag{49}$$

In this circumstance, quite often, it is the covariance operators $\mathbf{C}_M$ and $\mathbf{C}_D$ *that are used to define the metrics over the spaces $\mathcal{M}$ and $\mathcal{D}$. Then, $\mathbf{g}_m = \mathbf{C}_M^{-1}$ and $\mathbf{g}_d = \mathbf{C}_D^{-1}$. Grouping some of the constant factors in the factor k , Eq. (45) becomes here*

$$\sigma_m(\mathbf{m}) = k\ \exp\left[-\frac{1}{2} \left((\mathbf{m} - \mathbf{m}_{\text{prior}})^t\, \mathbf{C}_M^{-1} (\mathbf{m} - \mathbf{m}_{\text{prior}}) + (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}})^t\, \mathbf{C}_D^{-1} (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}}) \right) \right] \times \frac{\sqrt{\det\left(\mathbf{C}_M^{-1} + \mathbf{F}^T(\mathbf{m})\ \mathbf{C}_D^{-1} \mathbf{F}(\mathbf{m})\right)}}{\sqrt{\det \mathbf{C}_M^{-1}}} \tag{50}$$

(the constant factor $\sqrt{\det \mathbf{C}_M^{-1}}$ has been left for subsequent simplifications). Defining the misfit

$$S(\mathbf{m}) = -2 \log \frac{\sigma_m(\mathbf{m})}{\sigma_0} \ , \tag{51}$$

where σ_0 is an arbitrary value of $\sigma_m(\mathbf{m})$, gives, up to an additive constant,

$$S(\mathbf{m}) = S_1(\mathbf{m}) - S_2(\mathbf{m}) \ , \tag{52}$$

where $S_1(\mathbf{m})$ is the usual least-squares misfit function

$$S_1(\mathbf{m}) = (\mathbf{m} - \mathbf{m}_{\text{prior}})^t \, \mathbf{C}_M^{-1} \, (\mathbf{m} - \mathbf{m}_{\text{prior}}) + (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}})^t \, \mathbf{C}_D^{-1} \, (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}}) \qquad (53)$$

and where[14]

$$S_2(\mathbf{m}) = \log \det \left(\mathbf{I} + \mathbf{C}_M \mathbf{F}^t(\mathbf{m}) \, \mathbf{C}_D^{-1} \, \mathbf{F}(\mathbf{m}) \right) . \qquad (54)$$

Example 11. *If, in the context of Example 10, we have*[15] $\mathbf{C}_M \mathbf{F}^t \mathbf{C}_D^{-1} \mathbf{F} \ll \mathbf{I}$ *, we can use the low order approximation for $S_2(\mathbf{m})$, that is*[16]

$$S_2(\mathbf{m}) \approx \text{trace} \, \mathbf{C}_M \mathbf{F}^t(\mathbf{m}) \mathbf{C}_D^{-1} \, \mathbf{F}(\mathbf{m}) \, . \qquad (55)$$

Example 12. *If in the context of Example 10 we assume that the nonlinearities are weak, then the matrix of partial derivatives* $\mathbf{F}$ *is approximately constant, and Eq. (50) simplifies to*

$$\sigma_m(\mathbf{m}) = k \exp \left[-\frac{1}{2} \left((\mathbf{m} - \mathbf{m}_{\text{prior}})^t \, \mathbf{C}_M^{-1} (\mathbf{m} - \mathbf{m}_{\text{prior}}) + (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}})^t \, \mathbf{C}_D^{-1} \, (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}}) \right) \right] , \qquad (56)$$

and the function $S_2(\mathbf{m})$ is just a constant.

Example 13. *If the 'relation solving the forward problem'* $\mathbf{d} = \mathbf{f}(\mathbf{m})$ *happens to be a linear relation,* $\mathbf{d} = \mathbf{F}\mathbf{m}$, *then one gets the standard equations for linear problems (see Appendix F).*

Example 14. *We examine here the simplifications that we arrive at when assuming that the 'input' probability densities are Laplacian*:

$$\rho_m(\mathbf{m}) = k \exp \left(- \sum_\alpha \frac{|m^\alpha - m^\alpha_{\text{prior}}|}{\sigma_\alpha} \right) \qquad (57)$$

$$\rho_d(\mathbf{d}) = k \exp \left(- \sum_i \frac{|d^i - d^i_{\text{obs}}|}{\sigma_i} \right) . \qquad (58)$$

Equation (45) becomes here

$$\sigma_m(\mathbf{m}) = k \exp \left[- \left(\sum_\alpha \frac{|m^\alpha - m^\alpha_{\text{prior}}|}{\sigma_\alpha} + \sum_i \frac{|\, f^i(\mathbf{m}) - d^i_{\text{obs}}|}{\sigma_i} \right) \right] \Psi(\mathbf{m}) \, , \qquad (59)$$

where $\Psi(\mathbf{m})$ is a complex term containing, in particular, the matrix of partial derivatives $\mathbf{F}$ *. If this term is approximately constant (weak nonlinearities, constant metrics), then*

$$\sigma_m(\mathbf{m}) = k \exp \left[- \left(\sum_\alpha \frac{|m^\alpha - m^\alpha_{\text{prior}}|}{\sigma_\alpha} + \sum_i \frac{|f^i(\mathbf{m}) - d^i_{\text{obs}}|}{\sigma_i} \right) \right] . \qquad (60)$$

The formulas in the examples above give expressions that contain analytic parts (like the square roots containing the matrix of partial derivatives $\mathbf{F}$) . What we write as $\mathbf{d} = \mathbf{f}(\mathbf{m})$ may sometimes correspond to an explicit expression; sometimes it may correspond to the solution of an implicit equation.[17] Should $\mathbf{d} = \mathbf{f}(\mathbf{m})$ be an explicit expression, and should the 'prior probability densities' $\rho_m(\mathbf{m})$ and $\rho_d(\mathbf{d})$ (or the joint $\rho(\mathbf{m}, \mathbf{d})$) also be given by explicit expressions (as when we have Gaussian probability densities), then the formulas of this section would give explicit expressions for the posterior probability density $\sigma_m(\mathbf{m})$.

If the relation $\mathbf{d} = \mathbf{f}(\mathbf{m})$ is a linear relation, then the expression giving $\sigma_m(\mathbf{m})$ can sometimes be simplified easily (as with the linear Gaussian case to be examined below). More often than not the relation $\mathbf{d} = \mathbf{f}(\mathbf{m})$ is a complex nonlinear relation, and the expression we are left with for $\sigma_m(\mathbf{m})$ is explicit, but complex.

Once the probability density $\sigma_m(\mathbf{m})$ has been defined, there are different ways of 'using' it. If the 'model space' $\mathcal{M}$ has a small number of dimensions (say between one and four) the values of $\sigma_m(\mathbf{m})$ can be computed at every point of a grid and a graphical representation of $\sigma_m(\mathbf{m})$ can be attempted. A visual inspection of such a representation is usually worth a thousand 'estimators' (central estimators or estimators of dispersion). But, of course, if the values of $\sigma_m(\mathbf{m})$ are known at all points where $\sigma_m(\mathbf{m})$ has a significant value, these estimators can also be computed.

If the 'model space' $\mathcal{M}$ has a large number of dimensions (say from five to many millions or billions), then an exhaustive exploration of the space is not possible, and we must turn to Monte Carlo sampling methods to extract information from $\sigma_m(\mathbf{m})$. We discuss the application of Monte Carlo methods to inverse problems, and optimization techniques, in Section 6 and 7, respectively.

4.6 Physical Laws as Probabilistic Correlations

4.6.1 Physical Laws

We return here to the general case where it is not assumed that the total space $(\mathcal{M}, \mathcal{D})$ is the Cartesian product of two spaces.

In Section 4.5 we have examined the situation where the physical correlation between the parameters of the problem are expressed using an exact, analytic expression $\mathbf{d} = \mathbf{g}(\mathbf{m})$. In this case, the notion of conditional probability density has been used to combine the 'physical theory' with the 'data' and the 'a priori information' on model parameters.

But, we have seen that in order to properly define the notion of conditional probability density, it has been necessary to introduce a metric over the space, and to take a limit using the metric of the space. This is equivalent to put some 'thickness' around the theoretical relation $\mathbf{d} = \mathbf{g}(\mathbf{m})$, and to take the limit when the thickness tends to zero.

But actual theories have uncertainties, and, for more generality, it is better to explicitly introduce these uncertainties. Assume, then, that the physical correlations between the model parameters $\mathbf{m}$ and the data parameters $\mathbf{d}$ are not represented by an analytical expression like $\mathbf{d} = \mathbf{f}(\mathbf{m})$ but by a probability density

$$\vartheta(\mathbf{m}, \mathbf{d}) \ . \tag{61}$$

Example: Realistic 'Uncertainty Bars' around a Functional Relation. In the approximation of a constant gravity field, with acceleration $\mathbf{g}$, the position at time t of an apple in free fall is $\mathbf{r}(t) = \mathbf{r}_0 + \mathbf{v}_0 t + \frac{1}{2}\mathbf{g} t^2$, where $\mathbf{r}_0$ and $\mathbf{v}_0$ are, respectively, the position and velocity of the object at time $t = 0$. More simply, if the movement is 1D ,

$$x(t) = x_0 + v_0 t + \frac{1}{2} g t^2 \ . \tag{62}$$

Of course, for many reasons this equation can never be exact: air friction, wind effects, inhomogeneity of the gravity field, effects of the Earth rotation, forces from the Sun and the Moon (not to mention Pluto), relativity (special and general), and so on.

It is not a trivial task, given very careful experimental conditions, to estimate the size of the leading uncertainty. Although one might think of an equation $x = x(t)$ as a line, infinitely thin, there will always be sources of uncertainty (at least due to the unknown limits of validity of general relativity): looking at the line with a magnifying glass should reveal a fuzzy object of finite thickness. As a simple example, let us examine here the mathematical object we arrive at when assuming that the leading sources of uncertainty in the relation $x = x(t)$ are the uncertainties in the initial position and velocity of the falling apple. Let us assume that:

- the initial position of the apple is random, with a Gaussian distribution centered at x_0 , and with standard deviation σ_x ;
- the initial velocity of the apple is random, with a Gaussian distribution centered at v_0 , and with standard deviation σ_v .

Then it can be shown that at a given time t, the possible positions of the apple are random, with probability density

$$\vartheta(x|t) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_x^2 + \sigma_v^2 t^2}} \times \exp\left(-\frac{1}{2} \frac{\left(x - (x_0 + v_0 t + \frac{1}{2} g t^2)\right)^2}{\sigma_x^2 + \sigma_v^2 t^2} \right) \ . \tag{63}$$

This is obviously a conditional probability density for x , given t . Should we have any reason to choose some marginal probability density $\vartheta_t(t)$, then, the 'law' for the fall of the apple would be

$$\vartheta(x, t) = \vartheta(x|t)\, \vartheta_t(t) \ . \tag{64}$$

See Appendix C for more details.

4.6.2 Inverse Problems

We have seen that the result of measurements can be represented by a probability density $\rho_d(\mathbf{d})$ in the data space. We have also seen that the a priori information on the model parameters can be represented by another probability density $\rho_m(\mathbf{m})$ in the model space. When we talk about 'measurements' and about 'a priori information on model parameters,' we usually mean that we have a joint probability density in the $(\mathcal{M}, \mathcal{D})$ space, that is $\rho(\mathbf{m}, \mathbf{d}) = \rho_m(\mathbf{m}) \rho_d(\mathbf{d})$. Let us consider the more general situation where for the whole set of parameters $(\mathcal{M}, \mathcal{D})$ we have some information that can be represented by a joint probability density $\rho(\mathbf{m}, \mathbf{d})$. Having well in mind the interpretation of this information, let us use the simple term 'experimental information' for it:

$$\rho(\mathbf{m}, \mathbf{d}) \qquad \text{(experimental information)} \ . \tag{65}$$

We have also seen that we have information coming from physical theories, that predict correlations between the parameters, and it has been argued that a probabilistic description of these correlations is well adapted to the resolution of inverse problems.[18] Let $\vartheta(\mathbf{m}, \mathbf{d})$ be the probability density representing this 'theoretical information':

$$\vartheta(\mathbf{m}, \mathbf{d}) \qquad \text{(theoretical information)} \ . \tag{66}$$

A quite fundamental assumption is that in all the spaces we consider, there is a notion of volume that allows us to give meaning to the notion of a 'homogeneous probability distribution' over the space. The corresponding probability density is not constant, but is proportional to the volume element of the space (see Section 2.2):

$$\mu(\mathbf{m}, \mathbf{d}) \qquad \text{(homogeneous probability distribution)} \ . \tag{67}$$

Finally, we have seen examples suggesting that the conjunction of the experimental information with the theoretical information corresponds exactly to the AND operation defined over the probability densities, to obtain the 'conjunction of information,' as represented by the probability density

$$\sigma(\mathbf{m},\, \mathbf{d}) = k\, \frac{\rho(\mathbf{m},\, \mathbf{d})\; \vartheta(\mathbf{m},\, \mathbf{d})}{\mu(\mathbf{m},\, \mathbf{d})} \qquad (68)$$

$$\text{(conjunction of information)}\ ,$$

with marginal probability densities

$$\sigma_m(\mathbf{m}) = \int_{\mathcal{D}} d\mathbf{d}\, \sigma(\mathbf{m},\, \mathbf{d})\ ; \qquad \sigma_d(\mathbf{d}) = \int_{\mathcal{M}} d\mathbf{m}\, \sigma(\mathbf{m},\, \mathbf{d})\ . \qquad (69)$$

Example 15. *We may assume that the physical correlations between the parameters* **m** *and* **d** *are of the form*

$$\vartheta(\mathbf{m},\, \mathbf{d}) = \vartheta_{D|M}(\mathbf{d}|\mathbf{m})\, \vartheta_M(\mathbf{m})\ , \qquad (70)$$

this expressing that a 'physical theory' gives, on the one hand, the conditional probability for **d** *, given* **m** *, and, on the other hand, the marginal probability density for* **m** *. See Appendix C for more details.*

Example 16. *Many applications concern the special situation where we have*

$$\mu(\mathbf{m},\, \mathbf{d}) = \mu_m(\mathbf{m})\, \mu_d(\mathbf{d})\ ; \qquad \rho(\mathbf{m},\, \mathbf{d}) = \rho_m(\mathbf{m})\, \rho_d(\mathbf{d})\ . \qquad (71)$$

In this case, Eqs. (68) and (69) give

$$\sigma_m(\mathbf{m}) = k\, \frac{\rho_m(\mathbf{m})}{\mu_m(\mathbf{m})} \int_{\mathcal{D}} d\mathbf{d}\, \frac{\rho_d(\mathbf{d})\, \vartheta(\mathbf{m},\, \mathbf{d})}{\mu_d(\mathbf{d})}\ . \qquad (72)$$

If Eq. (70) holds, then

$$\sigma_m(\mathbf{m}) = k\, \rho_m(\mathbf{m})\, \frac{\vartheta_m(\mathbf{m})}{\mu_m(\mathbf{m})} \int_{\mathcal{D}} d\mathbf{d}\, \frac{\rho_d(\mathbf{d})\; \vartheta_{D|M}(\mathbf{d}|\mathbf{m})}{\mu_d(\mathbf{d})}\ . \qquad (73)$$

Finally, if the simplification $\vartheta_M(\mathbf{m}) = \mu_m(\mathbf{m})$ *arises (see Appendix C for an illustration), then*

$$\sigma_m(\mathbf{m}) = k\, \rho_m(\mathbf{m}) \int_{\mathcal{D}} d\mathbf{d}\, \frac{\rho_d(\mathbf{d})\; \vartheta(\mathbf{d}|\mathbf{m})}{\mu_d(\mathbf{d})}\ . \qquad (74)$$

Example 17. *In the context of the previous example, assume that observational uncertainties are Gaussian:*

$$\rho_d(\mathbf{d}) = k \exp\left(-\frac{1}{2} (\mathbf{d} - \mathbf{d}_{\text{obs}})^t\ \mathbf{C}_D^{-1} (\mathbf{d} - \mathbf{d}_{\text{obs}}) \right)\ . \qquad (75)$$

Note that the limit for infinite variances gives the homogeneous probability density $\mu_d(\mathbf{d}) = k$. *Furthermore, assume that uncertainties in the physical law are also Gaussian:*

$$\vartheta(\mathbf{d}|\mathbf{m}) = k \exp\left(-\frac{1}{2} (\mathbf{d} - \mathbf{f}(\mathbf{m}))^t\ \mathbf{C}_T^{-1} (\mathbf{d} - \mathbf{f}(\mathbf{m})) \right)\ . \qquad (76)$$

Here 'the physical theory says' that the data values must be 'close' to the 'computed values' $\mathbf{f}(\mathbf{m})$, *with a notion of closeness defined by the 'theoretical covariance matrix'* $\mathbf{C}_T$. *As demonstrated in Tarantola (1987, p. 158), the integral in Eq. (74) can be analytically evaluated, and gives*

$$\int_{\mathcal{D}} d\mathbf{d}\, \frac{\rho_d(\mathbf{d})\, \vartheta(\mathbf{d}|\mathbf{m})}{\mu_d(\mathbf{d})} = k \exp\left(-\frac{1}{2} (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}})^t (\mathbf{C}_D + \mathbf{C}_T)^{-1} (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}}) \right)\ . \qquad (77)$$

This shows that when using the Gaussian probabilistic model, observational and theoretical uncertainties combine through addition of the respective covariance operators (a nontrivial result).

Example 18. *In the 'Galilean law' example developed in Section 4.61, we described the correlation between the position* x *and the time* t *of a free falling object through a probability density* $\vartheta(x, t)$ *. This law says that falling objects describe, approximately, a space–time parabola. Assume that in a particular experiment the falling object explodes at some point of its space–time trajectory. A plain measurement of the coordinates* (x, t) *of the event gives the probability density* $\rho(x, t)$ *. By 'plain measurement' we mean here that we have used a measurement technique that is not taking into account the particular parabolic character of the fall (i.e., the measurement is designed to work identically for any sort of trajectory). The conjunction of the physical law* $\vartheta(x, t)$ *and the experimental result* $\rho(x, t)$ *, using expression (68), gives*

$$\sigma(x,\, t) = k\, \frac{\rho(x,\, t)\, \vartheta(x,\, t)}{\mu(x,\, t)}\ , \qquad (78)$$

where, as the coordinates (x, t) *are 'Cartesian,'* $\mu(x, t) = k$ *. Taking the explicit expression given for* $\vartheta(x, t)$ *in Eqs. (63) and (64), with* $\vartheta_t(t) = k$ *,*

$$\vartheta(x,\, t) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_x^2 + \sigma_v^2\, t^2}} \times \exp\left(-\frac{1}{2} \frac{\left(x - (x_0 + v_0\, t + \frac{1}{2}\, g\, t^2) \right)^2}{\sigma_x^2 + \sigma_v^2\, t^2} \right)\ , \qquad (79)$$

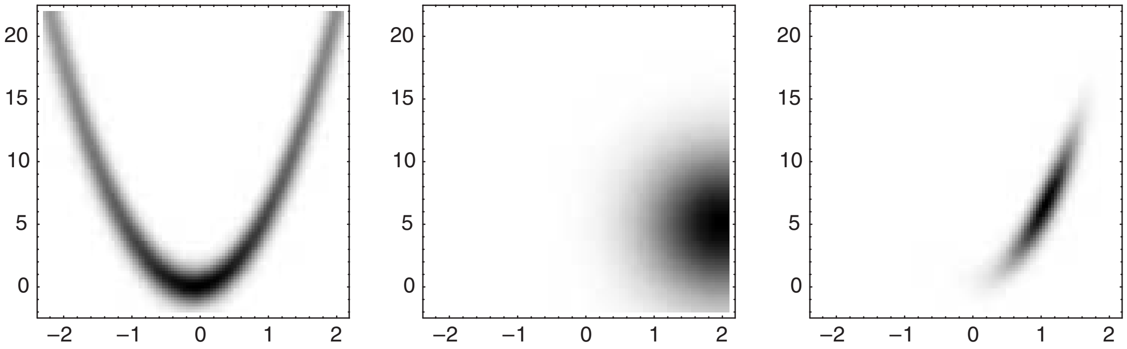

FIGURE 7 This figure has been made with the numerical values mentioned in Figure 17 (see Appendix C) with, in addition, $x_{\text{obs}} = 5.0\,\text{m}$, $\Sigma_x = 4.0\,\text{m}$, $t_{\text{obs}} = 2.0\,\text{sec}$ and $\Sigma_t = 0.75\,\text{sec}$.

and assuming the Gaussian form[19] *for* $\rho(x, t)$,

$$\rho(x,\,t) = \rho_x(x)\,\rho_t(t)$$
$$= k\,\exp\left(-\frac{1}{2}\frac{(x - x_{\text{obs}})^2}{\Sigma_x^2}\right)\,\exp\left(-\frac{1}{2}\frac{(t - t_{\text{obs}})^2}{\Sigma_t^2}\right) \tag{80}$$

we obtain the combined probability density

$$\sigma(x,\,t) = \frac{k}{\sqrt{\sigma_x^2 + \sigma_v^2 t^2}}\,\exp\left(-\frac{1}{2}\left(\frac{(x - x_{\text{obs}})^2}{\Sigma_x^2} + \frac{(t - t_{\text{obs}})^2}{\Sigma_t^2} + \frac{\left(x - (x_0 + v_0 t + \frac{1}{2}\,g t^2)\right)^2}{\sigma_x^2 + \sigma_v^2 t^2}\right)\right)\,. \tag{81}$$

Figure 7 illustrates the three probability densities $\vartheta(x,t)$, $\rho(x,t)$, *and* $\sigma(x,t)$. *See Appendix C for a more detailed examination of this problem.*

5. Solving Inverse Problems (I): Examination of the Probability Density

The next two sections deal with Monte Carlo and optimization methods. The implementation of these methods takes some programming effort that is not required when we face problems with fewer degrees of freedom (say, between one and five).

When we have a small number of parameters we should directly ‘plot’ the probability density.

In Appendix K the problem of estimation of a seismic hypocenter is treated, and it is shown there that the examination of the probability density for the location of the hypocenter offers a much better possibility for analysis than any other method.

6. Solving Inverse Problems (II): Monte Carlo Methods

6.1 Basic Equations

The starting point could be the explicit expression (Eq. (46)) for $\sigma_m(\mathbf{m})$ given in Section 4.5.2:

$$\sigma_m(\mathbf{m}) = k\,\rho_m(\mathbf{m})\,L(\mathbf{m})\,. \tag{82}$$

where

$$L(\mathbf{m}) = \left.\left(\frac{\rho_d(\mathbf{d})}{\sqrt{\det \mathbf{g}_d(\mathbf{d})}}\,\frac{\sqrt{\det\left(\mathbf{g}_m(\mathbf{m}) + \mathbf{F}^T(\mathbf{m})\,\mathbf{g}_d(\mathbf{d})\,\mathbf{F}(\mathbf{m})\right)}}{\sqrt{\det \mathbf{g}_m(\mathbf{m})}}\right)\right|_{\mathbf{d}=\mathbf{f}(\mathbf{m})}\,. \tag{83}$$

In this expression the matrix of partial derivatives $\mathbf{F} = \mathbf{F}(\mathbf{m})$, with components $D_{i\alpha} = \partial f_i / \partial m_\alpha$, appears. The ‘slope’ $\mathbf{F}$ enters here because the steeper the slope for a given $\mathbf{m}$, the greater the accumulation of points we will have with this particular $\mathbf{m}$. This is because we use explicitly the analytic expression $\mathbf{d} = \mathbf{f}(\mathbf{m})$. One should realize that using the more general approach based on Eq. (68) of Section 4.6.2, the effect is automatically accounted for, and there is no need to explicitly consider the partial derivatives.

Equation (82) has the standard form of a conjunction of two probability densities, and is therefore ready to be integrated in a Metropolis algorithm. But one should note that, contrary to many ‘nonlinear’ formulations of inverse problems, the partial derivatives $\mathbf{F}$ are needed even if we use a Monte Carlo method.

In some weakly nonlinear problems, we have $\mathbf{F}^T(\mathbf{m})\;\mathbf{g}_d(\mathbf{d})\;\mathbf{F}(\mathbf{m}) \ll \mathbf{g}_m(\mathbf{m})$, and then Eq. (83) becomes

$$L(\mathbf{m}) = \left.\frac{\rho_d(\mathbf{d})}{\mu_d(\mathbf{d})}\right|_{\mathbf{d}=\mathbf{f}(\mathbf{m})}\,, \tag{84}$$

where we have used $\mu_d(\mathbf{d}) = k\,\sqrt{\det \mathbf{g}_d(\mathbf{d})}$ (see Rule 2).

This expression is also ready for use in the Metropolis algorithm. In this way, sampling of the prior $\rho_m(\mathbf{m})$ is modified into a sampling of the posterior $\sigma_m(\mathbf{m})$, and the Metropolis Rule uses the 'likelihood function' $L(\mathbf{m})$ to calculate acceptance probabilities.

6.2 Sampling the Homogeneous Probability Distribution

If we do not have an algorithm that samples the prior probability density directly, the first step in a Monte Carlo analysis of an inverse problem is to design a random walk that samples the model space according to the homogeneous probability distribution $\mu_m(\mathbf{m})$. In some cases this is easy, but in other cases only an algorithm (a *primeval random walk*) that samples an arbitrary (possibly constant) probability density $\psi(\mathbf{m}) \neq \mu_m(\mathbf{m})$ is available. Then the Metropolis rule can be used to modify $\psi(\mathbf{m})$ into $\mu_m(\mathbf{m})$ (see Section 3.4). This way of generating samples from $\mu_m(\mathbf{m})$ is efficient if $\psi(\mathbf{m})$ is close to $\mu_m(\mathbf{m})$, otherwise it may be very inefficient.

Once $\mu(\mathbf{m})$ can be sampled, the Metropolis Rule allows us to modify this sampling into an algorithm that samples the prior.

6.3 Sampling the Prior Probability Distribution

The first step in the Monte Carlo analysis is to temporarily 'switch off' the comparison between computed and observed data, thereby generating samples of the prior probability density. This allows us to verify statistically that the algorithm is working correctly, and it allows us to understand the prior information we are using. We will refer to a large collection of models representing the prior probability distribution as the 'prior movie' (in a computer screen, when the models are displayed one after the other, we have a 'movie'). The more models present in this movie, the more accurate the representation of the prior probability density.

6.4 Sampling the Posterior Probability Distribution

If we now switch on the comparison between computed and observed data using, e.g., the Metropolis rule for the actual Eq. (82), the random walk sampling the prior distribution is modified into a walk sampling the posterior distribution.

Since data rarely put strong constraints on the Earth, the 'posterior movie' typically shows that many different models are possible. But even though the models in the posterior movie may be quite different, all of them predict data that, within experimental uncertainties, are models with high likelihood. In other words, we must accept that data alone cannot have a preferred model.

The posterior movie allows us to perform a proper resolution analysis that helps us to choose between different interpretations of a given data set. Using the movie we can answer complicated questions about the correlations between several model parameters. To answer such questions, we can view the posterior movie and try to discover structure that is well resolved by data. Such structure will appear as 'persistent' in the posterior movie.

The 'movie' can be used to answer quite complicated questions. For instance, to answer the question '*Which is the probability that the Earth has this special characteristic, but not having this other special characteristic?*' we can just count the number n of models (samples) satisfying the criterion, and the probability is $P = n/m$, where m is the total number of samples.

Once this 'movie' is generated, it is, of course, possible to represent the 1D or 2D marginal probability densities for all or for some selected parameters: it is enough to concentrate one's attention on those selected parameters in each of the samples generated. Those marginal probability densities may have some pathologies (like being multimodal, or having infinite dispersions), but those are the general characteristics of the joint probability density. Our numerical experience shows that these marginals are, quite often, 'stable' objects, in the sense that they can be accurately determined with only a small number of samples.

If the marginals are, essentially, beautiful bell-shaped distributions, then, one may proceed to merely computing mean values and standard deviations (or median values and mean deviations), using each of the samples and the elementary statistical formulas.

Another, more traditional, way of investigating resolution is to calculate covariances and higher-order moments. For this we need to evaluate integrals of the form

$$R_f = \int_{\mathcal{A}} d\mathbf{m}\ f(\mathbf{m})\ \sigma_m(\mathbf{m}) \tag{85}$$

where $f(\mathbf{m})$ is a given function of the model parameters and $\mathcal{A}$ is an event in the model space $\mathcal{M}$ containing the models we are interested in. For instance,

$$\mathcal{A} = \{\mathbf{m}\,|\text{a given range of parameters in } \mathbf{m} \text{ is } \textit{cyclic}\}\ . \tag{86}$$

In the special case when $\mathcal{A} = \mathcal{M}$ is the entire model space, and $f(\mathbf{m}) = m_i$, the R_f in Eq. (85) equals the mean $\langle m_i \rangle$ of the ith model parameter m_i . If $f(\mathbf{m}) = (m_i - \langle m_i \rangle)(m_j - \langle m_j \rangle)$, R_f becomes the covariance between the ith and jth model parameters. Typically, in the general inverse problem we cannot evaluate the integral in Eq. (85) analytically because we have no analytical expression for $\sigma(\mathbf{m})$. However, from the samples of the posterior movie $\mathbf{m}_1, \dots, \mathbf{m}_n$ we can approximate R_f by the simple average

$$R_f \approx \frac{1}{\text{total number of models}} \sum_{\{i \,|\, \mathbf{m}_i \in \mathcal{A}\}} f(\mathbf{m}_i) \ . \qquad (87)$$

7. Solving Inverse Problems (III): Deterministic Methods

As we have seen, the solution of an inverse problem essentially consists of a probability distribution over the space of all possible models of the physical system under study. In general, this 'model space' is high-dimensional, and the only general way to explore it is by using the Monte Carlo methods developed in Section 3.

If the probability distributions are 'bell-shaped' (i.e., if they look like a Gaussian or like a generalized Gaussian), then one may simplify the problem by calculating only the point around which the probability is maximum, with an approximate estimation of the variances and covariances. This is the problem addressed in this section. Among the many methods available to obtain the point at which a scalar function reaches its maximum value (relaxation methods, linear programming techniques, etc.) we limit our scope here to the methods using the gradient of the function, which we assume can be computed analytically, or at least, numerically. For more general methods, the reader may have a look at Fletcher (1980, 1981), Powell (1981), Scales (1985), Tarantola (1987) or Scales *et al.* (1992).

7.1 Maximum Likelihood Point

Let us consider a space $\mathcal{X}$, with a volume element dV defined. If the coordinates $\mathbf{x} \equiv \{x^1, x^2, \dots, x^n\}$ are chosen over the space, the volume element has an expression $dV(\mathbf{x}) = v(\mathbf{x})\, d\mathbf{x}$, and each probability distribution over $\mathcal{X}$ can be represented by a probability density $f(\mathbf{x})$. For any fixed small volume ΔV we can search for the point $\mathbf{x}_{\mathrm{ML}}$ such that the probability dP of the small volume, when centered around $\mathbf{x}_{\mathrm{ML}}$, attains a maximum. In the limit $\Delta V \to 0$ this defines the *maximum likelihood point*. The maximum likelihood point may be unique (if the probability distribution is unimodal), may be degenerate (if the probability distribution is 'chevron-shaped'), or may be multiple (as when we have the sum of a few bell-shaped functions).

The maximum likelihood point is *not* the point at which the probability density is maximum. Our definition implies that a maximum must be attained by the ratio between the probability density and the function $v(\mathbf{x})$ defining the volume element:[20]

$$\mathbf{x} = \mathbf{x}_{\mathrm{ML}} \iff F(\mathbf{x}) = \frac{f(\mathbf{x})}{v(\mathbf{x})} \quad (\text{maximum}) \ . \qquad (88)$$

As the homogeneous probability density is $\mu(\mathbf{x}) = k\, v(\mathbf{x})$ (see Rule 2), we can equivalently define the maximum likelihood point by the condition

$$\mathbf{x} = \mathbf{x}_{\mathrm{ML}} \iff \frac{f(\mathbf{x})}{\mu(\mathbf{x})} \quad (\text{maximum}) \ . \qquad (89)$$

The point at which a probability density has its maximum is, in general, not $\mathbf{x}_{\mathrm{ML}}$. In fact, the maximum of a probability density does not correspond to an intrinsic definition of a point: a change of coordinates $\mathbf{x} \mapsto \mathbf{y} = \psi(\mathbf{x})$ would change the probability density $f(\mathbf{x})$ into the probability density $g(\mathbf{y})$ (obtained using the Jacobian rule), but the point of the space at which $f(\mathbf{x})$ is maximum is not the same as the point of the space where $g(\mathbf{y})$ is maximum (unless the change of variables is linear). This contrasts with the maximum likelihood point, as defined by Eq. (89), which is an intrinsically defined point: no matter which coordinates we use in the computation we always obtain the same point of the space.

7.2 Misfit

One of the goals here is to develop gradient-based methods for obtaining the maximum of $F(\mathbf{x}) = f(\mathbf{x})/\mu(\mathbf{x})$. As a quite general rule, gradient-based methods perform quite poorly for (bell-shaped) probability distributions, as when one is far from the maximum the probability densities tend to be quite flat, and it is difficult to get, reliably, the direction of steepest ascent. Taking a logarithm transforms a bell-shaped distribution into a paraboloid-shaped distribution on which gradient methods work well.

The logarithmic volumetric probability, or *misfit*, is defined as $S(\mathbf{x}) = -\log(F(\mathbf{x})/F_0)$, where p' and F_0 are two constants, and is given by

$$S(\mathbf{x}) = -\log \frac{f(\mathbf{x})}{\mu(\mathbf{x})} \ . \qquad (90)$$

The problem of maximization of the (typically) bell-shaped function $f(\mathbf{x})/\mu(\mathbf{x})$ has been transformed into the problem of minimization of the (typically) paraboloid-shaped function $S(\mathbf{x})$:

$$\mathbf{x} = \mathbf{x}_{\mathrm{ML}} \iff S(\mathbf{x}) \quad (\text{minimum}) \ . \qquad (91)$$

Example 19. *The conjunction $\sigma(\mathbf{x})$ of two probability densities $\rho(\mathbf{x})$ and $\vartheta(\mathbf{x})$ was defined (Eq. (13)) as*

$$\sigma(\mathbf{x}) = p\, \frac{\rho(\mathbf{x})\; \vartheta(\mathbf{x})}{\mu(\mathbf{x})} \ . \qquad (92)$$

Then,

$$S(\mathbf{x}) = S_\rho(\mathbf{x}) + S_\vartheta(\mathbf{x}) \ , \qquad (93)$$

where

$$S_\rho(\mathbf{x}) = -\log\frac{\rho(\mathbf{x})}{\mu(\mathbf{x})}\,; \qquad S_\vartheta(\mathbf{x}) = -\log\frac{\vartheta(\mathbf{x})}{\mu(\mathbf{x})}\ . \tag{94}$$

Example 20. *In the context of Gaussian distributions we have found the probability density (see Example 12)*

$$\sigma_m(\mathbf{m}) = k\exp\left[-\frac{1}{2}\big((\mathbf{m}-\mathbf{m}_{\text{prior}})^t\ \mathbf{C}_M^{-1}(\mathbf{m}-\mathbf{m}_{\text{prior}}) + (\mathbf{f}(\mathbf{m})-\mathbf{d}_{\text{obs}})^t\ \mathbf{C}_D^{-1}(\mathbf{f}(\mathbf{m})-\mathbf{d}_{\text{obs}})\big)\right]\ . \tag{95}$$

The limit of this distribution for infinite variances is a constant, so in this case $\mu_m(\mathbf{m}) = k$. The misfit function $S(\mathbf{m}) = -\log(\sigma_m(\mathbf{m})/\mu_m(\mathbf{m}))$ is then given by

$$2\,S(\mathbf{m}) = (\mathbf{m}-\mathbf{m}_{\text{prior}})^t\ \mathbf{C}_M^{-1}(\mathbf{m}-\mathbf{m}_{\text{prior}}) + (\mathbf{f}(\mathbf{m})-\mathbf{d}_{\text{obs}})^t\,\mathbf{C}_D^{-1}(\mathbf{f}(\mathbf{m})-\mathbf{d}_{\text{obs}})\ . \tag{96}$$

The reader should remember that this misfit function is valid only for weakly nonlinear problems (see Examples 10 and 12). The maximum likelihood model here is the one that minimizes the sum of squares (96). This corresponds to the least-squares criterion.

7.3 Gradient and Direction of Steepest Ascent

One must not consider as synonymous the notions of 'gradient' and 'direction of steepest ascent.' Consider, for instance, an *adimensional* misfit function[21] $S(P,T)$ over a pressure P and a temperature T. Any sensible definition of the gradient of S will lead to an expression like

$$\text{grad } S = \begin{pmatrix} \frac{\partial S}{\partial P} \\ \\ \frac{\partial S}{\partial T} \end{pmatrix} \tag{97}$$

and this by no means can be regarded as a 'direction' in the (P,T) space (for instance, the components of this 'vector' does not have the dimensions of pressure and temperature, but of inverse pressure and inverse temperature).

Mathematically speaking, *the gradient of a function $S(\mathbf{x})$ at a point $\mathbf{x}_0$ is the linear function that is tangent to $S(\mathbf{x})$ at* $\mathbf{x}_0$. This definition of gradient is consistent with the more elementary one, based on the use of the first-order expansion

$$S(\mathbf{x}_0+\delta\mathbf{x}) = S(\mathbf{x}_0) + \widehat{\boldsymbol{\gamma}}_0^T\,\delta\mathbf{x} + \cdots \tag{98}$$

Here $\widehat{\boldsymbol{\gamma}}_0$ is called the gradient of $S(\mathbf{x})$ at point $\mathbf{x}_0$. It is clear that $S(\mathbf{x}_0) + \widehat{\boldsymbol{\gamma}}_0^T\,\delta\mathbf{x}$ is a linear function, and that it is tangent to $S(\mathbf{x})$ at $\mathbf{x}_0$, so the two definitions are in fact equivalent. Explicitly, the components of the gradient at point $\mathbf{x}_0$ are

$$(\widehat{\boldsymbol{\gamma}}_0)_p = \frac{\partial S}{\partial x^p}(\mathbf{x}_0)\ . \tag{99}$$

Everybody is well trained in computing the gradient of a function (event if the interpretation of the result as a direction in the original space is wrong). How can we pass from the gradient to the direction of steepest ascent (a bona fide direction in the original space)? In fact, the gradient (at a given point) of a function defined over a given space $\mathcal{E}$) is an element of the dual of the space. To obtain a direction in $\mathcal{E}$ we must pass from the dual to the primal space. As usual, it is the metric of the space that maps the dual of the space into the space itself. So if $\mathbf{g}$ is the metric of the space where $S(\mathbf{x})$ is defined, and if $\widehat{\boldsymbol{\gamma}}$ is the gradient of S at a given point, the *direction of steepest ascent* is

$$\boldsymbol{\gamma} = \mathbf{g}^{-1}\widehat{\boldsymbol{\gamma}}\ . \tag{100}$$

The direction of steepest ascent must be interpreted as follows: if we are at a point $\mathbf{x}$ of the space, we can consider a very small hypersphere around $\mathbf{x}_0$. The direction of steepest ascent points toward the point of the sphere at which $S(\mathbf{x})$ attains its maximum value.

Example 21. *In the context of least squares, we consider a misfit function $S(\mathbf{m})$ and a covariance matrix $\mathbf{C}_M$* . *If $\widehat{\boldsymbol{\gamma}}_0$ is the gradient of S , at a point $\mathbf{x}_0$, and if we use $\mathbf{C}_M$ to define distances in the space, the direction of steepest ascent is*

$$\boldsymbol{\gamma}_0 = \mathbf{C}_M\,\widehat{\boldsymbol{\gamma}}_0\ . \tag{101}$$

7.4 The Steepest Descent Method

Consider that we have a probability distribution defined over an n-dimensional space $\mathcal{X}$. Having chosen the coordinates $\mathbf{x} \equiv \{x^1,\ x^2,\dots,x^n\}$ over the space, the probability distribution is represented by the probability density $f(\mathbf{x})$ whose homogeneous limit (in the sense developed in Section 2.2) is $\mu(\mathbf{x})$. We wish to calculate the coordinates $\mathbf{x}_{\text{ML}}$ of the maximum likelihood point. By definition (Eq. (89)),

$$\mathbf{x} = \mathbf{x}_{\text{ML}} \iff \frac{f(\mathbf{x})}{\mu(\mathbf{x})}\quad \text{(maximum)}\ , \tag{102}$$

that is,

$$\mathbf{x} = \mathbf{x}_{\text{ML}} \iff S(\mathbf{x})\quad \text{(minimum)}\ , \tag{103}$$

where $S(\mathbf{x})$ is the misfit [Eq. (90)]

$$S(\mathbf{x}) = -k\log\frac{f(\mathbf{x})}{\mu(\mathbf{x})}\ . \tag{104}$$

Let us denote by $\widehat{\boldsymbol{\gamma}}(\mathbf{x}_k)$ the gradient of $S(\mathbf{x})$ at point $\mathbf{x}_k$, i.e. (Eq. (99)),

$$(\widehat{\boldsymbol{\gamma}}_0)_p = \frac{\partial S}{\partial x^p}(\mathbf{x}_0) \ . \tag{105}$$

We have seen above that $\widehat{\boldsymbol{\gamma}}(\mathbf{x})$ should not be interpreted as a direction in the space $\mathcal{X}$ but as a direction in the dual space. The gradient can be converted into a direction using a metric $\mathbf{g}(\mathbf{x})$ over $\mathcal{X}$. In simple situations the metric $\mathbf{g}$ will be the one used to define the volume element of the space, i.e., we will have $\mu(\mathbf{x}) = k\,v(\mathbf{x}) = k\sqrt{\det \mathbf{g}(\mathbf{x})}$, but this is not a necessity, and iterative algorithms may be accelerated by astute introduction of ad-hoc metrics.

Given, then, the gradient $\widehat{\boldsymbol{\gamma}}(\mathbf{x}_k)$ (at some particular point $\mathbf{x}_k$) to any possible choice of metric $\mathbf{g}(\mathbf{x})$ we can define the direction of steepest ascent associated with the metric $\mathbf{g}$, by (Eq. (101))

$$\boldsymbol{\gamma}(\mathbf{x}_k) = \mathbf{g}^{-1}(\mathbf{x}_k)\,\widehat{\boldsymbol{\gamma}}(\mathbf{x}_k) \ . \tag{106}$$

The algorithm of steepest descent is an iterative algorithm passing from point $\mathbf{x}_k$ to point $\mathbf{x}_{k+1}$ by making a 'small jump' along the local direction of steepest descent,

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \varepsilon_k\,\mathbf{g}_k^{-1}\,\widehat{\boldsymbol{\gamma}}_k \ , \tag{107}$$

where ε_k is an ad-hoc (real, positive) value adjusted to force the algorithm to converge rapidly (if ε_k is chosen too small, the convergence may be too slow; it is chosen too large, the algorithm may even diverge).

Many elementary presentations of the steepest descent algorithm just forget to include the metric $\mathbf{g}_k$ in expression (107). These algorithms are not consistent. Even the physical dimensionality of the equation is not assured. 'Numerical' problems in computer implementations of steepest descent algorithms can often be traced to the fact that the metric has been neglected.

Example 22. *In the context of Example 20, where the misfit function $S(\mathbf{m})$ is given by*

$$\begin{aligned} 2\,S(\mathbf{m}) = & (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}})^t\ \mathbf{C}_D^{-1}\,(\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}}) \\ & + (\mathbf{m} - \mathbf{m}_{\text{prior}})^t\ \mathbf{C}_M^{-1}\,(\mathbf{m} - \mathbf{m}_{\text{prior}}) \ , \end{aligned} \tag{108}$$

the gradient $\widehat{\boldsymbol{\gamma}}$, whose components are $\widehat{\boldsymbol{\gamma}}_\alpha = \partial S/\partial m^\alpha$, is given by the expression

$$\widehat{\boldsymbol{\gamma}}(\mathbf{m}) = \mathbf{F}^t(\mathbf{m})\,\mathbf{C}_D^{-1}\,(\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}}) + \mathbf{C}_M^{-1}\,(\mathbf{m} - \mathbf{m}_{\text{prior}}) \ , \tag{109}$$

where $\mathbf{F}$ is the matrix of partial derivatives

$$F^{i\alpha} = \frac{\partial f^i}{\partial m^\alpha} \ . \tag{110}$$

An example of computation of partial derivatives is given in Appendix M.

Example 23. *In the context of Example 22 the model space $\mathcal{M}$ has an obvious metric, namely, that defined by the inverse of the 'a priori' covariance operator $\mathbf{g} = \mathbf{C}_M^{-1}$. Using this metric and the gradient given by Eq. (109), the steepest descent algorithm (107) becomes*

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \varepsilon_k\left(\mathbf{C}_M\,\mathbf{F}_k^t\,\mathbf{C}_D^{-1}\,(\mathbf{f}_k - \mathbf{d}_{\text{obs}}) + (\mathbf{m}_k - \mathbf{m}_{\text{prior}})\right) \ , \tag{111}$$

where $\mathbf{F}_k \equiv \mathbf{F}(\mathbf{m}_k)$ and $\mathbf{f}_k \equiv \mathbf{f}(\mathbf{m}_k)$. The real positive quantities ε_k can be fixed after some trial and error by accurate linear search, or by using a linearized approximation.

Example 24. *In the context of Example 22 the model space $\mathcal{M}$ has a less obvious metric, namely, that defined by the inverse of the 'posterior' covariance operator, $\mathbf{g} = \widetilde{\mathbf{C}}_M^{-1}$.*[23] *Using this metric and the gradient given by Eq. (109), the steepest descent algorithm (107) becomes*

$$\begin{aligned} \mathbf{m}_{k+1} = \mathbf{m}_k - \varepsilon_k & \left(\mathbf{F}_k^t\,\mathbf{C}_D^{-1}\,\mathbf{F}_k + \mathbf{C}_M^{-1}\right)^{-1}\left(\mathbf{F}_k^t\,\mathbf{C}_D^{-1}\,(\mathbf{f}_k - \mathbf{d}_{\text{obs}}) \right. \\ & \left. + \,\mathbf{C}_M^{-1}\,(\mathbf{m}_k - \mathbf{m}_{\text{prior}})\right) \ , \end{aligned} \tag{112}$$

where $\mathbf{F}_k \equiv \mathbf{F}(\mathbf{m}_k)$ and $\mathbf{f}_k \equiv \mathbf{f}(\mathbf{m}_k)$. The real positive quantities ε_k can be fixed, after some trial and error, by accurate linear search, or by using a linearized approximation that simply gives[24] *$\varepsilon_k \approx 1$.*

The algorithm (112) is usually called a 'quasi-Newton algorithm.' This name is not well chosen: a Newton method applied to minimization of a misfit function $S(\mathbf{m})$ would be a method using the second derivatives of $S(\mathbf{m})$, and thus the derivatives $H^i_{\alpha\beta} = \frac{\partial^2 f^i}{\partial m^\alpha \partial m^\beta}$, that are not computed (or not estimated) when using this algorithm. It is just a steepest descent algorithm with a nontrivial definition of the metric in the working space. In this sense it belongs to the wider class of 'variable metric methods,' not discussed in this article.

7.5 Estimating Posterior Uncertainties

In the Gaussian context, the Gaussian probability density that is tangent to $\sigma_m(\mathbf{m})$ has its center at the point given by the iterative algorithm

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \varepsilon_k\left(\mathbf{C}_M\,\mathbf{F}_k^t\,\mathbf{C}_D^{-1}\,(\mathbf{f}_k - \mathbf{d}_{\text{obs}}) + (\mathbf{m}_k - \mathbf{m}_{\text{prior}})\right) \ , \tag{113}$$

(Eq. (111)) or, equivalently, by the iterative algorithm

$$\begin{aligned} \mathbf{m}_{k+1} = \mathbf{m}_k - \varepsilon_k & \left(\mathbf{F}_k^t\mathbf{C}_D^{-1}\,\mathbf{F}_k + \mathbf{C}_M^{-1}\right)^{-1}\left(\mathbf{F}_k^t\,\mathbf{C}_D^{-1}\,(\mathbf{f}_k - \mathbf{d}_{\text{obs}}) \right. \\ & \left. +\mathbf{C}_M^{-1}\,(\mathbf{m}_k - \mathbf{m}_{\text{prior}})\right) \end{aligned} \tag{114}$$

(Eq. (112)). The covariance of the tangent Gaussian is

$$\widetilde{\mathbf{C}}_M \approx \left(\mathbf{F}_\infty^t \, \mathbf{C}_D^{-1} \, \mathbf{F}_\infty + \mathbf{C}_M^{-1} \right)^{-1} , \qquad (115)$$

where $\mathbf{F}_\infty$ refers to the value of the matrix of partial derivatives at the convergence point.

7.6 Some Comments on the Use of Deterministic Methods

7.6.1 Linear , Weakly Nonlinear and Nonlinear Problems

There are different degrees of nonlinearity. Figure 8 illustrates four domains of nonlinearity, calling for different

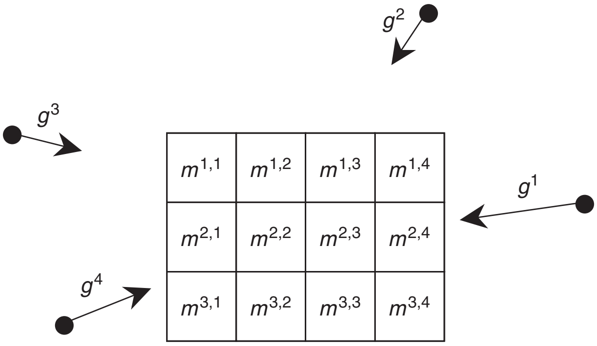
g^2
g^3
g^1
g^4
$m^{1,1}$ $m^{1,2}$ $m^{1,3}$ $m^{1,4}$
$m^{2,1}$ $m^{2,2}$ $m^{2,3}$ $m^{2,4}$
$m^{3,1}$ $m^{3,2}$ $m^{3,3}$ $m^{3,4}$

FIGURE 8 A simple example where we are interested in predicting the gravitational field **g** generated by a 2D distribution of mass.

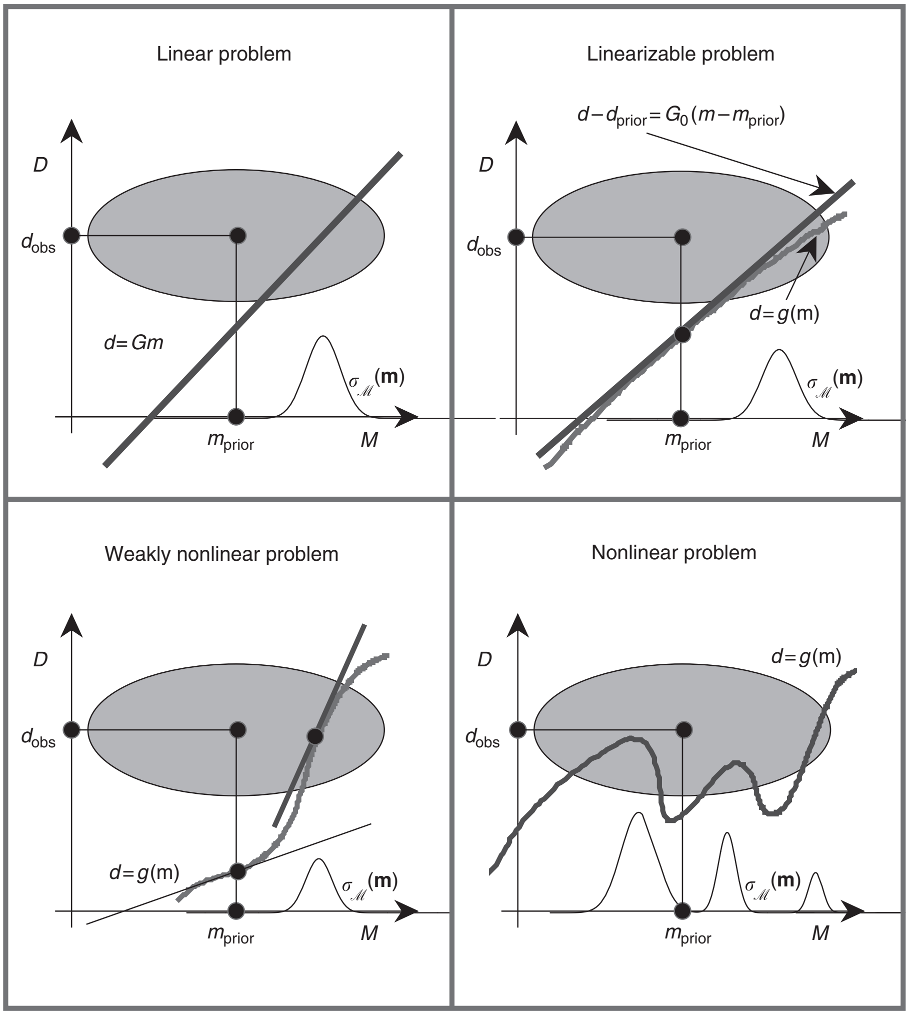

FIGURE 9 Illustration of the four domains of nonlinearity, calling for different optimization algorithms. The model space is symbolically represented by the abscissa, and the data space is represented by the ordinate. The gray oval represents the combination of prior information on the model parameters and information from the observed data. What is important is not an intrinsic nonlinearity of the function relating model parameters to data, but how linear the function is *inside the domain of significant probability*.

optimization algorithms. In this figure the abscissa symbolically represents the model space, and the ordinate represents the data space. The gray oval represents the combination of prior information on the model parameters, and information from the observed data.[25] It is the probability density $\rho(\mathbf{d}, \mathbf{m}) = \rho_d(\mathbf{d})\rho_m(\mathbf{m})$ seen elsewhere.

To fix ideas, the oval suggests here a Gaussian probability, but our distinction between problems according to their nonlinearity will not depend fundamentally on this.

First, there are strictly linear problems. For instance, in the example illustrated by Figure 8 the gravitational field $\mathbf{g}$ depends linearly on the masses inside the blocks.[26]

Strictly linear problems are illustrated at the top left of Figure 9. The linear relationship between data and model parameters, $\mathbf{d} = \mathbf{G}\,\mathbf{m}$, is represented by a straight line. The prior probability density $\rho(\mathbf{d}, \mathbf{m})$ 'induces' on this straight line the posterior probability density[27] $\sigma(\mathbf{d}, \mathbf{m})$ whose 'projection' over the model space gives the posterior probability density over the model parameter space, $\sigma_m(\mathbf{m})$. Should the prior probability densities be Gaussian, then the posterior probability distribution would also be Gaussian: this is the simplest situation.

Quasi-linear problems are illustrated at the bottom left of Figure 9. If the relationship linking the observable data $\mathbf{d}$ to the model parameters $\mathbf{m}$,

$$\mathbf{d} = \mathbf{g}(\mathbf{m}) \ , \qquad (116)$$

is approximately linear *inside the domain of significant prior probability* (i.e., inside the gray oval of the figure), then the posterior distribution is just as simple as the prior distribution. For instance, if the prior is Gaussian the posterior is also Gaussian.

In this case also, the problem can be reduced to the computation of the mean and the covariance of the Gaussian. Typically, one begins at some 'starting model' $\mathbf{m}_0$ (typically, one takes for $\mathbf{m}_0$ the 'a priori model' $\mathbf{m}_{\text{prior}}$),[28] linearizing the function $\mathbf{d} = \mathbf{g}(\mathbf{m})$ around $\mathbf{m}_0$, and one looks for a model $\mathbf{m}_1$ 'better than $\mathbf{m}_0$.'

Iterating such an algorithm, one tends to the model $\mathbf{m}_\infty$ at which the 'quasi-Gaussian' $\sigma_m(\mathbf{m})$ is maximum. The linearizations made in order to arrive to $\mathbf{m}_\infty$ are so far not an approximation: the point $\mathbf{m}_\infty$ is perfectly defined, independently of any linearization and any method used to find it. But once the convergence to this point has been obtained, a linearization of the function $\mathbf{d} = \mathbf{g}(\mathbf{m})$ around this point,

$$\mathbf{d} - \mathbf{g}(\mathbf{m}_\infty) = \mathbf{G}_\infty(\mathbf{m} - \mathbf{m}_\infty) \ , \qquad (117)$$

allows to obtain a good approximation to the posterior uncertainties. For instance, if the prior distribution is Gaussian this will give the covariance of the 'tangent Gaussian.'

Between linear and quasi-linear problems there are the 'linearizable problems.' The scheme at the top right of Figure 9 shows the case where the linearization of the function $\mathbf{d} = \mathbf{g}(\mathbf{m})$ around the prior model,

$$\mathbf{d} - \mathbf{g}(\mathbf{m}_{\text{prior}}) = \mathbf{G}_{\text{prior}}(\mathbf{m} - \mathbf{m}_{\text{prior}}) \ , \qquad (118)$$

gives a function that, inside the domain of significant probability, is very similar to the true (nonlinear) function.

In this case, there is no practical difference between this problem and the strictly linear problem, and the iterative procedure necessary for quasi-linear problems is here superfluous.

It remains to analyze the true nonlinear problems that, using a pleonasm, are sometimes called *strongly nonlinear problems*. They are illustrated at the bottom right of Figure 9.

In this case, even if the prior distribution is simple, the posterior distribution can be quite complicated. For instance, it can be multimodal. These problems are in general quite complex to solve, and only a Monte Carlo analysis, as described in the previous chapter, is feasible.

If full Monte Carlo methods cannot be used, because they are too expensive, then one can mix a random part (for instance, to choose the starting point) and a deterministic part. The optimization methods applicable to quasi-linear problems can, for instance, allow us to go from the randomly chosen starting point to the 'nearest' optimal point. Repeating these computations for different starting points, one can arrive at a good idea of the posterior distribution in the model space.

7.6.2 The Maximum Likelihood Model

The *most likely model* is, by definition, that at which the volumetric probability (see Appendix A) $\sigma_\beta(\mathbf{m})$ attains its maximum. As $\sigma_\beta(\mathbf{m})$ is maximum when $S(\mathbf{m})$ is minimum, we see that the most likely model is also the 'best model' obtained when using a 'least-squares criterion.' Should we have used the double exponential model for all the

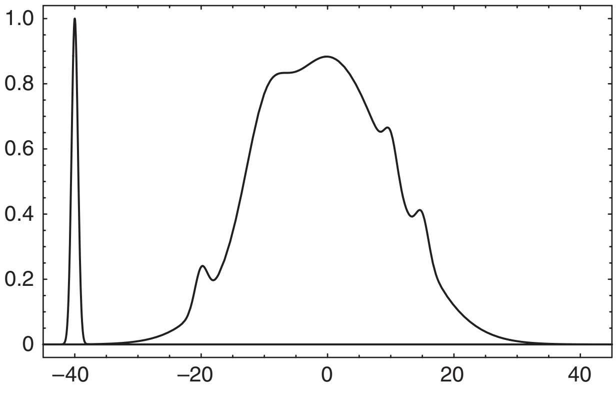

FIGURE 10 One of the circumstances where the 'maximum likelihood model' may not be very interesting is when it corresponds to a narrow maximum with small total probability, as the peak in the left part of this probability distribution.

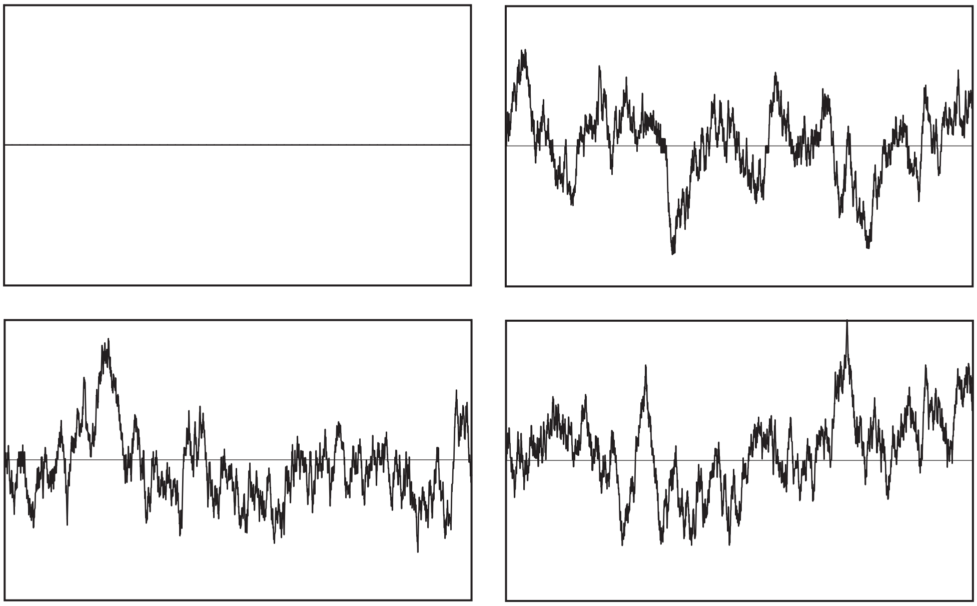

FIGURE 11 At the right, three random realizations of a Gaussian random function with zero mean and (approximately) exponential correlation function. The most likely function, i.e., the center of the Gaussian, is shown at the top left. We see that the most likely function is not a representative of the probability distribution.

uncertainties, then the most likely model would be defined by a 'least absolute values' criterion.

There are many circumstances where the most likely model is not an interesting model. One trivial example is when the volumetric probability has a 'narrow maximum, ' with small total probability (see Fig. 10). A much less trivial situation arises when the number of parameters is very large, as for instance when we deal with a random function (that, strictly speaking, corresponds to an infinite number of random variables). Figure 11, for instance, shows a few realizations of a Gaussian function with zero mean and an (approximately) exponential correlation. The most likely function is the center of the Gaussian, i.e., the null function shown at the top left. But this is not a representative sample of the probability distribution, as any realization of the probability distribution will have, with a probability very close to one, the 'oscillating' characteristics of the three samples shown at the right.

8. Conclusions

Probability theory is well adapted to the formulation of inverse problems, although its formulation must be rendered intrinsic (introducing explicitly the definition of distances in the working spaces, by redefining the notion of conditional probability density, and by introducing the notion of conjunction of states of information). The Metropolis algorithm is well adapted to the solution of inverse problems, as its inherent structure allows us to sequentially combine prior information, theoretical information, etc., and allows us to take advantage of the 'movie philosophy.' When a general Monte Carlo approach cannot be afforded, one can use simplified optimization techniques (like least squares). However, this usually requires strong simplifications that can only be made at the cost of realism.

Acknowledgements

We are very indebted to our colleagues (Bartolomé Coll, Miguel Bosch, Guillaume Évrard, John Scales, Christophe Barnes, Frédéric Parrenin, and Bernard Valette) for illuminating discussions. We are also grateful to the students of the Geophysical Tomography Group, and the students at our respective institutes (in Paris and Copenhagen).

References

Aki, K. and W.H.K. Lee (1976). Determination of three-dimensional velocity anomalies under a seismic array using first P arrival times from local earthquakes. *J. Geophys. Res.* **81**, 4381–4399.

Aki, K., A. Christofferson, and E.S. Husebye (1977). Determination of the three-dimensional seismic structure of the lithosphere. *J. Geophys. Res.* **82**, 277–296.

Backus, G. (1970a). Inference from inadequate and inaccurate data: I. *Proc. Natl. Acad. Sci. USA* **65**(1), 1–105.

Backus, G. (1970b). Inference from inadequate and inaccurate data: II. *Proc. Natl. Acad. Sci. USA* **65**(2), 281–287.

Backus, G. (1970c). Inference from inadequate and inaccurate data: III. *Proc. Natl. Acad. Sci. USA* **67**(1), 282–289.

Backus, G. (1971). Inference from inadequate and inaccurate data. 'Mathematical Problems in the Geophysical Sciences,' Lectures in Applied Mathematics **14**. American Mathematical Society, Providence, Rhode Island.

Backus, G. and F. Gilbert (1967). Numerical applications of a formalism for geophysical inverse problems. *Geophys. J. R. Astron. Soc.* **13**, 247–276.

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

Backus, G. and F. Gilbert (1970). Uniqueness in the inversion of inaccurate gross Earth data. *Philos. Trans. R. Soc. London* **266**, 123–192.

Dahlen, F.A. (1976). Models of the lateral heterogeneity of the Earth consistent with eigenfrequency splitting data. *Geophys. J. R. Astron. Soc.* **44**, 77–105.

Dahl-Jensen, D., K. Mosegaard, N. Gundestrup, G.D. Clow, S.J. Johnsen, A.W. Hansen, and N. Balling (1998). Past temperatures directly from the Greenland Ice Sheet. *Science* (Oct. 9), 268–271.

Fisher, R.A. (1953). Dispersion on a sphere. *Proc. R. Soc. London, A* **217**, 295–305.

Fletcher, R. (1980). 'Practical Methods of Optimization,' Vol. 1: Unconstrained Optimization. Wiley, New York.

Fletcher, R. (1981). 'Practical Methods of Optimization,' Vol. 2: Constrained Optimization. Wiley, New York.

Franklin, J.N. (1970). Well posed stochastic extensions of ill posed linear problems. *J. Math. Anal. Appl.* **31**, 682–716.

Gauss, C.F. (1809). 'Theoria Motus Corporum Cœlestium in Sectionis Conicis Solem Ambientum,' Hamburg. (Also in 'Werke,' Vol. 7. Olmc-Verlag, 1981.)

Geiger, L. (1910). Herdbestimmung bei Erdbeben aus den Ankunftszeiten. *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen* **4**, 331–349.

Geman, S. and D. Geman (1984). *IEEE Trans. Pattern Anal. Mach. Int.* **PAMI-6**(6), 721.

Gilbert, F. (1971). Ranking and winnowing gross Earth data for inversion and resolution. *Geophys. J. R. Astron. Soc.* **23**, 215–128.

Hadamard, J. (1902). Sur les problémes aux dérivées partielles et leur signification physique. *Bull. Univ. Princeton* **13**.

Hadamard, J. (1932). 'Le problème de Cauchy et les équations aux dérivées partielles linéaires hyperboliques.' Hermann, Paris.

ISO (1993). 'Guide to the Expression of Uncertainty in Measurement.' International Organization for Standardization, Switzerland.

Jackson, D.D. (1979). The use of a priori data to resolve nonuniqueness in linear inversion. *Geophys. J. R. Astron. Soc.* **57**, 137–157.

Jaynes, E.T. (1968). Prior probabilities. *IEEE Trans. Syst. Sci. Cybern.* **SSC-4**(3), 227–241.

Jeffreys, H. (1939). 'Theory of Probability.' Clarendon Press, Oxford. (Reprinted in 1961 by Oxford University Press.)

Kandel, A. (1986). 'Fuzzy Mathematical Techniques with Applications.' Addison-Wesley, Reading, MA.

Keilis-Borok, V.J. and T.B. Yanovskaya (1967). Inverse problems in seismology (structural review). *Geophys. J. R. Astron. Soc.* **13**, 223–234.

Kennett, B.L.N. and G. Nolet (1978). Resolution analysis for discrete systems. *Geophys. J. R. Astron. Soc.* **53**, 413–425.

Khan, A., K. Mosegaard, and K.L. Rasmussen (2000). A new seismic velocity model for the Moon from a Monte Carlo inversion of the Apollo lunar seismic data. *Geophys. Res. Lett.* **37**(11), 1591–1594.

Kimeldorf, G. and G. Wahba (1970). A correspondence between Bayesian estimation of stochastic processes and smooting by splines. *Ann. Math. Stat.* **41**, 495–502.

Kullback, S. (1967). The two concepts of information. *J. Am. Stat. Assoc.* **62**, 685–686.

Lehtinen, M.S., L. Päivärinta, and E. Somersalo (1989). Linear inverse problems for generalized random variables. *Inverse Prob.* **5**, 599–612.

Levenberg, K. (1944). A method for the solution of certain nonlinear problems in least-squares. *Q. Appl. Math.* **2**, 164–168.

Marquardt, D.W. (1963). An algorithm for least squares estimation of nonlinear parameters, SIAM J., **11**, 431–441.

Marquardt, D.W. (1970). Generalized inverses, ridge regression, biased linear estimation and non-linear estimation. *Technometrics* **12**, 591–612.

Menke, W. (1984). 'Geophysical Data Analysis: Discrete Inverse Theory.' Academic Press, New York.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1**, 1087–1092.

Minster, J.B. and T.M. Jordan (1978). Present-day plate motions. *J. Geophys. Res.* **83**, 5331–5354.

Mosegaard, K. and C. Rygaard-Hjalsted (1999). Bayesian analysis of implicit inverse problems. *Inverse Probl.* **15**, 573–583.

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

Mosegaard, K., S.C. Singh, D. Snyder, and H. Wagner (1997). Monte Carlo analysis of seismic reflections from Moho and the W-reflector. *J. Geophys. Res. B* **102**, 2969–2981.

Nolet, G. (1990). Partitioned wave-form inversion and 2D structure under the NARS array. *J. Geophys. Res.* **95**, 8499–8512.

Nolet, G., J. van Trier, and R. Huisman (1986). A formalism for nonlinear inversion of seismic surface waves. *Geophys. Res. Lett.* **13**, 26–29.

Parker, R.L. (1994). 'Geophysical Inverse Theory.' Princeton University Press, Princeton, NJ.

Parzen, E., K. Tanabe, and G. Kitagawa (Eds.) (1998). 'Selected Papers of Hirotugu Akaike,' Springer Series in Statistics. Springer-Verlag, New York.

Powell, M.J.D. (1981). 'Approximation Theory and Methods.' Cambridge University Press, Cambridge.

Press, F. (1968). Earth models obtained by Monte Carlo inversion. *J. Geophys. Res.* **73**, 5223–5234.

Rietsch, E. (1977). The maximum entropy approach to inverse problems. *J. Geophys.* **42**, 489–506.

Scales, L.E. (1985). 'Introduction to Non-Linear Optimization.' Macmillan, London.

Scales, J.A., M.L. Smith, and T.L. Fischer (1992). Global optimization methods for multimodal inverse problems. *J. Comput. Phys.* **102**, 258–268.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423.

Su, W.-J., R.L. Woodward, and A.M. Dziewonski (1992). Deep origin of mid-oceanic ridge velocity anomalies. *Nature* **360**, 149–152.

Tarantola, A. and B. Valette (1982a). Inverse problems = quest for information. *J. Geophys.* **50**, 159–170.

Tarantola, A. and B. Valette (1982b). Generalized nonlinear inverse problems solved using the least-squares criterion. *Rev. Geophys. Space Phys.* **20**, 219–232.

Tarantola, A. (1984). Inversion of seismic reflection data in the acoustic approximation. *Geophysics* **49**, 1259–1266.

Tarantola, A. (1986). A strategy for nonlinear elastic inversion of seismic reflection data. *Geophysics* **51**, 1893–1903.

Tarantola, A. (1987). 'Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation.' Elsevier, Amsterdam.

Taylor, A.E. and Lay, D.C. (1980). 'Introduction to Functional Analysis', John Wiley and Sons, New York.

Taylor, B.N. and C.E. Kuyatt (1994). Guidelines for evaluating and expressing the uncertainty of NIST measurement results. NIST Technical note 1297.

Tikhonov, A.N. (1963). Resolution of ill-posed problems and the regularization method (in Russian). *Dokl. Akad. Nauk SSSR* **151**, 501–504.

van der Hilst, R.D., S. Widiyantoro, and E.R. Engdahl (1997). Evidence for deep mantle circulation from global tomography. *Nature* **386**, 578–584.

Wiggins, R.A. (1969). Monte Carlo inversion of body-wave observations. *J. Geophys. Res.* **74**, 3171–3181.

Wiggins, R.A. (1972). The general linear inverse problem: implication of surface waves and free oscillations for Earth structure. *Rev. Geophys. Space Phys.* **10**, 251–285.

Woodhouse, J.H. and F.A. Dahlen (1978). The effect of general aspheric perturbation on the free oscillations of the Earth. *Geophys. J. R. Astron. Soc.* **53**, 335–354.

Notes

1. For instance, we could fit our observations with a heterogeneous but isotropic Earth model or, alternatively, with a homogeneous but anisotropic Earth.
2. Preliminary Earth Reference Model (PREM), Dziewonski and Anderson, PEPI, 1981. Inversion for Centroid Moment Tensor (CMT), Dziewonski, Chou and Woodhouse, JGR, 1982. First global tomographic model, Dziewonski, JGR, 1984.
3. The capacity element associated to the vector elements $d\mathbf{r}_1$, $d\mathbf{r}_2, \ldots d\mathbf{r}_n$ is defined as $d\tau = \varepsilon_{ij\ldots k}\, dr_1^i\, dr_2^j\, \ldots\, dr_n^k$, where $\varepsilon_{ij\ldots k}$ is the Levi-Civita capacity (whose components take the values $\{0, \pm 1\}$). If the metric tensor of the space is $\mathbf{g}(\mathbf{x})$, then $\eta_{ij\ldots k} = \sqrt{\det \mathbf{g}}\, \varepsilon_{ij\ldots k}$ is a true tensor, as it is the product of a density $\sqrt{\det \mathbf{g}}$ by a capacity $\varepsilon_{ij\ldots k}$. Then, the volume element, defined as $dV = \eta_{ij\ldots k}\, dr_1^i dr_2^j \ldots dr_n^k = \sqrt{\det \mathbf{g}}\, d\tau$, is a (true) scalar.
4. This is a property that is valid for any coordinate system that can be chosen over the space.
5. As a counterexample, the distance defined as $ds = |dx| + |dy|$ is not of the L_2 type (it is L_1).
6. This solves the complete problem for isotropic tensors only. It is beyond the scope of this text to propose rules valid for general anisotropic tensors: the necessary mathematics have not yet been developed.
7. The definition of the elastic constants was made before the tensorial structure of the theory was understood. Seismologists today should not use, at a theoretical level, parameters like the first Lamé coefficient λ or the Poisson ratio. Instead they should use κ and μ (and their inverses). In fact, our suggestion in this IASPEI volume is to use the true eigenvalues of the stiffness tensor, $\lambda_\kappa = 3\kappa$, and $\lambda_\mu = 2\mu$, which we propose to call the *eigen-bulk-modulus* and the *eigen-shear-modulus*, respectively.
8. Assume that $p(\mathbf{x})$ and $q(\mathbf{x})$ are normalized by $\int_{\mathcal{X}} d\mathbf{x}\, p(\mathbf{x}) = \mathbf{1}$ and $\int_{\mathcal{X}} d\mathbf{x}\, \mathbf{q}(\mathbf{x}) = \mathbf{1}$. Then, irrespective of the normalizability of $\mu(\mathbf{x})$ (as explained above, $p(\mathbf{x})$ and $q(\mathbf{x})$ are assumed to be absolutely continuous with respect to the homogeneous distribution), $(p \wedge q)(\mathbf{x})$ is normalizable, and its normalized expression is $(p \wedge q)(\mathbf{x}) = \dfrac{p(\mathbf{x})\, q(\mathbf{x})/\mu(\mathbf{x})}{\int_{\mathcal{X}} d\mathbf{x}\, p(\mathbf{x})\, q(\mathbf{x})/\mu(\mathbf{x})}$.
9. As a counter example, working at the surface of the sphere with geographical coordinates $(\mathbf{u}, \mathbf{v}) = (u, v) = (\vartheta, \varphi)$ this condition is **not** fulfilled, as $g_\varphi = \sin\theta$ is a function of ϑ: the surface of the sphere is not the Cartesian product of two 1D spaces.
10. That is, series of numbers that appear random if tested with any reasonable statistical test.
11. To see this, put $f(\mathbf{x}) = \mathbf{1}$, $\mu(\mathbf{x}) = \mathbf{1}$, and $g(\mathbf{x}) = \dfrac{\exp(-E(\mathbf{x})/T)}{\int \exp(-E(\mathbf{x})/T) d\mathbf{x}}$, where $E(\mathbf{x})$ is an 'energy' associated to the point $\mathbf{x}$, and T is a 'temperature'. The summation in the denominator is over the entire space. In this way, our acceptance rule becomes the classical Metropolis rule: point $\mathbf{x}_i$ is always accepted if $E(\mathbf{x}_i) \leq E(\mathbf{x}_j)$, but if $E(\mathbf{x}_i) > E(\mathbf{x}_j)$, it is only accepted with probability $p_{ij}^{\text{acc}} = \exp\big(-\big(E(\mathbf{x}_i) - E(\mathbf{x}_j)\big)/T\big)$.
12. A numerical method is called robust if it is not sensitive to a small number of large errors.
13. It would be violated, for instance, if we use the pair of elastic parameters longitudinal wave velocity – shear wave velocity, as the volume element in the space of elastic wave velocities does not factorize (see Appendix H).
14. We use here the properties $\log\sqrt{\mathbf{A}} = \frac{1}{2}\log\mathbf{A}$, and $\det\mathbf{AB} = \det\mathbf{BA}$
15. Typically, this may happen because the derivatives $\mathbf{F}$ are small or because the variances in $\mathbf{C}_M$ are large.
16. We first use $\log\det\mathbf{A} = \operatorname{trace}\log\mathbf{A}$, and then the series expansion of the logarithm of an operator, $\log(\mathbf{I} + \mathbf{A}) = \mathbf{A} - \frac{1}{2}\mathbf{A}^2 + \cdots$
17. Practically, it may correspond to the output of some 'black box' solving the 'forward problem'.
18. Remember that, even if we wish to use a simple method based on the notion of conditional probability density, an analytic expression like $\mathbf{d} = \mathbf{f}(\mathbf{m})$ needs some 'thickness' before going to the limit defining the conditional probability density. This limit crucially depends on the 'thickness', i.e., on the type of uncertainties the theory contains.
19. Note that taking the limit of $\vartheta(x, t)$ or of $\rho(x, t)$ for infinite variances we obtain $\mu(x, t)$, as we should.
20. The ratio $F(\mathbf{x}) = f(\mathbf{x})\, v(\mathbf{x})$ is what we refer to as *the volumetric probability* associated to the probability density $f(\mathbf{x})$. See Appendix A.
21. We take this example because typical misfit functions are adimensional (have no physical dimensions) but the argument has general validity.

22. As shown in Tarantola (1987), if $\boldsymbol{\gamma}_k$ is the direction of steepest ascent at point $\mathbf{m}_k$, i.e., $\gamma_k = \mathbf{C}_M\,\mathbf{F}_k^t\,\mathbf{C}_D^{-1}\,(\mathbf{f}_k - \mathbf{d}_{\mathrm{obs}}) + (\mathbf{m}_k - \mathbf{m}_{\mathrm{prior}})$, then, a local linearized approximation for the optimal ε_k gives

$$\frac{\varepsilon_k = \boldsymbol{\gamma}_k^t \mathbf{C}_M^{-1} \boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_k^t(\mathbf{F}_k^t\,\mathbf{C}_D^{-1}\,\mathbf{F}_k + \mathbf{C}_M^{-1})\,\boldsymbol{\gamma}_k}.$$

23. The 'best estimator' of $\tilde{\mathbf{C}}_M$ is

$$\tilde{\mathbf{C}}_M \approx \left(\mathbf{F}_k^t \mathbf{C}_D^{-1} \mathbf{F}_k + \mathbf{C}_M^{-1}\right)^{-1} . \qquad (119)$$

See, e.g., Tarantola (1987).

24. While a sensible estimation of the optimal values of the real positive quantities ε_k is crucial for the algorithm 111, they can in many usual circumstances be dropped from the algorithm 113.
25. The gray oval is the product of the probability density over the model space, representing the prior information, and the probability density over the data space representing the experimental results.
26. The gravitational field at point $\mathbf{x}_0$ generated by a distribution of volumetric mass $\rho(\mathbf{x})$ is given by

$$\mathbf{g}(\mathbf{x}_0) = \int dV(\mathbf{y})\, \frac{\mathbf{x}_0 - \mathbf{y}}{\|\mathbf{x}_0 - \mathbf{x}\|^3}\ \rho(\mathbf{x}) \ .$$

When the volumetric mass is constant inside some predefined (2D) volumes, as suggested in Figure 8, this gives

$$\mathbf{g}(\mathbf{x}_0) = \sum_A \sum_B \mathbf{G}^{\mathrm{A,B}}(\mathbf{x}_0)\, m^{\mathrm{A,B}} \ .$$

This is a strictly linear equation between data (the gravitational field at a given observation point) and the model parameters (the masses inside the volumes). Note that if instead of choosing as model parameters the total masses inside some predefined volumes one chooses the geometrical parameters defining the sizes of the volumes, then the gravity field is not a linear function of the parameters. More details can be found in Tarantola and Valette (1982b, page 229).

27. Using the 'orthogonal-limit' method described in Section 2.4.
28. The term 'a priori model' is an abuse of language. The correct term is 'mean a priori model'.

Editor's Note

Appendixes A–P are placed on the attached Handbook CD, under the directory \16Mosegaard. An introduction to probability concepts is given in Chapter 82, Statistical Principles for Seismologists, by Vere-Jones and Ogata. See also Chapter 52, Probing the Earth's Interior with Seismic Tomography, by Curtis and Snieder.