

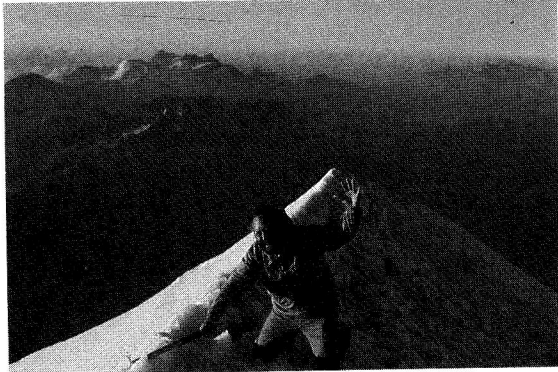
COURSE 1

**PROBABILISTIC FOUNDATIONS  
OF INVERSE THEORY**

ALBERT TARANTOLA

*I.P.G. Paris*

*Y. Desaubies, A. Tarantola and J. Zinn-Justin, eds.  
Les Houches, Session L, 1988  
Tomographie Océanographique et Géophysique /  
Oceanographic and Geophysical Tomography  
© Elsevier Science Publishers B.V., 1990*



## Contents

1. Abstract	5
2. Introduction to probability densities and volumetric probabilities	5
3. The notions of capacity element and of volume element	8
4. Information content	13
5. The state of null information	14
6. The state of perfect knowledge	15
7. Combination of informations	15
8. The data space, the model space, and the joint data $\times$ model space	17
9. Information given by physical theories	18
10. The inverse problem as a problem of combination of information	19
11. Example 1: theoretical uncertainties neglected	21
12. Example 2: all uncertainties are Gaussian	22
13. Robust inversion	24
14. Bibliographical comments	25
References	25

## 1. Abstract

The most general formulation of an inverse problem is as a problem of combination of information: information coming from measurements (and prior information) and information coming from a physical theory. The theory thus obtained is valid for fully nonlinear problems. Information is handled using probability densities, and states of information are combined by multiplying the corresponding probability densities. As special cases of the theory one can obtain standard methods, like least-squares or least-absolute-values.

The essentials of Inverse Problem Theory have already been exposed in my book (Tarantola, 1987), but as now I understand better some of the basic concepts, I hope that this exposition will be easier to follow. For instance, in my 1987 book I based all the theory on the concept of probability density. Here I show that a related concept, that of volumetric probability, simplifies the presentation.

## 2. Introduction to probability densities and volumetric probabilities

Physical parameters can either take continuous or discrete values. For instance, a temperature  $T$  is a continuously variable parameter, while the variable  $x \in \{x^1, x^2, x^3\} = \{rock, apple, cat\}$  is a discrete parameter.

To fix ideas, I will develop the theory corresponding to continuous variables. The reader interested with discrete variables can either use the guidelines given here and develop the corresponding theory, either simply derive the corresponding formulas from the formulas given here by an ad-hoc introduction of Dirac's "delta" functions.

The common intuition of probability is sufficient for our purposes, and I will not give here its abstract definition.

It is possible to introduce the notion of integral over a domain of a multidimensional space without the need of introducing the notion of volume element. Doing so, the function we are integrating will correspond to a *density*, like a "probability density" or a "mass density". That concept, to

be defined below, corresponds to a function that, in a coordinate change, has its values multiplied by the Jacobian of the transformation.

Alternatively, if the notion of volume element is defined, then we can introduce another concept: that of “volumetric probability” or of “specific mass”, whose values are invariant under a coordinate change.

For instance, let  $(r, \theta, \phi)$  be spherical coordinates in the euclidean 3D space. A *probability density*  $\bar{f}(r, \theta, \phi)$  is defined by imposing that the probability of a domain  $\mathcal{A}$  is obtained by

$$P(\mathcal{A}) = \underbrace{\int dr \int d\theta \int d\phi}_{(r, \theta, \phi) \in \mathcal{A}} \bar{f}(r, \theta, \phi) \quad (\text{for a probability density}) . \quad (1)$$

Instead, a *volumetric probability*  $f(r, \theta, \phi)$  is defined by imposing that the probability of a domain  $\mathcal{A}$  is obtained by

$$P(\mathcal{A}) = \underbrace{\int \int \int dV}_{(r, \theta, \phi) \in \mathcal{A}}(r, \theta, \phi) f(r, \theta, \phi) \quad (\text{for a specific probability}) , \quad (2)$$

where

$$dV(r, \theta, \phi) = r^2 \sin \theta \, dr \, d\theta \, d\phi .$$

Notice that I use an upper bar in  $\bar{f}$  to denote a “density”. Later, I will use a lower bar to denote a “capacity”. Essentially, in a coordinate change, a density is multiplied by the Jacobian of the transformation, while a capacity is divided by it. A true scalar, like  $f$ , is invariant under a coordinate change.

The relationship between the probability density and the specific probability is, in this example,

$$\bar{f}(r, \theta, \phi) = f(r, \theta, \phi) \frac{dV(r, \theta, \phi)}{dr \, d\theta \, d\phi} = f(r, \theta, \phi) \, r^2 \sin \theta .$$

It is because we want the expression (1) to be valid (keeping the same form) in every coordinate system (see equation (10)) that a probability density (or a mass density) must be multiplied by the Jacobian. Alternatively, in (2) it is the volume element  $dV$  which is multiplied by the Jacobian in a coordinate change, so that this expression has also the same form after a change of variables, in spite of the fact that the volumetric probability remains invariant.

Usually, probability theory is developed using probability densities rather than volumetric probabilities. This is OK, but one must then not forget that, as the value at a point of a probability density is not invariant ( i.e., it depends on the coordinate choice), notions such a “the point at which the probability density is maximum” are not invariant (changing the coordinates usually changes the point). Rather, “the point at which the specific probability is maximum” is an invariant notion, independent of the coordinate choice. Also, the “mean value” has intrinsic sense if it corresponds to the intuitive notion of “center of mass”. It can only be defined using specific probabilities. The usual definition through probability densities does not define intrinsically a point.

The only advantage of working with probability densities is that we do not need to give sense to the concept of “volume element”  $dV$  in order to be able to integrate.

Furthermore, assume that a random phenomenon produces points  $(r_i, \theta_i, \phi_i)$  with probability density  $\bar{f}(r, \theta, \phi)$ , and assume we wish to estimate  $\bar{f}(r, \theta, \phi)$  by making a three-dimensional histogram of a great number of realizations. We could then, for instance, consider a regular grid in an abstract space with axis  $(r, \theta, \phi)$  and simply count the number of realizations in each (small enough) “prism”. We do not need to care about which is the “actual volume” or each prism. Alternatively, introducing the concept of volume and dividing the number of realizations in each small “prism” by its volume, leads to an estimation of the specific probability  $f(r, \theta, \phi)$ .

In the example where  $(r, \theta, \phi)$  are spherical coordinates, an uniform probability inside a sphere of volume  $V = \frac{4}{3}\pi R^3$ , for instance, gives

$$f(r, \theta, \phi) = \frac{1}{V} \quad (\text{for the volumetric probability}) ,$$

and

$$\bar{f}(r, \theta, \phi) = \frac{1}{V} r^2 \sin \theta \quad (\text{for the probability density}) .$$

Should we have chosen cartesian coordinates  $(x, y, z)$ , then

$$f^*(x, y, z) = \frac{1}{V} \quad (\text{for the volumetric probability}) ,$$

and

$$\bar{f}^*(x, y, z) = \frac{1}{V} \quad (\text{for the probability density}) .$$

### 3. The notions of capacity element and of volume element

I am going now to introduce two different concepts that correspond, in an example where we use spherical coordinates, to the two notations

$$d\underline{V} = dr \, d\theta \, d\phi$$

and

$$dV(r, \theta, \phi) = r^2 \sin \theta \, dr \, d\theta \, d\phi .$$

The former will correspond to the “capacity element”; the latter to “volume element”. Differential objects like those above need not be taken too seriously. Although mathematicians have given them a rigorous definition, we only need to understand what we mean when we associate them to the integral symbol  $\int$ , as the limit of a discrete sum.

Let  $\mathcal{X}$  represent a  $n$ -dimensional manifold. We need  $n$  coordinates to identify a point in the  $n$ -dimensional manifold  $\mathcal{X}$ . Let these coordinates be  $(x^1, x^2, \dots, x^n)$ .

Examples: The surface of a sphere is a 2-dimensional manifold. The euclidean (3D) space is a 3-dimensional manifold. The real line is a 1-dimensional manifold. In thermodynamics we consider variables depending on pressure  $P$  and temperature  $T$ : the “quarter of the plane” defined by  $(P \geq 0, T \geq 0)$  is a 2-dimensional manifold.

The *capacity element*  $d\underline{V}$  corresponding to  $n$  vectors  $d\mathbf{r}_1, d\mathbf{r}_2, \dots, d\mathbf{r}_n$  at a given point of the space is defined by

$$d\underline{V} = \varepsilon_{ij\dots k} \, dr_1^i \, dr_2^j \dots \, dr_n^k , \quad (3)$$

where  $\varepsilon_{ij\dots k}$  is the  $n$ -dimensional Levi-Civita capacity, defined by

$$\begin{aligned} \varepsilon_{ij\dots k} &= +1 && \text{if } ij\dots k \text{ is an even permutation of } 12\dots n, \\ &= 0 && \text{if some indices are identical,} \\ &= -1 && \text{if } ij\dots k \text{ is an odd permutation of } 12\dots n . \end{aligned} \quad (4)$$

To define an integral, we take the  $n$  vectors  $d\mathbf{r}_1, d\mathbf{r}_2, \dots, d\mathbf{r}_n$  tangent to the coordinate lines at the considered point. This gives

$$d\underline{V} = \varepsilon_{12\dots n} \, dx^1 \, dx^2 \dots \, dx^n .$$

i.e.,

$$d\underline{V} = dx^1 \, dx^2 \dots \, dx^n . \quad (5)$$

Expressions (3) and (5) are equivalent. The first has the advantage of showing explicitly the covariant status of  $d\underline{V}$ . The second has the advantage of being “ready to use”.

Let  $\mathcal{A}$  be a subdomain of the working space  $\mathcal{X}$ . An expression like

$$P(\mathcal{A}) = \underbrace{\int \int \dots \int}_{(x^1, x^2, \dots, x^n) \in \mathcal{A}} d\underline{V} \, \bar{f}(x^1, x^2, \dots, x^n) , \quad (6)$$

where  $d\underline{V}$  is the capacity element defined in (3), can be rewritten, using (5), as

$$P(\mathcal{A}) = \underbrace{\int dx^1 \int dx^2 \dots \int dx^n}_{(x^1, x^2, \dots, x^n) \in \mathcal{A}} \bar{f}(x^1, x^2, \dots, x^n) , \quad (7)$$

and this can now easily be defined as the limit of a discrete sum, as usual.

If we have a probability over  $\mathcal{X}$ , then  $P(\mathcal{A})$  is defined. If, for any  $\mathcal{A}$  we can write expressions (6) (7), then we say that  $\bar{f}(x^1, x^2, \dots, x^n)$  is the *probability density* representing the probability under consideration.

We accept without further demonstration that if in a change of coordinates

$$x^i = x^i(y^1, y^2, \dots, y^n) \quad (i = 1, 2, \dots, n) , \quad (8)$$

the probability density is multiplied by the Jacobian  $J$  of the transformation,

$$\bar{f}^*(y^1, y^2, \dots, y^n) = J \bar{f}(x^1, x^2, \dots, x^n) . \quad (9)$$

then  $P(\mathcal{A})$  can be computed by any of the two expressions

$$\begin{aligned} P(\mathcal{A}) &= \underbrace{\int dx^1 \int dx^2 \dots \int dx^n}_{(x^1, x^2, \dots, x^n) \in \mathcal{A}} \bar{f}(x^1, x^2, \dots, x^n) \\ &= \underbrace{\int dy^1 \int dy^2 \dots \int dy^n}_{(y^1, y^2, \dots, y^n) \in \mathcal{A}} \bar{f}^*(y^1, y^2, \dots, y^n) , \end{aligned} \quad (10)$$

which means that (6) keeps the same form under any coordinate change.

It should be noticed that our definition of the capacity element  $d\underline{V}$  corresponds to the "exterior product" of "differential forms":

$$d\underline{V} = dx^1 \wedge dx^2 \wedge \dots \wedge dx^n, \quad (11)$$

but this notation and terminology, although compact, is less practical than the tensor notation used here.

This way of doing is general in that we can integrate over the space without caring about what the concept of "volume" may be. Doing so, we have introduced the notion of probability density. Let us now turn to the case where the concept of volume can be used: this will allow the introduction of the notion of volumetric probability.

Should we have a metric tensor  $g_{ij}(x^1, x^2, \dots, x^n)$  in our space, then it can be shown that the *volume element*  $dV(x^1, x^2, \dots, x^n)$  is given by

$$dV(x^1, x^2, \dots, x^n) = \bar{g}(x^1, x^2, \dots, x^n) d\underline{V}, \quad (12)$$

where  $\bar{g}(x^1, x^2, \dots, x^n)$  denotes the square root of the determinant of the matrix formed by the components  $g_{ij}(x^1, x^2, \dots, x^n)$ . This gives, explicitly,

$$dV(x^1, x^2, \dots, x^n) = \bar{g}(x^1, x^2, \dots, x^n) dx^1 dx^2 \dots dx^n. \quad (13)$$

The expression

$$\begin{aligned} V &= \underbrace{\int \int \dots \int dV(x^1, x^2, \dots, x^n)}_{(x^1, x^2, \dots, x^n) \in \mathcal{X}} \\ &= \underbrace{\int \int \dots \int d\underline{V}}_{(x^1, x^2, \dots, x^n) \in \mathcal{X}} \bar{g}(x^1, x^2, \dots, x^n) \end{aligned} \quad (14)$$

allows to interpret  $\bar{g}(x^1, x^2, \dots, x^n)$  as the *volume density* of the space. It is our mean to define how much volume  $dV$  is inside the capacity element  $d\underline{V} = dx^1 dx^2 \dots dx^n$ . In general we will not deal with metric spaces (i.e., spaces where the concept of length or angle is introduced), but in order to develop our theory, we will need spaces where the concept of volume makes sense. Then, we will take (12) as the definition of  $\bar{g}$ , irrespectively of any metric tensor  $g_{ij}$ . The introduction of the capacity element  $d\underline{V}$  allowed us

to define a probability density. Now, given a probability over  $\mathcal{X}$ , if, for any subdomain  $\mathcal{A}$  of  $\mathcal{X}$  we can write

$$P(\mathcal{A}) = \underbrace{\int \int \dots \int dV(x^1, x^2, \dots, x^n) f(x^1, x^2, \dots, x^n)}_{(x^1, x^2, \dots, x^n) \in \mathcal{A}}. \quad (15)$$

we say that  $f(x^1, x^2, \dots, x^n)$  is the *volumetric probability* representing the probability under consideration.

I said before that in a change of coordinates a probability density is multiplied by the Jacobian  $J$  of the transformation. This is why the capacity element keeps the same form (see equation (10)). In such a change of coordinates, the volume element is multiplied by the Jacobian. This allows to equation (15) to keep a form which is also invariant under a coordinate change:

$$\begin{aligned} P(\mathcal{A}) &= \underbrace{\int \int \dots \int}_{(x^1, x^2, \dots, x^n) \in \mathcal{A}} dV(x^1, x^2, \dots, x^n) f(x^1, x^2, \dots, x^n) \\ &= \underbrace{\int \int \dots \int}_{(y^1, y^2, \dots, y^n) \in \mathcal{A}} dV^*(y^1, y^2, \dots, y^n) f^*(y^1, y^2, \dots, y^n). \end{aligned} \quad (16)$$

but, now, the volumetric probability is a true scalar (i.e., its values are invariant under a coordinate change):

$$f^*(y^1, y^2, \dots, y^n) = f(x^1, x^2, \dots, x^n). \quad (17)$$

From (8), (14), and (15), it follows the relationship between probability density and volumetric probability:

$$\bar{f}(x^1, x^2, \dots, x^n) = \bar{g}(x^1, x^2, \dots, x^n) f(x^1, x^2, \dots, x^n). \quad (18)$$

Example: In the euclidean 3-dimensional space with spherical coordinates, the volume element  $dV(r, \theta, \phi)$  is given by

$$dV(r, \theta, \phi) = r^2 \sin \theta dr d\theta d\phi.$$

If  $\bar{f}(r, \theta, \phi)$  is a probability density, the probability of a domain  $\mathcal{A} \subset \mathcal{X}$  is given by

$$P(\mathcal{A}) = \underbrace{\int dr \int d\theta \int d\phi}_{(r, \theta, \phi) \in \mathcal{A}} \bar{f}(r, \theta, \phi). \quad (19)$$

The capacity element  $d\underline{V}$  is

$$d\underline{V} = \epsilon_{123} dx^1 dx^2 dx^3 = \epsilon_{123} dr d\theta d\phi = dr d\theta d\phi . \quad (20)$$

Definition (12) then gives

$$\bar{g}(r, \theta, \phi) = r^2 \sin \theta . \quad (21)$$

while (18) gives

$$\bar{f}(r, \theta, \phi) = \bar{g}(r, \theta, \phi) f(r, \theta, \phi) = r^2 \sin \theta f(r, \theta, \phi) . \quad (22)$$

In this example, if the physical dimension of  $r$  is  $L$ , the dimension of the volume element  $dV(r, \theta, \phi)$  is  $L^3$ , and the dimension of the capacity element  $d\underline{V}$  is  $L$ , which gives for the volume density  $\bar{g}(r, \theta, \phi)$  the dimension  $L^2$ . As the probability  $P$  is adimensional, the specific probability  $f(r, \theta, \phi)$  has dimension  $L^{-3}$  and the probability density  $\bar{f}(r, \theta, \phi)$  has dimension  $L^{-1}$ . We often encounter in physics pairs of variables like frequency-period, conductivity-resistivity, velocity-slowness, temperature- $(\beta$  parameter) ( $\beta = 1/kT$ ), . . . , with two common properties: they are all positive and one element of the pair equals the inverse of the other element. Choosing one parameter or its inverse is arbitrary, and it appears that the *logarithm* of those parameters has much simpler properties than the parameter itself.

Example: Consider a probability defined over a single axis, i.e., over a one dimensional space. Assume for instance that the axis corresponds to a positive physical parameter like an absolute temperature  $T$ . It is not obvious what is the right element of volume (here, in fact, element of length)  $d\ell(T)$  to be chosen along the axis. The euclidean choice  $d\ell(T) = dT$  is not right because this is not compatible with the euclidean choice on the reciprocal parameter  $\beta = 1/kT$ . The right choice is the euclidean length over the *logarithm of the temperature*,  $\Theta = \log(T/T_0)$ . This gives

$$d\ell(T) = \frac{dT}{T} . \quad (23)$$

Then, in this example, the volume density (here, a length density) is

$$\bar{g}(T) = \frac{1}{T} . \quad (24)$$

The probability of an interval  $(T_{\text{MIN}}, T_{\text{MAX}})$  is computed via any of the equations

$$\begin{aligned} P(T_{\text{MIN}} < T < T_{\text{MAX}}) &= \int_{T_{\text{MIN}}}^{T_{\text{MAX}}} d\ell(T) f(T) = \int_{T_{\text{MIN}}}^{T_{\text{MAX}}} dT \frac{1}{T} f(T) \\ P(T_{\text{MIN}} < T < T_{\text{MAX}}) &= \int_{T_{\text{MIN}}}^{T_{\text{MAX}}} dT \bar{f}(T), \end{aligned} \quad (25)$$

where  $f(T)$  is the volumetric probability and  $\bar{f}(T)$  the probability density, related through

$$\bar{f}(T) = \bar{g}(T)f(T) = \frac{1}{T}f(T) . \quad (26)$$

In this example, the volume element  $d\ell(T)$  is adimensional, the capacity element  $dT$  has dimension  $T$ , the volume density  $\bar{g}(T)$  has dimension  $T^{-1}$ , the volumetric probability  $f(T)$  is adimensional, and the probability density  $\bar{f}(T)$  has dimension  $T^{-1}$ .

The reader should verify that a change of variables  $\beta = 1/kT$  leads to a volume (here, length) element identical in form to (23):

$$d\ell^*(\beta) = \frac{d\beta}{\beta} . \quad (27)$$

It is this symmetry between (23) and (27) that justifies the choice of the logarithm  $\Theta$  of the temperature as "euclidean" variable: The choice  $d\ell(T) = dT$  as volume element would *not* have lead, under the change of variables  $\beta = 1/kT$  to the compatible form  $d\ell^*(\beta) = d\beta$ . See Tarantola(1987) for a more detailed discussion.

Here above,  $P(\mathcal{A})$  represented a probability. Should it represent a mass,  $\bar{f}$  would be a mass density, and  $f$  the *volumetric* (or *specific*) *mass*. Unfortunately, these two concepts are confused in many textbooks (see Brillouin (1960) and Weinberg (1972) for noteworthy exceptions) . It is a matter of taste to work with a probability density (resp. mass density) or with a volumetric probability (resp. a volumetric mass). What is important is to know what we are talking about. The concept of probability density (resp. mass density) is more general in that it does not require the definition of a measure of volume. The concept of volumetric probability (resp. specific mass) allows invariant definitions (i.e., definitions independent of the coordinate choice).

From now on, instead of explicitly writing  $(x^1, x^2, \dots, x^n)$  for our coordinates, I will symbolically write  $\mathbf{x}$ .

#### 4. Information content

Shannon's (1948) original definition of information content was for discrete probabilities:

$$I = \sum_{\alpha} \pi_{\alpha} \log \pi_{\alpha} . \quad (28)$$

The right generalization in the continuous case turns out to be

$$I = \int_{\mathcal{X}} dV(\mathbf{x}) f(\mathbf{x}) \log(V f(\mathbf{x})) . \quad (29)$$

where  $V$  is the (finite) volume of the space:

$$V = \int_{\mathcal{X}} dV(\mathbf{x}) .$$

Notice that, as the dimension of a volumetric probability is the inverse of that of  $V$ , we take, as we should, the logarithm of an adimensional number, and we obtain for  $I$  an adimensional number too.

## 5. The state of null information

It is defined by a constant volumetric probability:

$$\mu(\mathbf{x}) = \frac{1}{V} . \quad (30)$$

where  $V$  is the (finite) volume of the space into consideration, defined above.

The information content of the state of null information is null:

$$\begin{aligned} I &= \int_{\mathcal{X}} dV(\mathbf{x}) \mu(\mathbf{x}) \log(V \mu(\mathbf{x})) \\ &= \int_{\mathcal{X}} dV(\mathbf{x}) \frac{1}{V} \log(1) = \log(1) = 0 . \end{aligned} \quad (31)$$

Using the general relationship (18) between probability densities and volumetric probabilities shows that the probability density representing the state of null information is simply proportional to the volume density:

$$\bar{\mu}(\mathbf{x}) = \bar{g}(\mathbf{x}) \mu(\mathbf{x}) = \frac{1}{V} \bar{g}(\mathbf{x}) . \quad (32)$$

Then, in turn, (18) can be rewritten as

$$\bar{f} = V \bar{\mu} f . \quad (33)$$

We can then easily see that the information content is also given by

$$I = \int_{\mathcal{X}} dV \bar{f}(\mathbf{x}) \log(\bar{f}(\mathbf{x})/\bar{\mu}(\mathbf{x})) . \quad (34)$$

## 6. The state of perfect knowledge

If we have the certainty that the only possible value for the coordinate set  $\mathbf{x}$  is some given value  $\mathbf{x}_0$ , we can represent this state of information by the volumetric probability

$$f(\mathbf{x}) = \delta(\mathbf{x} - \mathbf{x}_0) , \quad (35)$$

where  $\delta(\mathbf{x})$  is the usual Dirac's "delta" function, defined by its action over an arbitrary "test function"  $\psi(\mathbf{x})$ :

$$\int_{\mathcal{X}} dV(\mathbf{x}) \delta(\mathbf{x}_0 - \mathbf{x}) \psi(\mathbf{x}) = \psi(\mathbf{x}_0) . \quad (36)$$

The information content of the state of perfect knowledge is infinite:

$$\begin{aligned} I &= \int_{\mathcal{X}} dV(\mathbf{x}) f(\mathbf{x}) \log(V f(\mathbf{x})) \\ &= \int_{\mathcal{X}} dV(\mathbf{x}) \delta(\mathbf{x} - \mathbf{x}_0) \log(V \delta(\mathbf{x} - \mathbf{x}_0)) = \log(V \delta(\mathbf{0})) = \infty . \end{aligned} \quad (37)$$

## 7. Combination of informations

Let  $f_1(\mathbf{x})$  and  $f_2(\mathbf{x})$  be two volumetric probabilities, representing two states of information on the parameters  $\mathbf{x}$ . I define the *intersection* (or *product*) of the two states of information by

$$f(\mathbf{x}) = \frac{f_1(\mathbf{x}) f_2(\mathbf{x})}{\int_{\mathcal{X}} dV(\mathbf{x}') f_1(\mathbf{x}') f_2(\mathbf{x}')} . \quad (38)$$

and the *union* (or *sum*) by

$$f(\mathbf{x}) = \frac{f_1(\mathbf{x}) + f_2(\mathbf{x})}{\int_{\mathcal{X}} dV(\mathbf{x}') (f_1(\mathbf{x}') + f_2(\mathbf{x}'))} . \quad (39a)$$

i.e.,

$$f(\mathbf{x}) = \frac{f_1(\mathbf{x}) + f_2(\mathbf{x})}{2} . \quad (39b)$$

We will later see that the intersection of states of information allows to combine information obtained from measurements with information from physical theories, and is at the very center of this probabilistic inverse theory.

I should mention that in fuzzy set theory (e.g., Kauffman, 1977) a fuzzy set is described by a function  $f(\mathbf{x})$  representing “the degree of belonging of  $\mathbf{x}$  to some ensemble”. Obviously,  $f(\mathbf{x})$  could also represent the “probability density” (in fact, volumetric probability) for  $\mathbf{x}$  to belong to the ensemble. In this theory, the classical notions of intersection or union of sets are generalized to fuzzy sets. The fathers of fuzzy set theory have chosen, for the intersection, the definition

$$f(\mathbf{x}) = \frac{\text{INF}(f_1(\mathbf{x}), f_2(\mathbf{x}))}{\int_{\mathcal{X}} dV(\mathbf{x}') \text{INF}(f_1(\mathbf{x}'), f_2(\mathbf{x}'))}, \quad (40)$$

and for the union,

$$f(\mathbf{x}) = \frac{\text{SUP}(f_1(\mathbf{x}), f_2(\mathbf{x}))}{\int_{\mathcal{X}} dV(\mathbf{x}') \text{SUP}(f_1(\mathbf{x}'), f_2(\mathbf{x}'))} \quad (41)$$

(normalization factors are mine). Those definitions and my definitions (38)–(39) have similar objectives, and, in many common situations, give similar results. I have found that the former are usually simpler to handle and give “finer” results.

Let me finally give the formula for the intersection of states of information when we use probability densities instead of specific probabilities. We have seen that a probability density  $\bar{f}(\mathbf{x})$  and the corresponding volumetric probability  $f(\mathbf{x})$  are related by

$$\bar{f}(\mathbf{x}) = \bar{g}(\mathbf{x}) f(\mathbf{x}) = V \bar{\mu}(\mathbf{x}) f(\mathbf{x}),$$

where  $\bar{g}(\mathbf{x})$  is the volume density of the space and where  $\bar{\mu}(\mathbf{x})$  is the probability density representing the state of null information.

Replacing in (38) gives then, for the intersection of states of information in terms of probability densities,

$$\bar{f}(\mathbf{x}) = \frac{\bar{f}_1(\mathbf{x}) \bar{f}_2(\mathbf{x}) / \bar{\mu}(\mathbf{x})}{\int_{\mathcal{X}} dV \bar{f}_1(\mathbf{x}') \bar{f}_2(\mathbf{x}') / \bar{\mu}(\mathbf{x}')}, \quad (42)$$

which was the form first suggested by Tarantola and Valette (1982a) and Tarantola (1987).

## 8. The data space, the model space, and the joint data $\times$ model space

Usually, the solution of the “forward problem” is written

$$\mathbf{d} = \mathbf{g}(\mathbf{m}), \quad (43)$$

this meaning that if the values of the *model parameters*  $\mathbf{m}$  are given, we can compute the *data values*  $\mathbf{d}$  through a (generally nonlinear) function  $\mathbf{g}(\mathbf{m})$ . In some sense, to be detailed below, solving the inverse problem means estimating  $\mathbf{m}$  from some knowledge on  $\mathbf{d}$ .

For instance, in the problem of estimating the location of an earthquake hypocenter from the arrival times of the seismic waves at some seismic observatories, a data set  $\mathbf{d}$  may consist of a series of arrival times, while the hypocenter coordinates are in  $\mathbf{m}$ .

I will call each possible value of the data set  $\mathbf{d}$  (resp. the model parameter set  $\mathbf{m}$ ) a *point* and not a “vector” because  $\mathbf{d}$  and  $\mathbf{m}$  may not belong to a linear space, but to a more general affine space, i.e., a space of “points” where the sum of points may not be defined).

The *data space*  $\mathcal{D}$  (resp. the *model space*  $\mathcal{M}$ ) is the set of all conceivable values of the data point  $\mathbf{d}$  (resp. the model parameter point  $\mathbf{m}$ ).

Now comes the fundamental assumption of the theory: that any “state of information” on the data or model parameter space may be described using a probability. As an example, a measurement may furnish some “observed data values”  $\mathbf{d}_{\text{obs}}$  and a variance-covariance matrix  $\mathbf{C}_D$  describing experimental uncertainties and uncertainty correlations. The assumption of Gaussian uncertainties, for instance, defines in the data space the specific probability

$$\rho_D(\mathbf{d}) = \frac{1}{(2\pi)^{n/2} \det^{1/2} \mathbf{C}_D} \exp\left(-\frac{1}{2} (\mathbf{d} - \mathbf{d}_{\text{obs}})^t \mathbf{C}_D^{-1} (\mathbf{d} - \mathbf{d}_{\text{obs}})\right).$$

This Gaussian case is only one example. More generally, we postulate that a result of a measurement of the data parameters  $\mathbf{d}$  can always be described by defining a general volumetric probability  $\rho_D(\mathbf{d})$  over the data space  $\mathcal{D}$ .

We usually have some a priori information on the values of the model parameters  $\mathbf{m}$ . By “a priori” I mean information which is independent on the results of our measurements on  $\mathbf{d}$ . This also can be described by defining a volumetric probability  $\rho_M(\mathbf{m})$  over the model space  $\mathcal{M}$ . Should we not have any a priori information at all, then we should take for  $\rho_M(\mathbf{m})$  the null information volumetric probability  $\mu_M(\mathbf{m}) = \text{const}$ .

Considering now the joint space  $\mathcal{X} = \mathcal{D} \times \mathcal{M}$ , these two states of information define the joint volumetric probability  $\rho(\mathbf{x}) = \rho(\mathbf{d}, \mathbf{m}) = \rho_{\mathcal{D}}(\mathbf{d})\rho_{\mathcal{M}}(\mathbf{m})$ . (That the joint volumetric probability is the product of the two marginal probabilities is the very definition of “a priori” information for the model parameters).

More generally, which are “model parameters” and which are “data parameters”? Consider again the problem of inverting arrival times of seismic waves in order to obtain the coordinates of an hypocenter. We all agree that the arrival times are directly measured, and are “data”, that the hypocentral coordinates are not directly measurable, and are “model parameters”, but what about the coordinates of the seismic stations? They are directly measured, and should be named “data parameters”, but as they must be in the right hand side of equation  $\mathbf{d} = \mathbf{g}(\mathbf{m})$ , they could also be named “model parameters”.

In fact, the distinction between “data parameters” and “model parameters” is arbitrary. We have a parameter space  $\mathcal{X}$  on which we have some information described by a volumetric probability  $\rho(\mathbf{x})$ . That information may come from direct measurements for some parameters, from more subjective arguments for other parameters. That “a priori” information has to be combined with information given by physical theories (as we will now see) in order to obtain the “a posteriori” information. It is only each time that we will need to write an equation like  $\mathbf{d} = \mathbf{g}(\mathbf{m})$ , or that we will need to assume independency of information (i.e., of uncertainties) through an hypothesis like  $\rho(\mathbf{x}) = \rho_{\mathcal{D}}(\mathbf{d})\rho_{\mathcal{M}}(\mathbf{m})$ , that talking about the “data parameters”  $\mathbf{d}$  and the “model parameters”  $\mathbf{m}$ , and, thus, artificially separating  $\mathcal{X}$  into  $\mathcal{D}$  and  $\mathcal{M}$ , may be useful.

## 9. Information given by physical theories

We have just seen that a physical theory is often considered to impose a functional constraint between data and model parameters:

$$\mathbf{d} = \mathbf{g}(\mathbf{m}) . \quad (43 \text{ again})$$

but this is too restrictive. In fact, physical theories always have approximations. For instance, Newton’s theory of gravitation is an approximation of Einstein’s theory, who never claimed his theory to be the ultimate truth. More commonly, for any given theory, we usually make approximations to make the theory numerically tractable. That is to say that from a given value of the model parameter set  $\mathbf{m}$  we cannot exactly predict what  $\mathbf{d}$  will

be. This suggests that instead of equation (43) we can use a conditional volumetric probability  $\theta(\mathbf{d}|\mathbf{m})$  describing the information and uncertainties we have on the values of the data parameters for each conceivable value of the model parameters. If our physical theory does not give any information on the values of the model parameters, then the overall information we have on the space  $\mathcal{X} = \mathcal{D} \times \mathcal{M}$  is represented by the joint volumetric probability

$$\theta(\mathbf{x}) = \theta(\mathbf{d}, \mathbf{m}) = \theta(\mathbf{d}|\mathbf{m})\mu_{\mathcal{M}}(\mathbf{m}) . \quad (44)$$

where  $\mu_{\mathcal{M}}(\mathbf{m}) = \text{const}$  represents the null information volumetric probability.

For more generality, we will allow the information provided by a physical theory to be represented by an arbitrary joint volumetric probability  $\theta(\mathbf{x})$  in the parameter space.

## 10. The inverse problem as a problem of combination of information

We have just seen that the results of our measurements and/or the a priori information we may have should be described using a specific probability  $\rho(\mathbf{x})$ , and that the information provided by physical theories could also be described using another volumetric probability  $\theta(\mathbf{x})$ . The main postulate of our theory is that those states of information may be combined to obtain an a posteriori state of information, described by  $\sigma(\mathbf{x})$ , and that the a posteriori state of information corresponds to the *intersection* (in the sense defined in Section 8) of the two previous states of information. This gives

$$\sigma(\mathbf{x}) = \frac{\rho(\mathbf{x})\theta(\mathbf{x})}{\int_{\mathcal{X}} dV(\mathbf{x}')\rho(\mathbf{x}')\theta(\mathbf{x}')} . \quad (45)$$

Usual applications correspond to the special case where the following assumptions are made:

- (i) We can partition  $\mathcal{X}$  into  $\mathcal{X} = \mathcal{D} \times \mathcal{M}$  such that

$$\rho(\mathbf{x}) = \rho(\mathbf{d}, \mathbf{m}) = \rho_{\mathcal{D}}(\mathbf{d}) \rho_{\mathcal{M}}(\mathbf{m}) . \quad (46)$$

where, usually,  $\rho_D(\mathbf{d})$  represents the result of some measurements, and  $\rho_M(\mathbf{m})$  corresponds to some a priori information.

(ii) The physical theory gives information on  $\mathbf{d}$  but not on  $\mathbf{m}$  :

$$\theta(\mathbf{x}) = \theta(\mathbf{d}, \mathbf{m}) = \theta(\mathbf{d}|\mathbf{m}) \mu_M(\mathbf{m}) , \quad (47)$$

where  $\mu_M(\mathbf{m}) = \text{const}$  represents the state of null information.

Then we usually get interested in the marginal a posteriori specific probabilities

$$\sigma_M(\mathbf{m}) = \int_{\mathcal{D}} dV(\mathbf{d}) \sigma(\mathbf{d}, \mathbf{m}) \quad (48)$$

and

$$\sigma_D(\mathbf{d}) = \int_{\mathcal{M}} dV(\mathbf{m}) \sigma(\mathbf{d}, \mathbf{m}) . \quad (49)$$

This gives

$$\sigma_M(\mathbf{m}) = \text{const} \rho_M(\mathbf{m}) \int_{\mathcal{D}} dV(\mathbf{d}) \rho_D(\mathbf{d}) \theta(\mathbf{d}|\mathbf{m}) \quad (50)$$

and

$$\sigma_D(\mathbf{d}) = \text{const} \rho_D(\mathbf{d}) \int_{\mathcal{M}} dV(\mathbf{m}) \theta(\mathbf{d}|\mathbf{m}) \rho_M(\mathbf{m}) . \quad (51)$$

On the model parameters  $\mathbf{m}$  we had an a priori information described by  $\rho_M(\mathbf{m})$ . We have combined this information with the theoretical information as described by  $\theta(\mathbf{d}|\mathbf{m})$  and with the information on the values of data parameters  $\mathbf{d}$  as described by  $\rho_D(\mathbf{d})$ . The result is  $\sigma_M(\mathbf{m})$ . We had  $\rho_M(\mathbf{m})$  and now we have  $\sigma_M(\mathbf{m})$ : we have solved the “inverse problem”.

The posterior volumetric probability  $\sigma_M(\mathbf{m})$  may have some pathologies: be multimodal, have infinite variances, ... but it is the solution of the inverse problem.

If we have a very small number of model parameters (say, up to four) we can directly “look” into the posterior volumetric probability in the model space by simply computing its values in a dense grid in the model space and plotting as many sections of the volumetric probability as we may need to understand what posterior information we really have.

For problems with high dimensionality, this is unpractical. Then, we can, for instance, use an algorithm generating pseudorandom models according to the posterior volumetric probability  $\sigma_M(\mathbf{m})$  and plot enough of them until our brain gets a good idea on what’s likely in the model space. Mosegaard and Tarantola (1989) suggest using a Monte Carlo method (simulated annealing) to generate the pseudorandom models.

If available computer technology does not allow the exploration of the posterior volumetric probability, then we can limit ourselves to much more modest objectives, as, for instance, computing a few estimators (mean parameter values, median parameter values, standard deviations, correlations, mean deviations, ...). The easiest central estimator to compute is usually the maximum likelihood point (point maximizing the volumetric probability) because we are then faced with a classical optimization problem. However, for truly multimodal posterior volumetric probabilities, all these estimators are dangerous, and sometimes meaningless.

## 11. Example 1: theoretical uncertainties neglected

The function  $\theta(\mathbf{d}|\mathbf{m})$  represents the information carried by a physical theory. If uncertainties in the theory are small compared with experimental uncertainties, we can assume that from a model  $\mathbf{m}$  we can exactly compute the “predicted data”  $\mathbf{d}_{\text{cal}}$  by some (generally nonlinear) function  $\mathbf{m} \rightarrow \mathbf{g}(\mathbf{m})$  :

$$\mathbf{d}_{\text{cal}} = \mathbf{g}(\mathbf{m}) .$$

This corresponds to the assumption that  $\theta(\mathbf{d}|\mathbf{m})$  has the form

$$\theta(\mathbf{d}|\mathbf{m}) = \delta(\mathbf{d} - \mathbf{g}(\mathbf{m})) \quad (52)$$

i.e., that in the space  $\mathcal{X} = \mathcal{D} \times \mathcal{M}$  all values of  $\mathbf{d}$  are forbidden excepted when  $\mathbf{d} = \mathbf{g}(\mathbf{m})$  .

Replacing (52) in (50) gives

$$\sigma_M(\mathbf{m}) = \text{const} \rho_M(\mathbf{m}) \rho_D(\mathbf{d}) \Big|_{\mathbf{d} = \mathbf{g}(\mathbf{m})} \quad (53)$$

Understanding this formula, is important. It states that the posterior volumetric probability at any point in the model space equals the prior volumetric probability times the volumetric probability of the data corresponding to that model.

## 12. Example 2: all uncertainties are Gaussian

Assume here that our measurements have furnished the “observed data values”  $\mathbf{d}_{\text{obs}}$  with uncertainties that can adequately be represented with a Gaussian volumetric probability with a variance-covariance matrix  $\mathbf{C}_D$  :

$$\rho_D(\mathbf{d}) = \frac{1}{(2\pi)^{n/2} \det^{1/2} \mathbf{C}_D} \exp\left(-\frac{1}{2}(\mathbf{d} - \mathbf{d}_{\text{obs}})^\dagger \mathbf{C}_D^{-1} (\mathbf{d} - \mathbf{d}_{\text{obs}})\right). \quad (54)$$

Assume also that the a priori information we have on the model parameters is adequately described by the a priori model  $\mathbf{m}_{\text{prior}}$  with attached Gaussian uncertainties as described by the variance-covariance matrix  $\mathbf{C}_M$  :

$$\rho_M(\mathbf{m}) = \frac{1}{(2\pi)^{m/2} \det^{1/2} \mathbf{C}_M} \times \exp\left(-\frac{1}{2}(\mathbf{m} - \mathbf{m}_{\text{prior}})^\dagger \mathbf{C}_M^{-1} (\mathbf{m} - \mathbf{m}_{\text{prior}})\right). \quad (55)$$

Finally, assume that our physical theory predicts for the data  $\mathbf{d}$  corresponding to the model  $\mathbf{m}$  the value  $\mathbf{d} = \mathbf{g}(\mathbf{m})$  with some uncertainties that are also Gaussian and are described by the variance-covariance matrix  $\mathbf{C}_T$  :

$$\theta(\mathbf{d}|\mathbf{m}) = \frac{1}{(2\pi)^{n/2} \det^{1/2} \mathbf{C}_T} \times \exp\left(-\frac{1}{2}(\mathbf{d} - \mathbf{g}(\mathbf{m}))^\dagger \mathbf{C}_T^{-1} (\mathbf{d} - \mathbf{g}(\mathbf{m}))\right). \quad (56)$$

$$\sigma_M(\mathbf{m}) = \text{const.} \times \exp\left(-\frac{1}{2}\left((\mathbf{g}(\mathbf{m}) - \mathbf{d}_{\text{obs}})^\dagger \mathbf{C}^{-1} (\mathbf{g}(\mathbf{m}) - \mathbf{d}_{\text{obs}}) + (\mathbf{m} - \mathbf{m}_{\text{prior}})^\dagger \mathbf{C}_M^{-1} (\mathbf{m} - \mathbf{m}_{\text{prior}})\right)\right). \quad (57)$$

Then, the integral over the data space needed in (50) in order to evaluate  $\sigma_M(\mathbf{m})$  can be analytically performed (Tarantola, 1987) and we obtain where

$$\mathbf{C} = \mathbf{C}_D + \mathbf{C}_T. \quad (58)$$

Notice that unless  $\mathbf{g}(\mathbf{m})$  is a linear function,  $\sigma_M(\mathbf{m})$  is not Gaussian.

From  $\sigma_M(\mathbf{m})$  we can obtain, by numerical integration, estimates as (a posteriori) mean values or (a posteriori) variances-covariances. Should we be interested only in the maximum likelihood point  $\mathbf{m}_{ML}$  (maximizing the volumetric probability), it is obvious that it minimizes the *nonlinear least squares expression*

$$S(\mathbf{m}) = \frac{1}{2}\left((\mathbf{g}(\mathbf{m}) - \mathbf{d}_{\text{obs}})^\dagger \mathbf{C}^{-1} (\mathbf{g}(\mathbf{m}) - \mathbf{d}_{\text{obs}}) + (\mathbf{m} - \mathbf{m}_{\text{prior}})^\dagger \mathbf{C}_M^{-1} (\mathbf{m} - \mathbf{m}_{\text{prior}})\right), \quad (59)$$

and it can be obtained using standard optimization techniques. For instance, a steepest descent algorithm gives (Tarantola, 1987):

$$\mathbf{m}_{n+1} = \mathbf{m}_n - \alpha (\mathbf{C}_M \mathbf{G}_n^\dagger \mathbf{C}^{-1} (\mathbf{g}(\mathbf{m}_n) - \mathbf{d}_{\text{obs}}) + (\mathbf{m}_n - \mathbf{m}_{\text{prior}})), \quad (60)$$

where  $\mathbf{G}$  is the linear tangent operator (derivative)

$$\mathbf{g}(\mathbf{m}_n + \delta \mathbf{m}) = \mathbf{g}(\mathbf{m}_n) + \mathbf{G}_n \delta \mathbf{m} + \dots$$

$\mathbf{G}^\dagger$  its transpose, and where  $\alpha$  is a constant small enough.

Finally, if the function solving the forward problem is linear,

$$\mathbf{d} = \mathbf{g}(\mathbf{m}) = \mathbf{Gm}. \quad (61)$$

then, it is not difficult to see (Tarantola, 1987) that the a posteriori volumetric probability  $\sigma_M(\mathbf{m})$  is Gaussian:

$$\sigma_M(\mathbf{m}) = \frac{1}{(2\pi)^{m/2} \det^{1/2} \mathbf{C}'_M} \times \exp\left(-\frac{1}{2}(\mathbf{m} - \mathbf{m}_{\text{posterior}})^\dagger \mathbf{C}'_M^{-1} (\mathbf{m} - \mathbf{m}_{\text{posterior}})\right), \quad (62)$$

with

$$\begin{aligned} \mathbf{m}_{\text{posterior}} &= \mathbf{m}_{\text{prior}} - (\mathbf{G}^\dagger \mathbf{C}^{-1} \mathbf{G} + \mathbf{C}_M^{-1})^{-1} \mathbf{G}^\dagger \mathbf{C}^{-1} (\mathbf{Gm}_{\text{prior}} - \mathbf{d}_{\text{obs}}) \\ &= \mathbf{m}_{\text{prior}} - \mathbf{C}_M \mathbf{G}^\dagger (\mathbf{G} \mathbf{C}_M \mathbf{G}^\dagger + \mathbf{C})^{-1} (\mathbf{Gm}_{\text{prior}} - \mathbf{d}_{\text{obs}}), \end{aligned} \quad (63)$$

and

$$\begin{aligned} \mathbf{C}'_M &= (\mathbf{G}^\dagger \mathbf{C}^{-1} \mathbf{G} + \mathbf{C}_M^{-1})^{-1} \\ &= \mathbf{C}_M - \mathbf{C}_M \mathbf{G}^\dagger (\mathbf{G} \mathbf{C}_M \mathbf{G}^\dagger + \mathbf{C})^{-1} \mathbf{G} \mathbf{C}_M. \end{aligned} \quad (64)$$

### 13. Robust inversion

The previous example, based on the Gaussian assumption leads to the least squares criterion. It is easy to implement numerically, and this is the main reason for the wide use of the Gaussian assumption, even if it is sometimes difficult to justify from an objective analysis of uncertainties.

The main practical defect of the Gaussian assumption (and, thus, of the whole least squares method) is that the estimates it gives for the model parameters are hypersensitive to a small amount of large errors (outliers) in a data set. It is well known that the method of least absolute values is not sensitive to outliers (Claerbout and Muir, 1973) : it is more *robust*.

From the point of view developed here this corresponds to choosing volumetric probabilities whose tails do not tend to zero as rapidly as the Gaussian tails. For instance, replacing the Gaussian function

$$f(x) = \frac{1}{\sqrt{2\pi} \sigma} \exp\left(-\frac{1}{2} \frac{(x - x_0)^2}{\sigma^2}\right) \quad (65)$$

by the double exponential

$$f(x) = \frac{1}{2\sigma} \exp\left(-\frac{|x - x_0|}{\sigma}\right) \quad (66)$$

leads to the least absolute values criterion, rather than to the least squares one (Tarantola, 1987) .

Recently, Crase et al. (1989) obtained nice results using an hyperbolic secant to model data uncertainties:

$$f(x) = \frac{1}{\pi\sigma} \operatorname{sech}\left(\frac{x - x_0}{\sigma}\right) \quad (67)$$

This bell-shaped function behaves like a double exponential for large values of  $|x - x_0|/\sigma$ , and like a Gaussian function for small values. This implies that data with small errors will contribute to the solution as do data in the least squares method, while data with large errors will not contribute, as do those data in the least absolute values method. Using gradient methods, it is not more difficult to use this criterion than the least squares one: where a least squares algorithm uses the weighted residuals

$$\frac{1}{\sigma^i} \left( \frac{d^i - g^i(\mathbf{m})}{\sigma^i} \right),$$

this recipe uses the expression

$$\frac{1}{\sigma^i} \tanh\left(\frac{d^i - g^i(\mathbf{m})}{\sigma^i}\right),$$

which dampes large residuals (see Crase et al., 1989): at no extra cost, we can transform any least squares gradient method into a robust one.

### 14. Bibliographical comments

The “foundations” I have described in this course give a very personal view of the theory. A pioneering description was made by Backus (1970, 1971), and the reader may get some pleasure in reading his papers. Parker (1975, 1977) has made some additions to the theory.

Historically important papers on Monte Carlo inversion are Keilis-Borok & Yanovskaya (1967) and Press (1968, 1971).

Franklin (1970) is interesting because he was the first to set properly least squares for (linear) functional problems (not addressed here).

Besides working on the foundations of Inverse Theory, I have been trying to find Oil (by inversion of recorded seismic waveforms). Should the reader wish to know my work on that domain, he may refer for instance to Tarantola (1986). Tarantola and Nercessian (1984) give an example of inversion of seismic travel times.

Anyone willing to face the optimization problem of finding maximum likelihood points may consult Fletcher (1980, 1981), Polack and Ribière (1969), Powell (1981), Scales (1985), Walsh (1975), and Watson (1980).

### Acknowledgements

A lot of thanks to the students of the School. Their questions and remarks helped me to improve my understanding of the theory.

### References

- Backus, G., 1970. Inference from inadequate and inaccurate data, Proceedings of the National Academy of Sciences, 65, 1, 1-105; 65, 2, 281-287; 67, 1, 282-289.  
 Backus, G., and Gilbert, F., 1967. Numerical applications of a formalism for geophysical inverse problems. Geophys. J. R. astron. Soc., 13, 247-276.

- Backus, G., and Gilbert, F., 1968. The resolving power of gross Earth data. *Geophys. J. R. astron. Soc.*, 16, 169-205.
- Backus, G., and Gilbert, F., 1970. Uniqueness in the inversion of inaccurate gross Earth data, *Philos. Trans. R. Soc. London*, 266, 123-192.
- Backus, G., 1971. Inference from inadequate and inaccurate data, *Mathematical problems in the Geophysical Sciences: Lecture in applied mathematics*, 14. American Mathematical Society, Providence, Rhode Island.
- Brillouin, L., 1960. *Les tenseurs en Mécanique et en Elasticité*. Masson et Cie., Paris.
- Claerbout, J.F., and Muir, F., 1973. Robust modelling with erratic data. *Geophysics*, 38, 5, 826-844.
- Crase, E., Pica, A., and Tarantola, A., 1989. Robust elastic nonlinear inversion of seismic waveforms: theoretical aspects and numerical results. *Geophysics*, in press.
- Fletcher, R., 1980. *Practical methods of optimization, Volume 1: Unconstrained optimization*, Wiley.
- Fletcher, R., 1981. *Practical methods of optimization, Volume 2: Constrained optimization*, Wiley.
- Franklin, J.N., 1970. Well posed stochastic extensions of ill posed linear problems. *J. Math. Anal. Applic.*, 31, 682-716.
- Kauffman, A., 1977. *Introduction à la théorie des sous-ensembles flous*. Masson, Paris.
- Keilis-Borok, V.J., and Yanovskaya, T.B., 1967. Inverse Problems of Seismology (Structural Review). *Geophys. J.R. astron. Soc.*, 13, 223-234.
- Licknerowicz, A., 1960. *Eléments de Calcul Tensoriel*, Armand Collin, Paris.
- Mosegaard, K., and Tarantola, A., 1989. Inverse problems and simulated annealing. (submitted to *Journal of Geophysical Research*).
- Parker, R.L., 1975. The theory of ideal bodies for gravity interpretation. *Geophys. J. R. astron. Soc.*, 42, 315-334.
- Parker, R.L., 1977. Understanding inverse theory. *Ann. Rev. Earth Plan. Sci.*, 5, 35-64.
- Polack, E. et Ribière, G., 1969. Note sur la convergence de méthodes de directions conjuguées. *Revue Fr. Inf. Rech. Oper.*, 16-R1, 35-43.
- Powell, M.J.D., 1981. *Approximation theory and methods*, Cambridge University Press.
- Press, F., 1968. Earth models obtained by Monte Carlo inversion. *J. Geophys. Res.*, Vol. 73, No. 16, 5223-5234.
- Press, F., 1971. An introduction to Earth structure and seismotectonics. *Proceedings of the International School of Physics Enrico Fermi. Course L. Mantle and Core in Planetary Physics*. J. Coulomb and M. Caputo (editors), Academic Press.

- Scales, J. A., and Gersztenkorn, A., 1988. Robust methods in inverse theory. *Inverse Problems*, 4, 1071-1-91.
- Scales, L. E., 1985. *Introduction to non-linear optimization*. Macmillan.
- 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.
- Tarantola, A. and Nercessian, A., 1984. Three-dimensional inversion without blocks. *Geophys. J. R. astron. Soc.*, 76, 299-306.
- Tarantola, A., and Valette, B., 1982a. Inverse Problems = Quest for Information. *J. Geophys.*, 50, 159-170.
- Tarantola, A., and Valette, B., 1982b. Generalized nonlinear inverse problems solved using the least-squares criterion. *Rev. Geophys. Space Phys.*, 20, No. 2, 219-232.
- Walsh, G.R., 1975. *Methods of optimization*, Wiley.
- Watson, G.A., 1980. *Approximation theory and numerical methods*, Wiley.