Inverse Problems = Quest for Information

Albert Tarantola and Bernard Valette

Institut de Physique du Globe de Paris, Université Pierre et Marie Curie, 4 place Jussieu, 75005 Paris, France

Note: The 1982 paper has been rewritten, to produce a searchable PDF file. This text is essentially identical to the original.

Abstract

We examine the general non-linear inverse problem with a finite number of parameters. In order to permit the incorporation of any a priori information about parameters and any distribution of data (not only of gaussian type) we propose to formulate the problem not using single quantities (such as bounds, means, etc.) but using probability density functions for data and parameters. We also want our formulation to allow for the incorporation of theoretical errors, i.e. non-exact theoretical relationships between data and parameters (due to discretization, or incomplete theoretical knowledge); to do that in a natural way we propose to define general theoretical relationships also as probability density functions. We show then that the inverse problem may be formulated as a problem of *combination of information:* the experimental information about data, the a priori information about parameters, and the theoretical information. With this approach, the general solution of the non-linear inverse problem is unique and consistent (solving the same problem, with the same data, but with a different system of parameters does not change the solution).

Key words: Information — Inverse problems — Pattern recognition — Probability

1 Introduction

Inverse problem theory was essentially developed in geophysics, to deal with largely underdetermined problems. The most important approaches to the solution of this kind of problem are well known to today's geophysicists (Backus and Gilbert 1967, 1968, 1970; Keilis-Borok and Yanovskaya 1967; Franklin 1970; Backus 1971; Jackson 1972; Wiggins 1972; Parker 1975; Rietsch 1977; Sabatier 1977).

The minimal constraints necessary for the formulation of an inverse problem are:

1. The formulation must be valid for linear as well as for strongly non linear problems.

2. The formulation must be valid for overdetermined as well as for underdetermined problems.

3. The formulation of the problem must be consistent with respect to a change of variables. (This is not the case with ordinary approaches: solving an inverse problem with a given parameter, e.g. a velocity v, leads to a solution v_0; solving the same problem with the same data but with another parameter, e.g. the slowness $n = 1/v$, leads to a solution n_0. There is no natural relation between v_0 and n_0 in ordinary approaches).

4. The formulation must be general enough to allow for general error distributions in the data (which may be not gaussian, asymmetric, multimodal, etc.).

5. The formulation must be general enough to allow for the formal incorporation of any a priori assumption (positivity constraints, smoothness, etc.).

6. The formulation must be general enough to incorporate theoretical errors in a natural way. As an example, in seismology, the theoretical error made by solving the forward travel time problem is often one order of magnitude *larger* that the experimental error of reading the arrival time on a seismogram. A coherent hypocenter computation must take into account experimental as well as theoretical errors. Theoretical errors may be due, for example, to the existence of some random parameters in the theory, or to theoretical simplifications, or to a wrong parameterization of the problem.

None of the approaches by previous authors satisfies this set of constraints. The main task of this paper is to demonstrate that all these constraints may be fulfilled when formulating the inverse problem using a simple extension of probability theory and information theory.

To do this we will limit ourselves to the study of systems which can be described with a *finite set of parameters.* This limitation is twofold: first, we will only be able to handle *quantitative* characteristics of systems. All qualitative aspects are beyond the scope of this paper. The second limitation is that to describe some of the characteristics of the systems we should employ functions rather than discrete parameters, as for example for the output of a continuously recording seismograph, or the velocity of seismic waves as a function of depth. In such cases we decide to sample the corresponding function.

The problem of adequate sampling is not a trivial one. For example, if the sampling interval of a seismogram is greater than the correlation length of the noise (seismic noise, finite pass-band filter, etc.), errors in data may be assumed to be independent; this will not be true when densifying the sampling. We explicitly assume in this paper that the discretization has been made carefully enough, so that densifying the sampling will only negligibly alter the results.

In the next section we will define precisely concepts such as parameter, probability density function, information, and combination of information; in Section 3 we discuss the concept of null information; in Section 4 we define the a priori information on a system; in Section 5 we define general theoreticaln relationships between data and parameters; finally in Section 6 we give the solution of the inverse problem. In Sections 7–10 we discuss this solution, we examine particular cases, and give a seismological illustration with actual data.

2 Parameters and Information

Let $\mathscr{S}$ be a *physical system* in a wide sense. By wide sense we mean that $\mathscr{S}$ consists of a physical system strictu senso, plus a family of measuring instruments and their outputs. We assume that $\mathscr{S}$ is a discrete system or that it has been discretized (for convenience of description or because the mathematical model which describes the physical system is discrete). In that case $\mathscr{S}$ can be described using a finite (perhaps large) set of parameters $\mathbf{X} = \{X_1, \dots, X_m\}$; any set of specific values of this set of parameters will be denoted $\mathbf{x} = \{x_1, \dots, x_m\}$. Each point $\mathbf{x}$ may be named a *model* of $\mathscr{S}$. The m-dimensional space $\mathscr{E}^m$ where the parameters $\mathbf{X}$ take their values may be named the *model space,* or the *parameter space.*

When a physical system $\mathscr{S}$ can be described by a set $\mathbf{X}$ of parameters, we say that $\mathscr{S}$ is *parametrizable.*

It should be noted that the parameterization of a system is not unique. We say that two parameterizations are *equivalent* if they are related by a bijection. Let

$$\mathbf{X} = \mathbf{X}(\mathbf{X}') \quad \mathbf{X}' = \mathbf{X}'(\mathbf{X}) \tag{2.1}$$

be two equivalent parameterizations of $\mathscr{S}$. We emphasize that equation (2.1) represent a transformation between mathematically equivalent parameters and that they *do not* represent any relationship between physically correlated parameters. An example of equivalent parameters is a velocity v and the corresponding slowness *defined* by $n = 1/v$. Let us remark that two equivalent parameterizations of $\mathscr{S}$ can also be seen as two different choices of *system of coordinates* in $\mathscr{E}^m$.

The degree of knowledge that we have about the values of the parameters of our system may range from total knowledge to total ignorance. A first postulate of this paper is that *any* state of knowledge on the values of $\mathbf{X}$ can be described using a *measure density function* $f(\mathbf{x})$; i.e. a real, positive, locally Lebesgue integrable function such that the positive measure defined by

$$m(A) = \int_A f(\mathbf{x})\, d\mathbf{x} \qquad A \subset \mathscr{E}^m \tag{2.2}$$

is absolutely continuous with respect to the Lebesgue measure defined over $\mathscr{E}^m$. The quantity $m(A)$ is named the *measure* of A. If $m(\mathscr{E}^m)$ is finite then $f(\mathbf{x})$ can be normalized in such a way that $m(\mathscr{E}^m) = 1$; in that case $f(\mathbf{x})$ is named a *probability density function,* $m(A)$ is then noted $p(A)$ and is named the *probability* of A.

All through this paper a measure density function, non normalized or non normalizable, will simply be named a density function.

Of course, the form of $f(\mathbf{x})$ depends on the chosen parameterization. Let $\mathbf{X}$ and $\mathbf{X}'$ be two equivalent parameterizations As ,we want the measure $m(A)$ to be invariant, it is easy to see that there exists a density function $f'(\mathbf{x}')$, which is related to $f(\mathbf{x})$ by the usual formula:

$$f'(\mathbf{x}') = f(\mathbf{x}) \cdot \left| \frac{\partial \mathbf{x}}{\partial \mathbf{x}'} \right| , \tag{2.3}$$

where the symbol $|\frac{\partial \mathbf{x}}{\partial \mathbf{x}'}|$ stands for the Jacobian of the transformation. (It never vanishes for equivalent parameterizations).

Let us define a particular density function $\mu(\mathbf{x})$ representing the state of total ignorance (Jaynes 1968; Rietsch 1977). Often the state of total ignorance will correspond to a uniform function $\mu(\mathbf{x}) = \text{const.}$, sometimes it will not, as discussed in Section 3. We will assume

$$\mu(\mathbf{x}) \neq 0 \tag{2.4}$$

everywhere in $\mathscr{E}^m$ (In fact this means that we restrict the space of parameters to the region not excluded by the state of total ignorance).

We should need a density function, rather than a probability density function when we are not able to define the *absolute* probability of a subset A, but we can define the *relative* probabilities of two subsets A and B. The most trivial example is when $f(\mathbf{x}) = \text{const.}$ and the space is not bounded.

Two density functions which differ only by a multiplicative constant will give the same *relative* probabilities, and all through this paper they will be considered identical:

$$f(\mathbf{x}) \equiv \text{const.}\, f(\mathbf{x}) \tag{2.5}$$

If the state of total ignorance corresponds to a probability density $\mu(\mathbf{x})$, then the content of *information* of any probability density $f(\mathbf{x})$ is defined by (Shannon 1948)

$$I(f;\mu) = \int f(\mathbf{x}) \operatorname{Log} \frac{f(\mathbf{x})}{\mu(\mathbf{x})}\, d\mathbf{x} . \tag{2.6}$$

This definition has the following properties, which are easily verified:

a) I is invariant with respect to a change of variables:

$$I(f;\mu) = I(f';\mu') . \tag{2.7}$$

(In Shannon's original definition of information for continuous variables the term $\mu(\mathbf{x})$ in equation 6 is missing, so that Shannon's definition is not invariant.)

b) Information cannot be negative:

$$I(f;\mu) \geqq 0 \tag{2.8}$$

c) the information of the state of total ignorance is null:

$$I(\mu;\mu) = 0 \tag{2.9}$$

the reciprocal being also true:

$$I(f;\mu) = 0 \qquad \Longrightarrow \qquad f = \mu. \tag{2.10}$$

We will say that each probability density (or, by extension, each density function) $f_i(\mathbf{x})$ represents a *state of information,* which will be noted s_i.

Let us now set up a problem which appears very often under different aspects. Its general formulation may be: Let $\mathbf{X}$ be a set a parameters describing some system $\mathscr{S}$. Let $\mu(\mathbf{x})$ be a density function representing the state of null information on the system. If we receive two pieces of information on our system, represented by the density functions $f_i(\mathbf{x})$ and $f_j(\mathbf{x})$, how do we combine f_i and f_j to obtain a density function $f(\mathbf{x})$ representing the final state of information?

We must first state *which* kind of combination we wish. To do this, let us first recall the way used in classical logic to define the combination of *logical propositions.* If p_i is a logical proposition, one defines its *value of truth* $v(p_i)$ by taking the values 1 or 0 when p_i is respectively certain or impossible (true or false). Let p_i and p_j be two logical propositions. It is usual to combine them in order to obtain new propositions, as for example by defining the *conjunction* of two propositions, $p_i \wedge p_j (p_i \textit{ and } p_j)$, or by defining the *disjunction* of two propositions $p_i \vee p_j$ (p_i *or* p_j), and so on. The usual way for defining the result of these combinations is by establishing their values of truth. For example, the conjunction $p_i \wedge p_j$ is defined by:

$$\left.\begin{array}{l} v(p_i) = 0 \\ \text{or} \\ v(p_j) = 0 \end{array}\right\} \Longleftrightarrow v\ p_i \wedge p_j\ = 0\ . \tag{2.11}$$

For our purposes, we need the *definition* of the *conjunction of two states of information*, $s_i \wedge s_j$. This definition must be the generalization to the concept of states of information of the properties of the conjunction of logical propositions.

We will see later that this definition will allow the solution of many seemingly different problems, in particular it contains the solution of the general inverse problem as it has been stated in the preceding section.

Let us note

$$f = f_i \wedge f_j \tag{2.12}$$

the operation which combines f_i and f_j (representing two states of information s_i and s_j) to obtain f (representing the conjunction $s = s_i \wedge s_j$). The definition must satisfy the following conditions:

a) $f_i \wedge f_j$ must be a density function. In particular, the content of information, of $f_i \wedge f_j$ must be invariant with respect to a change of parameters, i.e. equation (2.3) must be verified.

b) The operation must be commutative, i.e., for any f_i and f_j:

$$f_i \wedge f_j = f_j \wedge f_i\ . \tag{2.13}$$

c) The operation must be associative; i.e. for any f_i, f_j and f_k:

$$f_i \wedge\ \ f_j \wedge f_k\ \ =\ \ f_i \wedge f_j\ \ \wedge f_k \tag{2.14}$$

d) the conjunction of any state of information f_i with the null information μ must give f_i, i.e. must not result in any loss of information:

$$f_i \wedge \mu = f_i\ . \tag{2.15}$$

This equation means that μ is the neutral element for the operation.

e) The final condition corresponds to an extension of the defining property of the conjunction of logical propositions equation (2.11). For any measurable $A \subset \mathscr{E}^m$

$$\left.\begin{array}{l} \displaystyle\int_A f_i\, d\mathbf{x} = 0 \\ \text{or} \\ \displaystyle\int_A f_j\, d\mathbf{x} = 0 \end{array}\right\} \Longleftrightarrow \int_A f_i \wedge f_j\ d\mathbf{x} = 0\ , \tag{2.16}$$

which means that a necessary and sufficient condition for $f_i \wedge f_j$ to give a null probability to a subset A is that either f_i or f_j give a null probability for A.

This last condition implies that the measure engendered by $f_i \wedge f_j$ is absolutely continuous with respect to the measures engendered by f_i and f_j respectively. Using Nikodym's theorem it can be shown (Descombes 1972) that $f_i \wedge f_j$ may then necessarily be written in the form

$$f_i \wedge f_j = f_i \cdot f_j \cdot \Phi\ f_i, f_j\ \ , \tag{2.17}$$

where $\Phi(f_i, f_j)$ is a locally Lebesgue integrable, positive function. This last condition is strong, because it is valid everywhere in $\mathscr{E}^m$, in particular where f_i or f_j are null (otherwise this equation would be trivial).

The simplest choice for $\Phi(f_i, f_j)$ in order to satisfy conditions b) and c) is to take it as independent of f_i or f_j. Condition d) then imposes

$$\Phi\ f_i, f_j\ = \frac{1}{\mu}\ . \tag{2.18}$$

Condition a) is then automatically verified.

The above discussion suggests then the following definition:

Let s_i and s_j be two states of information represented respectively by the density functions f_i and f_j, let μ be a density function representing the state of null information. By definition, the *conjunction* of s_i and s_j, denoted $s = s_i \wedge s_j$ is a state of information represented by the density function $f(\mathbf{x})$ given by:

$$\boxed{f(\mathbf{x}) = \frac{f_i(\mathbf{x}) \cdot f_j(\mathbf{x})}{\mu(\mathbf{x})} \quad .} \tag{2.19}$$

The density function $f(\mathbf{x})$ is not necessarily normalizable, but except in some ad hoc examples, in most actual problems when one (or both) of the density function $f_i(\mathbf{x})$ or $f_j(\mathbf{x})$ is normalizable, the density function $f(\mathbf{x})$ is also normalizable, i.e. it is, in fact, a probability density.

In the following sections we will show that the definition (2.19) allows for a simple solution of general inverse problems. In the appendix we recall the definition of *marginal* probability densities, we show that the *conditional* p.d.f can be defined as a particular case of conjunction of states of information, and demonstrate the Bayes theorem (which is not used in this work because it is too restrictive for our purposes).

Let us emphasize that, given two states of information s_i and s_j on a system $\mathscr{S}$, the resulting state of information does not necessarily correspond to the conjunction $s_i \wedge s_j$. The conjunction, as defined above, must be used to combine two states of information only if these states of information have been obtained *independently,* as for example, for two independent physical measurements on a given set of parameters, or for combining experimental and theoretical information (see section 6).

Let us conclude this section by the remark that in our use of probability calculus we do not use concepts such as *random, variable, realization* of a random variable, *true value* of a parameter, and so on. Our density functions are interpreted in terms of *human knowledge,* rather than in terms of *statistical properties* of the Earth. Of course, we accept statistics, and we use the language of statistics when statistical computations are possible, but this is generally not the case in geophysical experiments.

3 The Null Information on a System

As the concept of null information is not straightforward, let us discuss it in some detail and start with some examples. Assume that our problem consists in the location of an earthquake focus from some set of data. Assume also that we are using Cartesian coordinates (X, Y, Z). The question is: which will be the density function $\mu(x, y, z)$ which is *least* informative on the location of the focus? The intuitive answer is that the least informative density function will be the one that assigns the same probability dP to all regions of equal volume dV:

$$dP = \text{const.}\, dV \ . \tag{3.1}$$

Since in Cartesian coordinates $dV = dx \cdot dy \cdot dz$, equation (2.2) gives the solution:

$$\mu(x, y.z) = \text{const.} \tag{3.2}$$

If instead of Cartesian coordinates we use spherical coordinates (R, Θ, Φ), the null information density function $\mu'(r, \theta, \phi)$ may be obtained from equation (3.1) writing the elementary volume dV in spherical coordinates, $dV = r^2 \cdot \sin\theta \cdot dr \cdot d\theta \cdot d\phi$ or from equation (3.2) by means of a change of variables. We arrive at

$$\mu(r, \theta, \phi) = \text{const} \cdot r^2 \cdot \sin\theta \tag{3.3}$$

which is far from a constant function. We see in this example that the density function representing the null information need not be constant.

We will now try to solve a less trivial question. Let $V = |\mathbf{V}|$ be some velocity. Could the null information density function $\mu(v) = \text{const.}$? To those who are tempted to answer *yes*, we ask another question. Let N be some slowness $(N = 1/V)$. Could the null information density function $\mu'(n) = \text{const.}$? Obviously, if $\mu(v)$ is constant, $\mu'(n)$ cannot be, and vice-versa.

To properly define the null information density function μ, we will follow Jaynes (1968), who suggested that suitable density functions are obtained under the condition of *invariance of form* of the function μ under the transformgroups which leave invariant the equations of physics (see also Rietsch, 1977). Clearly, the form of μ must be invariant under a change of space-time origin and under a change of space-time scale. To see its consequences let $\mathbf{O}$ and $\hat{\mathbf{O}}$ be two observers and let (X, Y, Z, T) and $(\hat{X}, \hat{Y}, \hat{Z}, \hat{T})$ be their coordinate system. The fact that observer $\hat{\mathbf{O}}$ has chosen a different space-time origin and scale is easily written in Cartesian coordinates:

$$\begin{aligned} &\hat{X} = X_0 + a \cdot X \\ &\hat{Y} = Y_0 + a \cdot Y \quad \hat{T} = T_0 + b \cdot T \\ &\hat{Z} = Z_0 + a \cdot Z \ , \end{aligned} \tag{3.4}$$

where a and b are constants. Thus, by the definition of velocity:

$$\hat{v} = \frac{|d\hat{r}|}{d\hat{t}} = \frac{a \cdot |dr|}{b \cdot dt} = c \cdot v \ , \tag{3.5}$$

where $c = a/b$ is a new constant. Let $\mu(v)$ be the null information density function for $\mathbf{O}$ and $\hat{\mu}(\hat{v})$ be the one of $\hat{\mathbf{O}}$. From equation (3.5) and (2.4) we must have:

$$\mu(v) = \hat{\mu}(\hat{v}) \cdot \frac{d\hat{v}}{dv} = c \cdot \hat{\mu}(c \cdot v) \ . \tag{3.6}$$

The invariance under transformations (3.4) will be realized if μ and $\hat{\mu}$ are *the same function,* that is:

$$\mu(w) = \hat{\mu}(w) \tag{3.7}$$

for all w.

From equations (3.6) and (3.7) it follows

$$\mu(v) = c \cdot \mu(c \cdot v) \ , \tag{3.8}$$

i.e.

$$\mu(v) = \frac{\text{const.}}{v}. \tag{3.9}$$

This result may appear puzzling to some. Let us ask which is the form of the null information density function for the slowness $N = 1/V$. We readily find

$$\mu'(n) = \mu(v) \cdot \frac{dv}{dn} = \frac{\text{const.}}{n} \ . \tag{3.10}$$

We see that the equivalent parameters V and N have null information density function of exactly the same form. In fact, it was in order to warrant this type of symmetry between all the powers of a parameter that Jeffreys (1939, 1957) suggested assigning to all continuous parameters X *known to be positive* a density function, representing the null information, of the form const./x .

Some formalisms of inverse problems attempt a definition of some probabilistic properties in parameters space (computation of standard deviations, etc.). We claim that these kind of problems cannot be consistently posed without explicitly stating the null information density function μ.

In most ordinary cases the choice

$$\mu = \text{const.} \tag{3.11}$$

will give reasonable results. Nevertheless we must emphasize that the solution of the same problem using a different set of parameters will be inconsistent with the choice of equation (3.11) for representing the state of null information in the new set of parameters, unless the change of parameters is linear.

4 Data and A Priori Information

Among the set of parameters $\mathbf{X}$ describing a system $\mathscr{S}$, the parameters describing the outputs of the measuring instrument are named *data* and written $\mathbf{D} = (D_1, \ldots, D_r)$. The rest of the parameters are then named parameters *strictu senso,* or, briefly, parameters, and are written $\mathbf{P} = (P_1, \ldots, P_s)$. If a partition of $\mathbf{X}$ into $\mathbf{X} = (\mathbf{D}, \mathbf{P})$ is made, then any density function on $\mathbf{X}$ may be equivalently written:

$$f(\mathbf{x}) = f(\mathbf{d}, \mathbf{p}) \ . \tag{4.1}$$

Let us consider a particular geophysical measurement, for example, the act of reading the arrival time of a particular phase on a seismogram. In the simplest case the seismologist puts all the information he has obtained from his measurement in the form of a given value, say t, and an "uncertainty", say σ_t. In more difficult cases, he may hesitate between two or more values. What he may do, more generaly, is to define, for each time interval Δt on the seismogram, the probability ΔP which he assigns to the arrival time t to be in the interval Δt. Doing this, he is putting the information which he obtains from his measurement into the form of a probability density $\rho(t) = \Delta P / \Delta t$. This probability density can be asymmetric, multimodal, etc. Extracting from this probability density a few estimators, such as mean or variance, would certainly lead to a loss of information, thus we have to take as an elementary datum the probability density $\rho(t)$ itself.

Let us now consider a non-directly measurable parameter P_α. Some examples of a priori information are:

a) We know only that P_α, is bounded by two values a $a \leqq p_\alpha \leqq b$. We will obviously represent this a priori information by a density function which is null outside the interval and which coincides with the null information density function inside the interval.

b) Inequality constraint $P_\alpha \leqq P_\beta$: we take a density function null for $p_\alpha > p_\beta$ and equal to the null information density function for $p_\alpha \leqq p_\beta$.

c) Some parameters $P_{\alpha+1}$, $P_{\alpha+2}, \ldots, P_\beta$ are spatially (or temporally) distributed, and we know that their variation is smooth. Accordingly, we will represent this a priori information by using a joint density function $\rho(p_{\alpha+1}, p_{\alpha+2}, \ldots, p_\beta)$ with the corresponding non-null assumed correlations (covariances).

d) We have some diffuse a priori information about some parameters. In that case we will define a priori density function with weak limits and large variances.

e) We have no a priori information at all. This a priori information is then represented by the null information function μ.

We see then that we may assume the existence of a density function

$$\rho(\mathbf{x}) = \rho(\mathbf{d}, \mathbf{p}) \tag{4.2}$$

named the a priori density function, representing both, the results of measurements and all a priori information on parameters.

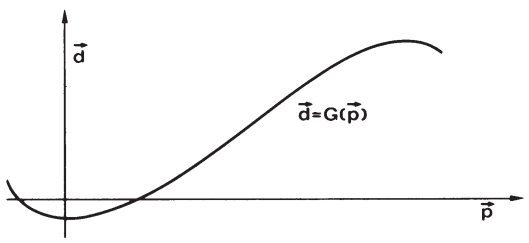

Figure 1: An exact theory, viewed as a functional relationship.

5 Theoretical Relationships

A theoretical relationship is usually viewed as a functional relation between the values of the parameters:

$$\mathbf{F}(\mathbf{x}) = \mathbf{F}(\mathbf{d},\mathbf{p}) = 0 \ . \tag{5.1}$$

Often the form (5.1) of a functional relationship may be simplified and may be written (see figure 1):

$$\mathbf{d} = \mathbf{G}(\mathbf{p}) \ . \tag{5.2}$$

This view is too restrictive. In most cases, even if the value $\mathbf{p}$ is given we are not able to *exactly* compute the corresponding value of $\mathbf{d}$, because our theory is incomplete, or because the theory contains some random parameters, or because we have roughly parametrized the system under study. In such cases, to be rigorous, we may exhibit not the value $\mathbf{d} = \mathbf{G}(\mathbf{p})$ but the probability density for $\mathbf{d}$, given $\mathbf{p}$, i.e. the conditional probability density (see figure 2)

$$\theta(\mathbf{d} \mid \mathbf{p}) \ . \tag{5.3}$$

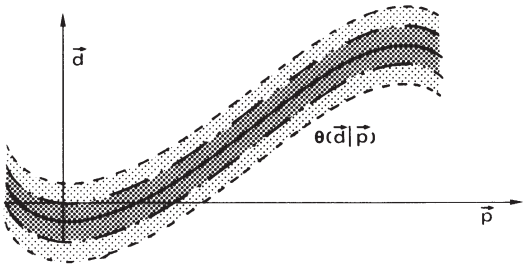

Figure 2: Putting "error-bars" on the theoretical relation $\mathbf{d} = \mathbf{G}(\mathbf{p})$.

We will see in Section 10 how to display such a conditional probability density for actual problems.

In all generality, we will assume that any theoretical relationship may be represented by a joint density function

$$\theta(\mathbf{x}) = \theta(\mathbf{d},\mathbf{p}) \ . \tag{5.4}$$

From the definition of conditional probability (see Appendix),

$$\theta(\mathbf{d},\mathbf{p}) = \theta(\mathbf{d} \mid \mathbf{p}) \cdot \theta_{\mathbf{p}}(\mathbf{p}) \ , \tag{5.5}$$

where $\theta_{\mathbf{p}}(\mathbf{p})$ is the marginal density function for $\mathbf{P}$. In the class of problems where the simplification (5.2) is used, the theory does not impose any constraint on $\mathbf{P}$ but only in $\mathbf{D}$. Equation (5.5) may then be rewritten

$$\theta(\mathbf{d},\mathbf{p}) = \theta(\mathbf{d} \mid \mathbf{p}) \cdot \mu_{\mathbf{p}}(\mathbf{p}) \ , \tag{5.6}$$

where $\mu_{\mathbf{p}}(\mathbf{p})$ is the null information density function.

The particular case of an exact theory (equation (5.2)) obviously corresponds to $\theta(\mathbf{d} \mid \mathbf{p}) = \delta(\mathbf{d} - \mathbf{G}(\mathbf{p}))$ where δ is the Dirac distribution. So, for an exact theory:

$$\theta(\mathbf{d},\mathbf{p}) = \delta(\mathbf{d} - \mathbf{G}(\mathbf{p})) \cdot \mu_{\mathbf{p}}(\mathbf{p}) \ . \tag{5.7}$$

In cases where a rigorous computation of $\theta(\mathbf{d} \mid \mathbf{p})$ cannot be made, but where we have an idea of the theoretical "errorbar" $\sigma_{\mathbf{T}}$, choices of $\theta(\mathbf{d},\mathbf{p})$ of a form similar to

$$\theta(\mathbf{d} \mid \mathbf{p}) = \text{const.} \exp \ -\frac{1}{2} \frac{\|\mathbf{d} - \mathbf{G}(\mathbf{p})\|^2}{\sigma_{\mathbf{T}}^2} \tag{5.8}$$

may be good enough to take into account this theoretical error.

In any case, we assume that theoretical relationships are in general represented by the joint density function of equation (5.4) which will be named the *theoretical* density function.

6 Statement and Solution of Inverse Problems

Let $\mathscr{S}$ be a physical system, and let $\mathbf{X}$ be a parametrization of $\mathscr{S}$. In Section 3 we have defined the density function $\mu(\mathbf{x})$ representing the state of null information on the system. In Section 4 we have defined the density function $\rho(\mathbf{x})$ representing all a priori information on the system, in particular the results of the measurements and a priori constraints on parameters. In Section 5 we have defined the density function $\theta(\mathbf{x})$ representing the theoretical relationships between parameters.

The conjunction of $\rho(\mathbf{x})$ and $\theta(\mathbf{x})$ gives a new state of information, which will be named the *a posteriori state of information.* The corresponding density function will be denoted $\sigma(\mathbf{x})$ and is given, using equation (2.19), by

$$\boxed{\sigma(\mathbf{x}) = \frac{\rho(\mathbf{x}) \cdot \theta(\mathbf{x})}{\mu(\mathbf{x})} \quad .} \tag{6.1}$$

To examine inverse problems we separate our set of parameters $\mathbf{X}$ into the subsets $\mathbf{X} = (\mathbf{D},\mathbf{P})$ representing data and parameters strictu sensu respectively. Equation (6.1) may then be rewritten

$$\sigma(\mathbf{d},\mathbf{p}) = \frac{\rho(\mathbf{d},\mathbf{p}) \cdot \theta(\mathbf{d},\mathbf{p})}{\mu(\mathbf{d},\mathbf{p})} \ . \tag{6.2}$$

From this equation we may compute the a posteriori marginal density functions:

$$\sigma_{\mathbf{d}}(\mathbf{d}) = \int \frac{\rho(\mathbf{d},\mathbf{p})\cdot\theta(\mathbf{d},\mathbf{p})}{\mu(\mathbf{d},\mathbf{p})}\, d\mathbf{p} \ , \tag{6.3}$$

$$\boxed{\sigma_{\mathbf{p}}(\mathbf{p}) = \int \frac{\rho(\mathbf{d},\mathbf{p})\cdot\theta(\mathbf{d},\mathbf{p})}{\mu(\mathbf{d},\mathbf{p})}\, d\mathbf{d}} \ . \tag{6.4}$$

Equation (6.4) performs the task of transferring to parameters, via theoretical correlations, the information contained in the data set. This is, by definition, the solution to an *inverse problem* (see figure 3).

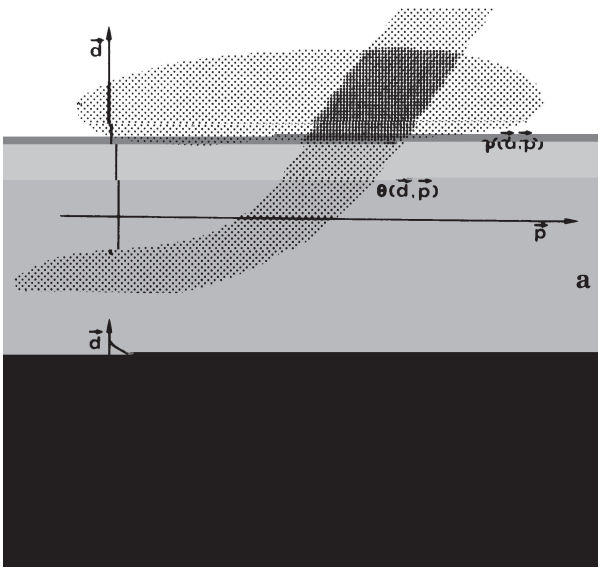

Figure 3: a) The theoretical model is often non exact (simplified, rough parameterization, etc.). We can then introduce the theoretical relationship between parameters as a density function $\theta(\mathbf{d},\mathbf{p})$ (see Section 5). b) The solution of the problem is then defined by $\sigma(\mathbf{x}) = \frac{\rho(\mathbf{x})\cdot\theta(\mathbf{x})}{\mu(\mathbf{x})}$. If the a priori density function contains small variances for data and great variances for parameters, the marginal density function $\sigma_{\mathbf{p}}(\mathbf{p})$ solves an "inverse problem". On the contrary, if in $\rho(\mathbf{x})$ the data have large variances and the parameters have small variances, $\sigma_{\mathbf{d}}(\mathbf{d})$ solves the "forward problem."

Equation (6.3) solves what could be named a generalized forward problem.

In most cases the a priori information on $\mathbf{D}$ is independent from the a priori information on $\mathbf{P}$,

$$\rho(\mathbf{d},\mathbf{p}) = \rho_{\mathbf{d}}(\mathbf{d})\cdot\rho_{\mathbf{p}}(\mathbf{p}) \tag{6.5}$$

and the theoretical density function is obtained in the form of a conditional density function (equation (5.6)):

$$\theta(\mathbf{d},\mathbf{p}) = \theta(\mathbf{d}\mid\mathbf{p})\cdot\mu_{\mathbf{p}}(\mathbf{p}) \ . \tag{6.6}$$

Equation (6.4) may then be simplified to

$$\boxed{\sigma_{\mathbf{p}}(\mathbf{p}) = \rho_{\mathbf{p}}(\mathbf{p})\cdot\int \frac{\rho_{\mathbf{d}}(\mathbf{d})\cdot\theta(\mathbf{d}\mid\mathbf{p})}{\mu_{\mathbf{d}}(\mathbf{d})}\, d\mathbf{d}} \ . \tag{6.7}$$

If, furthermore, the theoretical relationship may be considered as exact (i.e. we can write $\mathbf{d} = \mathbf{G}(\mathbf{p})$), then using equation (5.7).

$$\theta(\mathbf{d}\mid\mathbf{p}) = \delta(\mathbf{d}-\mathbf{G}(\mathbf{p})) \ . \tag{6.8}$$

Equation (6.7) may be easily integrated to:

$$\boxed{\sigma_{\mathbf{p}}(\mathbf{p}) = \rho_{\mathbf{p}}(\mathbf{p})\cdot\frac{\rho_{\mathbf{d}}(\mathbf{G}(\mathbf{p}))}{\mu_{\mathbf{d}}(\mathbf{G}(\mathbf{p}))}} \ . \tag{6.9}$$

This last equation solves the inverse problem for an exact, non-linear theory with arbitrary a priori constraints on parameters ($\rho_{\mathbf{p}}$), and an arbitrary probabilistic distribution of data ($\rho_{\mathbf{d}}$).

Returning to the general solution (6.4) let us answer the question of displaying the information contained in $\sigma_{\mathbf{p}}(\mathbf{p})$. If we are interested in a particular parameter, say P_1, all the information on P_1 is contained in the marginal density function (equation (2.5)):

$$\sigma_1(p_1) = \int \sigma_{\mathbf{p}}(\mathbf{p})\cdot dp_2\cdot dp_3\cdot\ldots\cdot dp_s \ . \tag{6.10}$$

As far as we are interested in the parameter P_1 and not in the correlations between this and other parameters, $\sigma_1(p_1)$ exhibits *all* the available information about P_1. For example, from $\sigma_1(p_1)$ we can precisely answer questions such as the probability that P_1 lies between two values. Alternatively, from $\sigma_1(p_1)$ it is possible to extract the mean value, the median value, the maximum-likelihood value, the standard deviation, the mean deviation, or any estimator we need.

Let us remark that it is possible to compute from the general solution $\sigma_{\mathbf{p}}(\mathbf{p})$ the *a posteriori* mathematical expectation:

$$E(\mathbf{p}) = \int \mathbf{p}\cdot\sigma_{\mathbf{p}}(\mathbf{p})\cdot d\mathbf{p} \tag{6.11}$$

or the a posteriori covariance matrix:

$$\begin{aligned}\mathbf{C} &= E\left\{\left(\mathbf{p}-E(\mathbf{p})\right)\cdot\left(\mathbf{p}-E(\mathbf{p})\right)^{\mathbf{T}}\right\} \\ &= \int \left(\mathbf{p}-E(\mathbf{p})\right)\cdot\left(\mathbf{p}-E(\mathbf{p})\right)^{\mathbf{T}}\sigma_{\mathbf{p}}(\mathbf{p})d\mathbf{p} \\ &= \int \mathbf{p}\cdot\mathbf{p}^{\mathbf{T}}\sigma_{\mathbf{p}}(\mathbf{p})d\mathbf{p} - E(\mathbf{p})\cdot E(\mathbf{p})^{\mathbf{T}} \ .\end{aligned} \tag{6.12}$$

Estimators such as $E(\mathbf{P})$ and $\mathbf{C}$ are similar to what is obtained in traditional approaches to inverse problems, but here they can be obtained without any linear approximation.

7 Existence, Uniqueness, Consistency, Robustness, Resolution

In inverse problem theory, it is not always possible to prove the existence or the uniqueness of the solution. With our approach, the existence of the solution is merely the existence of the a posteriori density function $\sigma(\mathbf{x})$ of equation (6.1) and results trivially from the assumption of the existence of $\rho(\mathbf{x})$ and $\theta(\mathbf{x})$. Of course we can obtain a solution $\sigma(\mathbf{x})$ with some pathologies (non-normalizable, infinite variance, not unique maximum likelihood point, etc.), but the solution is *the* a posteriori density function, with all the pathologies it may present.

Consistency is warranted because equation (6.1) is consistent, i.e. the function $\sigma(\mathbf{x})$ *is* a density function. Let us suppose again that we use velocity as a parameter and obtain as solution the density function $\sigma(v)$, then using the same data in the same problem (with the same discretization) but using slowness as parameter we obtain as a solution the density function $\sigma'(n)$. Since equation (6.1) is consistent, $\sigma'(n)$ will be related to $\sigma(v)$ by the usual formula (2.3) $\sigma'(n) = \sigma(v) \cdot |\frac{dv}{dn}|$: $\sigma'(n)$ and $\sigma(v)$ represent exactly the same state of information.

In those approaches where all the information contained in the data is condensed into the form of central estimators (mean, median, etc.), the notion of robustness must be carefully examined if we suspect that there may be blunders in the data set (Clearbout and Muir, 1973). In our approach, the suspicion of the presence of blunders in a data set may be introduced using long-tailed density functions in $\rho_{\mathbf{d}}(\mathbf{d})$, decreasing much more slowly than gaussian functions as, for example, exponential functions. Our experience shows that with such long-tailed functions, the solution $\sigma_{\mathbf{p}}(\mathbf{p})$ is rather insensitive to one blunder.

The concept of resolution must be considered under two different aspects: to what extent a given parameter has been "resolved" by the data? and what is the "spatial resolution" attained with our data set?

For the first aspect, let us consider a parameter P_i whose value does not influence the values of the data. This means that the theoretical density function $\theta(\mathbf{d}, \mathbf{p})$ does not depend on P_i. Even in this case we can obtain a certain amount of information on this parameter, if the other parameters are resolved by the data, and if the a priori density function $\rho_{\mathbf{p}}(\mathbf{p})$ introduces some correlation between parameters. Furthermore, let us assume that no correlation is introduced by $\rho_{\mathbf{p}}(\mathbf{p})$ between P_i and the other parameters. This is the worst case of non-resolution we can imagine for a parameter. Under these assumptions equation (6.7) can be written

$$\sigma_{\mathbf{p}}(\mathbf{p}) = \rho_i(P_i) \cdot \rho_{\mathbf{q}}(\mathbf{q}) \cdot \int \rho(\mathbf{d}) \cdot \frac{\theta(\mathbf{d} \mid \mathbf{q})}{\mu_{\mathbf{d}}(\mathbf{d})} d\mathbf{d} \ , \qquad (7.1)$$

where $\mathbf{q}$ is the vector $(p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_s)$. After integration over the set of $\mathbf{q}$, we find (dropping the multiplicative constant):

$$\sigma_i \ p_i \ = \rho_i \ p_i \ \ , \qquad (7.2)$$

which means that for a completely unresolved parameter, the a posteriori marginal density function equals the a priori one. The more $\sigma_i(p_i)$ differs from $\rho_i(p_i)$, the more the parameter P_i has been resolved by the data set.

The concept of spatial resolution applies to a different problem: Assume that $P_i, \ldots, P_j$ form a set of parameters spatially (or temporally) distributed as for example when the parameters represent the seismic velocities of successive geologic layers (or values from the sampling of some continuous geophysical record). Assume that we are not interested in obtaining the a posteriori density function for each parameter, but only the a posteriori mean values, as given by equation (6.11). There are two reasons for $E(\mathbf{P})$ to be a smooth vector (i.e. to have small variations between consecutive values $E(P_i)$ and $E(P_{i+1})$. The first reason may be the type of data used; it is well known for instance that long period surface wave data only give a smoothed vision of the Earth. The second reason for obtaining a smoothed solution may simply be that we decide a priori to impose such smoothness introducing non null a priori covariances in $\rho_{\mathbf{p}}(\mathbf{p})$ (see section 4).

This question of spatial resolution has been clearly pointed out, and extensively studied by Backus and Gilbert (1970). From our point of view, this problem must be solved by studying the a posteriori correlations between parameters. From equation (6.12):

$$C_{ij} = \int p_i \cdot p_j \cdot \sigma \ p_i, p_j \ \cdot d\, p_i \cdot d\, p_j - E \ p_i \ \cdot E \ p_j \ . \qquad (7.3)$$

If, for a given j, we plot the "curve" C_{ij} versus i we simultaneously obtain information on the a posteriori variance of the parameter (C_{ii}) and the spatial resolution (the length of correlation) (see figure 4).

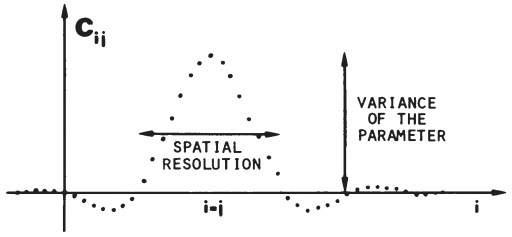

Figure 4: Rows (or columns) of the a posteriori covariance matrix showing both the length of spatial resolution and the a posteriori variance of the parameter.

In applications of Backus and Gilbert's point of view on inverse problems it is usual to study the *trade-off* be-

tween variance and resolution in order to choose the desired solution. In our approach, such a trade-off also exists: modifying the a priori variances or a priori correlations of parameters (in $\rho(\mathbf{p})$) results in a change of the a posteriori variances and resolutions. But, in our opinion, the a priori information on the values of parameters, as contained in $\rho(\mathbf{p})$, must not be stated in order to obtain a pleasant solution, but in order to closely correspond to the actual a priori information.

8 Computational Aspects

For linear problems, all the integrations of section 6 can often be performed analytically, and the most general solution can sometimes be reached easily (see for example the next section). Linear procedures may of course be used to obtain adequate approximations for the solution of weakly nonlinear problems.

For non-linear problems, the solution is less straightforward. Often the integrations in equation (6.4) or (6.7) can be performed analytically, no matter what the degree of nonlinearity (see for example section 10). The computation of the density of probability $\sigma_{\mathbf{P}}(\mathbf{p})$ at a given point $\mathbf{p}$ then involves mainly the solution of a forward problem. If the number of parameters is small we can then explicitly compute the marginal probability density for each one of the parameters, using agrid in the parameter space ordinary methods of numerical integration. If the number of parameters is great, Monte-Carlo methods of numerical integration should be used. The possibility of conveniently solving non-linear inverse problems will then depend on the possibility of solving the forward problem a large enough number of times. Let us remark that if we are not able to compute the marginal probability density for each one of the parameters of interest, we can limit ourselves to the computation of mean values and covariances (equations (6.11) and (6.12).

In problems where the solution of the forward problem is so costly that either the explicit computation of the density of probability in the parameter space and the computation of mean values and covariances cannot be performed in a reasonable computer time, we suggest to restrict the problem to the search of the maximum likelihood point in the parameter space (point at which the density of probability is maximum). This computation is often very easy to perform, and classical methods can be used for particular assumptions about the form of the probability densities representing experimental data and a priori assumptions on parameters. For example, it is easy to see that with gaussian assumptions, the search of the maximum likelihood point simply becomes a classical leastsquares problem (Tarantola and Valette, 1982). With exponential assumptions, the search of the maxima likelihood point becomes a L^1-norm problem (which shows that the exponential assumption gives a result more *robust* than the gaussian one). With the use of step functions, the a posteriori probability density in the parameter space is constant inside a given bounded domain. The point of this domain the maximizes some function of the parameters can be reached using the linear (or non-linear) programming techniques. We see thus that ordinary methods for solving parameterized inverse problems can be deduced as particular cases of our approach, and we want to emphasize that such methods should only be used when the explicit computation of the density of probability in the parameter space or the non-linear computation of mean values and covariances would be too much time consuming.

9 Gaussian Case

Since the gaussian linear problem is widely used, we will show how the usual formulae may be derived from our results.

By gaussian problem we mean that the a priori density function has a gaussian form for all parameters:

$$\rho(\mathbf{x}) = \exp\ -\frac{1}{2}\ \mathbf{x}-\mathbf{x}_0\ ^{\mathbf{T}}\cdot\mathbf{C}_0^{-1}\cdot\ \mathbf{x}-\mathbf{x}_0\ \quad , \tag{9.1}$$

where $\mathbf{x}_0$ is the a priori expected value and $\mathbf{C}_0$, is the a priori covariance matrix.

By the linear problem we mean that if the theory may be assumed to be exact, the theoretical relationship between parameters takes the general linear form:

$$\mathbf{F}\cdot\mathbf{x} = 0\ . \tag{9.2}$$

On the other hand, if theroretical errors may not be neglected, we assume that the theoretical density function also has a gaussian form:

$$\theta(\mathbf{x}) = \exp\ -\frac{1}{2}(\mathbf{F}\cdot\mathbf{x})^{\mathbf{T}}\cdot\mathbf{C}_{\mathbf{T}}^{-1}\cdot(\mathbf{F}\cdot\mathbf{x})\quad , \tag{9.3}$$

where the covariance matrix $\mathbf{C_T}$ describes theoretical errors (and tends to vanish for an exact theory).

We finally assume that *for the parameters chosen,* the null information is represented by a constant function:

$$\mu(\mathbf{x}) = \text{const.} \tag{9.4}$$

The a posteriori density function (equation (6.1) is then given by:

$$\begin{aligned}\sigma(\mathbf{x}) &= \rho(\mathbf{x})\cdot\theta(\mathbf{x})\\ &= \exp\ -\frac{1}{2}\Big[\ \mathbf{x}-\mathbf{x}_0\ ^{\mathbf{T}}\cdot\mathbf{C}_0^{-1}\cdot\ \mathbf{x}-\mathbf{x}_0\\ &\qquad\qquad +\mathbf{x}^{\mathbf{T}}\cdot\mathbf{F}^{\mathbf{T}}\cdot\mathbf{C}_{\mathbf{T}}^{-1}\cdot\mathbf{F}\cdot\mathbf{x}\Big]\end{aligned} \tag{9.5}$$

and after some matrix manipulations, (see appendix) we obtain

$$\sigma(\mathbf{x}) = \exp\ -\frac{1}{2}\ \mathbf{x}-\mathbf{x}_*\ ^{\mathbf{T}}\cdot\mathbf{C}_*^{-1}\cdot\ \mathbf{x}-\mathbf{x}_*\ \quad , \tag{9.6}$$

where

$$\mathbf{x}_* = \mathbf{P}\cdot\mathbf{x}_0, \tag{9.7}$$

$$\mathbf{C}_* = \mathbf{P}\cdot\mathbf{C}_0 \tag{9.8}$$

and

$$\begin{aligned}&\mathbf{P} = \mathbf{I}-\mathbf{Q}\\ &\mathbf{Q} = \mathbf{C}_0\cdot\mathbf{F}^{\mathbf{T}}\cdot\ \mathbf{F}\cdot\mathbf{C}_0\cdot\mathbf{F}^{\mathbf{T}}+\mathbf{C}_{\mathbf{T}}\ ^{-1}\cdot\mathbf{F}\ .\end{aligned} \tag{9.9}$$

Equation (9.6) shows that the a posteriori density function is gaussian, centered in $\mathbf{x}_*$ and with the covariance matrix $\mathbf{C}_*$.

If theoretical errors may be neglected, i.e. if equation (9.2) holds, we just drop the term $C_{\mathbf{T}}$ in equation (9.9) to obtain the corresponding solution.

To compare our results to those published in the literature, let us assume that the separation of $\mathbf{X}$ into the sets $\mathbf{D}$ and $\mathbf{P}$ is made:

$$\begin{aligned}\mathbf{x} &= \begin{matrix}\mathbf{d}\\ \mathbf{p}\end{matrix} \qquad \mathbf{x}_0 = \begin{matrix}\mathbf{d}_0\\ \mathbf{p}_0\end{matrix}\\ \mathbf{x}_* &= \begin{matrix}\mathbf{d}_*\\ \mathbf{p}_*\end{matrix} \qquad \mathbf{C}_0 = \begin{matrix}\mathbf{C}_{dd} & \mathbf{C}_{dp}\\ \mathbf{C}_{pd} & \mathbf{C}_{pp}\end{matrix}\ .\end{aligned} \tag{9.10}$$

We also assume that equation (9.2) simplifies to:

$$\mathbf{F}\cdot\mathbf{x} = [\mathbf{I}-\mathbf{G}]\cdot\ \begin{matrix}\mathbf{d}\\ \mathbf{p}\end{matrix}\ = \mathbf{d}-\mathbf{G}\cdot\mathbf{p} = 0\ . \tag{9.11}$$

Substituting equations (9.10), (9.11) in equations (9.7), (9.8), (9.9) we obtain the solution published by Franklin (1970) for the parametrized problem, which was obtained using the classical least squares approach. Our equations (9.7), (9.8), (9.9) are more compact than those of *Franklin* because we use the parameter space $\mathscr{E}^m$, and more general because we allow theoretical errors $\mathbf{C}_{\mathbf{T}}$.

Let us emphasize that in traditional approaches $\mathbf{x}_*$, is interpreted as *the best estimator of the "true" solution* and $\mathbf{C}_*$ is interpreted as the covariance matrix of the estimator. Our approach demonstrates that the a posteriori density function is gaussian, and that $\mathbf{x}_*$ and $\mathbf{C}_*$ are, respectively, the center and the disnersion of the density function.

The results shown here only apply to the linear least squares problem. For the non-linear problem, the reader should refer to Tarantola and Valette (1982).

station	x	y	z	t	σ_t
1	49.58	9.54	−0.80	13.35	0.02
2	48.07	7.74	−0.80	13.30	0.02
3	49.67	4.22	−1.50	13.79	0.02
4	52.34	14.37	−0.30	13.70	0.02
5	43.17	5.70	−0.80	13.90	0.02
6	46.79	17.87	−0.30	14.35	0.02
7	48.82	−1.87	−0.90	14.20	0.10
8	33.31	8.72	−0.60	15.41	0.02
9	23.10	9.54	−0.50	17.09	0.02
10	13.24	15.30	−0.40	19.00	0.10
11	−2.19	11.41	−1.20	18.73	0.02

Table 1: Coordinates of stations (km), arrival times and errors (s)

10 Example with Actual Data: Hypocenter Location

The data for a hypocenter location are the arrival times of phases at stations. The basic unknowns of the problem are the spatiotemporal coordinates of the focus. Some of the parameters which may be relevant to the problem are: the coordinates of seismic stations, the parameters describing the velocity model, etc. We will assume that the coordinates of the stations are accurately enough known to treat them as constants and not as parameters. The parameters describing the velocity model would be taken into account if we were performing a simultaneous inversion for hypocenter location and velocity determination, but this is not the case in this simple illustration.

In the example treated below we will then consider the parameters of the velocity model as constants, and we will assume that the only effect of our imprecise knowledge of the medium is not to allow an exact theoretical computation of the arrival times at the stations from a known source. Let

$$\mathbf{t} = \mathbf{g}(X,Y,Z,T) \tag{10.1}$$

be the theoretical (and not exact) relationship between arrival times and the spatio-temporal coordinates of the focus, derived from the wave propagation theory and the velocity model. Let $\mathbf{C}_{\mathbf{T}}$ be a covariance matrix which is a reasonable estimation of the errors made when theoretically computing the arrival time at stations from a source at (X,Y,Z). If we assume that the theoretical errors are gaussian, the theoretical relationship between data and parameters will be written

$$\theta(\mathbf{t}\mid X,Y,Z,T) = \exp\ -\frac{1}{2}\ \mathbf{t}-\mathbf{g}(X,Y,Z,T)\ ^{\mathbf{T}}$$

$$\cdot \mathbf{C_T^{-1}} \cdot \ \mathbf{t} - \mathbf{g}(X,Y,Z,T) \quad , \tag{10.2}$$

which correspond to equation (5.8).

The next simplifying hypothesis is to assume that our data possess a gaussian structure. Let $\mathbf{t}_0$ be their vector of mean values and $\mathbf{C_t}$, their covariance matrix:

$$\rho(\mathbf{t}) = \exp \quad -\frac{1}{2} \ \mathbf{t} - \mathbf{t}_0 \ ^{\mathbf{T}} \cdot \mathbf{C_t^{-1}} \cdot \ \mathbf{t} - \mathbf{t}_0 \quad . \tag{10.3}$$

As all our data and parameters consist in Cartesian spacetime coordinates, the null information function is constant and need not be considered (see section 3).

The a posteriori density function for parameters is directly given by equation (6.7) and after analytical integration we obtain (see appendix):

$$\begin{aligned} &\sigma(X,Y,Z,T) = \rho(X,Y,Z,T) \cdot \\ &\cdot \exp \quad -\frac{1}{2} \ \mathbf{t}_0 - \mathbf{g}(X,Y,Z,T) \ ^{\mathbf{T}} \cdot \ \mathbf{C_t} + \mathbf{C_T} \ ^{-1} \cdot \\ &\cdot \ \mathbf{t}_0 - \mathbf{g}(X,Y,Z,T) \quad . \end{aligned} \tag{10.4}$$

The a posteriori density function (10.4) gives the general solution for the problem of spatio-temporal hypocenter location in the case of gaussian data. We emphasize that this solution does not contain any "linear approximation".

We are sometimes interested in the *spatial* location of the quake focus, and not in its *temporal* location. The density function for the spatial coordinates is obtained, of course, by the marginal density function

$$\sigma(X,Y,Z) = \int_{-\infty}^{+\infty} \sigma(X,Y,Z,T)dT \ , \tag{10.5}$$

where we integrate over the range of the origin time T.

Classical least squares computations of hypocenter are based on the maximization of $\sigma(X,Y,Z,T)$. It is clear that if we are only interested in the spatial location we must maximize $\sigma(X,Y,Z)$ given by (10.5) instead of maximizing $\sigma(X,Y,Z,T)$. Let us show how the integration in (10.5) can be performed analytically.

We will assume that while we may sometimes have a priori information about the spatial location of the focus (from tectonic arguments, or from past experience in the region, etc.) it is generally impossible to have a priori information (independent from the data) about the origin time T. We will thus assume an a priori density function uniform on T,

$$\rho(X,Y,Z,T) = \rho(T) \cdot \rho(X,Y,Z) = \rho(X,Y,Z) \ . \tag{10.6}$$

The computed *arrival time* at a station i, $g_i(X,Y,Z,T)$ can be written

$$g_i(X,Y,Z,T) = h_i(X,Y,Z) + T \ , \tag{10.7}$$

where h_i is the *travel time* between the point (X,Y,Z) and the station i.

With (10.6) and (10.7), equation (10.5) can be integrated (see appendix) and gives

$$\begin{aligned} &\sigma(X,Y,Z) = K \cdot \rho(X,Y,Z) \\ &\cdot \exp \quad -\frac{1}{2} \ \tilde{\mathbf{t}}_0 - \tilde{\mathbf{h}}(X,Y,Z) \ ^{\mathbf{T}} \\ &\qquad \cdot \mathbf{P} \cdot \ \tilde{\mathbf{t}}_0 - \tilde{\mathbf{h}}(X,Y,Z) \quad . \end{aligned} \tag{10.8}$$

Here

$$\mathbf{P} = (\mathbf{C_t} + \mathbf{C_T})^{-1} \tag{10.9}$$

is a "weight matrix",

$$p_i = \sum_j P_{ij} \tag{10.10}$$

are "weights", and

$$K = \sum_i p_i = \sum_{ij} P_{ij}. \tag{10.11}$$

Moreover, $\tilde{t}_0^i$ is the observed arrival time minus the weighted mean of observed arrival times,

$$\tilde{t}_0^i = t_0^i - \frac{\sum_j p_j t_0^j}{\sum_j p_j} \tag{10.12}$$

and $\tilde{h}^i(X,Y,Z)$ is the computed travel time minus the weighted mean of computed travel times

$$\tilde{h}^i = h^i - \frac{\sum_j p_j h^j}{\sum_j p_j}. \tag{10.13}$$

(Note that $\mathbf{C_T}$ may depend on (X,Y,Z) and therefore P_{ij}, p_i, and K also.)

Equation (10.8) gives the general solution for the *spatial* location of a quake focus in the Gaussian case.

Table 1 shows the observed arrival times and their standard deviation. We have assumed that the theoretical errors are of the form :

$$\mathbf{C_T}(X,Y,Z)_{\ ij} = \sigma_{\mathbf{T}}^2 \cdot \exp \quad \left(-\frac{1}{2} \frac{D_{ij}^2}{\Delta^2} \right) \quad , \tag{10.14}$$

where D_{ij} is the distance between the station i and the station j, $\sigma_{\mathbf{T}}$ is some theoretical error, and Δ is the correlation length of errors (the wavelength or the length of lateral heterogeneities of the medium). By comparison of the layered model of velocities for the Western Pyrenees (Gagnepain et al., 1980) with data from refraction profiles (Gallart, 1980) we have chosen $\sigma_{\mathbf{T}} = 0.2$ s and $\Delta = 0.1$ km.

We also assumed that no a priori information is known about the epicenter, but that we know that the depth of the hypocenter is greater than -0.5 km (the mean topography):

$$\rho(X,Y,Z) = \rho(Z) = \begin{cases} 1 & \text{if } Z \geqq -0.5 \text{ km} \\ 0 & \text{if } Z < -0.5 \text{ km} . \end{cases} \tag{10.15}$$

We have then computed numerically from (10.8) the a posteriori marginal density functions for the epicenter and for the depth:

$$\sigma(X,Y) = \int_{-0.5\text{km}}^{+\infty} \sigma(X,Y,Z)dZ, \tag{10.16}$$

$$\sigma(Z) = \int_{-\infty}^{+\infty} dX \int_{-\infty}^{+\infty} dY \sigma(X,Y,Z) . \tag{10.17}$$

The results are shown in figures 5 and 6.

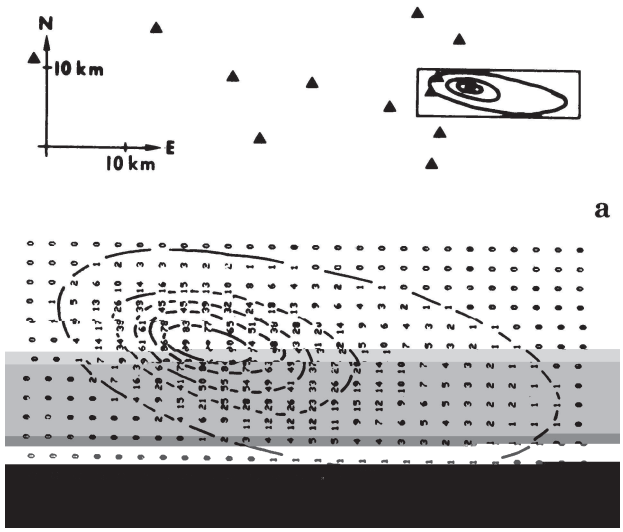

Figure 5: a. and b. Results of the inverse problem of hypocenter computation. a. shows the position of the stations and the probability density obtained for the epicentral coordinates. b. shows the computer output. Curves are visual interpolations.

We have also computed mean values and variances. The corresponding results are, in the local frame of figure 5:

$$\begin{aligned} E(X) &= 51.7 \text{ km} \quad & \sigma_X &= 1.5 \text{ km} \\ E(Y) &= 7.8 \text{ km} \quad & \sigma_Y &= 0.9 \text{ km} \\ E(Z) &= 5.6 \text{ km} \quad & \sigma_Z &= 2.1 \text{ km} \quad . \end{aligned}$$

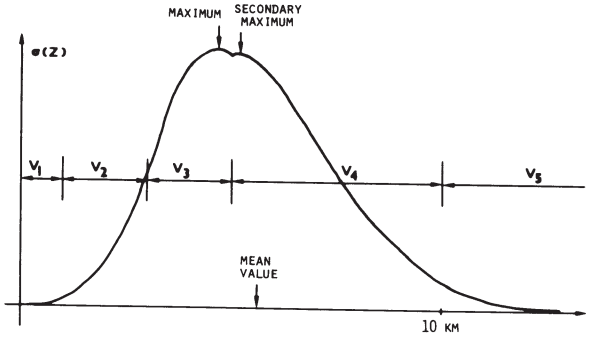

Figure 6: The probability density for depth. The layered velocity model is also shown. Note the existence of a secondary maximum likelihood point.

We wish to make the following remarks about these results.

First, they have been obtained exactly without the use of linear approximations. We have used numerical integration instead of matrix algebra and the computation of partial derivatives. The results shown in figures 5 and 6 represent the most general knowledge which can be obtained for the hypocenter coordinates from the arrival times, from the given velocity model, and from the given theoretical model (of wave propagation).

Since the velocity model is discontinuous in Z the a posteriori density functions have discontinuities in slope, as it is clearly seen in figure 6 for $\sigma(Z)$ To the extent that the discontinuities in the velocity model are artificial, the discontinuities of slope are of course also artificial. From figure 6 it is easy to visualize some of the problems which may affect the maximum likelihood approach. If the discontinuity of slope is similar to the one at 5 km depth, we will have secondary maxima. We can also have a discontinuity of slope of the type drawn in figure 7. In that case, algorithms searching for the maximum likelihood point will oscillate arount the point of slope discontinuity, leading to the well-known artificial situation in which hypocenters accumulate at the interface between layers of constant velocity.

11 Conclusion

Our informational approach to probability calculus allows us to formulate inverse problems in such a way that all necessary constraints (see Section 1) are satisfied. Essentially, we propose to work with the probability densities for parameters rather than with central estimators, as it is usually done.

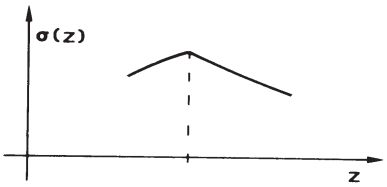

Figure 7: Example of discontinuity of slope leading to oscillations in maximum likelihood algorithms. The effect will be an artificial accumulation at the interfaces between layers

The general solution of inverse problems is expressed by the simple formula (6.1). We emphasize that inverse problems cannot be correctly stated until *the three* density functions $\rho(\mathbf{x})$ (data and a priori information about parameters), $\theta(\mathbf{x})$ (theory and theoretical errors), and $\mu(\mathbf{x})$ (null information) have been precisely defined.

We have demonstrated that the ideas developed in this paper give new insights into the oldest and best know inverse problem in geophysics: the hypocenter location. Of course out theory also applies to more difficult inverse problems. The only practical limitation comes from problems where the solution of the *forward* problem is very time-consuming *and* the number of parameters is high.

Acknowledgements. We would like to thank our colleagues of the I.P.G. for very helpful discussions, and Professor G. Jobert, Dr. A. Nercessian, T. Van der Pyl, Dr. J.C. Houard, Dr. Piednoir, and Professor C.J. Allegre for their suggestions. This work has been partially supported by the "Recherche Cooperative sur Programme no. 264, Etude Interdisciplinaire des Problemes Inverses". Contribution I.P.G. no. 363.

Appendix

Some Remarks on Probability Calculus

Let $\mathbf{X}$ be a parametrization of a physical system $\mathscr{S}$ and let $\mathbf{X_I} = \{X_1, \dots, X_r\}$ and $\mathbf{X_{II}} = \{X_{r+1}, \dots, X_m\}$ be a partition of $\mathbf{X}$. For any probability density $f(\mathbf{x}) = f(\mathbf{x_I}, \mathbf{x_{II}})$ we can define the *marginal* probability density

$$f_{\mathbf{I}}(\mathbf{x_I}) = \int f(\mathbf{x_I}, \mathbf{x_{II}}) d\mathbf{x_{II}} \ . \tag{A.1}$$

The interpretation of this definition is as follows: if we admit that $f(\mathbf{x_I}, \mathbf{x_{II}})$ represents all the knowledge that we possess on the whole set of parameters and if we disregard the parameters $\mathbf{X_{II}}$, then all the information on $\mathbf{X_I}$ is contained in $f_{\mathbf{I}}(\mathbf{x_I})$.

The conditional probability density for $\mathbf{X_I}$, given $\mathbf{X_{II}} = \mathbf{x^0_{II}}$ may be defined, in our approach, as the *conjunction* of a general state of information (represented by a probability density $f_i(\mathbf{x}) = f_i(\mathbf{x_I}, \mathbf{x_{II}})$) with the information $\mathbf{X_{II}} = \mathbf{x^0_{II}}$.

The information $\mathbf{X_{II}} = \mathbf{x^0_{II}}$ clearly corresponds to the probability density

$$f_j\ \mathbf{x_I}, \mathbf{x_{II}}\ = \mu_{\mathbf{I}} = big(\mathbf{x_I}\ \cdot \delta\ \mathbf{x_{II}} - \mathbf{x^0_{II}} \tag{A.2}$$

because f_j does not contain information on $\mathbf{X_I}$ and gives null probability for all values of $\mathbf{X_{II}}$ different from $\mathbf{x^0_{II}}$. $\mu(\mathbf{x_I})$ represents the null information on $\mathbf{X_I}$. Admitting that null informations are independent: $\mu(\mathbf{x_I}, \mathbf{x_{II}}) = \mu_I(\mathbf{x}_{II}) \cdot \mu_{\mathbf{II}}(\mathbf{x_{II}})$ and using equation (2.19) we obtain the combined probability:

$$f(\mathbf{x_I}, \mathbf{x_{II}}) = \frac{f_i(\mathbf{x_I}, \mathbf{x_{II}}) \cdot \mu_{\mathbf{I}}(\mathbf{x_I}) \cdot \delta(\mathbf{x_{II}} - \mathbf{x^0_{II}})}{\mu_{\mathbf{I}}(\mathbf{x_I}) \cdot \mu_{\mathbf{II}}(\mathbf{x_{II}})} \ . \tag{A.3}$$

Using definition (A.1) we obtain

$$f_{\mathbf{I}}(\mathbf{x_I}) = \mathbb{R} \frac{f_i(\mathbf{x_I}, \mathbf{x^0_{II}})}{f_i(\mathbf{x_I}, \mathbf{x^0_{II}}) d\mathbf{x_I}} \ , \tag{A.4}$$

which corresponds to what is ordinarily named the *conditional* probability density for $\mathbf{X_I}$ given $f_i(\mathbf{x_I}, \mathbf{x_{II}})$ and $\mathbf{X_{II}} = \mathbf{x^0_{II}}$. To follow the usual notation we will write this solution $f_i(\mathbf{x_I} \mid \mathbf{x^0_{II}})$ rather than $f_{\mathbf{I}}(\mathbf{x_I})$:

$$f_i\ \mathbf{x_I} \mid \mathbf{x^0_{II}}\ = \mathbb{R} \frac{f_i(\mathbf{x_I}, \mathbf{x^0_{II}})}{f_i(\mathbf{x_I}, \mathbf{x^0_{II}}) d\mathbf{x_I}} \ . \tag{A.5}$$

The Bayes problem may be states as follows: Let $f(\mathbf{x_I}, \mathbf{x_{II}})$ be the joint probability density representing all the available information on $\mathbf{X_I}$ and $\mathbf{X_{II}}$. If we learn that $\mathbf{X_{II}} = \mathbf{x_{II}}$ we obtain $g(\mathbf{x_I} \mid \mathbf{x_{II}})$ using equation (A.4). To the contrary, if we learn $\mathbf{X_I} = \mathbf{x_I}$ we obtain $g(\mathbf{x_{II}} \mid \mathbf{x_I})$. Which is the relation between $g(\mathbf{x_I} \mid \mathbf{x_{II}})$ and $g(\mathbf{x_{II}} \mid \mathbf{x_I})$?

We have

$$g\ \mathbf{x_I} \mid \mathbf{x_{II}}\ = \mathbb{R} \frac{f(\mathbf{x_I}, \mathbf{x_{II}})}{f\ \mathbf{x_I}, \mathbf{x_{II}}\ d\mathbf{x_I}} = \frac{f(\mathbf{x_I}, \mathbf{x_{II}})}{f_{\mathbf{II}}(\mathbf{x_{II}})}$$

$$g(\mathbf{x_{II}} \mid \mathbf{x_I}) = \mathbb{R} \frac{f(\mathbf{x_I}, \mathbf{x_{II}})}{f(\mathbf{x_I}, \mathbf{x_{II}}) d\mathbf{x_{II}}} \tag{A.6}$$

$$= \mathbb{R} \frac{f(\mathbf{x_I}, \mathbf{x_{II}})}{g(\mathbf{x_I} \mid \mathbf{x_{II}}) \cdot f_{\mathbf{II}}(\mathbf{x_{II}}) \cdot d\mathbf{x_{II}}}$$

and hence

$$g(\mathbf{x_{II}} \mid \mathbf{x_I}) = \mathbb{R} \frac{g(\mathbf{x_I} \mid \mathbf{x_{II}}) \cdot f_{\mathbf{II}}(\mathbf{x_{II}})}{g(\mathbf{x_I} \mid \mathbf{x_{II}}) \cdot f_{\mathbf{II}}(\mathbf{x_{II}}) \cdot d\mathbf{x_{II}}} \ , \tag{A.7}$$

which corresponds to Bayes theorem.

We have thus shown that well known theorems may be obtained using the concept of the conjunction of states of information. Many other problems may be solved using this concept. Consider for example n independent measurements of a given parameter X. In the particular case where the null information density is uniform

($\mu(x) = \text{const.}$), and each measurement gives a gaussian probability density $f_i(x)$ centered at x_i and with variance σ^2

$$f_i(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \quad -\frac{1}{2}\frac{(x-x_i)^2}{\sigma^2} \quad , \tag{A.8}$$

the iteration of equation (2.19) at each new measurement gives

$$f(x) = \frac{1}{\sqrt{2\pi}\Sigma} \exp \quad -\frac{1}{2}\frac{(x-\bar{x})^2}{\Sigma^2} \quad , \tag{A.9}$$

where

$$\bar{x} = \frac{\Sigma x_i}{n} \quad \Sigma = \frac{\sigma}{\sqrt{n}} \ , \tag{A.10}$$

which are well known results in statistics.

Demonstrations for Section 9

Let us first demonstrate two useful identities. If $\mathbf{C}_1$ and $\mathbf{C}_2$ are two positive definite matrices respectively of order $(n \times n)$ and $(m \times m)$, and $\mathbf{M}$ an arbitrary $(n \times m)$ matrix, then:

$$\begin{aligned} &\mathbf{M^T C_1^{-1} M} + \mathbf{C_2^{-1}}\ ^{-1} \mathbf{M^T C_1^{-1}} \\ &= \mathbf{C_2 M^T}\ \mathbf{C_1} + \mathbf{M C_2 M^T}\ ^{-1}, \end{aligned} \tag{A.11}$$

$$\begin{aligned} \mathbf{M^T C_1^{-1} M} + \mathbf{C_2^{-1}}\ ^{-1} &= \mathbf{C_2} - \mathbf{C_2 M^T} \\ &\times\ \mathbf{C_1} + \mathbf{M C_2 M^T}\ ^{-1} \mathbf{M C_2}. \end{aligned} \tag{A.12}$$

The first equation follows from the following obvious identities

$$\begin{aligned} &\mathbf{M^T} + \mathbf{M^T C_1^{-1} M C_2 M^T} \\ &= \mathbf{M^T C_1^{-1}}\ \mathbf{C_1} + \mathbf{M C_2 M^T} \\ &= \ \mathbf{M^T C_1^{-1} M} + \mathbf{C_2^{-1}}\ \mathbf{C_2 M^T} \end{aligned} \tag{A.13}$$

since $\mathbf{M^T C_1^{-1} M} + \mathbf{C_2^{-1}}$ and $\mathbf{C_1} + \mathbf{M C_2 M^T}$ are definite positive and thus regular matrices.

Furthermore (A.11) leads to

$$\begin{aligned} &\mathbf{C_2} - \mathbf{C_2 M^T}\ \mathbf{C_1} + \mathbf{M C_2 M^T}\ ^{-1} \mathbf{M C_2} \\ &= \mathbf{C_2} - \ \mathbf{M^T C_1^{-1} M} + \mathbf{C_2^{-1}}\ ^{-1} \mathbf{M^T C_1^{-1} M C_2} \\ &= \ \mathbf{M^T C_1^{-1} M} + \mathbf{C_2^{-1}})^{-1} \\ &\quad \times \quad \mathbf{M^T C_1^{-1} M} + \mathbf{C_2^{-1}}\ \mathbf{C_2} - \mathbf{M^T C_1^{-1} M C_2} \\ &= \ \mathbf{M^T C_1^{-1} M} + \mathbf{C_2^{-1}}\ ^{-1} \end{aligned} \tag{A.14}$$

which proves (A.12).

Now from equation (9.5) we obtain:

$$\begin{aligned} \sigma(\mathbf{x}) = \exp \quad &-\frac{1}{2} \Big[\ \mathbf{x} - \mathbf{x}_0\ ^{\mathbf{T}} \mathbf{C}_0^{-1}\ \mathbf{x} - \mathbf{x}_0 \\ &+ \mathbf{x}^{\mathbf{T}} \mathbf{F}^{\mathbf{T}} \mathbf{C}_{\mathbf{T}}^{-1} \mathbf{F} \mathbf{x} \Big] \\ = \exp \quad &-\frac{1}{2} \Big[\ \mathbf{x}^{\mathbf{T}}\ \mathbf{C}_0^{-1} + \mathbf{F}^{\mathbf{T}} \mathbf{C}_{\mathbf{T}}^{-1} \mathbf{F}\ \mathbf{x} \\ &- 2\mathbf{x}^{\mathbf{T}} \mathbf{C}_0^{-1} \mathbf{x}_0 + \mathbf{x}_0^{\mathbf{T}} \mathbf{C}_0^{-1} \mathbf{x}_0 \Big] \end{aligned} \tag{A.15}$$

then defining:

$$\mathbf{P} = \mathbf{I} - \mathbf{C}_0 \mathbf{F}^{\mathbf{T}}\ \mathbf{F} \mathbf{C}_0 \mathbf{F}^{\mathbf{T}} + \mathbf{C}_{\mathbf{T}}\ ^{-1} \mathbf{F}, \tag{A.16}$$

$$\mathbf{C}_* = \mathbf{P} \mathbf{C}_0, \tag{A.17}$$

$$\mathbf{x}_* = \mathbf{P} \mathbf{x}_0. \tag{A.18}$$

We obtain, using equation (A.12)

$$\begin{aligned} \mathbf{C}_* &= \mathbf{P} \mathbf{C}_0 \\ &= \mathbf{C}_0 - \mathbf{C}_0 \mathbf{F}^{\mathbf{T}}\ \mathbf{F} \mathbf{C}_0 \mathbf{F}^{\mathbf{T}} + \mathbf{C}_{\mathbf{T}}\ ^{-1} \mathbf{F} \mathbf{C}_0 \\ &= \ \mathbf{F}^{\mathbf{T}} \mathbf{C}_{\mathbf{T}}^{-1} \mathbf{F} + \mathbf{C}_0^{-1}\ ^{-1} \end{aligned} \tag{A.19}$$

and

$$\mathbf{x}_* = \mathbf{P} \mathbf{x}_0 = \mathbf{C}_* \mathbf{C}_0^{-1} \mathbf{x}_0. \tag{A.20}$$

Thus equation (A.15) becomes:

$$\begin{aligned} \sigma(\mathbf{x}) = \exp \quad &-\frac{1}{2} \Big[\ \mathbf{x}^{\mathbf{T}} \mathbf{C}_*^{-1} \mathbf{x} - 2\mathbf{x}^{\mathbf{T}} \mathbf{C}_*^{-1} \mathbf{x}_* \\ &+ \mathbf{x}_0^{\mathbf{T}} \mathbf{C}_*^{-1} \mathbf{x}_* \Big] \\ = \exp \quad &-\frac{1}{2} \Big[\ \mathbf{x} - \mathbf{x}_*\ ^{\mathbf{T}} \mathbf{C}_*^{-1}\ \mathbf{x} - \mathbf{x}_* \\ &- \ \mathbf{x}_* - \mathbf{x}_0\ ^{\mathbf{T}} \mathbf{C}_0^{-1} \mathbf{x}_0 \Big] \ . \end{aligned} \tag{A.21}$$

From (A.18), we deduce:

$$\mathbf{x}_* - \mathbf{x}_0 = -\mathbf{C}_0 \mathbf{F}^{\mathbf{T}}\ \mathbf{F} \mathbf{C}_0 \mathbf{F}^{\mathbf{T}} + \mathbf{C}_{\mathbf{T}}\ ^{-1} \mathbf{F} \mathbf{C}_0 \mathbf{x}_0 \tag{A.22}$$

and then:

$$\begin{aligned} \sigma(\mathbf{x}) &= \exp \quad -\frac{1}{2} \mathbf{x}_0^{\mathbf{T}} \mathbf{F}^{\mathbf{T}}\ \mathbf{F} \mathbf{C}_0 \mathbf{F}^{\mathbf{T}} + \mathbf{C}_{\mathbf{T}}\ ^{-1} \mathbf{F} \mathbf{x}_0 \\ &\cdot \exp \quad -\frac{1}{2}\ \mathbf{x} - \mathbf{x}_*\ ^{\mathbf{T}} \mathbf{C}_*^{-1}\ \mathbf{x} - \mathbf{x}_* \\ &= \text{const.} \exp \quad -\frac{1}{2}\ \mathbf{x} - \mathbf{x}_*\ ^{\mathbf{T}} \mathbf{C}_*^{-1}\ \mathbf{x} - \mathbf{x}_* \end{aligned} \tag{A.23}$$

which demonstrates equation (9.6).

Demonstrations for Section 10

Let us now evaluate the sum:

$$I = \int \exp \times \left[-\frac{1}{2} \left[\mathbf{d} - \mathbf{d}_0 \right]^{\mathbf{T}} \mathbf{C}_\mathbf{d}^{-1} \left[\mathbf{d} - \mathbf{d}_0 \right] + \left[\mathbf{d} - \mathbf{g}(\mathbf{p}) \right]^{\mathbf{T}} \mathbf{C}_\mathbf{T}^{-1} \left[\mathbf{d} - \mathbf{g}(\mathbf{p}) \right] \right] d\mathbf{d}. \quad \text{(A.24)}$$

The separation of the quadratic terms from the linear terms leads to:

$$I = \int \exp \left[-\frac{1}{2} \left[\mathbf{d}^{\mathbf{T}} \mathbf{A} \mathbf{d} - 2\mathbf{B}^{\mathbf{T}} \mathbf{d} + \mathbf{C} \right] \right] d\mathbf{d} \quad \text{(A.25)}$$

where:

$$\begin{aligned} \mathbf{A} &= \mathbf{C}_\mathbf{d}^{-1} + \mathbf{C}_\mathbf{T}^{-1} \\ \mathbf{B}^{\mathbf{T}} &= \mathbf{d}_0^{\mathbf{T}} \mathbf{C}_\mathbf{d}^{-1} + \mathbf{g}(\mathbf{p})^{\mathbf{T}} \mathbf{C}_\mathbf{T}^{-1} \\ \mathbf{C} &= \mathbf{d}_0^{\mathbf{T}} \mathbf{C}_\mathbf{d}^{-1} \mathbf{d}_0 + \mathbf{g}(\mathbf{p})^{\mathbf{T}} \mathbf{C}_\mathbf{T}^{-1} \mathbf{g}(\mathbf{p}). \end{aligned} \quad \text{(A.26)}$$

Since $\mathbf{A}$ is positive definite there follows:

$$\begin{aligned} I &= \int \exp \left[-\frac{1}{2} \left[\mathbf{d} - \mathbf{A}^{-1}\mathbf{B} \right]^{\mathbf{T}} \mathbf{A} \left[\mathbf{d} - \mathbf{A}^{-1}\mathbf{B} \right] + \left[\mathbf{C} - \mathbf{B}^{\mathbf{T}}\mathbf{A}^{-1}\mathbf{B} \right] \right] d\mathbf{d} \\ &= \exp \left[-\frac{1}{2} \left[\mathbf{C} - \mathbf{B}^{\mathbf{T}}\mathbf{A}^{-1}\mathbf{B} \right] \right] \\ &\quad \times \int \exp \left[-\frac{1}{2} \left[\mathbf{d} - \mathbf{A}^{-1}\mathbf{B} \right]^{\mathbf{T}} \mathbf{A} \left[\mathbf{d} - \mathbf{A}^{-1}\mathbf{B} \right] \right] d\mathbf{d} \\ &= (2\pi)^{n/2} (\det \mathbf{A})^{-1/2} \exp \left[-\frac{1}{2} \left[\mathbf{C} - \mathbf{B}^{\mathbf{T}}\mathbf{A}^{-1}\mathbf{B} \right] \right]. \end{aligned} \quad \text{(A.27)}$$

By substitution of (A.26) we obtain:

$$\begin{aligned} &\mathbf{C} - \mathbf{B}^{\mathbf{T}}\mathbf{A}^{-1}\mathbf{B} \\ &= \mathbf{d}_0^{\mathbf{T}} \left[\mathbf{C}_\mathbf{d}^{-1} - \mathbf{C}_\mathbf{d}^{-1} \left[\mathbf{C}_\mathbf{d}^{-1} + \mathbf{C}_\mathbf{T}^{-1} \right]^{-1} \mathbf{C}_\mathbf{d}^{-1} \right] \mathbf{d}_0 \\ &\quad + \mathbf{g}(\mathbf{p})^{\mathbf{T}} \left[\mathbf{C}_\mathbf{T}^{-1} - \mathbf{C}_\mathbf{T}^{-1} \left[\mathbf{C}_\mathbf{d}^{-1} + \mathbf{C}_\mathbf{T}^{-1} \right]^{-1} \mathbf{C}_\mathbf{T}^{-1} \right] \mathbf{g}(\mathbf{p}) \\ &\quad - 2\mathbf{g}(\mathbf{p}) \mathbf{C}_\mathbf{T}^{-1} \left[\mathbf{C}_\mathbf{d}^{-1} + \mathbf{C}_\mathbf{T}^{-1} \right]^{-1} \mathbf{C}_\mathbf{d}^{-1} \mathbf{d}_0. \end{aligned} \quad \text{(A.28)}$$

Thus, by using the two identities ((A.11)-(A.12)) we get:

$$\begin{aligned} &\mathbf{C} - \mathbf{B}^{\mathbf{T}}\mathbf{A}^{-1}\mathbf{B} \\ &= \mathbf{d}_0 \left[\mathbf{C}_\mathbf{d} + \mathbf{C}_\mathbf{T} \right]^{-1} \mathbf{d}_0 + \mathbf{g}(\mathbf{p})^{\mathbf{T}} \left[\mathbf{C}_\mathbf{d} + \mathbf{C}_\mathbf{T} \right]^{-1} \mathbf{g}(\mathbf{p}) \\ &\quad - 2\mathbf{g}(\mathbf{p}) \left[\mathbf{C}_\mathbf{d} + \mathbf{C}_\mathbf{T} \right]^{-1} \mathbf{d}_0 \\ &= \left[\mathbf{d}_0 - \mathbf{g}(\mathbf{p}) \right]^{\mathbf{T}} \left[\mathbf{C}_\mathbf{d} + \mathbf{C}_\mathbf{T} \right]^{-1} \left[\mathbf{d}_0 - \mathbf{g}(\mathbf{p}) \right]. \end{aligned} \quad \text{(A.29)}$$

Finally we obtain:

$$\begin{aligned} I = (2\pi)^{n/2} \cdot \left[\det \left[\mathbf{C}_\mathbf{d}^{-1} + \mathbf{C}_\mathbf{T}^{-1} \right] \right]^{-1/2} \\ \cdot \exp \left[-\frac{1}{2} \left[\mathbf{d}_0 - \mathbf{g}(\mathbf{p}) \right]^{\mathbf{T}} \left[\mathbf{C}_\mathbf{d} + \mathbf{C}_\mathbf{T} \right]^{-1} \right] \\ \times \left[\mathbf{d}_0 - \mathbf{g}(\mathbf{p}) \right] \end{aligned} \quad \text{(A.30)}$$

which demonstrates equation (10.4).

Let us define

$$\mathbf{P} = \left[\mathbf{C}_\mathbf{t} + \mathbf{C}_\mathbf{T} \right]^{-1}. \quad \text{(A.31)}$$

Using (10.4) and (10.7), the sum (10.5) becomes:

$$\begin{aligned} I &= \int \exp \left[-\frac{1}{2} \sum_{ij} \left[t_i^0 - h_i - T \right] \cdot P_{ij} \cdot \left[t_j^0 - h_j - T \right] \right] dT \\ &= \int \exp \left[-\frac{1}{2} \left[dT^2 - 2b\,T + c \right] \right] dT \end{aligned} \quad \text{(A.32)}$$

where :

$$\begin{aligned} a &= \sum_{ij} P_{ij} \\ b &= \sum_{ij} P_{ij} \cdot \left[t_j^0 - h_j \right] \\ c &= \sum_{ij} \left[t_j^0 - h_j \right] \cdot P_{ij} \cdot \left[t_j^0 - h_j \right]. \end{aligned} \quad \text{(A.33)}$$

This yields :

$$\begin{aligned} I &= \int \exp \left[-\frac{1}{2} \left[a \left[T - \frac{b}{a} \right]^2 + \left[c - \frac{b^2}{a} \right] \right] \right] dT \\ &= \left[\frac{2\pi}{a} \right]^{1/2} \exp \left[-\frac{1}{2} \left[c - \frac{b^2}{a} \right] \right]. \end{aligned} \quad \text{(A.34)}$$

By substitution of a, b and c given in (A.33) in the above expression, we obtain:

$$\begin{aligned} I = \left[\frac{2\pi}{\Sigma_{ij} P_{ij}} \right]^{1/2} \\ \cdot \exp \left[-\frac{1}{2} \left[\sum_{ij} \left[t_j^0 - h_j \right] \cdot P_{ij} \cdot \left[t_i^0 - h_i \right] - \frac{\left[\sum_{ij} P_{ij} \cdot (t_j^0 - h_j) \right]^2}{\Sigma_{ij} P_{ij}} \right] \right] \end{aligned} \quad \text{(A.35)}$$

which can also be written:

$$I = \left(\frac{2\pi}{\prod_{ij} P_{ij}} \right)^{1/2} \cdot \exp\left(-\frac{1}{2} \sum_{ij} \left(t_i^0 - h_i - \frac{\sum_{kl} P_{kl}(t_l^0 - h_l)}{\sum_{kl} P_{kl}} \right) \cdot P_{ij} \cdot \left(t_j^0 - h_j - \frac{\sum_{kl} P_{kl}(t_l^0 - h_l)}{\sum_{kl} P_{kl}} \right) \right) \tag{A.36}$$

or

$$I = \left(\frac{2\pi}{\prod_{ij} P_{ij}} \right)^{1/2} \cdot \exp\left\{ -\frac{1}{2} \sum_{ij} \left(\tilde{t}_i^0 - \tilde{h}_i \right) \cdot P_{ij} \cdot \left(\tilde{t}_j^0 - \tilde{h}_j \right) \right\} \tag{A.37}$$

where

$$\tilde{t}_i^0 = t_i^0 - \frac{\sum_{kl} P_{kl} \cdot t_l^0}{\sum_{kl} P_{kl}} \qquad \tilde{h}_i = h_i - \frac{\sum_{kl} P_{kl} \cdot h_l}{\sum_{kl} P_{kl}} \tag{A.38}$$

which is the expression (10.8).

References

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

Backus, G., Gilbert, F.: The resolving power of gross earth data - Geophys. J. R. Astron. soc. **16**, 169–205, 1968

Backus, G., Gilbert, F.: Uniqueness in the inversion of inaccurate gross earth data. Philos. Trans. R. soc. London **266**, 123–192, 1970

Backus, G.: Inference from inadequate and inaccurate data, Mathematical problems in the Geophysical Sciences: Lecture in applied Mathematics, **14**, American Mathematical Society, Providence, Rhode Island, 1971

Clearbout, J.F., Muir, F.: Robust modelling with erratic data. Geophysics **38**, 5, 826–844, 1973

Descombes, R.: Integration. Corollary 6.2, p. 106. Paris: Hermann 1972

Franklin, J.N.: Well posed stochastic extension of ill posed linear problems. J. Math. Anal. Applic. **31**, 682–716, 1970

Gagnepain, J., Modiano, T., Cisternas, A., Ruegg, J.C., Vadell, M., Hatzfeld, D., Mezcua, J.: Sismicitè de la règion d'Arette (Pyrènèes-Atlantiques) et mecanismes au foyer. Ann. Gèophys. **36**, 499–508, 1980

Gallart, J.: Structure crustale des Pyrènèes d'aprés les études de sismologie expèrimentale. Thesis of 3rd cycle. 132 p. Universite Paris VI, 1980

Jackson, D.D.: Interpretation of inaccurate, insufficient and inconsistent data. Geophys. J. R. Astron. soc. **28**, 97–110, 1972

Jaynes, E.T. : Prior probabilities. I.E.E.E. Transactions on systems, Science and cybernetics. Vol. SSC-4, No. 3, 227–241, 1968

Jeffreys, H.: Theory of probability, Oxford: Clarendon Press 1939

Jeffreys, H. : Scientific Inference, London: Cambridge University Press 1957

Keilis-Borok, V.I., Yanovskaya: Inverse problems in seismology, Geophys. J. R. Astron. soc. **13**, 223–234, 1967

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

Rietsch, E.: The maximum entropy approach to Inverse Problems. J. Geophys. **42**, 489–506, 1977

Sabatier, P.C.: On geophysical inverse problems and constraints. J. Geophys. **43**, 115–137, 1977

Shannon, C.E.: A mathematical theory of communication - Bell System Tech. J. **27**, 379–423, 1948

Tarantola, A., Valette, B.: Generalized non linear inverse problems solved using the least squares criterion. Rev. Geophys. Space Phys., **19**, No. 2, 1982 (in press)

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

Received August 21, 1981; Revised version February 1, 1982; Accepted February 2, 1982