Geophysical Prospecting **36**, 878–898, 1988

BAYESIAN ESTIMATION IN SEISMIC INVERSION. PART I: PRINCIPLES[1]

A. J. W. DUIJNDAM[2]

ABSTRACT

DUIJNDAM, A.J.W. 1988. Bayesian estimation in seismic inversion. Part I: Principles. *Geophysical Prospecting* **36**, 878–898.

This paper gives a review of Bayesian parameter estimation. The Bayesian approach is fundamental and applicable to all kinds of inverse problems. Its basic formulation is probabilistic. Information from data is combined with *a priori* information on model parameters. The result is called the *a posteriori* probability density function and it is the solution to the inverse problem. In practice an estimate of the parameters is obtained by taking its maximum. Well-known estimation procedures like least-squares inversion or l_1-norm inversion result, depending on the type of noise and *a priori* information given. Due to the *a priori* information the maximum will be unique and the estimation procedures will be stable except (in theory) for the most pathological problems which are very unlikely to occur in practice. The approach of Tarantola and Valette can be derived within classical probability theory.

The Bayesian approach allows a full resolution and uncertainty analysis which is discussed in Part II of the paper.

INTRODUCTION

The goal of seismics and other geophysical techniques is to provide information about the subsurface. The processing of data obtained should solve an inverse problem: the estimation of parameters describing the subsurface. For seismics the most straightforward and general formulation leads to a non-linear inverse problem for the complete multi-offset data set. It should incorporate all aspects of wave propagation. Attempts in this direction have been made (Tarantola 1986). The amount of computational power needed for this approach however is extremely large, considering today's hardware technology. The present processing method therefore still comprises several smaller steps, often involving a number of simplifying assumptions. Most of these simpler steps can be formulated as statistical or parametric inverse problems. Examples of processing steps formulated as parametric inverse problems are: (i) residual statics correction (Wiggins, Larner and Wisecup

[1] Received March 1987, revision accepted April 1988.

[2] Delft Geophysical B.V., P.O. Box 148, 2600 AC Delft, The Netherlands.

1976; Rothman 1985, 1986); (ii) estimation of velocity models (Gjöstdal and Ursin 1981; Van der Made 1988); (iii) wavelet estimation (Duijndam, Van Riel and Kaman 1984); (v) detailed inversion of post-stack seismic data (Kaman, Van Riel and Duijndam 1984; Cooke and Schneider 1983; Mendel 1983; amongst others).

In practice we always have to deal with uncertainties. Therefore an inverse problem should be formulated using probability theory. Well-known problems in inversion, when using only data, are the related items of non-uniqueness, ill-posedness and instability. In practice, these problems can be overcome by using *a priori* information about the parameters. The most fundamental and straightforward method is the so-called Bayesian approach to inversion.

Part I of this paper reviews the most important aspects of Bayesian inversion. From the probabilistic formulation the rationale of least-squares inversion, l_1-inversion and the use of constraints become clear. Part I concentrates on estimation fundamentals. It thereby provides the theoretical basis of two companion papers on the detailed inversion of post-stack seismic data.

An important aspect, nowadays recognized in seismics, is the need to provide an analysis of uncertainties in the estimates. This is easily appreciated. When a large number of parameter models (in a practical sense) has almost the same probability of corresponding to the true model one surely needs to know this. With only the estimated model available, this type of information is lacking. Uncertainty and resolution analysis is discussed in Part II (Duijndam, 1988).

Part I also introduces some new elements and a number of non-trivial and essential matters are discussed which, so far, have been ignored in geophysics.

The Inverse Problem

The general inverse problem can be stated as follows: a system generates output. From the output (observational data) and possibly other information, knowledge about the system can be inferred. In seismics one example of a system is the combination of the earth, a seismic data acquisition system and a seismic processing system. The corresponding output is a seismic section and the desired knowledge concerns the geology of the earth. To obtain quantitative knowledge, a parameterization of the system has to be chosen. A proper choice of parameters is very important, as shown in the two companion papers on the detailed inversion of post-stack seismic data. Hypothesis testing and related techniques may guide the choice. In Part I this item is not discussed further. The inverse problem is studied starting from when some parameterization has been chosen.

The solution of the inverse problem is obtained by combining information concerning data, theoretical relations and *a priori* information about the parameters in a suitable form. Because uncertainties in information play an important role, the mathematical tools to be used are those of probability theory. The reader is assumed to be familiar with the basics of probability theory. In Appendix A a concise overview is given of the basic concepts used. Introductory textbooks on the subject are e.g. Papoulis (1965) and Mood, Graybill and Boes (1974). An excellent book on parameter estimation is Bard (1974). An outstanding article on inverse

theory is Tarantola and Valette (1982a). See also Tarantola (1987). Tarantola's fundamental starting point is different from that of classical probability theory, a fact that seems to go unnoticed in geophysics. In Appendix C, however, it is shown that Tarantola's starting points and his method of combining states of information can be derived from classical probability theory.

Bayes' Rule

Now the most basic relation in Bayesian estimation theory is discussed. Let the vector **y** contain discretized data. The parameters are contained in the parameter vector **x**. Both **x** and **y** are vectors of variables. From their joint probability density function (pdf) $p(\mathbf{x}, \mathbf{y})$ and the concepts of marginal and conditional pdfs, Bayes' rule can easily be derived (see Appendix A),

$$p(\mathbf{x} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathbf{x}) \cdot p(\mathbf{x})}{p(\mathbf{y})}. \tag{1}$$

The function $p(\mathbf{x} \mid \mathbf{y})$ is the so-called *a posteriori* pdf. It is the conditional pdf of **x** after a realization (measurement) of the data vector **y** is obtained. The function $p(\mathbf{y} \mid \mathbf{x})$ is the conditional pdf of **y** given **x**. As discussed below, it contains information about the theoretical relations between parameters and data and it also contains noise properties. *A posteriori*, when a measurement result **d** can be substituted for **y** and the function is viewed as a function of **x**, it is also called the likelihood function. The second factor in the numerator is $p(\mathbf{x})$. It is the marginal pdf of $p(\mathbf{x}, \mathbf{y})$ for **x**. It reflects the information about **x** when disregarding the data and thus it should contain the *a priori* knowledge on the parameters. The denominator $p(\mathbf{y})$ does not depend on **x** and can be considered as a constant factor in the inverse problem.

It is important to realize that $p(\mathbf{x} \mid \mathbf{y})$ contains all the information available about **x** given the data **y** and, therefore, it is in fact the solution to the inverse problem. It is mostly due to the impossibility of displaying the function in a practical way for more than one or two parameters that a point estimate (discussed below) is derived from it.

Equation (1) can also be used without the restriction that all functions are strict pdfs in the sense that their integrals are one. Then constant factors are immaterial (see also Tarantola and Valette (1982a) and Bard (1974)) and the functions are simply called density functions.

A Priori Information

Information about the parameters which is available independent of the data can be used as *a priori* information and is formulated in $p(\mathbf{x})$. This type of information may come from general knowledge about the system under study, e.g. geological knowledge.

A priori knowledge about parameters often consists of an idea about the values and uncertainties in these values. A suitable probability density function to describe

this type of information is the Gaussian or normal distribution,

$$p(\mathbf{x}) = (2\pi)^{-n/2} |\mathbf{C}_x|^{-1/2} \exp\{-\tfrac{1}{2}(\mathbf{x} - \mathbf{x}^i)^T \mathbf{C}_x^{-1}(\mathbf{x} - \mathbf{x}^i)\}, \tag{2}$$

where n is the number of parameters, $\mathbf{x}^i$ is the mean of the distribution (the guessed values) and $\mathbf{C}_x$ is the covariance matrix, which specifies the uncertainties.

Another pdf sometimes used for specifying *a priori* information is the longer tailed exponential distribution (see Appendix B):

$$p(\mathbf{x}) = 2^{-n/2} |\mathbf{C}_x|^{-1/2} \exp\{-2\|\mathbf{C}_x^{-1/2}(\mathbf{x} - \mathbf{x}^i)\|_1\}, \tag{3}$$

where $\|\cdot\|_1$ denotes the l_1-norm of a vector,

$$\|\mathbf{x}\|_1 = \sum_i |\mathbf{x}_i|. \tag{4}$$

Often additional *a priori* information can be formulated by hard constraints, preventing parameters obtaining physically impossible values such as negative thicknesses and negative propagation velocities. Hard constraints can occur as bounds on the parameters but also as strict relations between parameters. Hard constraints define area where the *a priori* distribution is zero.

The Likelihood Function

The conditional pdf $p(\mathbf{y} | \mathbf{x})$ gives the probability of the data, given the parameters $\mathbf{x}$. Most inverse problems can be treated using the so-called standard reduced model (Bard 1974),

$$\mathbf{y} = \mathbf{g}(\mathbf{x}) + \mathbf{n}, \tag{5}$$

where $\mathbf{g}(\mathbf{x})$ is the forward model, used to create synthetic data. It can be non-linear. The vector $\mathbf{n}$ contains the errors or noise.

Let the result of a measurement be denoted by a vector of numbers $\mathbf{d}$. When $\mathbf{y} = \mathbf{d}$ is substituted in $p(\mathbf{y} | \mathbf{x})$, the result, interpreted as a function of $\mathbf{x}$, is called the likelihood function, denoted by $l(\mathbf{x})$,

$$l(\mathbf{x}) = p(\mathbf{y} = \mathbf{d} | \mathbf{x}). \tag{6}$$

In the literature a distinction is sometimes made between theoretical and observational errors. In seismics, for example, neglecting multiples and using an acoustic instead of an elastic theory would typically be regarded as theoretical errors. Noise on the data due to, e.g. traffic would be regarded as observational errors. The distinction, however, is arbitrary as is easily illustrated. Let $\mathbf{f}$ denote an ideal theory. The theoretical errors $\mathbf{n}_1$ are defined as

$$\mathbf{n}_1 = \mathbf{f}(\mathbf{x}) - \mathbf{g}(\mathbf{x}). \tag{7}$$

Substitution in (5) yields

$$\mathbf{y} = \mathbf{f}(\mathbf{x}) - \mathbf{n}_1 + \mathbf{n}. \tag{8}$$

The remaining error term on the right-hand side is denoted by $\mathbf{n}_2$.

$$\mathbf{n}_2 = \mathbf{y} - \mathbf{f}(\mathbf{x}), \tag{9}$$

and constitutes the observational errors. The theoretical and observational errors of course sum to the total error

$$\mathbf{n} = \mathbf{n}_1 + \mathbf{n}_2. \tag{10}$$

From (5) and (10) it is clear that both types of errors are treated in the same way. That the distinction *must* be arbitrary is confirmed when we consider how $\mathbf{f}(\mathbf{x})$ would be defined in practice. One may argue that an ideal theory fully explains the data. It takes every aspect of the system into account. Hence, $\mathbf{y} = \mathbf{f}(\mathbf{x})$ and therefore $\mathbf{n}_2 = 0$. All errors $\mathbf{n} = \mathbf{n}_1 = \mathbf{f}(\mathbf{x}) - \mathbf{g}(\mathbf{x})$ are then theoretical. The opposite way of reasoning is that since no theory is perfect but arbitrary to some extent, we might as well declare $\mathbf{g}(\mathbf{x})$ to be 'ideal'. We then have $\mathbf{f}(\mathbf{x}) = \mathbf{g}(\mathbf{x})$, $\mathbf{n}_1 = \boldsymbol{\phi}$ and hence all errors $\mathbf{n} = \mathbf{n}_2 = \mathbf{y} - \mathbf{f}(\mathbf{x})$ are observational! Neither viewpoint is wrong. The definition of the ideal theory and hence the distinction between theoretical and observational errors is simply arbitrary.

When the errors $\mathbf{n}$ are independent of $\mathbf{g}(\mathbf{x})$ and have a pdf p_n it follows:

$$p(\mathbf{y} \mid \mathbf{x}) = p_n(\mathbf{y} - \mathbf{g}(\mathbf{x})). \tag{11}$$

In practice, despite the above considerations, one will be inclined to call one type of error theoretical and another observational. Theoretical errors will usually have other distributions, e.g. they may be correlated with $\mathbf{g}(\mathbf{x})$. When the error terms, however, are independent, the pdf of the sum is the convolution of the individual pdfs:

$$p_n = p_{n1} * p_{n2}, \tag{12}$$

see Mood *et al.* (1974). Otherwise a more general formulation can be used, see Duijndam (1987).

Duijndam (1987) has shown that the approach of Tarantola and Valette (1982a, b), which distinguishes theoretical and observational errors on a more fundamental level, yields the same result as the Bayesian approach under different interpretations of theoretical and observational errors.

Point Estimation

Because it is impractical, if not impossible, to inspect the *a posteriori* pdf through the whole of parameter space, a so-called point estimate is usually computed. When the objective is to estimate the parameters as accurately as possible in a least-squares sense the mean of $p(\mathbf{x} \mid \mathbf{y})$ is obtained (see e.g. Bard 1974):

$$\hat{\mathbf{x}} = \int \mathbf{x} p(\mathbf{x} \mid \mathbf{y}) d\mathbf{x}. \tag{13}$$

This estimator is sometimes referred to as the least-mean-squared error or the Bayes estimator. For the properties of this estimator the reader is referred to the text-books. Unfortunately the evaluation of (13) requires the computation of $p(\mathbf{x} \mid \mathbf{y})$

through the whole of parameter space which makes it practically impossible in most cases. An alternative and more practical solution is to choose the maximum of the *a posteriori* density function, sometimes referred to as MAP estimation. When $p(\mathbf{x} \mid \mathbf{y})$ is symmetrical and unimodal, the mean coincides with the mode and the least-mean-squared estimator is equivalent to the MAP estimator. This estimator can be interpreted as yielding the most likely values of the parameters given data and *a priori* information.

For a uniform *a priori* distribution $p(\mathbf{x})$, which is often taken as the state of null information, it is easily seen that the maximum of the *a posteriori* density function coincides with the maximum of the likelihood function. MAP estimation is then equivalent to maximum likelihood estimation (MLE). The difference between MLE and MAP estimation in general is clear. MLE does not take *a priori* information into account. For a discussion on the asymptotic properties of MAP estimation and MLE see Bard (1974). The importance of asymptotic properties should not be overemphasized. In practice there is always a limited amount of data. Unfortunately, MAP estimation is also sometimes referred to as maximum likelihood estimation. The *a posteriori* density function is then called the unconditional likelihood function.

Analytical results of MAP estimation depend on the form of the pdf s involved. We shall first consider Gaussian distributions for noise and *a priori* information. The means and the covariance matrices are assumed to be given throughout this paper. The *a priori* distribution is then given by (2):

$$p(\mathbf{x}) = \text{const. exp } \{-\tfrac{1}{2}(\mathbf{x}^i - \mathbf{x})^T \mathbf{C}_x^{-1} (\mathbf{x}^i - \mathbf{x})\}. \tag{14}$$

When the noise is assumed to have zero mean and covariance matrix $\mathbf{C}_n$ its pdf is

$$p(\mathbf{n}) = \text{const. exp } \{-\tfrac{1}{2}\mathbf{n}^T \mathbf{C}_n^{-1} \mathbf{n}\}. \tag{15}$$

The likelihood function follows with (11)

$$p(\mathbf{y} = \mathbf{d} \mid \mathbf{x}) = \text{const. exp } \{-\tfrac{1}{2}(\mathbf{d} - \mathbf{g}(\mathbf{x}))^T \mathbf{C}_n^{-1} (\mathbf{d} - \mathbf{g}(\mathbf{x}))\}. \tag{16}$$

Maximizing the product of $p(\mathbf{x})$ and $p(\mathbf{y} = \mathbf{d} \mid \mathbf{x})$ is equivalent to minimizing the sum of the exponents, as given by the function F,

$$2F(\mathbf{x}) = (\mathbf{d} - \mathbf{g}(\mathbf{x}))^T \mathbf{C}_n^{-1} (\mathbf{d} - \mathbf{g}(\mathbf{x})) + (\mathbf{x}^i - \mathbf{x})^T \mathbf{C}_x^{-1} (\mathbf{x}^i - \mathbf{x}). \tag{17}$$

This is a weighted non-linear least-squares or l_2 norm. The factor 2 is introduced for notational convenience in the following theory. The first term of F is the energy of the weighted residuals or data mismatch $\mathbf{d} - \mathbf{g}(\mathbf{x})$. The second term is the weighted l_2 norm of the deviation of the parameters from their *a priori* mean values $\mathbf{x}^i$. From a non-Bayesian point of view this term stabilizes the solution. It is not present in maximum likelihood estimation. The relative importance of data mismatch and parameter deviations is determined by their uncertainties as specified in $\mathbf{C}_n$ and $\mathbf{C}_x$.

The minimum of (17) can be found with optimization methods. For the linear problem $\mathbf{g}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ an explicit solution of (17) is obtained:

$$\hat{\mathbf{x}} = (\mathbf{A}^T \mathbf{C}_n^{-1} \mathbf{A} + \mathbf{C}_x^{-1})^{-1} (\mathbf{A}^T \mathbf{C}_n^{-1} \mathbf{d} + \mathbf{C}_x^{-1} \mathbf{x}^i). \tag{18}$$

This solution, introduced in geophysics by Jackson (1979) (see also Franklin (1970) or Bard (1974)) is the least-mean-squared error estimator under Gaussian assumptions. A number of well-known estimators such as the Gauss-Markov (weighted least-squares), the linear least-squares estimator and the diagonally-stabilized least-squares estimator can be derived as special cases of (18).

The assumption of the double exponential distribution as given in (3) leads to the minimization of an l_1-norm:

$$F(\mathbf{x}) = \|\mathbf{C}_n^{-1/2}(\mathbf{d} - \mathbf{g}(\mathbf{x}))\|_1 + \|\mathbf{C}_x^{-1/2}(\mathbf{x}^i - \mathbf{x})\|_1. \tag{19}$$

The use of uniform distributions leads to linear constraints on data mismatch or parameter deviations, in general form given by

$$\mathbf{l}_1 \leq \mathbf{A}(\mathbf{d} - \mathbf{g}(\mathbf{x})) \leq \mathbf{u}_1, \tag{20a}$$

$$\mathbf{l}_2 \leq \mathbf{B}(\mathbf{x}^i - \mathbf{x}) \leq \mathbf{u}_2, \tag{20b}$$

where $\mathbf{l}_1$, $\mathbf{l}_2$, $\mathbf{u}_1$ and $\mathbf{u}_2$ represent vectors with lower and upper bounds, and $\mathbf{A}$ and $\mathbf{B}$ are suitable matrices determined by the specific problem.

Selection of the Type of PDF

In the application of inverse theory, the question rises which type of pdf is to be used for the noise and the *a priori* information. In Fig. 1 the three most often used pdf s are given for a 1D problem, each with a standard deviation of one. They are the Gaussian, the double exponential and the uniform distribution. They can be combined in practice with hard constraints which specify regions where the pdf s are zero.

The Gaussian pdf has the following advantages (Bard 1974):

1. It has been found to approximate closely the behaviour of many measurements in nature.

2. By the so-called central limit theorem, the distribution of the sum of a large number of identically-distributed independent random variables is Gaussian.

3. It is the pdf which, given the mean and the covariance, contains the least information as determined by Shannon's information measure. This was defined by Shannon (1948), for a dimensionless pdf as

$$I = \mathrm{E}(\log p) = \int p(\mathbf{x}) \log p(\mathbf{x})\, \mathrm{d}\mathbf{x}. \tag{21}$$

This means that when we have the mean and the covariance, we do not use more information than we legitimately know by choosing the Gaussian pdf.

4. It is mathematically most tractable.

Point 1 obviously only applies to the data. Point 2 may also apply to *a priori* information, when information from several sources is combined. Point 3 is a strong reason in favour of Gaussian pdf s, as Shannon's information measure has some properties which reflect those of the concept of information as used in everyday life.

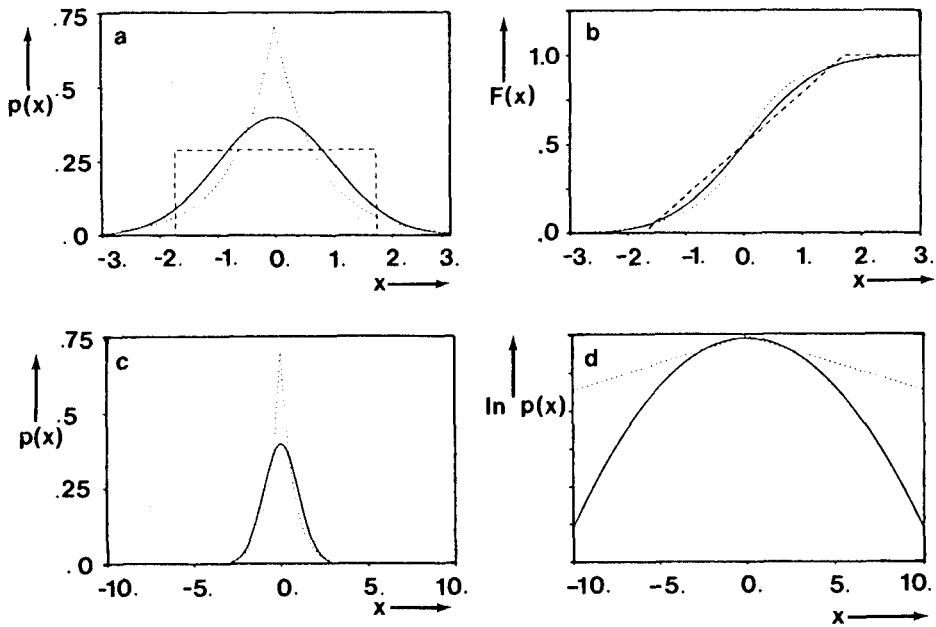

FIG. 1. 1D pdfs with zero mean and a standard deviation of one. (——) Gaussian pdf, (· · · · ·) double exponential pdf, (– – –) uniform pdf. (a) pdfs; (b) the corresponding distribution functions $F(x) = \int_{-\infty}^{x} p(x)\,dx$; (c) the Gaussian and double exponential pdf for a wide range; (d) as (c) but in logarithmic display.

One may wonder, however, when the covariance matrix is available in practice. Even with *a priori* information, where an idea of the uncertainty is simply given by someone working on the problem, it is questionable whether the uncertainty value is to be attributed to a standard deviation. Although standard deviations are often used to indicate uncertainties in practice, this usage must be based on the (implicit or explicit) assumption that the underlying pdf has a form close to the Gaussian one. For this pdf, the standard deviation is indeed a reasonable measure of uncertainty; the interval of $(\mu - \sigma, \mu + \sigma)$ corresponds with a 67% confidence interval.

Of these four points, the pragmatic one (4) is perhaps the strongest reason for using Gaussian pdfs. All the mathematics can be nicely resolved, and fast optimization schemes have been developed for the resulting least-squares problems. The author would like to augment the list with the simple statement that the Gaussian pdf often describes our knowledge reasonably. Especially with regard to *a priori* knowledge about parameters, one often wants the top of the pdf to be flat, with no strong preference around the mean. Further away from the mean, the pdf should gradually decrease and it should go rapidly to zero far away from the mean (say 3–4 times the standard deviation). Of course, this need not hold for *all* types of information! Sometimes there are reasons to choose another type of pdf. It is, for example, well known that least-squares schemes are not robust, i.e. are sensitive to large outliers. Noise realizations with large outliers are better described by the double exponential distribution. This distribution leads to the more robust l_1-norm

schemes, see e.g. Claerbout and Muir (1973). The uniform distribution has also (implicitly) been used for the inversion of seismic data. As far as I know the only type of errors that is described by the uniform distribution is quantization errors. In seismics, however, these errors are rarely large enough to be important.

The question concerning the type of pdf is often stated in the following form: 'What type of noise *is* in the data?' This question reflects a way of thinking typical for an objective interpretation of the concept of probability (see also 'Discussion'). In this interpretation a number of known (in the sense of identified) or unknown processes constitute a random generator corrupting the data. It has been suggested that we should try to find the pdf according to which the errors are generated. In the most general form, however, the dimension of the pdf is equal to the number of data points. We then only have one realization available, from which the form of the pdf can never be determined.

We need the assumption of repetitiveness in order to have the noise samples identically distributed, so that something can be said about the form of the pdf. This assumption, however, can never be tested for its validity and is therefore metaphysical rather than physical. In what is called the subjective Bayesian interpretation, another way of reasoning is followed. The noise reflects our uncertainties concerning the combination of data and theory. The solution of an inverse problem consists of combining *a priori* information and information from data and theory. The selection of another type of pdf is equivalent to asking another question. The inspection of residuals after inversion may give reason to modify the type or the parameters of the distribution chosen.

A Two-Parameter Example

The utilization and the benefits of Bayes' rule (1) can be illustrated with a simple synthetic example. It contains only two parameters and therefore allows the full visualization of the pdfs. The problem is a 1D seismic inverse problem and concerns the estimation of the acoustic impedance and the thickness in traveltime of a thin layer. The true acoustic impedance profile is given in Fig. 2a as a function of traveltime. The acoustic impedance Z above and below the layer, as well as the position of the upper boundary τ_1, are given. The values are 5×10^6 kg m^{-2} s^{-1} and 50 ms respectively. The first parameter to be estimated is the acoustic impedance of the thin layer. For the sake of clarity only the difference ΔZ with the background impedance is referred to. The second parameter is the thickness in traveltime $\Delta\tau$ of the layer. The true values of the parameters are $\Delta Z = 3 \times 10^6$ kg m^{-2} s^{-1} and $\Delta\tau = 1.5$ ms respectively. The forward model used is the convolutional model with primaries only. For this particular problem it can be written in the form:

$$s(t) = \frac{\Delta Z}{2Z + \Delta Z} \{w(t - \tau_1) - w(t - (\tau_1 + \Delta\tau))\}. \tag{22}$$

Using this expression and the zero-phase wavelet $w(t)$ as given in Figs 2b and c synthetic data is generated and is shown in Fig. 2d. Bandlimited noise with an energy of -3 dB relative to the noise-free data is added to it. The resulting noisy

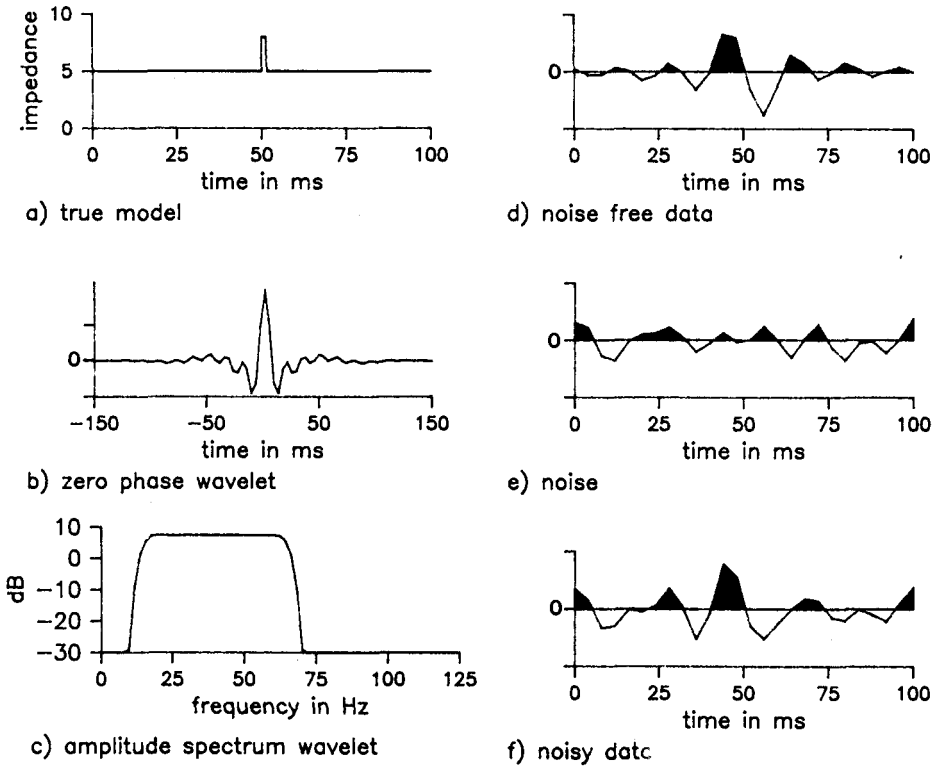

FIG. 2. Setup of the two-parameter example. The unit of impedance in (a) is 10^6 kg m^{-2} s^{-1}.

data as shown in Fig. 2f is used for inversion. The available *a priori* information on the parameters is given in the form of Gaussian pdfs, depicted in Figs 3a and b. The position of the peaks represent the *a priori* values and the standard deviations represent the uncertainties in these values. The values are:

$$\Delta Z^i = 3.4 \times 10^6 \text{ kg m}^{-2} \text{ s}^{-1}, \qquad \sigma_{\Delta Z} = 0.5 \times 10^6 \text{ kg m}^{-2} \text{ s}^{-1}, \tag{23}$$

$$\Delta\tau^i = 3 \text{ ms}, \qquad \sigma_{\Delta\tau} = 2 \text{ ms}.$$

For the thickness there is the additional hard constraint that its value cannot be less than zero. This is expressed by zero *a priori* probability for negative values. The true values of the parameters are also indicated in the figure. They are not equal to the *a priori* values, but they are within one standard deviation interval. The *a priori* information on the parameters is independent. The 2D pdf is therefore the product of the two 1D pdfs:

$$p(\Delta Z, \Delta\tau) = p(\Delta Z)p(\Delta\tau), \tag{24}$$

and is given by (2) for the region $\Delta\tau \geq 0$ with

$$\mathbf{x}^i = \begin{pmatrix} \Delta Z^i \\ \Delta\tau^i \end{pmatrix} \tag{25}$$

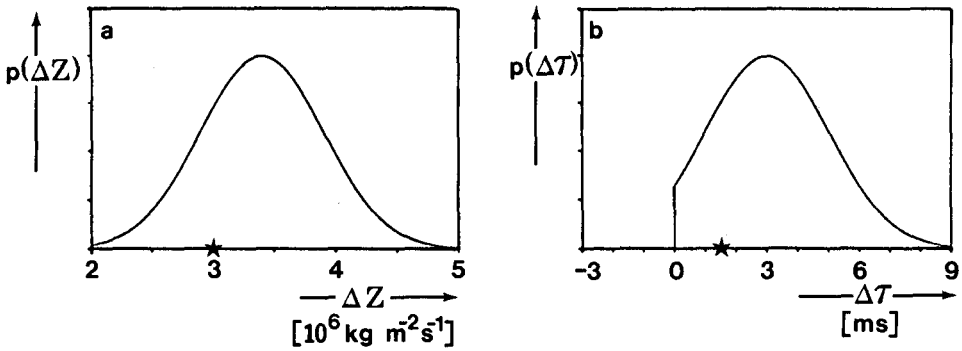

FIG. 3. *A priori* pdfs on the parameters ΔZ(a) and $\Delta\tau$(b). The true values are indicated on the x-axis with an $*$.

and

$$C_x = \begin{bmatrix} \sigma^2_{\Delta Z} & 0 \\ 0 & \sigma^2_{\Delta\tau} \end{bmatrix}. \tag{26}$$

In Fig. 4 the 2D pdfs for this problem are given. Figures 5a–c give the corresponding contour plots, with the values of the true parameters indicated. The contours of the *a priori* pdf show as ellipses because, in terms of *a priori* standard deviations, the ranges plotted for both parameters are not equal: ten for ΔZ versus six for $\Delta\tau$. The hard constraint is clearly visible. The likelihood function is computed under the assumption of white Gaussian noise with a power corresponding to a signal-to-noise ratio of 3 dB. The formula used is thus (16) with $C_n = \sigma_n^2 I$ and $g_i =$

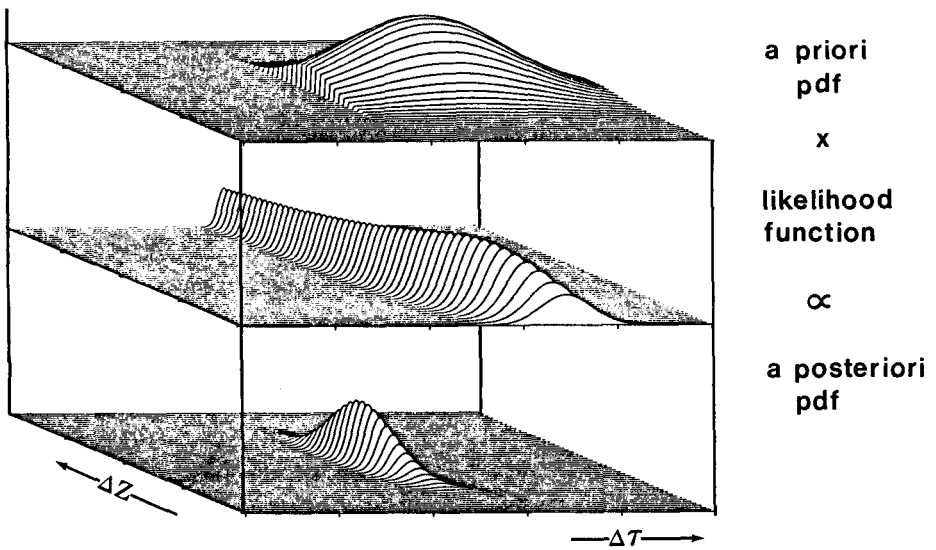

FIG. 4. The 2D density functions of the two-parameter example. The *a posteriori* pdf is proportional to the product of the *a priori* pdf and the likelihood function.

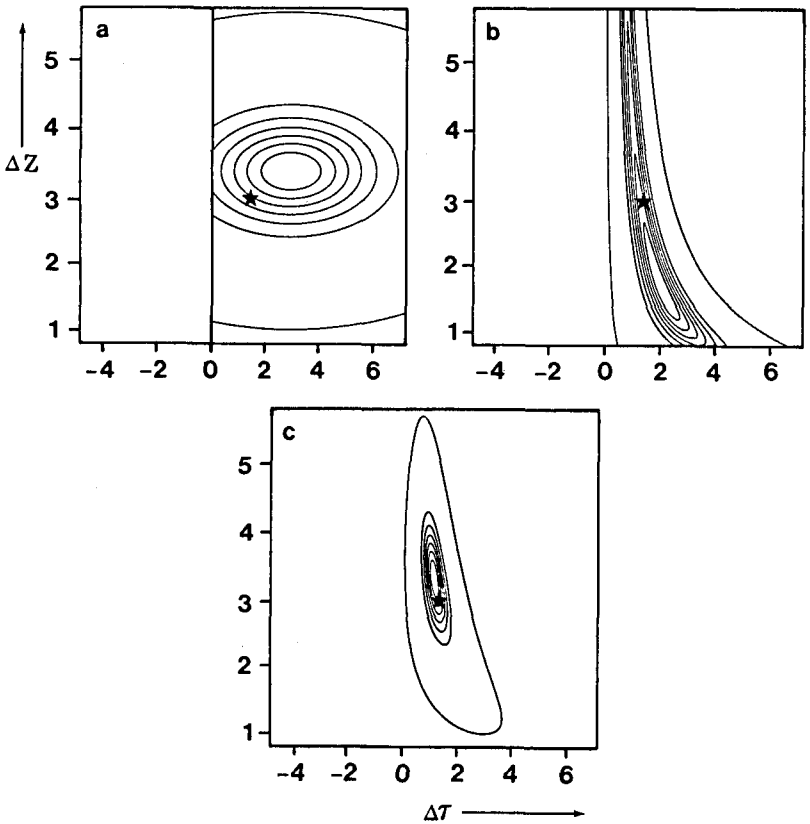

FIG. 5. Contours of the *a priori* pdf (a), the likelihood function (b) and the *a posteriori* pdf (c). The units of ΔZ and $\Delta\tau$ are 10^6 kg^{-2} s^{-1} and ms respectively. The location of the true model is indicated with an *.

$s(i\Delta t)$, with $s(t)$ defined in (22). The function has a unique maximum but a wide range of significantly different models, lying on the ridge, has almost equal likelihood. This arises because the response of a thin layer can be approximated by

$$s(t) = \frac{\Delta Z}{2Z + \Delta Z}\,\Delta\tau w'(t - \tau_m), \tag{27}$$

where $w'(t)$ is the time derivative of the wavelet and $\tau_m = \tau_1 + \Delta\tau/2$ is the position of the middle of the layer. This position can be considered fixed for the range of $\Delta\tau$ under consideration. The synthetic data depends on the product of $\Delta\tau$ and a (nearly linear) function of ΔZ. Therefore, an infinite number of combinations ΔZ and $\Delta\tau$ give equal synthetic data and hence equal data mismatch and likelihood values through relation (16).

The product of the *a priori* pdf and the likelihood function renders the solution of the inverse problem: the *a posteriori* pdf, given in Figs 4 and 5c. It is much more restricted than the likelihood function and has a unique maximum much closer to

TABLE 1. Numerical details of the two-parameter example.

Model	ΔZ (10^6 kg^{-2} s^{-1})	$\Delta\tau$ (ms)	$\lvert\Delta Z - \Delta Z_{\text{true}}\rvert$ (10^6 kg^{-2} s^{-1})	$\lvert\Delta\tau - \Delta\tau_{\text{true}}\rvert$ (ms)	$\lvert d - g(x)\rvert_2/\lvert d\rvert_2$ (dB)
True	3.0	1.5	0	0	—
A priori	3.4	3.0	0.4	1.5	—
MLE	2.0	1.8	1.0	0.3	−4.15
MAP	3.32	1.25	0.32	0.25	−4.0

the true model. In Table 1 the true model and the maxima of the pdfs are given, together with the deviations from the true model and the data mismatch for MLE and MAP estimation. The data mismatch for MAP estimation is higher because the parameters are restricted in their freedom to explain the data.

In Fig. 6 the data mismatch (residual) is given in comparison with the data and the noise realization. The residual strongly resembles the noise because the number of parameters is much smaller than the number of data points. In maximum likelihood estimation the residual energy is always lower than the noise energy. For MAP estimation this need not hold when the *a priori* model $\mathbf{x}^i$ is not equal to the true model. Tests on the residuals are important. They can indicate 'inconsistent

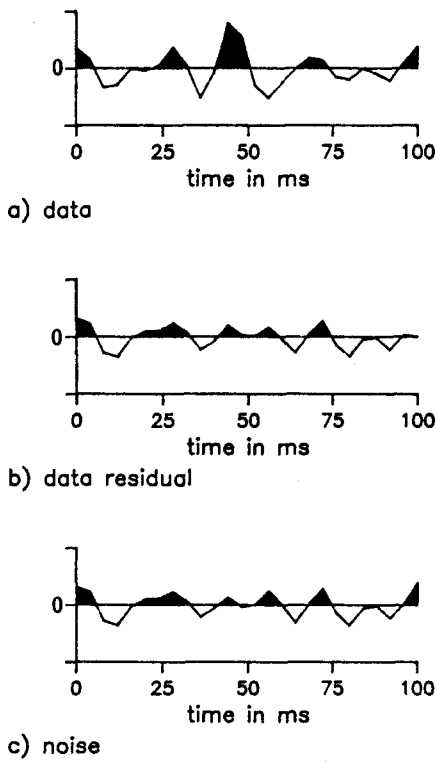

FIG. 6. The data mismatch (b) in comparison with the data (a) and the noise realization (c).

information'. The chosen noise level may be too low, the forward model may be incorrect, the parameter model may be too simple, etc. Procedures for tests on residuals are described in statistical textbooks. The issue is not pursued further here.

Optimization

In geophysical literature estimation and optimization are not always clearly distinguished. This may lead to confusion. Estimation or inversion theory is the combination of statistical and physical theories and arguments that may lead to a function that has to be minimized, e.g. (18). Optimization is the mathematical or numerical technique for the determination of that minimum. In principle, therefore, any optimization technique that finds the minimum will do. In practice, however, efficiency considerations usually make the proper choice of an optimization algorithm very important. Many textbooks consider this subject, e.g. Gill, Murray and Wright (1981) and Scales (1985).

The optimization methods used in the two-parameter example and in the problems in the companion papers belong to the Newton methods (quasi-Newton and special least-squares methods), because the number of parameters is moderate, and first and second derivatives can easily be provided.

Discussion and Conclusion

The specification and use of *a priori* distributions have long been the subject of dispute in statistics. The problem is strongly related to the interpretation of the concept of probability. For a long time statistical thought has been dominated by the frequency school, which interprets probabilities as relative frequencies. This school, in attempting to develop objective techniques, rejects the use of subjective *a priori* information. Its main antagonist is usually referred to as the Bayesian school, which advocates the interpretation of the probability concept as subjective knowledge. It can rightly be argued that both interpretations are possible and that they simply apply to different situations. The problem remains, however, which interpretation to choose for practical inverse problems. The frequentists reject Bayes' rule as a basis for inverse problems, although by itself, as a mathematical tautology, it is compatible with a frequency interpretation. For a discussion of the Bayesian point of view see Jeffreys (1939, 1957), Savage (1954) and De Finetti (1974). Tarantola (1987) adopted the same interpretation. Outspoken proponents of the Bayesian interpretation can be found amongst authors on maximum entropy techniques, e.g. Jaynes (1968, 1985). This interpretation has gained strongly in the past decade. Bayesian methods are used more and more, also in geophysics.

It is beyond the scope of this article to discuss these matters in detail. A comprehensive overview and comparison of the different approaches was given by Barnett (1982). I consider the Bayesian arguments more convincing and the results

are certainly more encouraging. The utilization of *a priori* information can strongly improve estimation results. It solves the problems of non-uniqueness and instability, as shown in this paper and the two companion papers.

ACKNOWLEDGEMENTS

I wish to thank Professor A. J. Berkhout and A. van der Schoot for critically reviewing the manuscript and Miss T. van Lier for patiently typing it. Thanks are also due to Delft Geophysical for permission to publish this paper.

APPENDIX A

BASICS OF PROBABILITY THEORY

This appendix briefly lists the probabilistic concepts used in this article. For a more thorough introduction or a more complete overview, the reader is referred to textbooks on statistics.

The probability of an event A is denoted by $P(A)$. The joint distribution function of a set of n variables X_i is defined by

$$F_X(\mathbf{x}) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n), \tag{A1}$$

where $\mathbf{x}$ is the vector containing the values x_i. In the sequel as well as in the body of the paper a less strict mathematical notation will be used to improve readability, except at the point where confusion may arise. Let $F(\mathbf{x})$ simply denote the distribution function of a vector of variables $\mathbf{x}$. Provided F is differentiable, the probability density function (pdf) $p(\mathbf{x})$ of $\mathbf{x}$ is defined by

$$p(\mathbf{x}) = \frac{\partial^n F(\mathbf{x})}{\partial x_1 \partial x_2, \ldots, \partial x_n}. \tag{A2}$$

The probability of $\mathbf{x}$ being in a certain region or volume A is given by

$$P(\mathbf{x} \in A) = \int_A p(\mathbf{x}) \mathrm{d}x_1 \mathrm{d}x_2, \ldots, \mathrm{d}x_n, \tag{A3}$$

or, in shorthand,

$$P(\mathbf{x} \in A) = \int_A p(\mathbf{x})\, \mathrm{d}\mathbf{x}. \tag{A4}$$

The integration is over the volume A. From the usual axioms and conventions of probability theory it follows

$$p(\mathbf{x}) \geq 0. \tag{A5}$$

A strict pdf is normalized

$$\int p(\mathbf{x})\, d\mathbf{x} = 1. \tag{A6}$$

The expectation of a function $g(\mathbf{x})$ is defined as

$$Eg(\mathbf{x}) = \int g(\mathbf{x})p(\mathbf{x})\, d\mathbf{x}. \tag{A7}$$

In particular, the mean or the expectation $\boldsymbol{\mu}$ of $\mathbf{x}$ is given by

$$\boldsymbol{\mu} = E\mathbf{x} = \int \mathbf{x}p(\mathbf{x})\, d\mathbf{x}. \tag{A8}$$

The integral is performed for each element of $\mathbf{x}$. The covariance matrix $\mathbf{C}$ is defined as

$$\mathbf{C} = E(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T p(\mathbf{x})\, d\mathbf{x}. \tag{A9}$$

The integral is performed for each element of $\mathbf{C}$. The diagonal of $\mathbf{C}$ contains the variances $\sigma_i^2 = E(x_i - \mu_i)^2$. The square root σ_i is called the standard deviation. The correlation coefficients ρ_{ij} are defined as

$$\rho_{ij} = \frac{c_{ij}}{\sigma_i \sigma_j}. \tag{A10}$$

They take values between -1 and 1. The matrix with the elements ρ_{ij} is called the correlation matrix. The diagonal contains values of 1.

The joint pdf of two vectors of variables $\mathbf{x}$ and $\mathbf{y}$ is denoted by $p(\mathbf{x}, \mathbf{y})$. The vectors $\mathbf{x}$ and $\mathbf{y}$ are said to be independent when their joint pdf can be written as

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}) \cdot p(\mathbf{y}). \tag{A11}$$

Here $\mathbf{x}$ and $\mathbf{y}$ are also uncorrelated. The converse does not always hold. Zero correlation, however, does imply independence for the often used Gaussian pdf. The marginal pdf for $\mathbf{x}$ is defined by

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{y})\, d\mathbf{y}. \tag{A12}$$

It can be interpreted as being the pdf of $\mathbf{x}$ when disregarding $\mathbf{y}$, or when averaging over $\mathbf{y}$. $p(\mathbf{x}, \mathbf{y})$ is therefore integrated over $\mathbf{y}$. The conditional pdf $p(\mathbf{x} \mid \mathbf{y})$ is the pdf for $\mathbf{x}$, given values for $\mathbf{y}$. It is defined as

$$p(\mathbf{x} \mid \mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{y})}, \tag{A13}$$

and is therefore proportional to the joint pdf $p(\mathbf{x}, \mathbf{y})$, with the fixed values for $\mathbf{y}$ substituted. The division by $p(\mathbf{y})$ renders $p(\mathbf{x}\,|\,\mathbf{y})$ a strict pdf (property (A6)). From (A13) and the similar relation

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y}\,|\,\mathbf{x})p(\mathbf{x}), \tag{A14}$$

Bayes' rule follows

$$p(\mathbf{x}\,|\,\mathbf{y}) = \frac{p(\mathbf{y}\,|\,\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}, \tag{A15}$$

which is the starting point for Bayesian inversion techniques. A useful result for more complex problems is the chain rule, which is obtained by repeatedly applying (A13) on the combination of vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$:

$$p(\mathbf{x}_n, \mathbf{x}_{n-1}, \ldots, \mathbf{x}_1) = \left[\prod_{i=2}^{n} p(\mathbf{x}_i\,|\,\mathbf{x}_{i-1}, \ldots, \mathbf{x}_1)\right] \cdot p(\mathbf{x}_1). \tag{A16}$$

Appendix B

Probability Density Functions

The most often used probability density function is the well-known Gaussian or normal pdf:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\,|\,\mathbf{C}\,|^{1/2}} \cdot \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}, \tag{B1}$$

where $\boldsymbol{\mu}$ and $\mathbf{C}$ are the mean and the covariance matrix respectively. The Gaussian pdf is mathematically most tractable. For an overview and derivations of its elegant properties, see Miller (1975). A less frequently used pdf is the double exponential or Laplace distribution. It leads to the more robust l_1-norm estimators. As far as I know this distribution is only used in geophysics for independently distributed parameters (or data). The joint pdf for n parameters with possibly different standard deviations σ_i is then the product of n 1D pdfs:

$$p(\mathbf{x}) = \prod_i \frac{\sqrt{2}}{\sigma_i} \exp\left\{-\sqrt{2}\,\frac{|\,x_i - \mu_i\,|}{\sigma_i}\right\}. \tag{B2}$$

The parameters are uncorrelated. This pdf can be generalized for non-zero correlations. Consider the multi-dimensional pdf

$$p(\mathbf{x}) = \frac{|\,\mathbf{W}\,|}{2^{n/2}} \exp\{-\sqrt{2}\|\mathbf{W}(\mathbf{x}-\boldsymbol{\mu})\|_1\}, \tag{B3}$$

where $\mathbf{W}$ is a non-singular square matrix and where $\|\cdot\|_1$ denotes the l_1-norm of a vector

$$\|\mathbf{x}\|_1 = \sum_i |\,x_i\,|. \tag{B4}$$

The following properties can be derived. $p(\mathbf{x})$ as given in (B3) is a strict pdf:

$$\int p(\mathbf{x})\, d\mathbf{x} = 1. \tag{B5}$$

The expectation of $\mathbf{x}$ is given by

$$\int \mathbf{x} p(\mathbf{x})\, d\mathbf{x} = \boldsymbol{\mu}. \tag{B6}$$

The convariance matrix $\mathbf{C}$ is given by

$$\mathbf{C} = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T p(\mathbf{x})\, d\mathbf{x} = (\mathbf{W}^T\mathbf{W})^{-1}. \tag{B7}$$

Property (B7) shows that $\mathbf{W}$ and thereby $p(\mathbf{x})$ as given in (B3) is not uniquely determined by the mean and the covariance matrix unlike the Gaussian case. Consider the particular choice $\mathbf{W} = \mathbf{C}^{-1/2}$. The resulting expression for $p(\mathbf{x})$ reads

$$p(\mathbf{x}) = \frac{1}{2^{n/2}|\mathbf{C}|^{1/2}} \cdot \exp\{-\sqrt{2}\|\mathbf{C}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})\|_1\}. \tag{B8}$$

This distribution lacks a number of favourable properties that the Gaussian pdf has. A linear transformation of parameters distributed according to e.g. (B8) leads to a distribution of the form (B3), but not necessarily to the form (B8). Like the Gaussian pdf, however, zero correlations imply independence. The specific linear transformation $y = \mathbf{C}^{-1/2}\mathbf{x}$ renders n independent identically-distributed (iid) parameters, with each parameter distributed according to a 1D Laplace distribution with unit standard deviation.

APPENDIX C

THE APPROACH OF TARANTOLA AND VALETTE

Inversion theory is fundamental as it describes how quantitative knowledge is obtained from experimental data. As such it has a scope that covers the whole of empirical science and is applicable in a much wider area than geophysics alone. Seen in this light it is interesting that Tarantola and Valette (1982a) (see also Tarantola 1987) formulated a theory for inverse problems to solve a number of alleged problems of the classical Bayesian approach. Their formulation can be seen as a generalization of the Bayesian one. Unlike Bayes' rule, it can handle inverse problems where the data is not obtained as a set of numbers, but rather as a pdf ('vague data'). This is, for example, more appropriate where the data for inversion is obtained by interpretation of analogue data or instruments, e.g. the reading of arrival times from a seismogram. Their theory seems to distinguish more strictly between theoretical and observational errors than the Bayesian approach. This

appendix shows that the basic concept and formulation of "conjunction of states of information", which is the cornerstone of their theory, can also be derived within classical probability theory.

In Tarantola's theory a state of information on a parameter vector $\mathbf{z}$ is represented by a density function $p(\mathbf{z})$. Two independently-obtained states of information (representing e.g. theoretical and observational knowledge) can be combined according to

$$p_i(\mathbf{z}) \wedge p_j(\mathbf{z}) = \frac{p_i(\mathbf{z}) p_j(\mathbf{z})}{\mu(\mathbf{z})}, \tag{C1}$$

where $\mu(z)$ denotes the state of complete ignorance. In classical probability theory an equivalent conjunction of states of information can be achieved as follows:

Let the set of propositions a represent a body of knowledge, e.g. *a priori* information. The conditional probability $P(z \mid a)$ gives the probability of z given the *a priori* information. Similarly $P(z \mid t)$ is the theoretical state of information about z when t denotes the body of theoretical knowledge. Combining theoretical and *a priori* knowledge about z is, of course, equivalent to deriving the probability of z conditional on the conjunction of a and t: $P(z \mid a \wedge t)$.

This can be derived, starting with the definition of conditional probability:

$$P(z \wedge a \wedge t) = P(t \mid z \wedge a) P(z \wedge a). \tag{C2}$$

When a and t are independent this can be resolved further to

$$P(z \wedge a \wedge t) = P(t \mid z) P(z \mid a) P(a). \tag{C3}$$

Using Bayes' rule,

$$P(z \mid a \wedge t) P(a) P(t) = \frac{P(z \mid t) P(t)}{P(z)} P(z \mid a) P(a) \tag{C4}$$

or

$$P(z \mid a \wedge t) = \frac{P(z \mid a) P(z \mid t)}{P(z)}, \tag{C5}$$

$P(z)$ is the marginal probability of z, i.e. the probability when disregarding all other knowledge (*a priori* and theoretical). Hence $P(z)$ represents the state of complete ignorance. The equivalent form for continuous vectors of variables is:

$$p(\mathbf{z} \mid a \wedge t) = \frac{p(\mathbf{z} \mid a) p(\mathbf{z} \mid t)}{p(\mathbf{z})}, \tag{C6}$$

which is (C1) in a different notation!

Note that the intuitive (?) demand of Tarantola and Valette (1982a, b) that states of information be independent in order to allow their conjunction by (C1) explicitly occurs in the derivation of (C5) from the basics of probability theory. However, one

should not conclude too hastily that the two theories are identical. After all, Tarantola's conjunction of states of information is derived without the concept of conditional probability. The latter is rather a result of the first. This situation is reversed in classical probability theory. It is questionable whether a formal definition of independence of states of information can be given without the concept of conditional probability. Nevertheless, it is interesting to see how equivalent results are obtained from different starting points and intuitive notions.

The consequence of the equivalence of (C1) and (C6) is, of course, that any result from Tarantola's theory can also be derived within classical probability theory. For more details on practical consequences and the relation to the maximum entropy formalism see Duijndam (1987).

REFERENCES

BARD, Y. 1974. *Nonlinear Parameter Estimation.* Academic Press Inc.

BARNETT, V. 1982. *Comparative Statistical Inference*, 2nd edn. John Wiley & Sons.

CLAERBOUT, J.F. and MUIR, F. 1973. Robust modelling with erratic data. *Geophysics* **38**, 826–844.

COOKE, D.A. and SCHNEIDER, W.A. 1983. Generalized linear inversion of reflection seismic data. *Geophysics* **48**, 665–676.

DE FINETTI, B. 1974. *Theory of Probability*, Vol. 1. John Wiley & Sons.

DUIJNDAM, A.J.W. 1987. Detailed Bayesian inversion of seismic data. Ph.D. thesis, Delft University of Technology, Delft.

DUIJNDAM, A.J.W. 1988. Bayesian estimation in seismic inversion. Part II: Uncertainty analysis. *Geophysical Prospecting* **36**, 899–918.

DUIJNDAM, A.J.W., VAN RIEL, P. and KAMAN, E.J. 1984. An iterative scheme for wavelet estimation and seismic section inversion in reservoir seismology. 54th SEG meeting, Atlanta, Expanded Abstracts 521–523.

FRANKLIN, J.N. 1970. Well-posed stochastic extensions of ill-posed linear problems. *Journal of Mathematical Analysis and Applications* **31**, 682–716.

GILL, P.E., MURRAY, W. and WRIGHT, M.H. 1981. *Practical Optimization.* Academic Press Inc.

GJÖSTDAL, H. and URSIN, B. 1981. Inversion of reflection times in three dimensions. *Geophysics* **46**, 972–983.

JACKSON, D.D. 1979. The use of a priori information to resolve non-uniqueness in linear inversion. *Geophysical Journal of the Royal Astronomical Society* **28**, 97–109.

JAYNES, E.T. 1968. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics* **SSC-4**, 227–241.

JAYNES, E.T. 1985. Where do we go from here? *Maximum-Entropy and Bayesian Methods in Inverse Problems* (ed. by C. R. Smit and W. T. Grandy), pp. 21–58. D. Reidel Publishing Company.

JEFFREYS, H. 1939. *Theory of Probability.* Clarendon Press.

JEFFREYS, H. 1957. *Scientific Inference.* Cambridge University Press.

KAMAN, E.J., VAN RIEL, P. and DUIJNDAM, A.J.W. 1984. Detailed inversion of reservoir data by constrained parameter estimation and resolution analysis, 54th SEG meeting, Atlanta, Expanded Abstracts, 652–655.

MENDEL, J.M. 1983. *Optimal Seismic Deconvolution.* Academic Press Inc.

MILLER, K.S. 1975. *Multidimensional Gaussian Distributions.* John Wiley & Sons.

MOOD, M., GRAYBILL, F.A. and BOES, D.C. 1974. *Introduction to the Theory of Statistics,* 3rd edn. McGraw-Hill Book Co.

PAPOULIS, A. 1965. *Probability, Random Variables and Stochastic Processes.* McGraw-Hill Book Co.

ROTHMAN, D.H. 1985. Nonlinear inversion, statistical mechanics, and residual statics estimation. *Geophysics* **50**, 2797–2807.

ROTHMAN, D.H. 1986. Automatic estimation of large residual statics corrections. *Geophysics* **51**, 332–346.

SAVAGE, L.J. 1954. *The Foundations of Statistics.* John Wiley & Sons.

SCALES, L.E. 1985. *Introduction to Nonlinear Optimization.* Macmillan.

SHANNON, C.E. 1948. A mathematical theory of communication. *Bell System Technical Journal* **27**, 379–423 and 623–656.

TARANTOLA, A. 1986. A strategy for nonlinear elastic inversion of seismic data. *Geophysics* **51**, 1893–1903.

TARANTOLA, A. 1987. *Inverse Problem Theory, Methods for Data Fitting and Model Parameter Estimation.* Elsevier Science Publishers.

TARANTOLA, A. and VALETTE, B. 1982a. Inverse problems = quest for information. *Journal of Geophysics* **50**, 159–170.

TARANTOLA, A. and VALETTE, B. 1982b. Generalized nonlinear inverse problems solved using the least squares criterion. *Reviews of Geophysics and Space Physics* **20**, 219–232.

VAN DER MADE, P.M. 1988. Determination of macro subsurface models by generalized inversion. Ph.D. thesis, Delft University of Technology, Delft.

WIGGINS, R., LARNER, K. and WISECUP, D. 1976. Residual statics analysis as a general linear inverse problem. *Geophysics* **41**, 922–938.