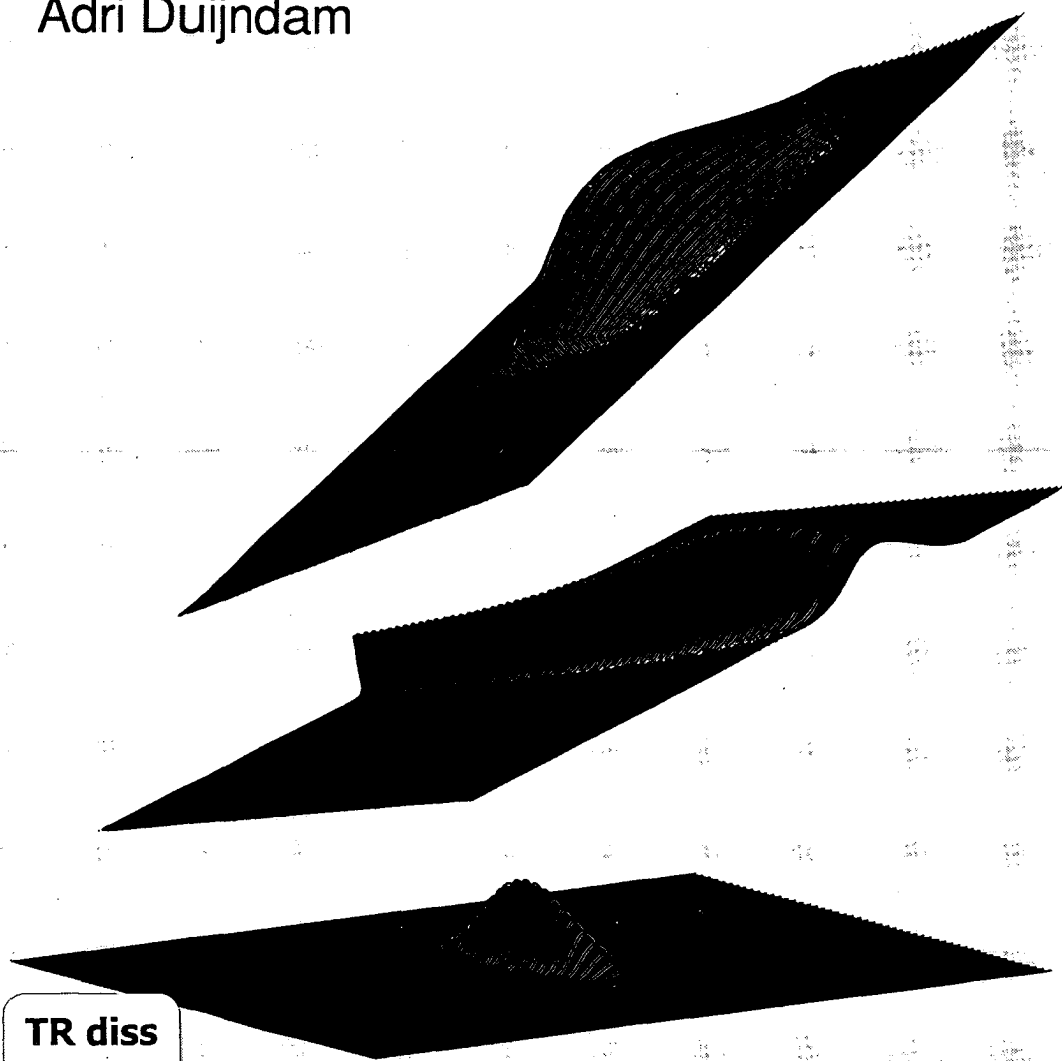


Detailed Bayesian inversion of seismic data

Adri Duijndam



**TR diss
1568**

436.0r6

Detailed Bayesian inversion of seismic data

317842

7A class 1568

DETAILED BAYESIAN INVERSION OF SEISMIC DATA



PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus,
prof.dr. J.M. Dirken,
in het openbaar te verdedigen
ten overstaan van een commissie,
aangewezen door het College van Dekanen
op donderdag 24 september 1987 te 16.00 uur door

ADRIANUS JOSEPH WILLIBRORDUS DUIJNDAM

geboren te Monster
natuurkundig ingenieur

Gebotekst Zoetermeer / 1987

TR diss
1568

Dit proefschrift is goedgekeurd door de promotor
prof.dr.ir. A.J. Berkhout

Copyright © 1987, by Delft Geophysical, Delft, The Netherlands.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of Delft Geophysical, P.O. Box 148, 2600 AC Delft, The Netherlands.

CIP-DATA KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Duijndam, Adrianus Joseph Willibrordus

Detailed Bayesian inversion of seismic data / Adrianus Joseph Willibrordus Duijndam. –

[S.l. : s.n.] (Zoetermeer : Gebotekst). – Ill.

Thesis Delft. – With ref.

ISBN 90-9001781-X

SISO 562 UDC 550.34(043.3)

Subject headings: seismology / Bayesian parameter estimation.

cover design: Adri Duijndam

typesetting and lay-out: Gerda Boone

printed in the Netherlands by: N.K.B. Offset bv, Bleiswijk

Aan mijn ouders

Preface

My involvement in research on detailed inversion of poststack data started with my participation in the industry sponsored Princeps project in the group of seismics and acoustics at Delft University. In this project trace inversion was approached from a parameter estimation point of view. In the beginning the project was very strongly directed towards proving the feasibility of initial versions of the detailed inversion scheme. Nonlinear least squares data fitting was used with hard constraints on the parameters to prevent implausible results. The accuracy of the scheme was determined by Monte Carlo experiments, using an inefficient optimization scheme at the time. The speed, main memory and disk space of the group's 16 bit minicomputer were grossly inadequate for such experiments, so an array processor at the group of signal processing was used in the evening hours. Data files were transported over the department network. A large amount of effort was spent on getting things computed and displayed, while relatively little information was gained. Gradually during the project things improved. A 32 bit minicomputer with much larger memory was purchased together with scientific and graphics software packages. A more efficient optimization algorithm speeded up the optimization with a factor of about 100, while the array processor was not even used anymore. Gradually therefore there came more time to think things over, especially in a period when, due to a fire in the building, the computer was not available for several weeks and a sponsor meeting had to be canceled. It was in this time that I found that the theoretical basis of what we were doing could only be provided by Bayesian parameter estimation and became aware of the controversy in the statistical world between the Bayesian and the classical approach. Out of a desire to really understand what we were practicing I studied a bit more but soon found out that this topic is a research field by itself. I guess that it is inevitable that we build our empires on quicksand (at least partly). When the Princeps 1 period expired I continued this research at Delft geophysical with tight contacts with the Princeps 2 project. It was in this period that the topic of uncertainty analysis became much clearer defined and that the software on multi trace inversion was finished and tested, together with the integration of wavelet estimation.

Many people helped and stimulated me in this research. I would like to thank the researchers involved in or closely connected to the Princeps project, Paul van Riel, Erik Kaman, Pieter van der Made, Gerd Jan Lörtzer and Johan de Haas for many enthusiastic and stimulating discussions. Piet Broersen of the group of signal processing of Delft University appeared to have developed a scheme for parameter selection that I was looking for. I thank him for discussions on this topic and for providing the software of his scheme. Thanks are also due to Ad van der Schoot and Alex Geerlings for reviewing parts of the manuscript. Special thanks are due to my promotor and supervisor of the Princeps project, Prof. Berkhout, who stimulated and promoted this research strongly. Prof. Ted Young co-supervised the project. I thank the oil company that provided the field data set used in chapter 8.

In preparing this thesis I had much help from Tinka van Lier and Tsjander Sardjoe of Delft Geophysical, who respectively typed part of the manuscript and finalized most of the figures. Rinus Boone gave me much advice concerning the form of this thesis and prepared a number of graphs. Gerda Boone of Gebotekst prepared the final version of the text.

I would like to express my sincere gratitude to the management of Delft Geophysical for having given me the opportunity to do this research and to write this thesis, especially in a time when the oil industry was not particularly booming. I thank my dear family and friends for actively stimulating me and for their faith that I would survive the period of working on this thesis with all my faculties intact.

Delft, August, 1987

Adri Duijndam

CONTENTS

INTRODUCTION

1

1 PROBABILITY THEORY AND PRINCIPLES OF BAYESIAN INVERSION

1.1.	Introduction	7
1.2.	Interpretation of the concept of probability	8
1.3.	Fundamentals of probability theory	9
1.4.	Joint probabilities and densities	11
1.5.	Marginal and conditional pdf's	12
1.6.	Expectation, covariance and correlation	15
1.7.	Often used probability density functions	16
1.8.	Bayes' rule as a basis for inverse problems	18
1.9.	The likelihood function	20
1.10.	A priori information	22
1.11.	Point estimation	23
1.12.	The selection of the type of pdf	26
1.13.	A two-parameter example	28
1.14.	Inversion for uncertain data	33
1.15.	The maximum entropy formalism	35
1.16.	The approach of Tarantola and Valette	38
1.17.	Discussion and conclusions on general principles	43

2 OPTIMIZATION

2.1.	Introduction	47
2.2.	Newton optimization methods	48
2.3.	Special methods for nonlinear least squares problems	49
2.4.	An efficiency indication	51
2.5.	Constraints	53
2.6.	Scaling.	53
2.7.	The NAG library	54

3 UNCERTAINTY ANALYSIS

3.1.	Introduction	55
3.2.	Inspection of the pdf's along principal components	56
3.2.1.	Principles	57
3.2.2.	Scaling or transformation of parameters	58
3.2.3.	The least squares problem	59
3.2.4.	The two-parameter example	62
3.3.	The a posteriori covariance matrix	64
3.3.1.	Computation	64
3.3.2.	The two-parameter example	66
3.3.3.	Linearity check	69
3.3.4.	Standard deviations and marginal pdf's	69
3.4.	Sampling distributions	70
3.5.	The resolution matrix	72
3.6.	Summary of uncertainty analysis	73

4 GENERAL ASPECTS OF DETAILED INVERSION OF POSTSTACK DATA

4.1.	Introduction	77
4.2.	Design considerations	78
4.3.	Some methods described in literature	79
4.4.	Outline of a procedure for detailed inversion of poststack data	81

5 SINGLE TRACE INVERSION GIVEN THE WAVELET

5.1.	Introduction	85
5.2.	Formulation of the inverse problem	86
5.3.	Optimization and derivatives	88
5.4.	Fast trace and derivative generation	89
5.5.	Linearity and sensitivity analysis	92
5.6.	Results for single trace inversion	95
5.7.	Inversion of a section by sequential single trace inversion	105

6 MULTI TRACE INVERSION GIVEN THE WAVELET

6.1.	Introduction	109
6.2.	Parameterization	109
6.3.	The estimator and derivatives	112
6.4.	Basic steps in multi trace inversion	113
6.5.	Practical incorporation of a priori information	115
6.6.	Results for multi trace inversion	117

7 WAVELET ESTIMATION

7.1.	Introduction	133
7.2.	The forward model and parameterization	135
7.3.	The Gauss-Markov estimator	136
7.4.	Incorporation of a priori information	141
7.5.	Ridge regression	142
7.6.	Stabilization using the SVD	146
7.7.	Parameter selection	150
7.8.	Reflectivity errors	158
7.9.	Conclusions	160

8 INTEGRAL ESTIMATION OF ACOUSTIC IMPEDANCE AND THE WAVELET

8.1.	Introduction	161
8.2.	A consistent Bayesian scheme	162
8.3.	An iterative scheme	164
8.4.	Results on synthetic data	166
8.5.	A real data example	171

9 EVALUATION OF THE INVERSION SCHEME

9.1.	Introduction	185
9.2.	Refinements and extensions of the scheme	186
9.3.	Deterministic versus statistical techniques	188
9.4.	Testing of the inversion procedure	189
9.5.	Final conclusions	195

APPENDIX A

Bayes' rule and uncertain data	197
--	-----

APPENDIX B

The response of a thin layer	200
--	-----

REFERENCES	202
----------------------	-----

SUMMARY	210
-------------------	-----

SAMENVATTING	212
------------------------	-----

CURRICULUM VITAE	214
----------------------------	-----

INTRODUCTION

DETAILED INVERSION IN SEISMIC PROCESSING

Seismics and other geophysical techniques aim to provide information about the subsurface. The processing of data obtained should solve an inverse problem: the estimation of parameters describing geology. In seismics one can distinguish two different approaches to the seismic inverse problem. The first approach is usually referred to as inverse scattering. In its most general formulation this approach attempts to estimate all relevant (elastic) subsurface parameters and accounts for all aspects of wave propagation. Strategies to solve this problem are being developed (Tarantola, 1986), leading to a nonlinear inverse problem for the complete multi offset data set. The amount of computational power needed for this approach however is extremely large, considering today's hardware technology. Even somewhat simplified formulations, with the parameters of density and propagation velocity in an acoustic formulation are computationally very demanding. Schemes for this approach are not operational yet and are rather topics of research.

The second, conventional, approach aims in the first place at obtaining an image of the subsurface in terms of reflectivity. Figure 1. (after Berkhout, 1985) depicts the processing flow used in this approach. The preprocessing involves steps like demultiplexing, trace editing, application of static corrections and predictive deconvolution. After the preprocessing three different branches can be followed, all leading to a bandlimited

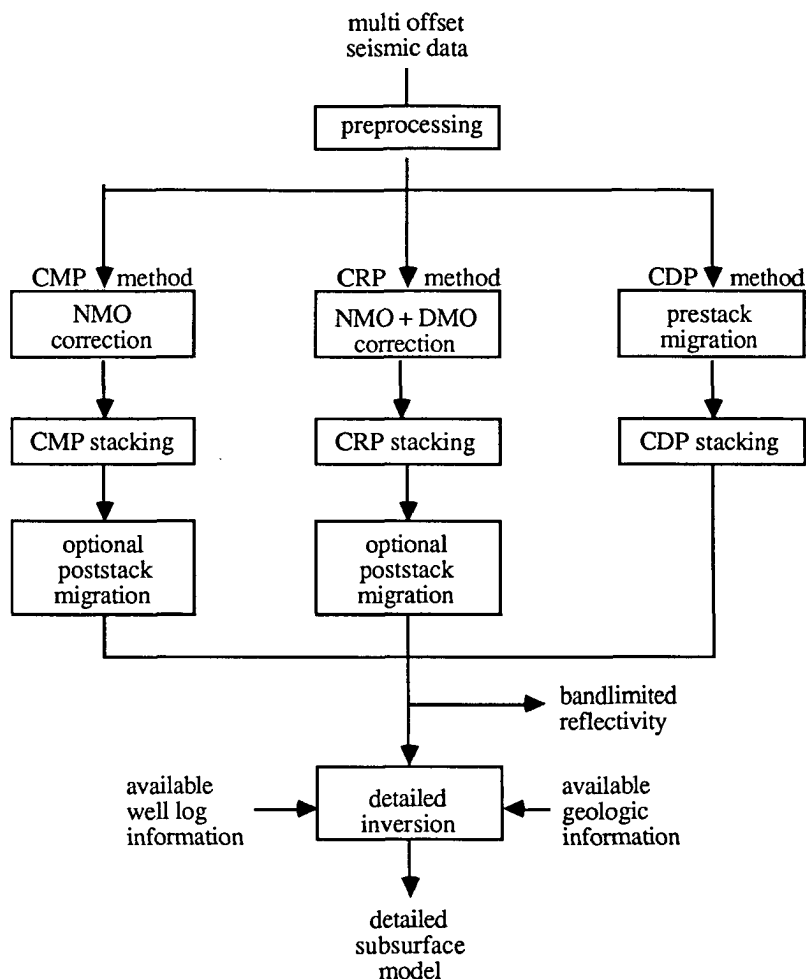


Figure 1 Processing flow to obtain a detailed subsurface model from multi offset seismic data (after Berkhout, 1985).

reflectivity image of the subsurface, either in time or in depth. The left branch is called the CMP (common midpoint) method and is industry standard. Normal moveout correction is applied to CMP gathers to correct for traveltime differences. The traces of the gathers are stacked. This results in a data reduction and an improvement of the signal to noise ratio. The data set now obtained can be regarded as a pseudo zero-offset section. The CRP (common reflection point) method, also often applied in processing nowadays, involves besides NMO correction the so-called DMO (dip moveout) correction, giving a more accurate result than the CMP method. In both methods optionally a poststack migration can be applied. Industry standard is time migration, in which the output is reflectivity as a

function of vertical time. The aim is mainly to focus diffraction energy. The other possibility, at this point in time still less often applied, is depth migration, which yields a reflectivity image as a function of depth. A more accurate so-called macro velocity model of the subsurface is required for depth migration. Both branches assume that the velocity distribution is not very complex. If this assumption is violated the results may be poor and one will have to use the CDP (common depth point) method of the right branch in order to obtain a good image. The CDP method involves migration of prestack data and true common depth point stacking. Because it is computationally much more demanding and requires an accurate macro velocity model it is not widely applied in the industry yet. Its superior results however for geologically complex media have been demonstrated (by e.g. V.d. Schoot et al., 1987).

As mentioned above all three branches yield a bandlimited reflectivity image (usually referred to as a section) as output. Common practice is that an interpreter tries to derive the geologic knowledge he requires from this section. Very often his main interest is in a specific target zone of the data, concerning a (possible) reservoir. Due to the bandlimitations in time or depth however it is often very difficult to derive detailed geology from the data. Because of the width of the main lobe of the wavelet (which can be regarded as the bandlimiting filter) and the interference of its sidelobes properties of thin layers, like thicknesses and e.g. acoustic impedances can not visually be determined from a poststack data set, even after optimal wavelet deconvolution. It is for this reason that techniques have been developed which aim at a detailed inversion for thin layer properties. Their position in the processing flow is depicted in figure 1. This detailed inversion is the main topic of this thesis. It can be of great help to the interpreter. Besides the seismic data well log data and available geological information, provided by the interpreter, should be used. Because of this last aspect detailed inversion is a good example of interactive and interpretive processing and is very well suited for workstation environments.

INVERSE THEORY

The problem of deriving detailed geology from seismic data is a typical example of a general and essential element in empirical science, viz. the drawing of inferences from observations. For any quantitative drawing of inferences from observational data a conceptual model of reality is prerequisite. The conceptual model of the part of reality under study can often partly be described in mathematical or physical/mathematical language. The mathematical model will contain free parameters that have to be estimated. An inverse problem can now be defined as the problem of estimating those parameters using observational data. The related theory is called (parametric) inverse theory.

Theoretical relations between parameters and (potential) data are of course essential in an inverse problem. The problem of the computation of synthetic data, given values for the parameters, is called the forward problem. It is in general substantially easier to solve than the corresponding inverse problem, in more than one aspect.

In a practical inverse problem we always have to deal with uncertainties. Therefore an inverse problem should be formulated using probability theory. Many inverse problems or processing steps in seismics are indeed formulated as statistical or parametric inverse problems. Examples are, apart from the topics covered in this thesis:

- the general inverse scattering problem (Tarantola, 1984, 1986)
- residual statics correction (Wiggins et al., 1976, Rothman, 1985, 1986)
- estimation of velocity models (Gjøstøl and Ursin, 1981, V.d. Made et al., 1984)
- wavelet and Q-factor estimation (Rosland and Sandvin, 1987)

In other geophysical disciplines statistical inverse theory is also very often used. A recent example in electrical sounding is Pous et al. (1987).

Well known problems in inversion, when using only data, are the related items of nonuniqueness, ill-posedness and instability. In practice, these problems can be overcome by using a priori information on the parameters, provided that this is available. The most fundamental and straightforward way to do so is to utilize the so-called Bayesian approach to inversion, which will be the cornerstone of the estimation problems discussed in this thesis.

As mentioned above the imaging of the earth's interior using seismic data is a typical inverse problem. Many efforts are made to improve the processing of seismic data and to extract information from it. This is not surprising regarding the economical interests involved. In processing and information extraction steps all kinds of tricks are used in order to get a "good result". From a scientific viewpoint these tricks must have a basis in certain (possibly limiting) assumptions or in the utilization of extra information. It is essential for a proper evaluation of results that the basic concepts and assumptions of a data processing or information extraction step are in the open and clear. It is for this reason mainly, and because of the fact that the results of Bayesian estimation are used more and more in geophysics while the basic principles are not very well known, that the fundamentals of probability theory and Bayesian estimation are fairly extensively discussed in this thesis. Although much of the material of chapters 1 and 2 will be well known to anyone with experience in nonlinear inversion it is nonetheless given so that any geophysicist will be able to understand the material from the starting points.

THE OUTLINE OF THIS THESIS

This thesis can be divided in two parts. The first part consists of the first three chapters, is devoted to inverse theory and as such is applicable to any inverse problem, inside or outside the realm of physics. Part 2, from chapter 4 onwards, discusses the application of Bayesian estimation to the detailed inversion of poststack seismic data.

The principles of Bayesian estimation are described in chapter 1. Some relations are discussed between the utilization of Bayes' rule when data is obtained in the form of numbers and more general formulations for "uncertain data", like Jeffrey's method of probability kinematics, the maximum entropy principle and the informational approach of Tarantola and Valette. There is a beautiful consistency between the various approaches. In practice usually the maximum of the so-called a posteriori probability density function is used as an estimator for the parameters. For nonlinear models this maximum has to be found by optimization algorithms. A brief overview of the methods most relevant for the estimation problems discussed in this thesis is given in chapter 2. It is more and more realized in geophysics that only an estimate for parameters is not a complete answer to an inverse problem. Like any physical measurement an idea about uncertainties is necessary. In chapter 3 a number of practical methods for uncertainty analysis are discussed.

Chapter 4 is an introduction to the detailed inversion of poststack seismic data. General aspects are considered, the literature briefly reviewed and the general setup of a strategy for inversion is given. Chapters 5 and 6 describe inversion in terms of acoustic impedances and travel times assuming that the wavelet is known. In chapter 5 a single trace approach is discussed. In chapter 6 this is extended to a multi trace approach, a quasi two-dimensional approach. Because the wavelet is never known in practice it has to be estimated as well. This topic is discussed in chapter 7, where it is assumed that the reflectivity of the target zone is known (from a well log for example). In chapter 8 the material of chapters 5, 6 and 7 are combined in a proposed scheme for the integral estimation of the wavelet and the acoustic impedance profile. Results on real data are shown. Chapter 9 finally gives concluding remarks and a critical analysis of the results obtained.

PROBABILITY THEORY AND PRINCIPLES OF BAYESIAN INVERSION

1.1 INTRODUCTION

It will be clear from the formulation of an inverse problem as given in the introduction that inverse theory has a scope much wider than geophysics or physics alone. This gives great interest to the fact that controversy concerning the fundamental way to tackle this type of problems has been raging during this century. The two major schools, stated somewhat oversimplified, are the classical or frequentist school and the Bayesian school. The differences of opinion are not merely philosophical but strongly influence statistical practice.

This thesis is based on Bayesian parameter estimation. Because most geophysicists will not be very familiar with statistics the basics of probability theory and Bayesian estimation are discussed in this chapter. Basic concepts are discussed in the first few sections. It is shown how Bayes' rule can be used to solve an inverse problem and what the relevant functions in the solution are. Point estimation is discussed as a practical procedure. The mathematical form of a point estimator depends on the type of probability density functions chosen and some considerations concerning this choice are therefore given. Bayesian estimation is then illustrated with a two-parameter example. In the last sections the relation to more general methods that allow inversion of "uncertain data" is discussed. It is shown that the different approaches are consistent with each other.

1.2 INTERPRETATION OF THE CONCEPT OF PROBABILITY

The parting of ways in statistical practice is to some extent due to a difference in the interpretation of the concept of probability. There is a vast amount of literature on the foundations of statistics and the interpretation of the concept of probability. In the literature different classifications are given. The following interpretations can be distinguished:

- a) the classical interpretation.
- b) the frequency interpretation.
- c) the Bayesian interpretation.
 - c.1) the logical interpretation.
 - c.2) the subjective interpretation.

The classical interpretation should not be confused with the "classical school", which adopts the frequency interpretation. In the classical *interpretation* the probability of an event A occurring in an experiment is defined to be the ratio of the number of outcomes which imply A to the total number of possible outcomes, provided the latter are equally likely. When for example the outcomes 1 to 6 of throwing a die are considered equally likely then the probability that a 3 will occur as the result of a throw is 1/6. The major criticism of this definition is that it is circular. "Equally likely" can only mean "equally probable". Furthermore this definition seriously limits the applicability of probability theory. For these reasons the classical interpretation is not considered as a serious contender. With respect to the relative frequency interpretation several definitions have been formulated (see Jeffreys (1939) for a critical discussion of them). The best known definition is associated with the name of Von Mises (1936, 1957). The probability of an event A occurring in an experiment is the limit of the relative frequency n_A/n of the occurrences of the event A:

$$P(A) = \lim_{n \rightarrow \infty} \frac{n_A}{n}, \quad (1.1)$$

where n_A is the number of trials in which A is obtained and n the total number of trials. With "trials" is meant repetitions of the experiment under identical circumstances. The definition aims at providing an objective and empirical tool for evaluating probabilities.

Fundamental objections raised concern amongst others the problem that the definition can never be used in practice because the number of trials is always finite. The limit can only be *assumed* to exist. Furthermore serious difficulties arise with the precise definition of "repetition under identical circumstances". The frequency interpretation is also limited in its application, see e.g. Cox (1946). It can give no meaning to the probability of a hypothesis (Jeffreys, 1939).

The Bayesian interpretation owes its name to Thomas Bayes, who first formulated the principle of inverse probability (in a paper published posthumously in 1763). Although in

principle it can be subdivided into two different subclasses, a common element is that probability is interpreted as a "degree of belief". As such probability theory can be seen as an extension of deductive logic and is also called inductive logic. Whereas in deductive logic a proposition can either be true or false, in inductive logic the probability of a proposition constitutes a degree of belief, with proof or disproof as extremes.

The Bayesian school can be subdivided corresponding to two different interpretations. In the so called logical interpretation probability is objective, an aspect of the "state of affairs". In the subjective interpretation the degree of belief is a personal degree of belief. Subjective probability is simply used to reflect a person's ideas or knowledge. The only restriction on the utilization of probabilities is that it is consistent, i.e. that the axioms of probability theory are not violated. Proponents of the logical interpretation are Keynes (1929), Jeffreys (1939, 1957), Carnap (1962), Jaynes (see e.g. his publications of 1968 and 1985), Box and Tiao (1973) and Rosenkrantz (1977). Outspoken proponents of the subjective interpretation are De Finetti (1974) and Savage (1954). Further Bayesian writings are Lindley (1974) and Tarantola (1987). An extensive overview and comparison of interpretations of probability and the resulting ways of practicing statistics is given by Barnett (1982).

Most authors are outspoken proponents of one of the interpretations but there are also authors like Carnap (1962) who state that more than one interpretation can rightly be held and that different situations simply ask for different interpretations.

The interpretation of the probability concept is not the only reason for the adoption of one approach or another. In later parts of this thesis further comparisons are made. To the author a Bayesian interpretation seems conceptually clearer. As will be demonstrated later, it also gives superior results. The rest of this chapter is developed in a Bayesian setting. The question whether the logical or the subjective interpretation is preferable is left aside. No practical consequences for seismic inversion are envisaged.

1.3 FUNDAMENTALS OF PROBABILITY THEORY

There is more than one way to erect an axiomatic structure of probability theory. It is beyond the scope of this thesis however to discuss these matters in detail. Nevertheless an outline of an axiomatic structure is discussed. This allows the reader to fully follow the line of reasoning from some very basic postulates to all the consequences of Bayesian estimation in seismic inversion. It is not the intention of the author to give a rigorous mathematical and logical treatment nor to present the best thought-out axiomatic structure. Instead a set of simple and often given postulates is given as a basis and the theory is worked out from there.

Let q_i denote a proposition. The conjunction of propositions q_i ($i=1, \dots, n$), denoted by $q_1 \wedge q_2 \wedge \dots \wedge q_n$ is the proposition that all q_i are true (logical "and"). The disjunction of the propositions q_i ($i=1, \dots, n$), denoted by $q_1 \vee q_2 \vee \dots \vee q_n$ is the proposition that at least one of the q_i is true (logical "or"). The disjunction is also called the logical sum and is denoted by $\sum_i q_i$. The set of propositions q_i ($i=1, \dots, n$) are jointly exhaustive if at least one of the q_i is true. The propositions are mutually exclusive if only one of them can be true.

The probability $P(q_i)$ assigns a number to the proposition q_i , representing the degree of belief that q_i is true. It is defined to satisfy the following axioms:

$$(1) \quad P(q_i) \geq 0 \quad . \quad (1.2)$$

The probability of a true proposition t is one:

$$(2) \quad P(t) = 1 \quad . \quad (1.3)$$

If the propositions q_i ($i=1, \dots, n$) are mutually exclusive then:

$$(3) \quad P\left(\sum_i q_i\right) = \sum_i P(q_i) \quad (1.4)$$

Axiom (3) is called the additivity rule. From these axioms it follows that for any proposition a :

$$0 \leq P(a) \leq 1 \quad , \quad (1.5)$$

$$P(a \wedge \neg a) = 0 \quad , \quad (1.6)$$

and

$$P(\neg a) = 1 - P(a) \quad , \quad (1.7)$$

where $\neg a$ denotes the negation of a ("not" a). Property (1.6) specifies that a false proposition has probability zero. Property (1.5) specifies that the probability of a proposition may take values between zero (for a false proposition) and 1 (for a true proposition).

Note that in the postulates the probability of a proposition is directly defined on a number scale. In the philosophically more fundamental theory of Jeffreys (1939) the axioms concern statements about probability only. Numbers are only a way to code or represent probabilities and are therefore introduced by conventions in his theory. In fact, the axioms (1.3) and (1.4) are conventions in his system and axiom (1.2) follows from these conventions and his more fundamental axioms.

As an example consider a discrete variable X that can have one and only one of the values x_i ($i=1, \dots, n$). It follows from the axioms that:

$$P(X=x_i) \geq 0 \quad , \quad i = 1, \dots, n \quad , \quad (1.8)$$

$$P(X=x_i \vee X=x_j) = P(X=x_i) + P(X=x_j) \quad , \quad i \neq j \quad (1.9)$$

and

$$\sum_{i=1}^n P(X=x_i) = 1 \quad . \quad (1.10)$$

Let us now consider a continuous variable X . The distribution function $F_X(x)$ is defined as:

$$F_X(x) = P(X \leq x) \quad . \quad (1.11)$$

From the axioms it has the properties:

$$F(-\infty) = 0 \quad , \quad (1.12)$$

$$F(\infty) = 1 \quad . \quad (1.13)$$

Furthermore:

$$P(x_1 < X \leq x_2) = F(x_2) - F(x_1) \quad . \quad (1.14)$$

It can be shown that F is a nondecreasing function of x :

$$F(x_1) \leq F(x_2) \quad \text{for } x_1 \leq x_2 \quad . \quad (1.15)$$

The probability density function (pdf) of X is defined as:

$$p(x) = \frac{dF(x)}{dx} \quad . \quad (1.16)$$

Its integral over a certain range is the probability that X takes a value in that range:

$$P(x_1 < X \leq x_2) = \int_{x_1}^{x_2} p(x) dx \quad . \quad (1.17)$$

From the axioms it follows:

$$p(x) \geq 0 \quad , \quad (1.18)$$

and

$$\int p(x) dx = 1 \quad . \quad (1.19)$$

1.4 JOINT PROBABILITIES AND DENSITIES

The concept of joint probabilities is entailed in the fundamentals of probability theory as sketched in the previous section. Its definition is: The joint probability of propositions is the probability of their conjunction:

$$P(q_1 \wedge q_2 \wedge \dots \wedge q_n) \quad (1.20)$$

The joint distribution function of a set of random variables X_i is:

$$P(X_1 \leq x_1 \wedge X_2 \leq x_2 \wedge \dots \wedge X_n \leq x_n) = F(x_1, x_2, \dots, x_n) \quad (1.21)$$

Let the sets of variables X_i and x_i be gathered in vectors $\mathbf{X}$ and $\mathbf{x}$ respectively and $F(\mathbf{x})$ be shorthand for $F(x_1, x_2, \dots, x_n)$. The joint pdf of $\mathbf{X}$ then is the straightforward generalization of the univariate case:

$$p(\mathbf{x}) = \frac{\partial^n F(\mathbf{x})}{\partial x_1 \partial x_2 \dots \partial x_n} \quad (1.22)$$

The probability of the vector $\mathbf{X}$ taking values in a volume A is:

$$P(\mathbf{X} \in A) = \int_A p(\mathbf{x}) dx_1 dx_2 \dots dx_n \quad (1.23)$$

or in shorthand:

$$P(\mathbf{X} \in A) = \int_A p(\mathbf{x}) d\mathbf{x} \quad (1.24)$$

From the axioms it follows:

$$p(\mathbf{x}) \geq 0 \quad (1.25)$$

and

$$\int p(\mathbf{x}) d\mathbf{x} = 1 \quad (1.26)$$

From this point onwards a more sloppy notation will be used in order to improve readability. When sets of random variables are partitioned over vectors $\mathbf{x}$ and $\mathbf{y}$ the notation $p(\mathbf{x}, \mathbf{y})$ denotes the joint pdf of $\mathbf{x}$ and $\mathbf{y}$ (and implicitly the joint pdf of all their elements).

1.5 MARGINAL AND CONDITIONAL PDF'S

The concepts of marginal and especially conditional pdf's play a vital role in inverse theory. Although mathematically the concepts can be introduced very quickly they are discussed at some length in order to provide a proper introduction and interpretation in a Bayesian context.

The marginal pdf occurs in the following situation. Suppose the joint pdf $p(\mathbf{x}, \mathbf{y})$ reflects the degree of belief on variables $\mathbf{x}$ and $\mathbf{y}$. The question may arise what the proper form of the pdf $p(\mathbf{x})$ is, when disregarding the variable $\mathbf{y}$. In order to derive an answer the

propositions $a = "x_1 < x \leq x_2"$ and the true proposition $t = "-\infty < y \leq \infty"$ are introduced. From the axioms of probability theory (1.2) – (1.4) and using some propositional calculus (symbolic logic) it can be derived that:

$$P(a \wedge t) = 1 - P(\neg a \vee \neg t) \quad , \quad (1.27)$$

using $(a \wedge t) \Leftrightarrow \neg(\neg a \vee \neg t)$ and (1.7). The propositions $\neg a$ and $\neg t$ are mutually exclusive so that, using (1.4):

$$P(a \wedge t) = 1 - P(\neg a) - P(\neg t) \quad . \quad (1.28)$$

The probability of a false proposition is 0 so that, using (1.7) again:

$$\begin{aligned} P(a \wedge t) &= 1 - (1 - P(a)) - 0 \\ &= P(a) \quad . \end{aligned} \quad (1.29)$$

This result means that the probability of a proposition is identical to the probability of the conjunction of that proposition and a true proposition. Substitution of the propositions a and t in (1.29) yields:

$$P(x_1 < x \leq x_2) = P(x_1 < x \leq x_2, -\infty < y < \infty) \quad , \quad (1.30)$$

or,

$$\int_{x_1}^{x_2} p(x) \, dx = \int_{x_1}^{x_2} \int_{-\infty}^{\infty} p(x, y) \, dy \, dx \quad . \quad (1.31)$$

This yields the definition of the marginal pdf $p(x)$:

$$p(x) = \int_{-\infty}^{\infty} p(x, y) \, dy \quad . \quad (1.32)$$

Apart from the logical derivation the interpretation is clear. Disregarding y means not imposing any bounds on its values. The probability of x taking values in some region, disregarding y , must be equal to the probability of x taking values in that region and y taking values anywhere between $-$ and $+$ infinity! Furthermore relation (1.32) reflects an averaging over y . The marginal pdf $p(x)$ reflects the state of information "on the average".

For the multivariate case with vectors $\mathbf{x}$ and $\mathbf{y}$ (1.32) is readily generalized to:

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{y}) \, d\mathbf{y} \quad . \quad (1.33)$$

In many textbooks the conditional pdf $p(\mathbf{x}|\mathbf{y})$ of $\mathbf{x}$, given $\mathbf{y}$, is only shortly defined:

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{y})} \quad (1.34)$$

Peterka (1980) gives a lucid Bayesian introduction of this concept, which roughly is as

follows. When the state of information on x and y is described by the pdf $p(x,y)$ and the information becomes available that values for y are obtained, how should the pdf of x in this new situation be calculated? Obviously this pdf should be proportional to $p(x,y)$ with the obtained values for y substituted. In order to render $p(x|y)$ a pdf that satisfies the axioms it has to be normalized. It can easily be shown that the result is (1.34).

The definition of conditional probability from which (1.34) can be derived through differentiation is:

$$P(b|a) = \frac{P(a,b)}{P(a)} \quad (1.35)$$

where a and b denote propositions. In all axiomatic descriptions of probability theory (1.35) or a similar expression is introduced through an axiom or a definition. R.T.Cox (1946, 1978) also takes it as an axiom but gives very compelling reasons for doing so. First he argues that the probability $P(a,b)$ should be given by some function G with arguments $P(b|a)$ and $P(a)$:

$$P(a,b) = G(P(a), P(b|a)) \quad (1.36)$$

using a simple example: The probability that a long-distance runner can run from one place to another (a) and can run back the same day (b), should depend on the probability $P(a)$ of his being able to run to that place and the probability $P(b|a)$ that he can run back the same day given the fact that he *has* run the first stretch. Now by demanding that probability theory should be consistent with symbolic logic (or Boolean algebra as he calls it), he derives that, without any loss of generality the simplest solution $G(x,y) = xy$ can be chosen. Equation (1.36) then turns into:

$$P(a,b) = P(b|a) P(a) \quad (1.37)$$

which is equivalent to (1.35).

Two vectors of random variables are defined to be independent when:

$$p(x,y) = p(x) p(y) \quad (1.38)$$

Consequently, for independent x and y :

$$p(x|y) = p(x) \quad (1.39)$$

and

$$p(y|x) = p(y) \quad (1.40)$$

From equation (1.34):

$$p(x|y) = \frac{p(x,y)}{p(y)} \quad (1.34)$$

and the similar relation:

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y}|\mathbf{x}) p(\mathbf{x}) \quad , \quad (1.41)$$

Bayes' rule follows:

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x}) p(\mathbf{x})}{p(\mathbf{y})} \quad . \quad (1.42)$$

In section 1.8 it is shown how this fundamental result can be used as a basis for inverse problems. A useful result for more complex problems is the chain rule which is obtained by repeatedly applying (1.34) on the combination of vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$:

$$p(\mathbf{x}_n, \mathbf{x}_{n-1}, \dots, \mathbf{x}_1) = \left(\prod_{i=2}^n p(\mathbf{x}_i | \mathbf{x}_{i-1}, \dots, \mathbf{x}_1) \right) \cdot p(\mathbf{x}_1) \quad . \quad (1.43)$$

1.6 EXPECTATION, COVARIANCE AND CORRELATION

The expectation of a function $g(\mathbf{x})$ with respect to the pdf $p(\mathbf{x})$ is defined as:

$$E g(\mathbf{x}) = \int g(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \quad . \quad (1.44)$$

In particular, the expectation or mean μ_i of an element x_i of the vector is defined as:

$$\mu_i = E x_i = \int x_i p(\mathbf{x}) d\mathbf{x} \quad . \quad (1.45)$$

Performing the integral first for all elements except x_i we have:

$$\mu_i = E x_i = \int x_i p(x_i) dx_i \quad , \quad (1.46)$$

where

$$p(x_i) = \int p(\mathbf{x}) dx_1 dx_2 \dots dx_{i-1} dx_{i+1} \dots dx_n \quad (1.47)$$

is the marginal pdf $p(x_i)$ for x_i . The mean μ_i for x_i of the multivariate pdf $p(\mathbf{x})$ is (by definition (1.45)) equivalent to the mean of the marginal pdf. The expectation $\boldsymbol{\mu}$ of the vector $\mathbf{x}$ is:

$$\boldsymbol{\mu} = E \mathbf{x} = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x} \quad , \quad (1.48)$$

which is shorthand for the combination of n equations of the form (1.45) for all elements.

The covariance matrix $\mathbf{C}$ is defined by its elements:

$$C_{ij} = E(x_i - \mu_i)(x_j - \mu_j) = \int (x_i - \mu_i)(x_j - \mu_j) p(\mathbf{x}) d\mathbf{x} \quad . \quad (1.49)$$

The combination for all elements can be written in shorthand:

$$\mathbf{C} = E(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T p(\mathbf{x}) d\mathbf{x} \quad , \quad (1.50)$$

where the superscript T denotes transposed. The diagonal of **C** contains the variances $\sigma_i^2 = E(x_i - \mu_i)^2$ of the variables x_i . Their square roots σ_i are the standard deviations of the variables. Note that, as with the mean, the variance and standard deviation of x_i are (by definition) equivalent to those of the marginal pdf:

$$\sigma_i^2 = C_{ii} = E(x_i - \mu_i)^2 = \int (x_i - \mu_i)^2 p(x_i) dx_i \quad (1.51)$$

The correlation coefficient ρ_{ij} of the variables x_i and x_j is defined as:

$$\rho_{ij} = \frac{C_{ij}}{\sigma_i \sigma_j} \quad (1.52)$$

It has the properties:

$$-1 \leq \rho_{ij} \leq 1 \quad (1.53)$$

and

$$\rho_{ii} = \frac{C_{ii}}{\sigma_i^2} = 1 \quad (1.54)$$

The matrix **P** containing the elements ρ_{ij} is called the correlation matrix. The diagonal, by (1.54), contains values of 1.

1.7 OFTEN USED PROBABILITY DENSITY FUNCTIONS

The most often used pdf is the well known Gaussian or normal pdf:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} |\mathbf{C}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\} \quad (1.55)$$

where $\boldsymbol{\mu}$ and **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).

The univariate uniform pdf is given by:

$$\begin{aligned} p(x) &= \frac{1}{2a} & \mu - a \leq x \leq \mu + a \\ &= 0 & x < \mu - a, x > \mu + a \end{aligned} \quad (1.56)$$

Its standard deviation is $a/\sqrt{3}$. A less frequently used distribution is the double exponential or the Laplace distribution. It leads to the so called l_1 -norm estimators as is discussed in section 1.11. This estimator frequently appears in geophysical literature. The expression for the univariate double exponential pdf is:

$$p(x) = \frac{1}{\sqrt{2} \sigma} \exp \left\{ -\sqrt{2} \frac{|x-\mu|}{\sigma} \right\} , \quad (1.57)$$

with μ and σ the mean and standard deviation respectively. In figure 1.1 the univariate forms of the three pdf's discussed above are shown for a zero mean and a standard deviation of 1. From figures 1.1.c,d it can be seen that the double exponential pdf has longer tails than the Gaussian one has.

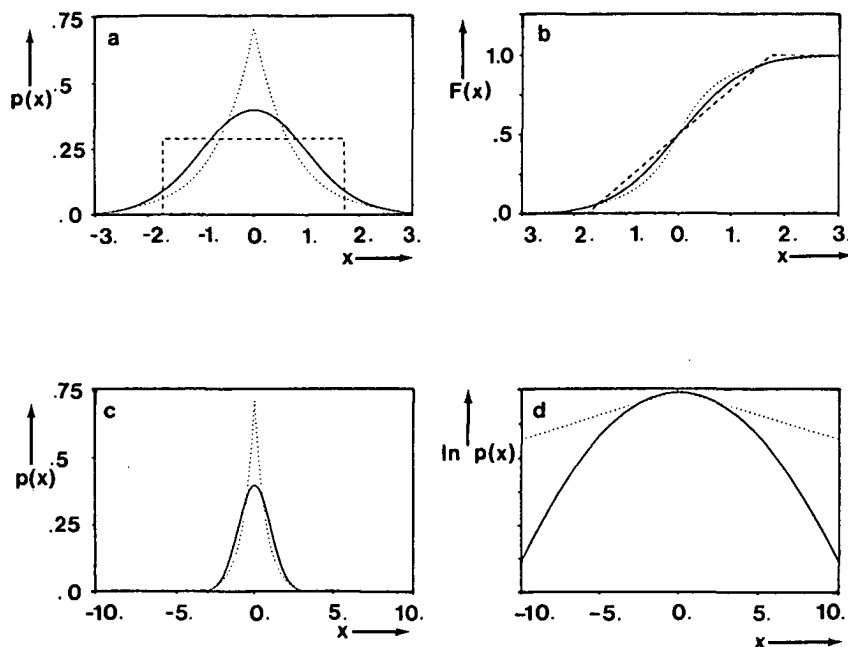


Figure 1.1 One dimensional pdf's with zero mean and a standard deviation of 1.

_____ Gaussian pdf; double exponential pdf; - - - - - uniform pdf
a) pdf's; b) the corresponding distribution functions $F(x) = \int p(x)dx$; c) the Gaussian and double exponential pdf for a wide range; d) as c) but in logarithmic display.

To the knowledge of the author the double exponential 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 one-dimensional pdf's:

$$p(x) = \prod_i \frac{1}{\sqrt{2} \sigma_i} \exp \left\{ -\sqrt{2} \frac{|x_i - \mu_i|}{\sigma_i} \right\} . \quad (1.58)$$

The parameters are uncorrelated. This pdf however can be generalized to the case with nonzero correlations. Consider the multi dimensional pdf:

$$p(\mathbf{x}) = \frac{|W|}{2^{n/2}} \exp \left\{ -\sqrt{2} \|\mathbf{W}(\mathbf{x} - \boldsymbol{\mu})\|_1 \right\} , \quad (1.59)$$

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

$$\|\mathbf{x}\|_1 = \sum_i |x_i| \quad (1.60)$$

The following properties can be derived: $p(\mathbf{x})$ as given in (1.59) is a strict pdf:

$$\int p(\mathbf{x}) d\mathbf{x} = 1 , \quad (1.61)$$

the expectation of $\mathbf{x}$ is given by:

$$\int \mathbf{x} p(\mathbf{x}) d\mathbf{x} = \boldsymbol{\mu} , \quad (1.62)$$

and the covariance matrix $\mathbf{C}$ is given by:

$$\mathbf{C} = \int (\mathbf{x} - \boldsymbol{\mu}) (\mathbf{x} - \boldsymbol{\mu})^T p(\mathbf{x}) d\mathbf{x} = (\mathbf{W}^T \mathbf{W})^{-1} . \quad (1.63)$$

Property (1.63) shows that $\mathbf{W}$ and thereby $p(\mathbf{x})$ as given in (1.63) 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:

$$p(\mathbf{x}) = \frac{1}{2^{n/2} |\mathbf{C}|^{1/2}} \cdot \exp \left\{ -\sqrt{2} \|\mathbf{C}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})\|_1 \right\} . \quad (1.64)$$

This distribution lacks a number of favourable properties that the Gaussian pdf has. A linear transformation of parameters distributed according to (1.64) for example leads to a distribution of the form (1.59), but not necessarily to the form (1.64). Like the Gaussian pdf however, zero correlations imply independence. The specific linear transformation $\mathbf{y} = \mathbf{C}^{-1/2}\mathbf{x}$ renders n independent identically distributed (iid) parameters, with each parameter distributed according to a one-dimensional Laplace distribution with unit standard deviation.

1.8 BAYES' RULE AS A BASIS FOR INVERSE PROBLEMS

A mathematical model describing an aspect of reality will often contain free parameters that have to be estimated. In seismics for example these parameters describe thicknesses and acoustic properties of geological layers in the subsurface of the earth. Let these parameters be gathered in the vector $\mathbf{x}$ and let the vector $\mathbf{y}$ contain discretised data. Suppose $p(\mathbf{x}, \mathbf{y})$

reflects the state of information on x and y before measurements for y are obtained. When data as a result of measurements determine the values of y then, as discussed in section 1.5, the state of information on x should be represented by $p(x|y)$, which is given by Bayes' rule (1.42):

$$p(x|y) = \frac{p(y|x) p(x)}{p(y)} \quad (1.42)$$

The pdf $p(x|y)$ is the so-called a posteriori pdf. The function $p(y|x)$ is the conditional pdf of y given x . As discussed in the next section it contains the theoretical relations between parameters and data including 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 $p(x)$. It is the marginal pdf of $p(x,y)$ for x . It reflects the information on x when disregarding the data and thus it should contain the a priori knowledge on the parameters. The denominator $p(y)$ does not depend on x and can be considered as a constant factor in the inverse problem.

It is important to realize that $p(x|y)$ contains all information available on x given the data y and therefore 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 in section 1.11) is derived from it. Formula (1.42) can also be used without the restriction that all functions are strict pdf's in the sense that their integrals are one. In that case constant factors are considered immaterial (see also Tarantola and Valette (1982a) and Bard (1974)). The functions are then called density functions.

Bayes' rule is especially appealing because it provides a mathematical formulation of how previous knowledge can be updated when new information becomes available. Starting from the prior knowledge $p(x)$ an update is obtained by Bayes' rule when data y_1 becomes available:

$$p(x|y_1) = \frac{p(y_1|x) p(x)}{p(y_1)} \quad (1.65)$$

When additional data y_2 becomes available the new a posteriori pdf is given by:

$$\begin{aligned} p(x|y_1, y_2) &= \frac{p(y_1, y_2|x) p(x)}{p(y_1, y_2)} \\ &= \frac{p(y_2|y_1, x)}{p(y_2|y_1)} \cdot \frac{p(y_1|x) p(x)}{p(y_1)} \end{aligned} \quad (1.66)$$

Note that the second factor on the right hand side is the a posteriori pdf on data y_1 as given in (1.65). When the data vectors y_1 and y_2 are independent (1.66) simplifies to:

$$p(\mathbf{x}|\mathbf{y}_1, \mathbf{y}_2) = \frac{p(\mathbf{y}_2|\mathbf{x})}{p(\mathbf{y}_2)} \cdot \frac{p(\mathbf{y}_1|\mathbf{x}) p(\mathbf{x})}{p(\mathbf{y}_1)} . \quad (1.67)$$

It turns out that the a posteriori pdf after the data $\mathbf{y}_2$ has been obtained is again computed with Bayes' rule with the a priori information given by the a posteriori information on data $\mathbf{y}_1$! This process can be repeated for each new set of data that becomes available. Bayes' theorem thus describes the process of learning from experience.

1.9 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 are treated using the standard reduced model (Bard, 1974):

$$\mathbf{y} = \mathbf{g}(\mathbf{x}) + \mathbf{n} , \quad (1.68)$$

where $\mathbf{g}(\mathbf{x})$ is the forward model, used to create synthetic data. It can be nonlinear. The vector $\mathbf{n}$ contains the errors or noise. When $\mathbf{n}$ is independent of $\mathbf{g}(\mathbf{x})$ and has a pdf p_n it follows:

$$p(\mathbf{y}|\mathbf{x}) = p_n(\mathbf{y} - \mathbf{g}(\mathbf{x})) . \quad (1.69)$$

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 $l(\mathbf{x})$:

$$l(\mathbf{x}) = p(\mathbf{y}=\mathbf{d}|\mathbf{x}) . \quad (1.70)$$

Using equation (1.69):

$$l(\mathbf{x}) = p_n(\mathbf{d} - \mathbf{g}(\mathbf{x})) . \quad (1.71)$$

In literature a distinction is sometimes made between theoretical and observational errors. In seismics for example the neglect of multiples and the utilization of 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 completely arbitrary. This is easily illustrated. Let $\mathbf{f}$ denote an ideal theory. The theoretical errors $\mathbf{n}_1$ are defined as:

$$\mathbf{n}_1 = \mathbf{f}(\mathbf{x}) - \mathbf{g}(\mathbf{x}) . \quad (1.72)$$

Substitution in (1.68) yields:

$$\mathbf{y} = \mathbf{f}(\mathbf{x}) - \mathbf{n}_1 + \mathbf{n} . \quad (1.73)$$

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

$$\mathbf{n}_2 = \mathbf{y} - \mathbf{f}(\mathbf{x}) , \quad (1.74)$$

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

$$\mathbf{n} = \mathbf{n}_1 + \mathbf{n}_2 \quad (1.75)$$

From (1.68) and (1.75) it is already clear that both types of errors are treated in the same way. That the distinction *must* be arbitrary becomes clear 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 and arbitrary to some extent, we might as well declare $\mathbf{g}(\mathbf{x})$ to be "ideal". We then have $\mathbf{f}(\mathbf{x}) = \mathbf{g}(\mathbf{x})$, $\mathbf{n}_1 = 0$ and hence all errors $\mathbf{n} = \mathbf{n}_2 = \mathbf{y} - \mathbf{f}(\mathbf{x})$ are observational! None of the two viewpoints is wrong. The definition of the ideal theory and hence the distinction between theoretical and observational errors is simply arbitrary.

In practice one will nevertheless be inclined to call one type of error theoretical and another observational. The likelihood function can then be derived by introducing $\mathbf{y}_1 = \mathbf{f}(\mathbf{x})$ as the ideal synthetic data, applying the chain rule (1.43) on $p(\mathbf{y}, \mathbf{y}_1, \mathbf{x})$ and integrating over $\mathbf{y}_1$. The result is:

$$p(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{y}|\mathbf{y}_1, \mathbf{x}) p(\mathbf{y}_1|\mathbf{x}) d\mathbf{y}_1 \quad (1.76)$$

Usually the observational errors are assumed to be independent on the parameters $\mathbf{x}$ so that $p(\mathbf{y}|\mathbf{y}_1, \mathbf{x}) = p(\mathbf{y}|\mathbf{y}_1)$. When the pdf p_{n1} of $\mathbf{n}_1$ is independent on $\mathbf{g}(\mathbf{x})$ we have:

$$p(\mathbf{y}_1|\mathbf{x}) = p_{n1}(\mathbf{y}_1 - \mathbf{g}(\mathbf{x})) \quad (1.77)$$

Similarly when the pdf p_{n2} of $\mathbf{n}_2$ is independent on $\mathbf{f}(\mathbf{x})$ we have:

$$p(\mathbf{y}|\mathbf{y}_1) = p_{n2}(\mathbf{y} - \mathbf{y}_1) \quad (1.78)$$

Substitution in (1.76) yields:

$$p(\mathbf{y}|\mathbf{x}) = \int p_{n2}(\mathbf{y} - \mathbf{y}_1) p_{n1}(\mathbf{y}_1 - \mathbf{g}(\mathbf{x})) d\mathbf{y}_1 \quad (1.79)$$

which is recognized as a convolution integral when introducing $\mathbf{z} = \mathbf{y}_1 - \mathbf{g}(\mathbf{x})$ so that:

$$p(\mathbf{y}|\mathbf{x}) = \int p_{n2}(\mathbf{y} - \mathbf{g}(\mathbf{x}) - \mathbf{z}) p_{n1}(\mathbf{z}) d\mathbf{z} \quad (1.80)$$

This result is equivalent to relation (1.69) when:

$$\mathbf{p}_n = \mathbf{p}_{n1} * \mathbf{p}_{n2} \quad (1.81)$$

stating nothing else than the well known fact that the pdf of the sum of two independent vectors of variables is given by the convolution of the pdf's of the terms (see e.g. Mood, Graybill and Boes, 1974). When both pdf's are Gaussian the covariance matrix of the sum

is the sum of the covariance matrices for the two components, a result often stated in inversion literature.

1.10 A PRIORI INFORMATION

Information about the parameters that is available independent of the data can be used as a priori information and is formulated in $p(x)$. This type of information may come from general knowledge about the system under study. An example is general geological knowledge.

A priori knowledge about parameters often consist 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:

$$p(x) = \frac{1}{(2\pi)^{n/2} |C_x|^{1/2}} \exp \left\{ -\frac{1}{2} (x-x^i)^T C_x^{-1} (x-x^i) \right\} \quad , \quad (1.82)$$

where n is the number of parameters, x^i is the mean of the distribution (the guessed values) and 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 section 1.7):

$$p(x) = \frac{1}{2^{n/2} |C_x|^{1/2}} \exp \left\{ -\sqrt{2} \|C_x^{-1/2} (x-x^i)\|_1 \right\} \quad . \quad (1.83)$$

Often additional a priori information can be formulated by hard constraints, preventing the parameters to obtain 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 areas where the a priori distribution is zero.

The specification and utilization of prior distributions have been the subject of dispute for a long time. In fact, besides the differences in interpretation of the concept of probability, it has been the major stumbling-block for adherents of the classical approach. It has been argued that no unique and objective prior distributions can be given. For subjective Bayesians there is no problem. The only formal restriction on the distribution is that the axioms of probability theory are not violated. An a priori distribution may simply reflect an expert's opinion. For a discussion of the practical usage of subjective probability in relation to expert opinion see Cooke (1986). Logical Bayesians consider specification of an a priori pdf inevitable and they try to determine objective priors. In general it holds for the Bayesian viewpoint that a priori knowledge is updated each time new data becomes available. The question of course arises with which distribution this process is started.

Starting distributions should reflect no preference and are therefore called noninformative priors. One would be inclined to think at first sight that a noninformative pdf should be uniform. A problem however is that this distribution is not invariant with respect to parameter transformations. A uniform distribution for e.g. a velocity parameter transforms into a nonuniform distribution for the corresponding slowness parameter so that it would seem that there *is* information on the slowness while there is no information on the velocity! Several authors address the problem of specifying suitable noninformative priors, see e.g. Jeffreys (1939), Jaynes (1968) and Box and Tiao (1973). A number of rules are given, an important one being that the noninformative prior should be invariant with respect to parameter transformations that leave the problem essentially unchanged. In section 1.15 Jaynes' proposal for using the maximum entropy principle when additional constraints are known is discussed.

Note that the specification of a prior distribution is not critical as long as it is locally flat in comparison to the likelihood function. The latter will then determine the a posteriori pdf. This is what we hope to get from an experiment. In the sequel of this thesis however it will be shown that often the likelihood function is not very pronounced for certain linear combinations of parameters. For these linear combinations the a priori information determines the a posteriori pdf.

In this thesis it is assumed that in seismic data inversion informative a priori knowledge concerning parameters is available. This may come from well logs, from information on related areas or from experts (interpreters). Some a priori knowledge is contained in our fundamental physical concepts. The thickness of a layer for example can not be less than zero. An objection to the utilization of a priori information sometimes given is that it may hide information coming from the data, often expressed by the adage "let the data speak for themselves". A counter-argument is that without a priori information sometimes absurd results may be obtained. For a devoted Bayesian moreover the conditional pdf $p(x|y)$ is the only meaningful measure of information on x , given the data y and therefore through Bayes' rule, the a priori pdf of x necessarily has to be given. To the author's opinion the objection can for a large part be circumvented by a proper uncertainty analysis in which the relative contributions of data and a priori information to the answer can be evaluated and compared, see chapter 3.

1.11 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 derived from it. A point estimate renders a set of numbers as estimates for the parameters. Ideally the point estimate is equal to the true values of the parameters but in general this will of course not be the case. In

order to obtain an optimal estimate one may specify a cost function $C(\mathbf{x}, \hat{\mathbf{x}})$ representing the cost of selecting an estimate $\hat{\mathbf{x}}$ when the true parameter values are given by $\mathbf{x}$. Often this will represent a true economical cost. The risk R is defined as the expectation of the cost C , when $\mathbf{x}$ is used as a point estimate:

$$R = E(C) = \int C(\mathbf{x}, \hat{\mathbf{x}}) p(\mathbf{x}|\mathbf{y}) d\mathbf{x} , \quad (1.84)$$

which of course only makes sense when the a posteriori pdf $p(\mathbf{x}|\mathbf{y})$ is accepted as the state of information on the true parameters $\mathbf{x}$. In scientific inference one is primarily interested in a point estimate that is as accurate as possible. There is more than one way to quantify this desire. An often used cost function is the quadratic one:

$$C = (\mathbf{x} - \hat{\mathbf{x}})^T \mathbf{W} (\mathbf{x} - \hat{\mathbf{x}}) , \quad (1.85)$$

where the weighting matrix $\mathbf{W}$ is positive definite. Minimizing the risk

$$R = \int (\mathbf{x} - \hat{\mathbf{x}})^T \mathbf{W} (\mathbf{x} - \hat{\mathbf{x}}) p(\mathbf{x}|\mathbf{y}) d\mathbf{x} , \quad (1.86)$$

with respect to the point estimate $\hat{\mathbf{x}}$ is equivalent to selection of the point where $\partial R / \partial \hat{\mathbf{x}}$ is zero. It follows that:

$$\frac{\partial R}{\partial \hat{\mathbf{x}}} = 2 \mathbf{W} \int (\mathbf{x} - \hat{\mathbf{x}}) p(\mathbf{x}|\mathbf{y}) d\mathbf{x} = 0 , \quad (1.87)$$

and therefore, using the fact that the integral over $p(\mathbf{x}|\mathbf{y})$ is one:

$$\hat{\mathbf{x}} = \int \mathbf{x} p(\mathbf{x}|\mathbf{y}) d\mathbf{x} . \quad (1.88)$$

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 textbooks. Unfortunately the evaluation of (1.88) requires the computation of $p(\mathbf{x}|\mathbf{y})$ 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 $p(\mathbf{x}|\mathbf{y})$ is symmetric and unimodal, the mean coincides with the mode and the least mean squared error estimator is equivalent to the MAP estimator. This estimator can be interpreted as giving the most likely value of the parameters given data, theory and a priori information.

For a uniform a priori distribution $p(\mathbf{x})$, which is often taken as the noninformative prior it is easily seen that the maximum of the a posteriori density function coincides with the maximum of the likelihood function. MAP estimation then is equivalent to maximum likelihood estimation (MLE). The difference between MLE and MAP estimation for the general case is clear. MLE does not take a priori information into account. For a discussion of the asymptotic properties of MAP estimation and MLE, see Bard (1974) a.o. The

importance of asymptotic properties should not be overemphasised. In practice there is always a limited amount of data. Note that 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 thesis. The a priori distribution is then given by (1.82):

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} |\mathbf{C}_x|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x}^i - \mathbf{x})^T \mathbf{C}_x^{-1} (\mathbf{x}^i - \mathbf{x}) \right\} \quad (1.82)$$

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

$$p(\mathbf{n}) = \text{const} \exp \left\{ -\frac{1}{2} \mathbf{n}^T \mathbf{C}_n^{-1} \mathbf{n} \right\} \quad (1.89)$$

The likelihood function follows with (1.69):

$$p(\mathbf{y}=\mathbf{d}|\mathbf{x}) = \text{const} \exp \left\{ -\frac{1}{2} (\mathbf{d} - \mathbf{g}(\mathbf{x}))^T \mathbf{C}_n^{-1} (\mathbf{d} - \mathbf{g}(\mathbf{x})) \right\} \quad (1.90)$$

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

$$2F(\mathbf{x}) = (\mathbf{d} - \mathbf{g}(\mathbf{x}))^T \mathbf{C}_n^{-1} (\mathbf{d} - \mathbf{g}(\mathbf{x})) + (\mathbf{x}^i - \mathbf{x})^T \mathbf{C}_x^{-1} (\mathbf{x}^i - \mathbf{x}) \quad (1.91)$$

This is a weighted nonlinear least squares or l_2 norm. The factor 2 is introduced for notational convenience later on. 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 nonBayesian 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 (1.91) can be found with optimization methods, discussed in chapter 2. For the linear forward model $\mathbf{g}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ an explicit solution of (1.91) is obtained:

$$\hat{\mathbf{x}} = (\mathbf{A}^T \mathbf{C}_n^{-1} \mathbf{A} + \mathbf{C}_x^{-1})^{-1} (\mathbf{A}^T \mathbf{C}_n^{-1} \mathbf{d} + \mathbf{C}_x^{-1} \mathbf{x}^i) \quad (1.92)$$

This solution, introduced in geophysics by Jackson (1979), but also to be found in 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) estimator, the linear least squares estimator and the diagonally stabilized least squares estimator can be derived as special cases of (1.92).

The assumption of the double exponential distribution as given in (1.64) leads to the minimization of a l_1 norm:

$$F(\mathbf{x}) = \|\mathbf{C}_n^{-1/2}(\mathbf{d} - \mathbf{g}(\mathbf{x}))\|_1 + \|\mathbf{C}_x^{-1/2}(\mathbf{x} - \mathbf{x}^i)\|_1 \quad (1.93)$$

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

$$\mathbf{l}_1 \leq \mathbf{D}(\mathbf{d} - \mathbf{g}(\mathbf{x})) \leq \mathbf{u}_1 \quad (1.94a)$$

$$\mathbf{l}_2 \leq \mathbf{E}(\mathbf{x}^i - \mathbf{x}) \leq \mathbf{u}_2 \quad (1.94b)$$

where $\mathbf{l}_1$, $\mathbf{l}_2$, $\mathbf{u}_1$ and $\mathbf{u}_2$ represent vectors with lower and upper bounds, and $\mathbf{D}$ and $\mathbf{E}$ are suitable matrices determined by the problem at hand.

1.12 THE SELECTION OF THE TYPE OF PDF

In the actual application of inverse theory, the question rises which type of pdf is to be used for the noise and the a priori information. In figure 1.1 the three most often used pdf's are given for a one-dimensional case, 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 as stated by 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, has the least informational content as determined by Shannon's information measure, which is defined by Shannon (1948) (see also section 1.15), for a dimensionless pdf:

$$I = E(\log p) = \int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} \quad (1.95)$$

This means that when we have the mean and the covariance, we do not use more information that 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 seems to make a strong argument in favour of Gaussian pdf's, as Shannon's information measure has some properties one would want to demand of such a measure. However, one may wonder when in practice the covariance matrix is available. Even with a priori information where an

idea of the uncertainty is simply given by an expert 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 argument for using Gaussian pdf's. All mathematics can be nicely worked out, and fast optimization schemes have been developed for the resulting least squares problems. The author would like to augment the list with the argument 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, having 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 (three to four times the standard deviation, say). 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. Clearbout and Muir (1973). A practical situation where this may be appropriate is the processing of a target zone of seismic data that contains a multiple that is not taken into account in the forward model. The uniform distribution has also (implicitly) been used for the inversion of seismic data. To the knowledge of the author the only type of errors that is described by the uniform distribution is quantization errors. In seismics however, these errors are seldom large enough to be of any importance.

The question concerning the type of pdf is often stated in the following form: 'Of what type is the noise on the data?' This question reflects a way of thinking typical for an objective interpretation of the concepts of probability (see also sections 1.2 and 1.17). In this interpretation a number of known (in the sense of identified) or unknown processes constitute a random generator corrupting the data. It is sometimes 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.

The assumption of repetitiveness is needed in order to get 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 the subjective Bayesian interpretation, an other 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.

1.13 A TWO-PARAMETER EXAMPLE

The utilization and the benefits of Bayes' rule (1.42) can be illustrated with a simple synthetic example. It contains only two parameters and therefore allows the full visualization of the relevant functions. The problem is a one-dimensional seismic inverse problem and concerns the estimation of the acoustic impedance and the thickness in traveltimes of a thin layer. The true acoustic impedance profile is given in figure 1.2.a as a function of traveltimes. 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 $5 \cdot 10^6 \text{ kgm}^{-2}\text{s}^{-1}$ and 50 ms respectively. The first parameter to be estimated is the acoustic impedance of the thin layer. For the sake of clarity the difference ΔZ with the background impedance is referred to only.

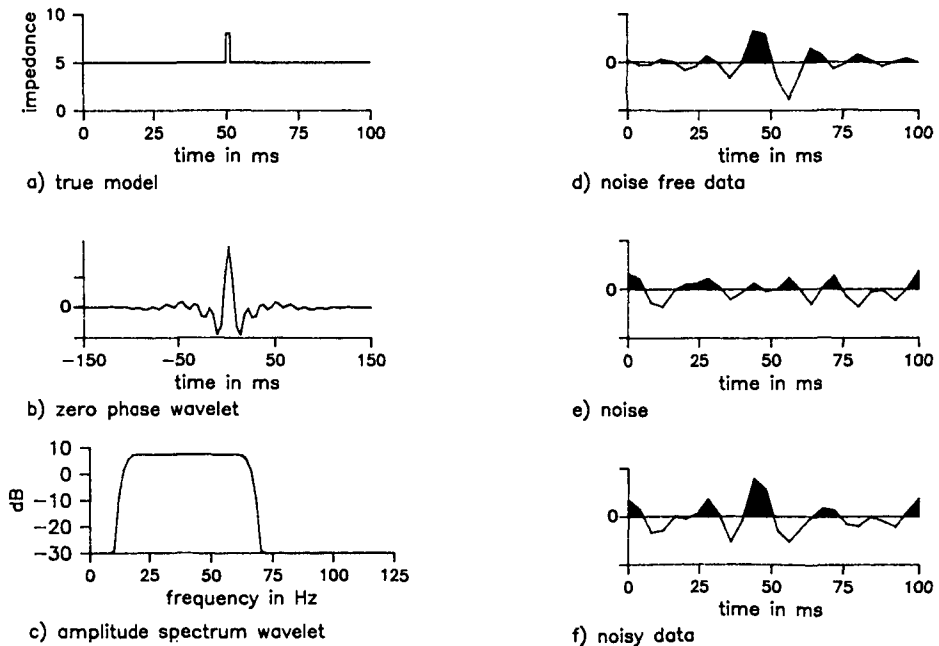


Figure 1.2 Setup of the two-parameter example.

The second parameter is the thickness in traveltime $\Delta\tau$ of the layer. The true values of the parameters are $\Delta Z = 3 \cdot 10^6 \text{ kgm}^{-2}\text{s}^{-1}$ and $\Delta\tau = 1.5 \text{ ms}$ respectively. The forward model used is the plane wave convolutional model with primaries only. For this particular problem it can be written in the form:

$$s(t) = \frac{\Delta Z}{2Z + \Delta Z} \{ w(t - \tau_1) - w(t - (\tau_1 + \Delta\tau)) \} . \quad (1.96)$$

Using this expression and the zero phase wavelet $w(t)$ as given in figures 1.2b, c synthetic data is generated and is shown in figure 1.2.d. Bandlimited noise with an energy of -3 dB relative to the noise free data is added to it. The resulting noisy data as shown in figure 1.2.f is used for inversion.

The available a priori information on the parameters is given in the form of Gaussian pdf's, depicted in figures 1.3a, b. The position of the peaks represent the a priori values and the standard deviations represent the uncertainties in these values. The values are:

$$\begin{aligned} \Delta Z^i &= 3.4 \cdot 10^6 \text{ kgm}^{-2}\text{s}^{-1} , & \sigma_{\Delta Z} &= .5 \cdot 10^6 \text{ kgm}^{-2}\text{s}^{-1} \\ \Delta\tau^i &= 3 \text{ ms} , & \sigma_{\Delta\tau} &= 2 \text{ ms} . \end{aligned} \quad (1.97)$$

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. Note that this implies that $\sigma_{\Delta\tau}$ as given in (1.97) is not exactly the standard deviation of the whole a priori pdf for the thickness, but it is only the standard deviation of the Gaussian part. In the sequel of this thesis the covariance matrix of the Gaussian part will nevertheless shortly be referred to as "a priori covariance matrix". The same convention is used for the a posteriori covariance matrix.

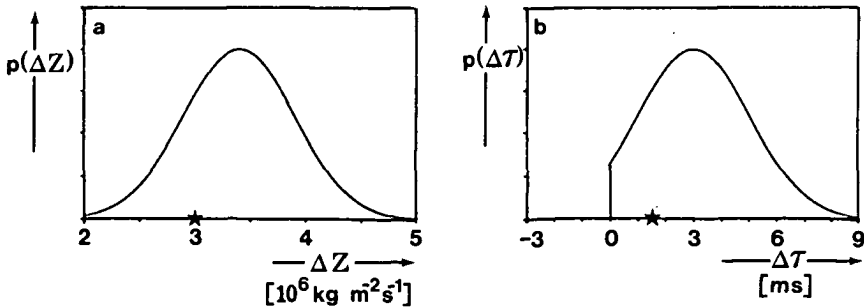


Figure 1.3 A priori pdf's on the parameters ΔZ (a) and $\Delta\tau$ (b). The true values are indicated on the x-axis with a "*".

The true values of the parameters are also indicated in figure 1.3. They are not equal to the a priori values of the parameters but within one standard deviation interval away from them. The a priori information on the parameters is independent. The two-dimensional pdf is therefore the product of the two one-dimensional pdf's:

$$p(\Delta Z, \Delta \tau) = p(\Delta Z) p(\Delta \tau) , \quad (1.98)$$

and is given by equation (1.82) for the region $\Delta \tau \geq 0$ with:

$$\mathbf{x}^i = \begin{bmatrix} \Delta Z^i \\ \Delta \tau_i \end{bmatrix} \quad (1.99)$$

and

$$\mathbf{C}_x = \begin{bmatrix} \sigma_{\Delta Z}^2 & 0 \\ 0 & \sigma_{\Delta \tau}^2 \end{bmatrix} . \quad (1.100)$$

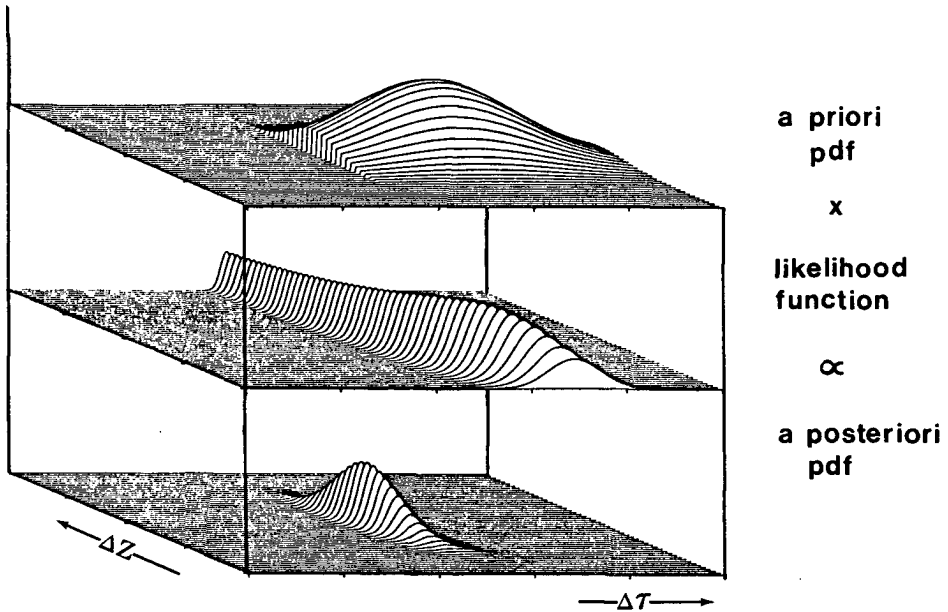


Figure 1.4 The two-dimensional density functions of the two-parameter example. The a posteriori pdf is proportional to the product of the a priori pdf and the likelihood function.

In figure 1.4 the two-dimensional pdf's for this problem are given. Figures 1.5a,b,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

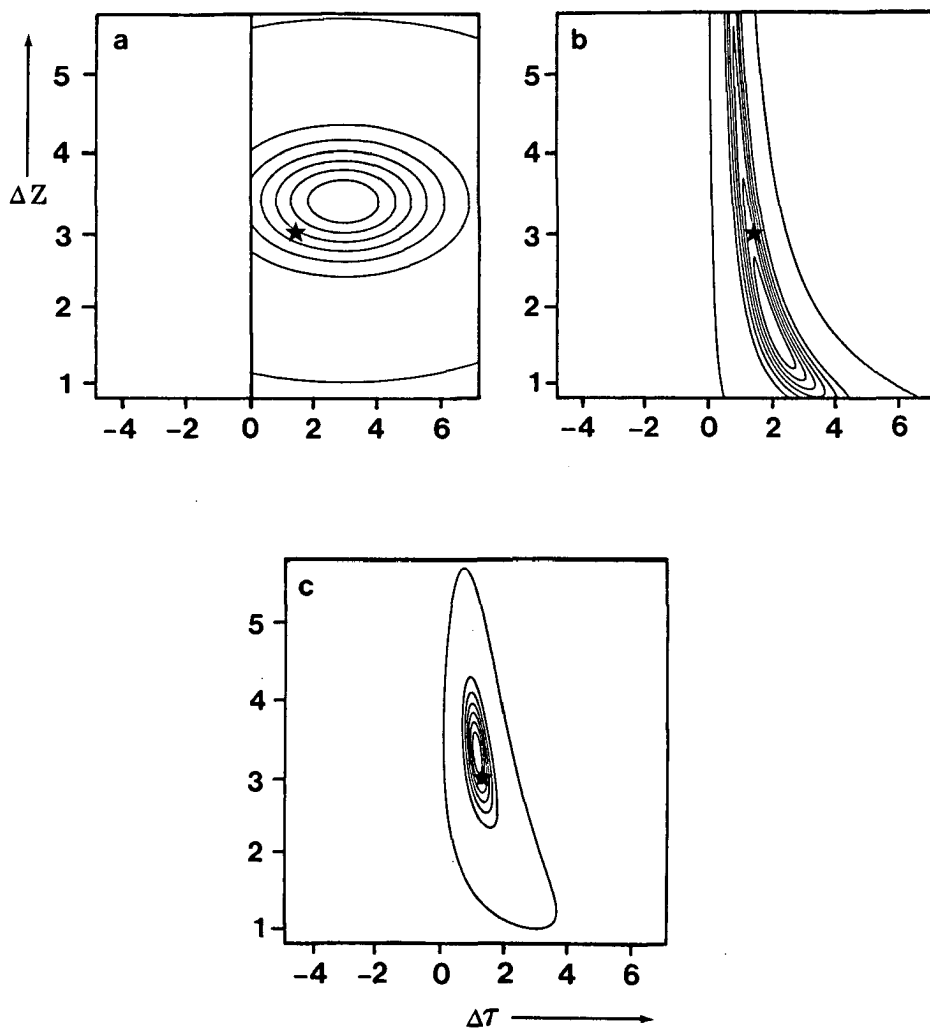


Figure 1.5 Contours of the a priori pdf (a), the likelihood function (b) and the a posteriori pdf (c). The units of ΔZ and $\Delta \tau$ are $10^6 \text{ kgm}^{-2}\text{s}^{-1}$ and ms respectively. The location of the true model is indicated with a "*".

deviations, the ranges plotted for both parameters are not equal: 10 for ΔZ vs. 6 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 S/N ratio of 3.dB. The formula used is thus (1.90) with $C_n = \sigma_n^2 \mathbf{I}$ and $g_i = s(i\Delta t)$, with $s(t)$ defined in (1.96). The function has a unique maximum, but a wide range of significantly different models, lying on the ridge, has almost equal likelihood. This stems from the fact that the response of a thin layer can be approximated by:

$$s(t) = \frac{\Delta Z}{2Z + \Delta Z} \cdot \Delta\tau \cdot w'(t - \tau_m) \quad , \quad (1.101)$$

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 thought as fixed for the range of τ_m 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 (1.90).

The product of the a priori pdf and the likelihood function renders the solution of the inverse problem: the a posteriori pdf, given in figures 1.4, and 1.5.c. It is much more restricted than the likelihood function and has a unique maximum closer to the true model on the scales of the picture. In table 1.1 the true model and the maxima of the pdf's are given, together with the deviations from the true model and the datamismatch for MLE and MAP estimation. A measure of the overall error in the estimated model depends of course on the weights assigned to each parameter, see e.g. the quadratic error norm (1.85). When the significant ranges e.g. are 1 ms for the thickness and .5 kgm⁻²s⁻¹ for the acoustic impedance then, according to the quadratic norm, the MAP estimate is much closer to the true model than the maximum likelihood estimate. The data mismatch for MAP estimation is higher because the parameters are restricted in their freedom to explain the data.

Table 1.1 Numerical details of the two-parameter example.

model	ΔZ [10 ⁶ kgm ⁻² s ⁻¹]	$\Delta\tau$ [ms]	$ \Delta Z - \Delta Z_{\text{true}} $ [10 ⁶ kgm ⁻² s ⁻¹]	$ \Delta\tau - \Delta\tau_{\text{true}} $ [ms]	residual energy [dB]
true	3.0	1.5	0.	0.	—
a priori	3.4	3.0	0.4	1.5	—
MLE	2.4	1.6	0.6	0.1	-4.15
MAP	3.32	1.25	0.32	0.25	-4.0

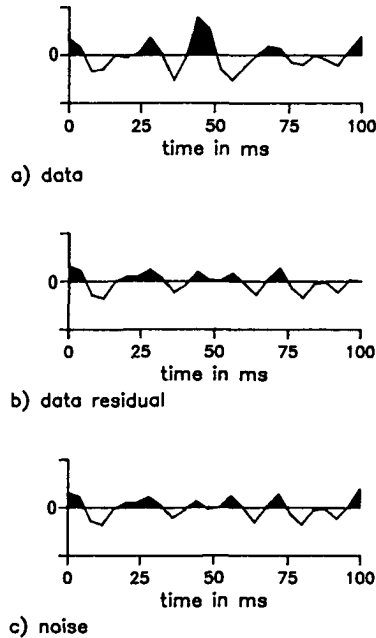


Figure 1.6 The data mismatch (b) in comparison with the data (a) and the noise realization (c).

In figure 1.6 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 be the case when the a priori model x^i does not equal the true model. Tests on the residuals are important. They can indicate 'inconsistent information'. The noise level may have been chosen 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 further pursued in this thesis.

1.14 INVERSION FOR UNCERTAIN DATA

Bayes' rule for probability densities (1.42) provides an answer to the inverse problem when a set of numbers (observations) becomes available for the data vector y . One may

say that the true values for y become known and that the probabilities for the parameters are recalculated. There are however practical inverse problems in which the data vector y is not exactly determined. The observations are given by a probability density. An example, given by Tarantola (1987), is that of the reading of an arrival time on a seismogram. Due to noise or other effects an interpreter cannot exactly determine the arrival time but he can specify a probability density function describing his degree of belief on it. Bayes' rule cannot directly be applied for this type of data. Note that in principle this means that neither Bayes' rule, nor the likelihood function can be used for parameter estimation in cases where analog instruments are read! That this type of data has nevertheless been processed with statistical techniques during the past centuries must be due to either the fact that (for some observations) the observation errors are negligible compared to other types of errors, or to a "trick" described in appendix A which still allows the utilization of Bayes' rule (or the likelihood function) to draw inferences on parameters.

In this section a more straightforward and elegant solution for this problem is presented. The basic principle is given by R.C.Jeffrey (1983, first published 1965) for probabilities. A formulation for probability densities can be derived from it but is here derived directly instead from basic considerations in a manner analogously to the reasoning of Jeffrey. Suppose that a priori degree of belief on parameters x and data y is given by the pdf $p_0(x,y)$. Suppose further that, as a result of observation the degree of belief on the marginal pdf of y changes to $p_1(y)$. The question is now how to propagate this change of belief on y over the rest of the probability structure. Note that this involves the calculation of a new pdf $p_1(x,y)$ which is an action usually not considered in classical statistics. There a vector of random variables has just one pdf: the pdf. The action of revising probabilities is called probability kinematics. The answer to the question lies in the establishment that whereas the information on y may have changed there is no reason to change the conditional degree of belief on x given y so that:

$$p_1(x|y) = p_0(x|y) . \quad (1.102)$$

This is sufficient to derive the solution to the inverse problem, the marginal pdf on x , $p_1(x)$:

$$\begin{aligned} p_1(x) &= \int p_1(x,y) dy \\ &= \int p_1(x|y) p_1(y) dy \\ &= \int p_0(x|y) p_1(y) dy . \end{aligned} \quad (1.103)$$

This is indeed the type of answer desired. It establishes the solution in terms of the existing conditional degree of belief $p_0(x|y)$ and the new information $p_1(y)$. Also the solution is

appealing when comparing it to the solution of Bayes' rule (1.34). The solution is an average of possible a posteriori pdf's, with weights as determined by $p_1(y)$. And, as a limiting case, when data becomes known exactly: $p_1(y) = \delta(y-d)$, the solution is the a posteriori pdf as derived from Bayes' rule:

$$p_1(x) = p_0(x|y=d) \quad (1.104)$$

In appendix A it is discussed how the same results can be derived from Bayes' rule in a less straightforward way.

1.15 THE MAXIMUM ENTROPY FORMALISM

In the past few decades maximum entropy techniques have drawn much attention. Jaynes (1968) proposed to use it as a basis for deriving objective a priori probability distributions. It is however also used as a tool or principle for inversion itself.

Shannon (1948) introduced the concept of entropy as a measure of uncertainty in information theory. When X is a discrete random variable with probabilities P_i of obtaining the values x_i its entropy is defined as:

$$H = - \sum_i P_i \log P_i \quad (1.105)$$

That H is a measure of uncertainty is attested by its properties, which are (following Rosenkrantz, 1977):

- (1) $H(P_1, \dots, P_m) = H(P_1, \dots, P_m, 0)$, the entropy is fully determined by the alternatives which are assigned nonzero probability.
- (2) When all the P_i are equal $H(P_1, \dots, P_m)$ is increasing in m , the number of equiprobable alternatives.
- (3) $H(P_1, \dots, P_m) = 0$, a minimum, when some $P_i = 1$.
- (4) $H(P_1, \dots, P_m) = \log m$, a maximum, when each $P_i = 1/m$.
- (5) Any averaging of the P_i (i.e. any flattening of the distribution) increases H .
- (6) H is nonnegative.
- (7) $H(P_1, \dots, P_m)$ is invariant under any permutation of the indices $1, \dots, m$.
- (8) $H(P_1, \dots, P_m)$ is continuous in its arguments.

Important concepts are the joint entropy $H_{X,Y}$ of the variables X and Y and the conditional entropy $H_{X|Y}$. The respective definitions are:

$$H_{X,Y} = - \sum_i \sum_j P_{ij} \log P_{ij} \quad , P_{ij} = P(X=x_i, Y=y_j) \quad (1.106)$$

and

$$H_{X|Y} = - \sum_j P_j \sum_i P_{ij} \log P_{ij} \quad , \quad P_{ij} = P(X=x_i | Y=y_j) \quad . \quad (1.107)$$

With these definitions the list of properties can be extended with:

$$(9) \quad H_{X,Y} = H_{Y|X} + H_X = H_{X|Y} + H_Y \quad (\text{additivity})$$

$$(10) \quad H_X - H_{X|Y} = H_Y - H_{Y|X} \quad (\text{from (9)})$$

$$(11) \quad H_{Y|X} \leq H_Y, \text{ with equality if } X \text{ and } Y \text{ are independent.}$$

$$(12) \quad H_{X,Y} \leq H_X + H_Y \text{ with equality if } X \text{ and } Y \text{ are independent (from (9) and (11)).}$$

It can be shown that properties (2), (8) and (9) uniquely characterise H . These are the properties which Shannon demanded of an uncertainty measure.

Shannon (1948) also defined a measure for continuous variables:

$$H = - \int p(x) \log p(x) dx \quad . \quad (1.108)$$

This measure however is not invariant under reparameterization. A transformation of parameters changes the value of H . A modification of (1.108) that is invariant is given by Jaynes (1963, 1968), see also Rietsch (1977):

$$H = - \int p(x) \log \frac{p(x)}{m(x)} dx \quad . \quad (1.109)$$

In the absence of constraints on $p(x)$ the entropy H is maximized by $p(x) = m(x)$. Hence $m(x)$ is the pdf that represents a state of complete ignorance on x .

Jaynes proposed to use the "principle of maximum entropy" for deriving prior probabilities when a number of constraints on the probabilities are known and nothing else. The principle is that the pdf should maximize the entropy measure (1.109) (for the continuous case) subject to the constraints. Then no more information than is legitimately available is put in the prior probabilities. The pdf sought is to be normalized:

$$\int p(x) dx = 1 \quad . \quad (1.110)$$

When we have information fixing the means of m different functions $f_k(x)$:

$$\int f_k(x) p(x) dx = F_k \quad (k=1, \dots, m) \quad , \quad (1.111)$$

where F_k are the numerical values, the problem is to maximize (1.119) subject to the constraints (1.110) and (1.111). The solution is (Jaynes (1968)):

$$p(x) = \frac{m(x)}{Z} \exp \{ \lambda_1 f_1(x) + \dots + \lambda_m f_m(x) \} \quad , \quad (1.112)$$

with the partition function:

$$Z(\lambda_1, \dots, \lambda_m) = \int m(x) \exp \{ \lambda_1 f_1(x) + \dots + \lambda_m f_m(x) \} dx \quad , \quad (1.113)$$

The Lagrange multipliers λ_k follow from the constraints (1.111) and are determined by:

$$\frac{\partial \ln Z}{\partial \lambda_k} = F_k \quad k=1, \dots, m \quad (1.114)$$

The problem remains what the pdf $m(x)$, representing complete ignorance should be. Jaynes (1968) argues that such a pdf can often be determined by specifying a set of parameter transformations recognized to transform the problem into an equivalent one. The desideratum of consistency then determines the form of the pdf $m(x)$. As shortly discussed in section 1.12 the principle of maximum entropy provides a rationale for the utilization of Gaussian pdf's, when only mean and covariance are known, and the state of complete ignorance can be described by a (locally) uniform pdf.

The principle of maximum entropy itself has also been used as a tool for inversion. Burg (1967) e.g. used it for spectral estimation. Other examples can be found in Ray Smith and Grandy (1985) and Rietsch (1977). Some authors (see e.g. Jaynes, 1985) regard the maximum entropy method as a limiting case of the full Bayesian procedure, viz. the noise free case. P.M. Williams (1980) on the other hand, derives Bayes' rule and Jeffrey's generalization of it (see section 1.14) as a special case of the "minimum information" principle. His analysis is for the discrete case. Here, in an analogous way, the continuous case is discussed. First the measure of information in $p(x,y)$ relative to $p_0(x,y)$ is defined as:

$$I_{X,Y}(p,p_0) = \int p(x,y) \log \frac{p(x,y)}{p_0(x,y)} dx dy \quad (1.115)$$

Note the differences of the minus sign with the entropy definition (1.109) and the fact that p_0 can be any pdf, that serves as a reference for the problem at hand, and not only the one expressing complete ignorance. Using the fact that for any positive real number x :

$$x \log x - x + 1 \geq 0 \quad , \quad \text{with equality if and only if } x=1 \quad , \quad (1.116)$$

it can be derived that

$$I_{X,Y}(p,p_0) = \int p_0(x,y) \left\{ \frac{p(x,y)}{p_0(x,y)} \log \frac{p(x,y)}{p_0(x,y)} - \frac{p(x,y)}{p_0(x,y)} + 1 \right\} dx dy \geq 0 \quad , \quad (1.117)$$

with equality if and only if $p=p_0$. This means that any $p(x,y) \neq p_0(x,y)$ has a higher information content than $p_0(x,y)$ according to this measure. Analogously to the discrete case we can define the conditional information measure:

$$I_{X|Y}(p,p_0) = \int p(y) p(x|y) \log \frac{p(x|y)}{p_0(x|y)} dx dy \quad (1.118)$$

and the marginal information measure:

$$I_Y(p, p_0) = \int p(y) \log \frac{p(y)}{p_0(y)} dy . \quad (1.119)$$

It can easily be derived that:

$$I_{X,Y}(p, p_0) = I_{X|Y}(p, p_0) + I_Y(p, p_0) . \quad (1.120)$$

Williams formulates the principle of minimum information as:

“Given the prior distribution p_0 , the probability p appropriate to a new state of information is the one that minimizes $I_{x,y}(p, p_0)$ subject to whatever constraints the new information imposes”.

If $p_0(x, y)$ is the prior distribution and if observation leads to a new marginal pdf $p_1(y)$ for y it follows from (1.120) when using (1.117) that $I_{x,y}$ is minimized by $p_1(x|y) = p_0(x|y)$, the increase being given by $I_Y(p_1, p_0)$. The solution for $p(x, y)$ is thus:

$$p_1(x, y) = p_0(x|y) p_1(y) , \quad (1.121)$$

and the solution for the parameters x is:

$$p_1(x) = \int p_0(x|y) p_1(y) dy . \quad (1.122)$$

This result is equivalent to (1.103), which was the extension to the continuous case of Jeffrey's generalization of Bayes' rule. Remember that Bayes' rule is a special case of (1.122), when $p_1(y) = \delta(y-d)$, i.e. when there is no uncertainty in the data.

Whereas P.M. Williams concludes that Bayes' rule as a tool for inversion can be derived from the minimum information principle, the author of this thesis is more inclined to state that the results of this principle merely shows that it is not inconsistent with (the generalization of) Bayes' rule and that the results yield another argument in favour of the interpretation of the information measure.

1.16 THE APPROACH OF TARANTOLA AND VALETTE

Inversion theory is very 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 especially interesting that Tarantola and Valette (1982a) (see also Tarantola, 1987) formulated an alternative theory for inverse problems in order to solve a number of alleged problems of the Bayesian approach. Unlike Bayes' rule it can handle problems where the data is not obtained in the form of a set of numbers (explicit data) but rather in the form of a pdf (uncertain data). This is for example more appropriate in cases where the data for inversion is obtained by interpretation of analog data or instruments, e.g. the reading of

arrival times from a seismogram. Their theory distinguishes between theoretical and observational errors, which gives rise to an interpretation problem in practice. In this section it is shown that:

- (1) The basic concept and formulation of "conjunction of states of information", which is the cornerstone of the approach of Tarantola and Valette is consistent with classical probability theory. In fact, a relation with identical interpretation can be derived within the limits of the latter.
- (2) For the case of explicit data the approach of Tarantola and Valette renders equivalent results under different interpretations of theoretical and observational errors. The results equal those of the Bayesian approach.
- (3) For the case of uncertain data their formulation leads to a different result than the extension of Jeffrey's probability kinematics as discussed in section 1.14. The results are not inconsistent but have different interpretation. The approach of Tarantola and Valette may be of more practical value.

First the theory of Tarantola and Valette is briefly set out. Probabilities and probability densities describe "states of information" in a typical Bayesian interpretation. It is emphasized that pdf's need not be normalizable. Nonnormalized pdf's are called "density functions". Their interpretation is in terms of *relative* probability. The cornerstone of their theory is the conjunction of states of information, which is a generalization of the conjunction of propositions in propositional logic. Let $p_i(z)$ and $p_j(z)$ denote probability density functions on z and let $\mu(z)$ denote the state of null information, i.e. the pdf describing the state of complete ignorance. The conjunction $p_i \wedge p_j$ of p_i and p_j is designed to have the following properties:

- (1) The conjunction should be commutative:

$$p_i \wedge p_j = p_j \wedge p_i \quad . \quad (1.123)$$

- (2) The conjunction of any state of information p_i with the state of null information should not result in any loss of information:

$$p_i \wedge \mu = p_i \quad . \quad (1.124)$$

- (3) For any A:

$$\int_A p_i(z) dz = 0 \quad \Rightarrow \quad \int_A p_i(z) \wedge p_j(z) dz = 0 \quad . \quad ((1.125)$$

From these three properties it is derived that (Tarantola, 1987):

$$p_i(z) \wedge p_j(z) = \frac{p_i(z) p_j(z)}{\mu(z)} \quad . \quad (1.126)$$

Note that this result is obtained without the utilization of the concept of conditional probability, which is rather derived as a special result of (1.126). It is stressed however by Tarantola and Valette (1982a) that this definition of conjunction of states of information can be used only when states of information have been obtained *independently*. A formal definition of independence however is not given and it is questionable whether this can be done without using the concept of conditional probability.

In their formalism an inverse problem is solved by combining a priori information $\rho(x,y)$ on parameters x and data y with information concerning their theoretical relations $\theta(x,y)$. The result is called the a posteriori state of information $\sigma(x,y)$ and is given by (1.126):

$$\sigma(x,y) = \frac{\rho(x,y) \theta(x,y)}{\mu(x,y)} \quad (1.127)$$

The marginal density function for x :

$$\sigma(x) = \int \frac{\rho(x,y) \theta(x,y)}{\mu(x,y)} dy \quad (1.128)$$

is the solution to the inverse problem. In many situations the a priori information on x and y will be independent:

$$\rho(x,y) = \rho(x) \rho(y) \quad (1.129)$$

the theoretical relations can be formulated in the form of a conditional density function:

$$\theta(x,y) = \theta(y|x) \mu(x) \quad (1.130)$$

and the states of null information on x and y are independent (for a motivation, see Tarantola (1987)):

$$\mu(x,y) = \mu(x) \mu(y) \quad (1.131)$$

Equation (1.128) then can be written as:

$$\sigma(x) = \rho(x) \int \frac{\rho(y) \theta(y|x)}{\mu(y)} dy \quad (1.132)$$

The writings of Tarantola and Valette are not specific concerning the interpretation of the concept of "data" and thereby the interpretation of theoretical and observational errors. Suppose that explicit data are obtained and that the errors are known in statistical terms. Formula (1.132) can then be worked out under two extreme interpretations:

- 1) The errors are considered theoretical. In this interpretation there is no uncertainty in the data:

$$\rho(y) = \delta(y-d) \quad (1.133)$$

Substitution in (1.132) yields:

$$\theta(x) = \frac{\rho(x) \theta(y=d|x)}{\mu(y=d)} . \quad (1.134)$$

When constant factors are considered immaterial this result is equal to Bayes' rule.

2) The errors are observational. In this interpretation the theoretical relations are error free:

$$\theta(y|x) = \delta(y - g(x)) . \quad (1.135)$$

In order to obtain $\rho(y)$ it is necessary to introduce another data vector y_1 for which the vector of numbers d is obtained. The a priori knowledge $\rho(y)$ is then the a posteriori solution to another inverse problem for which $y_1=d$ is the data and the a priori information on y is the null information $\mu(y)$. The solution for such a problem is (1.134), with an adapted notation:

$$\rho(y) = \frac{\mu(y) \theta(y_1=d|y)}{\mu(y_1=d)} . \quad (1.136)$$

Substitution of (1.136) and (1.135) in (1.132) yields:

$$\sigma(x) = \frac{\rho(x) \theta(y_1=d|x)}{\mu(y_1=d)} , \quad (1.137)$$

where the obvious identity $\theta(y_1=d|g(x)) = \theta(y_1=d|x)$ has been used. The result is equivalent to (1.131) and to Bayes' rule.

It can also be shown that the final result is identical to Bayes' rule when the errors are interpreted as partly theoretical and partly observational, provided they are independent. That these results are identical is of course essential. When it is arbitrary what interpretation we take, the final results under different interpretations should be identical.

It is now shown that a relation with essentially the same interpretation as the conjunction of states of information of Tarantola and Valette (1.126) can be derived within classical probability theory as sketched in sections 1.3 – 1.5. It is implicit in most Bayesian writings and explicit in most logical Bayesian writings that probabilities are always conditional on some or other body of knowledge or data (see e.g. Jeffreys (1939)). Let the set of propositions "a" represent a body of knowledge, for example a priori information. The conditional probability $P(z|a)$ gives the probability of z given the a priori information. Similarly $P(z|t)$ is the theoretical state of information on z when t denotes the body of theoretical knowledge. Combining theoretical and a priori knowledge on z is of course equivalent to deriving the probability of z conditional on the conjunction of a and t : $P(z|a \wedge t)$. This can be derived, starting with the definition of conditional probability:

$$P(z|a \wedge t) = P(t|z \wedge a) P(z|a) . \quad (1.138)$$

When a and t are independent this can be worked out further:

$$P(z \wedge a \wedge t) = P(t|z) P(z|a) P(a) \quad (1.139)$$

Using Bayes' rule:

$$P(z|a \wedge t) P(a) P(t) = \frac{P(z|t) P(t)}{P(z)} P(z|a) P(a) \quad (1.140)$$

or

$$P(z|a \wedge t) = \frac{P(z|a) P(z|t)}{P(z)} \quad (1.141)$$

$P(z)$ 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:

$$p(z|a \wedge t) = \frac{p(z|a) p(z|t)}{p(z)} \quad (1.142)$$

which actually is expression (1.126) in a different notation! Note that this result is readily extended to the situation with any number of bodies of knowledge. One may for example distinguish an a priori, observational and theoretical body of knowledge a, o and t respectively. Provided they are independent we have:

$$p(z|a \wedge o \wedge t) = \frac{p(z|a) p(z|o) p(z|t)}{p(z)^2} \quad (1.143)$$

Note also that the intuitive (?) demand of Tarantola and Valette that states of information be independent in order to allow their conjunction by (1.126) explicitly occurs in the derivation of (1.141) from the basics of probability theory. One should however 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. As mentioned above however 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.

A last point worth mentioning is a difference between the solution of Tarantola and Valette (1.132) and the extension of Jeffrey's result (1.103) for uncertain data. The latter can be rewritten using Bayes' rule as:

$$p_1(x) = p_0(x) \int \frac{p_1(y) p_0(y|x)}{p_0(y)} dy \quad (1.144)$$

All terms of (1.144) have equal interpretation as the corresponding ones in (1.132), except

for the denominators of the integrands and hence, the a posteriori results $\sigma(x)$ and $p_1(x)$. The denominator in (1.44), the marginal pdf $p_0(y)$:

$$p_0(y) = \int p_0(y|x) p_0(x) dx \quad , \quad (1.145)$$

represents a state of information on y , in which a priori knowledge $p_0(x)$ and the theoretical relations $p_0(y|x)$ play a role and is therefore not equal to the state of complete ignorance $\mu(y)$ that appears in (1.132). The apparent inconsistency resolves when we realize that (1.132) and (1.144) solve different problems. In Tarantola's approach the observational knowledge $p(y)$ is independent of the theoretical relations and usually independent of the a priori knowledge on the parameters x . In Jeffrey's approach $p_1(y)$ represents the marginal pdf of y in a new state of information and therefore also reflects a priori information, through the theoretical relations. Which approach is most practical is more a psychological question: Are observations free from a priori knowledge or not? The answer may strongly depend on the experiment and the observer in question.

1.17 DISCUSSION AND CONCLUSIONS ON GENERAL PRINCIPLES

In this chapter the fundamentals of the Bayesian theory for inversion or parameter estimation have been sketched. In the previous 3 sections an intimate relationship between some fundamental principles has been shown for the general case of uncertain data. It may be said that there is consensus concerning the basic formulas to be used in inversion. For explicit data Bayes' rule can be used as the basic principle for parameter estimation. In the rest of this thesis this is the situation applicable: Seismic data is assumed to be available in discretized form in a digital computer. There is no uncertainty with respect to the values.

In this thesis detailed inversion of seismic data will be approached from a Bayesian standpoint. It is important to realize that the classical or frequency approach to inversion, which has dominated statistical practice throughout this century is fundamentally different. In the Bayesian approach the a posteriori pdf is the answer to the inverse problem and a point estimate in the form of e.g. the maximum of this pdf can be seen as a result of information reduction. In the classical approach the situation is different. There the so called sample space of different outcomes of an experiment is defined. The process of observation means "drawing" at random an outcome from this space. A measured seismic data set e.g. is one out of an infinite number of data sets that were potential outcomes but simply not obtained or realized. The pdf describing the distribution of the outcome is $p(y|x)$. This distribution is known except for the unknown parameters x . In maximum likelihood estimation the parameters x are determined such that their values render the actual data $y=d$ the highest likelihood. Unlike the a posteriori pdf, the likelihood function should *not* be interpreted as a state of information on the parameters x . The procedure of

deriving an estimate from the data is called an estimator, the procedure described above is referred to as the maximum likelihood estimator. Like any other estimator in classical statistics its usefulness can only be determined by examining its statistical properties. Essential in such an analysis is the so called sampling distribution of the estimator. This is the distribution which describes the variations of the estimator due to variations of the data. Two important properties of the estimator are its bias and its (co)variance. The bias is the difference between the expectation of the estimator and the true values of the parameters. Although the latter are never known in practice this concept may still provide some insight in the estimation process under certain assumptions e.g. that of a linear forward model. It is a measure of the "systematic" error in an estimator. The (co)variance describes the random error in the estimator. With the bias it sums to the mean squared error which is a measure of the total inaccuracy of the estimator. The mean squared error, when accepted as a measure of accuracy should be as low as possible. This virtually always entails a trade-off between bias and variance. Nevertheless there is a tendency amongst authors writing in the spirit of the classical approach to more strongly reduce the bias out of some fear for making systematic errors.

It will be clear that the classical approach differs fundamentally from the Bayesian one. Apart from the incorporation of a priori information this reflects itself strongly in procedures for uncertainty analysis and estimator assessment. Concepts like bias and resolution matrix (see chapter 3) are not natural within a Bayesian context. Also the concept of confidence intervals is fundamentally different in the two approaches.

Proponents of the classical approach usually reject the utilization of a priori information for a number of reasons, see section 1.10. At this point it is merely stressed once more that:

- (1) Within the Bayesian paradigm there is no conceptual problem with a priori information.
- (2) Without a priori information absurd results may occur.
- (3) Superior results are obtained with a priori information (as can even be derived within a classical setting).
- (4) The influence of a priori information can always be assessed in a proper uncertainty analysis, see chapter 3.

To the author the concepts of Bayesian estimation are clearer and more natural. Together with the superior results this is the reason for selecting this approach for the detailed inversion of seismic data. We still remain however with the philosophical differences between the subjective and the logical interpretation of the concept of probability. The following statements reflect the opinion of the author: If the concept of probability is to reflect "degree of belief", then certainly this must be a subjective degree of belief since any knowledge we can speak about with some authority is human or subjective knowledge. It seems very difficult to give an account of objective degree of belief. Nevertheless in order

to keep science a coherent activity it is reasonable to demand that there is intersubjective agreement on how probabilities should be assigned when a certain type of knowledge is available. A principle like that of maximum entropy is an important tool in this respect. A further analysis would require a thorough philosophical study of both viewpoints. Since no practical consequences for practical inversion of seismic data are envisaged this issue is further left aside.

2

OPTIMIZATION**2.1 INTRODUCTION**

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, see chapter 1. Optimization is the mathematical or numerical technique for the actual 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 a very important one. Many textbooks have appeared on the subject, for example Gill, Murray and Wright (1981) and Scales (1985). A very rough classification of optimization methods is:

1. *function comparison or direct search methods*

These methods have very poor convergence rates and are only advantageous for highly discontinuous functions.

2. *steepest descend method*

Although for smooth functions this method is more efficient than direct search methods, it also has a poor convergence rate.

3. *conjugate gradient methods*

These methods have reasonable convergence rates and are most suited for large scale problems, because of modest storage requirements. Scales (1985) recommends these techniques for problems with more than 250 parameters.

4. Newton methods

This is a very broad class of methods. It has the best convergence rates (so-called quadratic convergence near the minimum).

All these methods will find a local minimum. A Monte Carlo algorithm that can find a global minimum is described by Kirkpatrick et al. (1983).

In this chapter a concise overview of Newton methods is given. They have been used for the inversion of poststack seismic data as described in chapter 4 onwards. Gaussian distributions for noise and a priori information lead to (nonlinear) least squares problems, when maximizing the a posteriori pdf. Special methods for nonlinear least squares problems have been given in optimization literature and are shortly discussed in section 2.3. An illustration of differences in efficiency of different schemes is given in section 2.4 for a least squares problem. The subjects of the implementation of constraints and the important practical problem of scaling are touched upon in sections 2.5 and 2.6. Most of the optimization problems in this thesis were solved using routines from the NAG (Numerical Algorithms Group) library, discussed in section 2.7.

2.2 NEWTON OPTIMIZATION METHODS

Let F denote the function to be minimized. Like most optimization methods the Newton methods generate a sequence of points $\mathbf{x}_k$ by:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k, \quad (2.1)$$

where $\mathbf{p}_k$ is the search direction and α_k is determined by a univariate minimization technique (a so called line search), such that the function value $F(\mathbf{x}_{k+1})$ is a minimum along the line which is defined by (2.1). Usually this line search does not need to be very accurate. The various Newton methods can be characterized by the way in which the search direction $\mathbf{p}_k$ is generated. When first and second derivatives of the function F exist it can be approximated by the first three terms of a Taylor series around $\mathbf{x}_k$:

$$F(\mathbf{x}_k + \mathbf{p}) \approx F(\mathbf{x}_k) + \mathbf{q}_k^T \mathbf{p} + \frac{1}{2} \mathbf{p}^T \mathbf{H}_k \mathbf{p}, \quad (2.2)$$

where $\mathbf{q}$ is the vector of first derivatives:

$$q_i = \frac{\partial F}{\partial x_i} \quad (2.3)$$

and $\mathbf{H}$ the so-called Hessian matrix of second derivatives:

$$H_{ij} = \frac{\partial^2 F}{\partial x_i \partial x_j}. \quad (2.4)$$

Both $\mathbf{q}_k$ and $\mathbf{H}_k$ are evaluated at $\mathbf{x}_k$. The quadratic function on the right hand side of (2.2)

is minimized by:

$$\mathbf{p} = -\mathbf{H}_k^{-1} \mathbf{q}_k \quad (2.5)$$

For a true quadratic function equation (2.2) is exact and equation (2.5) is the solution to the minimization problem. The minimum is found in one step, with α_k in (2.1) equal to one. For nonquadratic functions equation (2.5) generates the search direction. Close to the minimum the approximation (2.2) is good and therefore the Newton method converges to the minimum very fast.

The method that uses (2.5) to generate the search direction is simply called Newton's method. A number of problems may arise with it. One of them is e.g. that the Hessian matrix may become singular. These problems are solved in so-called 'modified Newton methods' where the Hessian matrix is modified such that it is positive definite.

Derivatives should be given whenever possible, see Scales (1985). When they are not available they can be generated with finite difference techniques. For second derivatives however, this will often be too expensive and attractive alternatives then are the quasi-Newton methods. In these methods the Hessian matrix or its inverse is approximated at each iteration, using an updating formula that uses the changes in the gradient vector at subsequent iterations. For the details the reader is referred to the textbooks mentioned above.

2.3 SPECIAL METHODS FOR NONLINEAR LEAST SQUARES PROBLEMS

Special methods can be devised when the function to be minimized is a sum of squares such as in equation (1.91), which defines the maximum of the a posteriori pdf under Gaussian assumptions. A sum of squares can be written as:

$$2F = \sum_i \mathbf{e}_i^2 = \mathbf{e}^T \mathbf{e} \quad (2.6)$$

where the factor 2 is introduced for notational convenience in later formulas. Equation (2.6) is related to the nonlinear least squares estimator (1.91) through:

$$\mathbf{e} = \begin{bmatrix} \mathbf{C}_n^{-1/2} (\mathbf{d} - \mathbf{g}(\mathbf{x})) \\ \mathbf{C}_x^{-1/2} (\mathbf{x}^i - \mathbf{x}) \end{bmatrix} \quad (2.7)$$

The vector $\mathbf{e}$ will be called the residual vector and consists of data and parameter mismatch, weighted by their respective uncertainties. It can easily be derived that the gradient $\mathbf{q}$ of F is given by:

$$\mathbf{q} = \mathbf{J}^T \mathbf{e} \quad , \quad (2.8)$$

where $\mathbf{J}$ is the so-called Jacobian matrix of $\mathbf{e}$:

$$J_{ij} = \frac{\partial e_i}{\partial x_j} \quad . \quad (2.9)$$

The Hessian matrix $\mathbf{H}$ of F is given by:

$$\mathbf{H} = \mathbf{J}^T \mathbf{J} + \sum_{i=1}^m e_i \mathbf{T}_i \quad , \quad (2.10)$$

where m is the number of elements of $\mathbf{e}$ and $\mathbf{T}_i$ is the Hessian of e_i . For low residuals e_i and/or for quasi-linear problems, in which the $\mathbf{T}_i$ matrices contain low values, the Hessian can be approximated by:

$$\mathbf{H} \approx \mathbf{J}^T \mathbf{J} \quad . \quad (2.11)$$

This means that the matrix of second derivatives $\mathbf{H}$ can be computed using only the first derivatives of the residual vector and hence of the forward model $\mathbf{g}(\mathbf{x})$. For the special case of a linear problem $\mathbf{g}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ equation (2.11) is exact and the solution can be derived using (2.5) and (2.8):

$$\mathbf{J}^T \mathbf{J} \mathbf{x} = -\mathbf{J}^T \mathbf{e}|_{\mathbf{x}=0} \quad . \quad (2.12)$$

Using, for the linear case,

$$\mathbf{J} = - \begin{bmatrix} \mathbf{C}_n^{-1/2} \mathbf{A} \\ \mathbf{C}_x^{-1/2} \end{bmatrix} \quad (2.13)$$

and provided $\mathbf{J}^T \mathbf{J}$ is nonsingular the closed form solution is the earlier given equation (1.92):

$$\hat{\mathbf{x}} = (\mathbf{A}^T \mathbf{C}_n^{-1} \mathbf{A} + \mathbf{C}_x^{-1})^{-1} (\mathbf{A}^T \mathbf{C}_n^{-1} \mathbf{d} + \mathbf{C}_x^{-1} \mathbf{x}^i) \quad . \quad (2.14)$$

When, for nonlinear problems, the second term of (2.10) is completely ignored the Gauss-Newton method results that generates a search direction at each point $\mathbf{x}_k$ through:

$$\mathbf{J}_k^T \mathbf{J}_k \mathbf{p}_k = -\mathbf{J}_k^T \mathbf{e}_k \quad . \quad (2.15)$$

This equation is in mathematical form equivalent to relation (2.12). This may have caused some authors to treat nonlinear least squares estimation problems as sequential linear least squares estimation problems. This is conceptually confusing and the way of thinking may easily lead to errors, as has been pointed out by Tarantola and Valette, (1982b). Equation (2.15) is nothing more than the generation of a search direction in an attempt to find the

minimum of (2.6). Indeed, other equations than (2.15) may be used as well. Problems for example may arise when $\mathbf{J}_k^T \mathbf{J}$ is singular or nearly singular. This is remedied in the Levenberg-Marquardt method, where (2.15) is replaced by:

$$(\mathbf{J}_k^T \mathbf{J}_k + \lambda_k \mathbf{I}) \mathbf{p}_k = -\mathbf{J}_k^T \mathbf{e}_k, \quad (2.16)$$

where the parameter λ_k should guarantee the generation of a good search direction $\mathbf{p}_k$.

For the problems where the second term of (2.10) cannot be neglected, a least squares quasi-Newton scheme can be used where the search direction is generated through:

$$(\mathbf{J}_k^T \mathbf{J}_k + \mathbf{B}_k) \mathbf{p}_k = -\mathbf{J}_k^T \mathbf{e}_k, \quad (2.17)$$

where $\mathbf{B}_k$ is updated by some updating formula at each iteration. Gill and Murray (1978) use the singular value decomposition of $\mathbf{J}$ in order to decide when the Gauss-Newton method is used or when (an approximation of) the full Hessian matrix (2.10) is to be used.

2.4 AN EFFICIENCY INDICATION

A simple Monte Carlo experiment illustrates the relative efficiencies of the above mentioned methods for least squares problems. The optimization problem chosen is the one that occurs in single trace inversion for reflection coefficients r_i and lag times τ_i . The forward model used is the convolutional model:

$$s(t) = \sum_i r_i w(t - \tau_i), \quad (2.18)$$

where $w(t)$ is the wavelet that is known in this problem. Two synthetic data sets are inverted, each containing 10 traces. The true models underlying the traces are shown in figure 2.1.a The reflector separation is 20 ms and the absolute value of the reflection coefficients is .2. A noise free data set is generated with a zero phase wavelet and is shown in figure 2.1.c. A second data set is generated from it by adding noise. The S/N ration is 1. The data set is shown in figure 2.1.d. The start models used are shown in figure 2.1.b. The differences between the start and the true models have zero mean and standard deviations of 3 ms for the lagtimes and .05 for the reflection coefficients. Each data set is optimized in a least squares sense. No a priori information is used. Two methods are used for optimization:

- (1) The Gauss-Newton method, which is especially designed for least squares problems.
- (2) A general modified Newton method where the full Hessian is calculated.

In table 2.1 the averaged computation times per trace are given in seconds. The computations were done on a Gould 32/67 minicomputer.

The conclusions to be drawn from the figures are clear. In the case of the noise free data set the objective function more strongly resembles a quadratic function close to the

Table 2.1 Computation times per trace [s].

	noise free data set	noisy data set
Gauss-Newton method	1.2	2.8
Modified Newton method	1.7	2.5

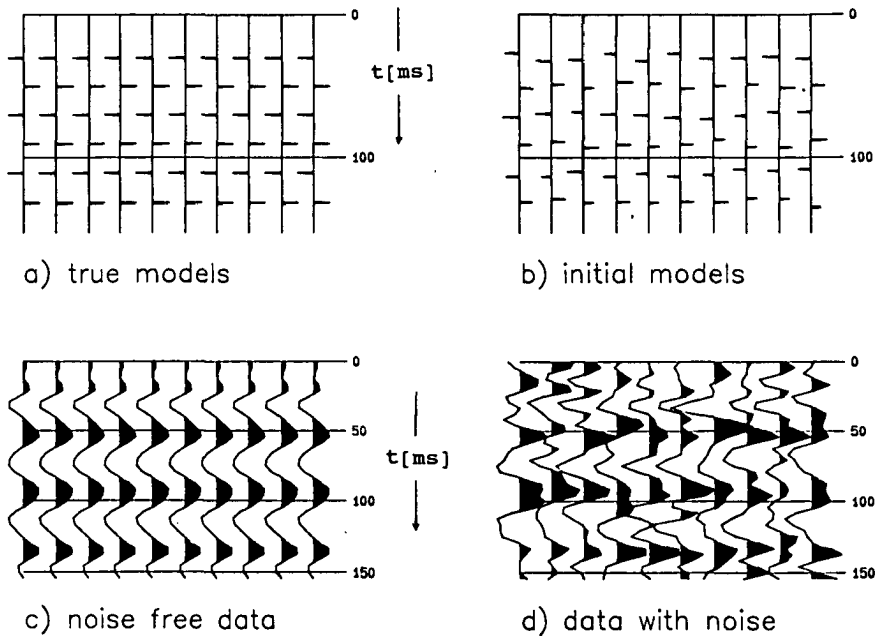


Figure 2.1 Setup of the efficiency experiment.

minimum and hence this data set is optimized faster than the noisy data set. Because the residuals are low (even 0 at the minimum), the Gauss-Newton method is faster than the more general modified Newton method as the Hessian is calculated more efficiently. Note that this difference would even be bigger when equation (2.15) is solved by a standard linear least squares algorithm instead of being solved by explicitly calculating the Hessian according to (2.11) as was necessary in this experiment (due to the available software). In the case of the noisy data set the residuals are higher. The modified Newton method then is faster because the Hessian is calculated more precise. This yields better search directions and the whole procedure takes fewer iterations. The extra computations spent on the Hessian matrix repay through. It is stressed that this example only gives rough indications

that confirm the expectations from theory. The absolute times are so short due to a very efficient computation of the function and its derivatives. This is discussed in chapter 4.

2.5 CONSTRAINTS

Hard constraints define regions where the a priori pdf and hence the a posteriori pdf is zero. The maximum can never lie in such a region. Constraints can occur in the form of equality or inequality constraints:

$$\begin{aligned} \mathbf{c}(\mathbf{x}) &= \mathbf{b} && \text{(equality)} \\ \mathbf{1} \leq \mathbf{c}(\mathbf{x}) \leq \mathbf{u} && \text{(inequality)} \end{aligned} \quad (2.19)$$

The functions $\mathbf{c}(\mathbf{x})$ can be nonlinear in general. Special cases are linear constraints, which will be applied in the inversion of poststack seismic data, see chapter 4:

$$\mathbf{1} \leq \mathbf{B}\mathbf{x} \leq \mathbf{u} \quad , \quad (2.20)$$

and bounds on the parameters:

$$\mathbf{1} \leq \mathbf{x} \leq \mathbf{u} \quad . \quad (2.21)$$

which are also applicable for the problems sketched in chapter 4. There are several ways along which constraints can be dealt with in optimization. There are special, more efficient, methods for linear and for bounds constraints. These matters however, are not further discussed here. The reader is referred to the textbooks on optimization e.g. Gill et.al. (1981) and Scales (1985).

2.6 SCALING

A very important practical item is that of scaling of the objective function and the parameters. The parameters should be scaled such that they are all in the same order of magnitude during the optimization procedure. Otherwise numerical problems may occur. Accuracy may be lost and the computation time may become unnecessarily large. A simple way of scaling (which nevertheless may strongly complicate the software) is the linear transformation:

$$\mathbf{x} = \mathbf{D}\tilde{\mathbf{x}} + \mathbf{c} \quad , \quad (2.22)$$

where $\mathbf{D}$ is a diagonal matrix, and $\tilde{\mathbf{x}}$ is the vector of transformed parameters. The values in $\mathbf{D}$ and $\mathbf{c}$ are chosen such that the optimization problem is well scaled with respect to the new parameters $\tilde{\mathbf{x}}$.

2.7 THE NAG LIBRARY

Most of the optimization problems in this thesis (all nonlinear least squares problems) were solved using routines from the NAG library, mark 10 and 11, running on a Gould minicomputer. Quasi Newton and modified Newton methods were used. The gradient and the Hessian were given, although, for most problems, the latter is approximated by $\mathbf{J}^T\mathbf{J}$, because of the large programming effort for implementing the full Hessian. The constraints were handled by an augmented Lagrangian method. Ideally for these problems would have been special nonlinear least squares routines with a special efficient implementation of the linear and bounds constraints. Regretfully these types of algorithms were not yet available in these releases of the NAG software.

The scaling was done in the application software (some optimization algorithms can perform automatic scaling), using the linear transformation as described in section 2.6. The stopping criteria were set such that the accuracy obtained was at least one order better than required in practice in order to avoid premature stopping with a resulting loss in accuracy.

UNCERTAINTY ANALYSIS

3.1 INTRODUCTION

The solution to an inverse problem when explicit data is available is the a posteriori pdf, computed according to Bayes' rule. Because it is impossible for all but the most trivial problems to inspect the a posteriori pdf through the whole of parameter space its maximum is taken as a point estimate of the parameters. Such a point estimate (which consists of a number for each parameter) is of course a limited amount of information. When the a posteriori pdf is not sharply peaked and a large number of models has almost equal probability one would surely want to know this. An analysis of uncertainties is therefore considered very important.

The two basic questions to be answered in such an analysis procedure are:

- (1) How well is the estimate determined by the a posteriori pdf, i.e. by the combination of data and a priori knowledge?
- (2) What are the respective contributions of data and a priori information?

The answer to the first question can be found by inspection of the form of the a posteriori pdf. The width of the peak determines how well the estimate is determined by the total information available. The second question can be answered by comparison of the a priori pdf and the likelihood function. Uncertainty analysis is very closely related to what in geophysical literature is called resolution analysis. In this thesis the terms "resolution from the total state of information", or shortly "resolution" and "resolution from the data" are used. The first term expresses how well the estimate is determined by the combination of a

priori, observational and theoretical knowledge. With the latter it is expressed to what extent the parameter estimates are determined by the combination of theoretical and observational information (or in more popular terms: "the data") in comparison with the a priori information. It should be realized that with these definitions parameters can be well resolved by the total state of information, while being poorly resolved from the data!

Because it is impossible to inspect the full multi-dimensional functions, the problem is essentially one of information reduction. This information reduction can be achieved in several ways depending on how much information is desired. First, the functions can be inspected along directions of interest. These can be directions along which the parameters are poorly resolved i.e. along which the a posteriori pdf is long-stretched. In practice these will virtually always be directions along which the data do not provide significant information and the a priori information determines the answer instead. The directions are found from the eigenvalue decomposition of the Hessian matrix at the estimated model. This method is discussed in section 3.2.

A stronger reduction of information can be obtained by computation of the a posteriori covariance matrix as discussed in section 3.3. For sufficiently linear models this matrix is the inverse of the Hessian matrix at the estimated model. From the covariance matrix the standard deviations can be computed. These can be interpreted as overall error bounds of the parameters and will often be the most convenient and simplest albeit limited form of information.

It is shown that these approaches to uncertainty or resolution analysis are closely related to methods for resolution analysis as proposed by other authors for linear or linearized forward models. It is shown that for Gaussian distributions and nonlinear models these analysis procedures can be performed more accurately when second derivatives are taken into account in the computation of the Hessian matrix.

Two concepts that are conceptually *not* closely related to the Bayesian approach to inversion are also discussed in this chapter. The first one is (the covariance matrix of) the sampling distribution which is (implicitly) used by Jackson (1979) and Jackson and Matsuura (1985). The second is the well known resolution matrix. Again it is shown that for the Gaussian case both can be computed more accurately when using second derivatives of the forward model in the Hessian matrix. It is also shown however that the interpretation of the resolution matrix is not without problems for the nonlinear case.

3.2 INSPECTION OF THE PDF'S ALONG PRINCIPAL COMPONENTS

The a posteriori pdf $p(\mathbf{x}|\mathbf{y})$ determines the resolution from the total state of information. The most relevant information can be obtained in a practical way and without approximations by determining essential directions in the parameter space and by plotting the relevant functions along these directions.

3.2.1 Principles

Suppose F is the objective function to be minimized in order to find the maximum of the a posteriori pdf. For the particular cases in this thesis $F(\mathbf{x}) = -\ln(p(\mathbf{x}|\mathbf{y}))$. We are interested in the region where $F(\mathbf{x})$ is close to $F(\hat{\mathbf{x}})$, with $\hat{\mathbf{x}}$ the maximum of the a posteriori pdf. In particular we may want to know in what region the difference $F(\mathbf{x}) - F(\hat{\mathbf{x}})$ is smaller than some specified number ε . This is also called an ε -indifference region (Bard, 1974). Following Bard (1974), $F(\mathbf{x})$ is expanded in a Taylor series around $\hat{\mathbf{x}}$, retaining only the first terms:

$$F(\mathbf{x}) \approx F(\hat{\mathbf{x}}) + \mathbf{q}^T \Delta \mathbf{x} + \frac{1}{2} \Delta \mathbf{x}^T \mathbf{H} \Delta \mathbf{x} \quad , \quad (3.1)$$

where

$$\Delta \mathbf{x} = \mathbf{x} - \hat{\mathbf{x}} \quad . \quad (3.2)$$

the gradient $\mathbf{q}$ and the Hessian matrix $\mathbf{H}$ are defined in (2.3) and (2.4) respectively. When $F(\hat{\mathbf{x}})$ is a minimum, $\mathbf{H}$ is (semi) positive definite. In the sequel only the case of an unconstrained minimum is discussed. The case of a constrained minimum is much more complex (Bard, 1974) and will not be treated here. At an unconstrained minimum the gradient vanishes ($\mathbf{q} = \mathbf{0}$) and (3.1) becomes:

$$F(\mathbf{x}) \approx F(\hat{\mathbf{x}}) + \frac{1}{2} \Delta \mathbf{x}^T \mathbf{H} \Delta \mathbf{x} \quad . \quad (3.3)$$

For the boundary of the ε -indifference region we find:

$$2\varepsilon = 2(F(\mathbf{x}) - F(\hat{\mathbf{x}})) = \Delta \mathbf{x}^T \mathbf{H} \Delta \mathbf{x} \quad , \quad (3.4)$$

This is the equation of an N -dimensional ellipsoid. The ellipsoids corresponding to different values of ε are concentric and similar in shape, see figure 1.5c. Far away from the maximum the approximation (3.3) is not valid anymore for nonlinear problems and the ellipses deform. This is clearly visible in figure 1.5c. The so-called canonical form of (3.4) can be obtained by using the eigenvalue decomposition of the (semi) positive definite matrix $\mathbf{H}$:

$$\mathbf{H} = \mathbf{V} \mathbf{\Lambda}^2 \mathbf{V}^T \quad , \quad (3.5)$$

where $\mathbf{V}$ is the matrix which columns are the normalized eigenvectors of $\mathbf{H}$ and the diagonal matrix $\mathbf{\Lambda}^2$ contains the nonnegative eigenvalues, denoted by λ_i^2 . The principle axes of the ellipsoids are in the direction of the so-called vector of canonical parameters which are defined as $\mathbf{r} = \mathbf{V}^T \Delta \mathbf{x}$. With this definition (3.4) can be written as:

$$2\varepsilon = \mathbf{r}^T \mathbf{\Lambda}^2 \mathbf{r} = \sum_i \lambda_i^2 r_i^2 \quad . \quad (3.6)$$

This is the so-called canonical form. The principle axes of the ellipsoids correspond with the coordinate axes in the $\mathbf{r}$ space, the eigenvectors of $\mathbf{H}$. From (3.6) it follows that the lengths of the principle axes are inversely proportional to the square roots λ_i of the eigenvalues. This means that the ellipsoids are long stretched in the directions corresponding to low eigenvalues. Along these directions the parameters are less well resolved. The corresponding concept of weak directions will be defined below.

A practical and still detailed uncertainty and resolution analysis can be carried out by scanning the parameter space along the principle axes and plotting the pdf's or other desired functions like the datamismatch etc. Especially the directions along which the parameters are poorly resolved will be important in practice.

3.2.2 Scaling or transformation of parameters

It should be realized that in general the eigenvalue decomposition of the Hessian matrix as expressed in equation (3.5) is physically meaningless. No physical units can be assigned to the elements of the matrices $\mathbf{V}$ and $\mathbf{\Lambda}$ such that equation (3.5) is consistent. This has been pointed out by Tarantola (1987) for the closely related eigenvalue decomposition of a covariance matrix. The decomposition can nevertheless be carried out as simply a numerical procedure acting on the values that are contained in the matrix $\mathbf{H}$ in some unit system. It is however in general not invariant with respect to linear transformations of the parameters. Simply expressing some of the parameters in different units will change the results of the eigenvalue decompositions. After such a change the eigenvectors correspond to different physical parameter models than before the change. Consider the linear transformation from $\Delta\mathbf{x}$ to $\Delta\tilde{\mathbf{x}}$:

$$\Delta\mathbf{x} = \mathbf{D}\Delta\tilde{\mathbf{x}} \quad , \quad (3.7)$$

where $\mathbf{D}$ is a nonsingular square matrix. Substitution of the parameter transformation in (3.3) yields:

$$F(\mathbf{D}\tilde{\mathbf{x}}) = F(\mathbf{D}\tilde{\mathbf{x}}) + \frac{1}{2} \Delta\tilde{\mathbf{x}}^T \mathbf{D}^T \mathbf{H} \mathbf{D} \Delta\tilde{\mathbf{x}} \quad , \quad (3.8)$$

Let the eigenvalue decomposition of the transformed Hessian matrix be given by:

$$\tilde{\mathbf{H}} = \mathbf{D}^T \mathbf{H} \mathbf{D} = \mathbf{W} \mathbf{\Lambda}^2 \mathbf{W}^T \quad . \quad (3.9)$$

It can be shown using (3.5) and (3.9) that the eigenvectors of $\tilde{\mathbf{H}}$ correspond to those of $\mathbf{H}$ according to the linear transformation (3.7):

$$\mathbf{V} = \mathbf{D} \mathbf{W} \quad (3.10)$$

if and only if $\mathbf{D}$ is a so-called unitary matrix defined by:

$$\mathbf{D}^T = \mathbf{D}^{-1} \quad . \quad (3.11)$$

The directions of the eigenvectors are preserved for the less restrictive condition:

$$\mathbf{D}\mathbf{D}^T = \mathbf{D}^T\mathbf{D} = c\mathbf{I} \quad , \quad (3.12)$$

with c an arbitrary constant. The difference between (3.12) and (3.11) is an irrelevant constant factor for all parameters.

The question arises in which units to express the parameters or, starting from some unit system, which linear transformation to apply such that the eigenvalue decomposition is most meaningful. Simply expressing the parameters in SI units can be a bad choice because parameters of different types can be of very different numerical orders. More useful options are:

1. to express the parameters in units that correspond with the parameter ranges of interest for the problem;
2. to take the a priori standard deviations as units and;
3. to statistically normalize the parameters with the transformation $\mathbf{D} = \mathbf{C}_x^{1/2}$, so that the transformed parameters have the identity matrix as a priori covariance matrix.

Each of these options renders the transformed parameters dimensionless. Option 3 is a very interesting one, especially when considering the least squares problem which is analyzed in the next section.

3.2.3 The least squares problem

In the least squares problem the objective function F is given by (1.91) and is written here in the form of equation (2.6):

$$F = \frac{1}{2} \mathbf{e}^T \mathbf{e} \quad , \quad (3.13)$$

with $\mathbf{e}$ as defined in (2.7) in the form of the partitioned vector:

$$\mathbf{e} = \begin{bmatrix} \mathbf{C}_n^{-1/2} (\mathbf{d} - \mathbf{g}(\mathbf{x})) \\ \mathbf{C}_x^{-1/2} (\mathbf{x}^i - \mathbf{x}) \end{bmatrix} . \quad (3.14)$$

The Jacobian $\mathbf{J}$ is partitioned accordingly:

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_y \\ \mathbf{J}_x \end{bmatrix} = - \begin{bmatrix} \mathbf{C}_n^{-1/2} \mathbf{A} \\ \mathbf{C}_x^{-1/2} \end{bmatrix} , \quad (3.15)$$

with $\mathbf{A} = \partial \mathbf{g} / \partial \mathbf{x}$. It can easily be verified that the Hessian $\mathbf{H}$ can be written as the sum

$$\mathbf{H} = \mathbf{H}_y + \mathbf{H}_x \quad . \quad (3.16)$$

where $\mathbf{H}_y$ is the contribution of the data. It stems from the likelihood function at this point $\mathbf{x}$, see equation (2.10):

$$\mathbf{H}_y = \mathbf{J}_y^T \mathbf{J}_y + \sum_{i=1}^{N_d} e_i \mathbf{T}_i \quad (3.17)$$

where N_d is the number of data points. $\mathbf{H}_x$ is due to the a priori information and reads:

$$\mathbf{H}_x = \mathbf{J}_x^T \mathbf{J}_x = \mathbf{C}_x^{-1} \quad (3.18)$$

After a linear transformation the transformed Hessian will be:

$$\begin{aligned} \tilde{\mathbf{H}} &= \mathbf{D}^T \mathbf{H} \mathbf{D} \\ &= \mathbf{D}^T \mathbf{H}_y \mathbf{D} + \mathbf{D}^T \mathbf{H}_x \mathbf{D} \end{aligned} \quad (3.19)$$

Using the transformation $\mathbf{D} = \mathbf{C}_x^{1/2}$ and substitution of (3.18) leads to:

$$\tilde{\mathbf{H}} = \mathbf{C}_x^{1/2} \mathbf{H}_y \mathbf{C}_x^{1/2} + \mathbf{I} \quad (3.20)$$

Let the eigenvalue decomposition of the first term be given by:

$$\begin{aligned} \tilde{\mathbf{H}}_y &= \mathbf{C}_x^{1/2} \mathbf{H}_y \mathbf{C}_x^{1/2} \\ &= \mathbf{V} \Lambda_y \mathbf{V}^T \end{aligned} \quad (3.21)$$

The eigenvalue decomposition of $\mathbf{H}$ then is:

$$\tilde{\mathbf{H}} = \mathbf{V}(\Lambda_y + \mathbf{I})\mathbf{V}^T \quad (3.22)$$

The eigenvalue spectrum of $\mathbf{H}$ is built up by two terms:

$$\lambda_i^2 = \lambda_{yi} + 1 \quad (3.23)$$

Because $\mathbf{H}$ is positive definite (semi-positive definite in exceptional cases) all eigenvalues are nonnegative and can therefore be written as squares. It follows from (3.23) that the diagonal values λ_{yi} of Λ_y cannot be less than -1 . The larger value on the righthand side of (3.23) most strongly determines the curvature of the objective function F along the direction of the eigenvector $\mathbf{v}_i$. Therefore, directions for which $\lambda_{yi} < 1$ can be defined as ill-resolved (with respect to the data!) or weak directions. In these directions the a priori information more strongly determines the a posteriori pdf than the likelihood function does.

For low residuals and/or for sufficiently linear models the second term of the right hand side of (3.17) vanishes. In that case we have for the first term of (3.20):

$$\begin{aligned} \tilde{\mathbf{H}}_y &= \mathbf{C}_x^{1/2} \mathbf{J}_y^T \mathbf{J}_y \mathbf{C}_x^{1/2} \\ &= \mathbf{C}_x^{1/2} \mathbf{A}^T \mathbf{C}_n^{-1} \mathbf{A} \mathbf{C}_x^{1/2} \\ &= \tilde{\mathbf{A}}^T \tilde{\mathbf{A}} \end{aligned} \quad (3.24)$$

where $\mathbf{A}$ is the sensitivity or forward matrix of the linear or linearized problem, weighted

with the uncertainties of noise and a priori information:

$$\tilde{\mathbf{A}} = \mathbf{C}_n^{-1/2} \mathbf{A} \mathbf{C}_x^{1/2} \quad (3.25)$$

The approximation of $\mathbf{H}_y$ as given in (3.24) is (semi) positive definite and its eigenvalue decomposition can therefore be written as:

$$\tilde{\mathbf{H}}_y = \mathbf{V} \mathbf{S}^2 \mathbf{V}^T \quad (3.26)$$

where the matrices $\mathbf{V}$ and $\mathbf{S}$ are (by definition) exactly the same as the ones occurring in the singular value decomposition (SVD) of:

$$\tilde{\mathbf{A}} = \mathbf{U} \mathbf{S} \mathbf{V}^T \quad (3.27)$$

These results clearly establish the link between this approach of uncertainty and resolution analysis and those as given for linear models using the SVD as given by e.g. Jackson (1973, 1976) and Van Riel and Berkhout (1985). Using relation (3.17) in full form is of course more accurate for nonlinear models, see also Kennett (1978).

In the examples shown in this thesis the problem will always be statistically normalized and the second term of the Hessian will be neglected. The spectra used are the singular value spectra of the relevant Jacobian matrices because singular values represent the reciprocal of a distance in the (statistically normalized) parameter space. For an easy overview the relevant equations for this situation are given here. The basic matrix is the statistically normalized Jacobian matrix which is built up of a data and a priori part:

$$\tilde{\mathbf{J}} = \begin{bmatrix} \tilde{\mathbf{J}}_y \\ \tilde{\mathbf{J}}_x \end{bmatrix} = - \begin{bmatrix} \mathbf{C}_n^{-1/2} \mathbf{A} \mathbf{C}_x^{1/2} \\ \mathbf{I} \end{bmatrix} \quad (3.28)$$

The SVD's of the matrices are given by:

$$\tilde{\mathbf{J}} = \mathbf{U} \mathbf{\Lambda} \mathbf{V}^T \quad (3.29)$$

$$\tilde{\mathbf{J}}_y = \mathbf{U}_y \mathbf{S} \mathbf{V}^T \quad (3.30)$$

$$\tilde{\mathbf{J}}_x = \mathbf{V} \mathbf{I} \mathbf{V}^T \quad (3.31)$$

The Hessian is given by:

$$\tilde{\mathbf{H}} = \tilde{\mathbf{H}}_y + \tilde{\mathbf{H}}_x \quad (3.32)$$

in approximated form:

$$\tilde{\mathbf{J}}^T \mathbf{J} = \tilde{\mathbf{J}}_y^T \tilde{\mathbf{J}}_y + \tilde{\mathbf{J}}_x^T \tilde{\mathbf{J}}_x \quad (3.33)$$

and, using (3.29) – (3.31):

$$\mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^T = \mathbf{V}\mathbf{S}^2\mathbf{V}^T + \mathbf{V}\mathbf{I}\mathbf{V}^T \quad (3.34)$$

In practice only the SVD of $\mathbf{J}_y$ needs to be computed. For the spectra it follows:

$$\mathbf{\Lambda}^2 = \mathbf{S}^2 + \mathbf{I} \quad (3.35)$$

The elements of the a posteriori spectrum follow from:

$$\lambda_i = \sqrt{s_i^2 + 1} \quad (3.36)$$

3.2.4 The two-parameter example

The procedure can be nicely illustrated for the two-parameter example as discussed in section 1.13. The parameters are statistically normalized as discussed above: $\mathbf{D} = \mathbf{C}_x^{1/2}$. The eigenvalue analysis is based on (3.28) - (3.36). The singular values s_i of the reweighted forward matrix are:

$$\begin{aligned} s_1 &= 9.083 \\ s_2 &= .127 \end{aligned} \quad (3.37)$$

The square roots of the eigenvalues of $\mathbf{H}$ follow from (3.36):

$$\begin{aligned} \lambda_1 &= 9.138 \\ \lambda_2 &= 1.008 \end{aligned} \quad (3.38)$$

The principal axis in direction 2 is approximately 9 times as long as the one in direction 1. This can be verified in figure 3.1, where the contours of the a posteriori pdf are once more given. The ranges plotted along the x and y axis are equal in terms of a priori standard deviations. The eigenvectors are given by the columns of the matrix $\mathbf{V}$:

$$\mathbf{V} = \begin{bmatrix} .0684 & -.9977 \\ .9977 & .0684 \end{bmatrix} \quad (3.39)$$

It can easily be verified that $\mathbf{V}$ is orthonormal. The directions of the eigenvectors are drawn in figure 3.1. The second direction, corresponding with the small eigenvalue is in the direction of decreasing impedance and increasing layer thickness. Along this direction the synthetic data $g(\mathbf{x})$ and hence the likelihood function varies only very slowly. The direction can be analytically verified by expansion of (1.101) in a Taylor series around the estimated model. When only the first order terms are retained, the linear combination of parameters for which the (approximated) function remains constant can directly be determined. The eigenvector 2 results.

The uncertainty analysis along eigenvectors can be done by plotting the a priori pdf, the likelihood function and the a posteriori pdf around the maximum of the a posteriori pdf.

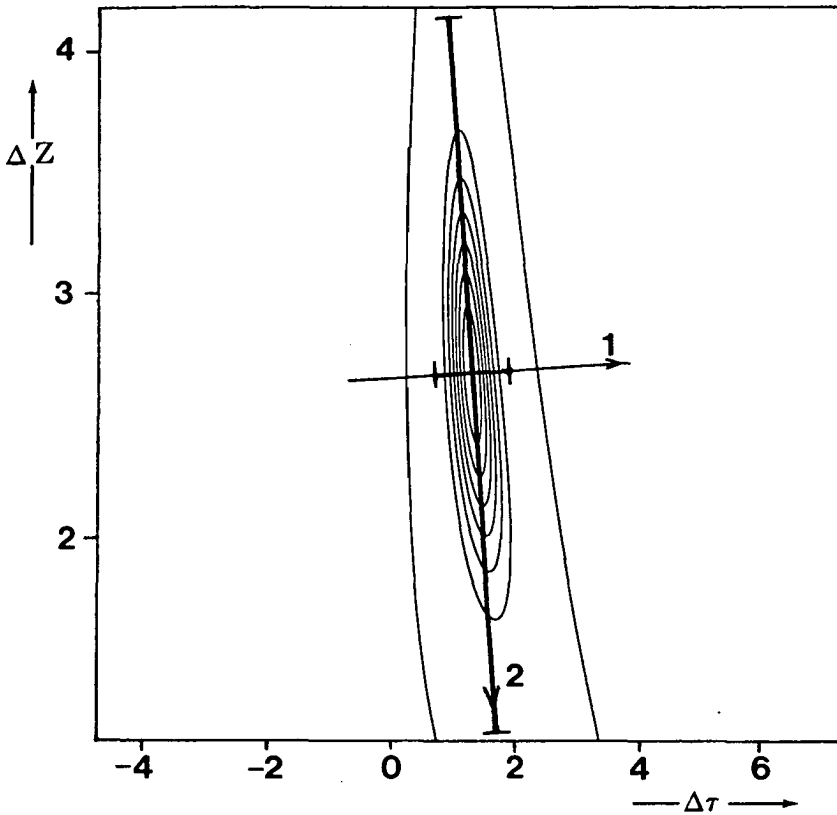


Figure 3.1 The eigenvectors in the two-parameter example. The ranges used for plotting of the pdf's are indicated.

The ranges that are visualized are indicated in figure 3.1 by thick parts. The plots of the pdf's themselves are given in figure 3.2. The ranges (6 units in the canonical system) are chosen such that the a posteriori pdf nicely fits in the plot. Along the x-axis the corresponding impedance models are plotted. For a more precise evaluation of the ranges of the models the middle (estimated) and the 'extreme' models are once more plotted at the bottom of the pictures. It is clear that along direction 1 the likelihood function determines the a posteriori pdf or, in more popular terms, the data determines the answer. Along direction 2 on the other hand, the a priori information determines the answer.

In this example the plots of course hardly add any information to what we already knew from the full 2-dimensional plots. In, for example, a 10-dimensional problem however, the full pdf's cannot be plotted while 10 plots along the eigenvectors are still practical and may give a good idea of the functions involved.

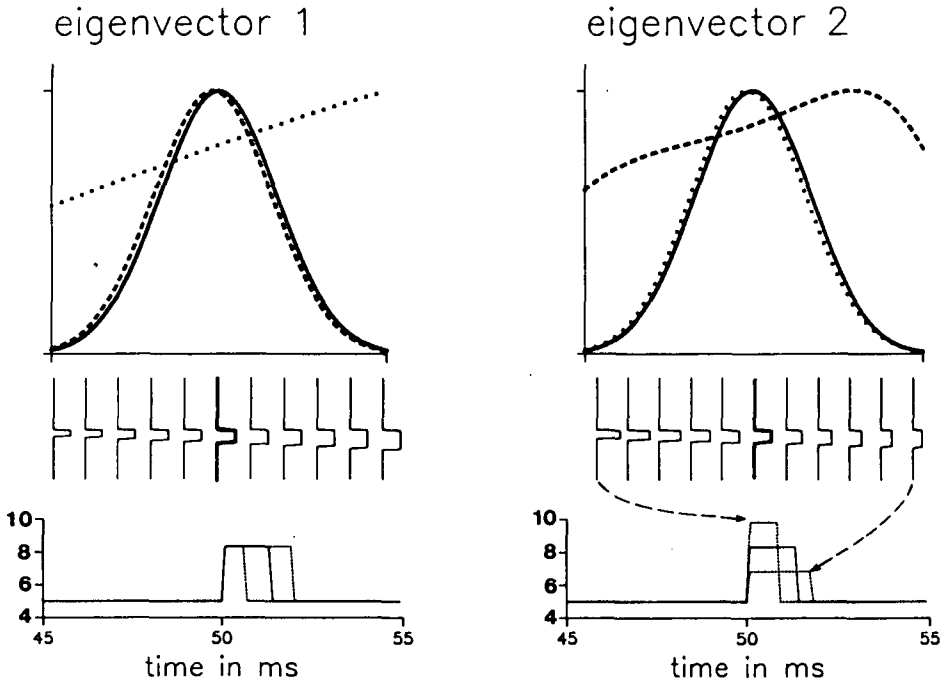


Figure 3.2 Pdf's along the eigenvectors. The estimated and the extreme models are plotted in the bottom picture.

..... a priori pdf; ----- likelihood function; _____ a posteriori pdf

3.3 THE A POSTERIORI COVARIANCE MATRIX

3.3.1 Computation

An analysis of uncertainties or resolution can be given in more concise form by covariance matrices. The definition of the covariance matrix is given in section 1.6. For nonlinear models an exact computation of the a posteriori covariance matrix would involve a full computation of the a posteriori pdf through the whole of parameter space. This is computationally too demanding for all but the most trivial cases. Fortunately a good approximation is possible when the forward model is linear enough. Equation (3.3) gives the approximation of the objective function F around the maximum $\hat{x}$:

$$F(\mathbf{x}) \approx F(\hat{\mathbf{x}}) + \frac{1}{2} (\mathbf{x} - \hat{\mathbf{x}})^T \mathbf{H}(\mathbf{x} - \hat{\mathbf{x}}) \quad (3.40)$$

When F is the negative logarithm of the a posteriori pdf, we have for the latter:

$$p(\mathbf{x}|\mathbf{y}) \approx p(\hat{\mathbf{x}}|\mathbf{y}) \cdot \exp \left\{ -\frac{1}{2} (\mathbf{x} - \hat{\mathbf{x}})^T \mathbf{H} (\mathbf{x} - \hat{\mathbf{x}}) \right\} . \quad (3.41)$$

This means that around the maximum the a posteriori pdf looks like a Gaussian distribution with mean $\mathbf{x} = \hat{\mathbf{x}}$ and covariance matrix $\mathbf{C}_{\hat{\mathbf{x}}} = \mathbf{H}^{-1}$. This holds irrespective of the distributions involved for noise and a priori information, provided of course that the approximation (3.40) holds. For linear models and Gaussian assumptions equations (3.40) and (3.41) are exact. The a posteriori covariance is then given by:

$$\begin{aligned} \mathbf{C}_{\hat{\mathbf{x}}} &= \mathbf{H}^{-1} \\ &= (\mathbf{J}^T \mathbf{J})^{-1} \\ &= (\mathbf{A}^T \mathbf{C}_n^{-1} \mathbf{A} + \mathbf{C}_x^{-1})^{-1} , \end{aligned} \quad (3.42)$$

where equation (3.15) has been used. This is a well known result, see e.g. Bard (1974). For nonlinear models relations (3.16) – (3.19) have to be used. It should be realized that only when the a posteriori pdf is low enough far from the maximum and approximation (3.41) holds for the whole of parameter space the inverse Hessian $\mathbf{H}^{-1}$ is a good approximation for the a posteriori covariance matrix.

It is illustrating to consider two extreme situations:

1. The data fully determines the answer; the shape of the a posteriori pdf is fully determined by the likelihood function. We then have:

$$\mathbf{C}_{\hat{\mathbf{x}}} = \mathbf{H}_y^{-1} , \quad (3.43)$$

with $\mathbf{H}_y$ given in (3.17). For linear models the covariance matrix equals the one of the well known Gauss-Markov estimator:

$$\mathbf{C}_{\hat{\mathbf{x}}} = \mathbf{A}^T \mathbf{C}_n^{-1} \mathbf{A} . \quad (3.44)$$

Note however that the a posteriori pdf conceptually differs from the sampling distribution of which in classical statistics the covariance matrix for the Gauss-Markov estimator is derived.

2. When the a priori information fully determines the answer we have:

$$\mathbf{C}_{\hat{\mathbf{x}}} = \mathbf{H}_x^{-1} = \mathbf{C}_x . \quad (3.45)$$

The eigenvalue decomposition can provide some more insight in the nature of the a posteriori covariance matrix. Consider the case where the parameters are weighted with $\mathbf{D} = \mathbf{C}_x^{-1/2}$ (statistically normalized), and where $\mathbf{H}_y$ is positive definite as is the case for linear models. We then have, using (3.20) and (3.26):

$$\begin{aligned}
 \tilde{C}_{\hat{x}} &= \tilde{H}^{-1} \\
 &= V(S^2 + I)^{-1} V^T \\
 &= \sum_i \frac{1}{s_i^2 + 1} v_i v_i^T .
 \end{aligned} \tag{3.46}$$

The values s_i are inversely proportional to the square root of the noise power. It can be seen from (3.46) that for the extreme case of the noise power going to zero, the a posteriori covariance matrix will still not vanish when there are s_i with value zero. There will always remain a rest term:

$$C_{\hat{x}} = \sum_{i \in 0} v_i v_i^T , \tag{3.47}$$

where 0 denotes the set of eigenvectors for which $s_i=0$. Along these directions the data does not provide any information, no matter how low the noise level.

3.3.2 The two-parameter example

Comparison of the a posteriori covariance matrix with the a priori covariance matrix reveals how much information has been gained by using the information from the data. An illustration is again given for the two-parameter example. The a priori and a posteriori covariance matrices are:

$$C_x = \begin{bmatrix} 2.5 \cdot 10^{11} & 0 \\ 0 & 4.10^{-6} \end{bmatrix} , \quad C_{\hat{x}} = \begin{bmatrix} 2.45 \cdot 10^{11} & -66. \\ -66. & 6.8 \cdot 10^{-8} \end{bmatrix} ,$$

in the (SI) units:

$$\begin{bmatrix} \text{kg}^2 \text{m}^{-4} \text{s}^{-2} & \text{kgm}^{-2} \text{s}^{-2} \\ \text{kgm}^{-2} \text{s}^{-2} & \text{s}^{-2} \end{bmatrix} .$$

An absolute interpretation of the a posteriori covariance in such a unit system is not easy. A direct comparison with the a priori covariance matrix provides information that is easier to interpret. This especially holds for the diagonal values which give the variances of the parameters. Their square roots, the standard deviations are given by:

$$\sigma_{\Delta Z} = .5 \cdot 10^6 \text{kgm}^{-2} \text{s}^{-1} , \quad \sigma_{\Delta \dot{Z}} = .495 \cdot 10^6 \text{kgm}^{-2} \text{s}^{-1} ,$$

$$\sigma_{\Delta \tau} = 2. \cdot 10^{-3} \text{s} \quad \sigma_{\Delta \dot{\tau}} = .26 \cdot 10^{-3} \text{s} .$$

They can be interpreted as overall uncertainty measures (although this is not always

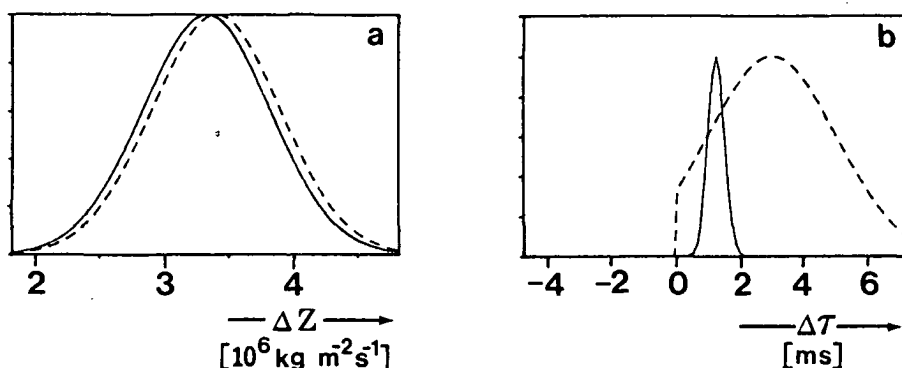


Figure 3.3 A priori and approximated a posteriori pdf's for the acoustic impedance ΔZ (a) and the time thickness $\Delta \tau$ (b).

..... a priori pdf
 ——— a posteriori pdf

appropriate when the form of the pdf is far from Gaussian). In figure 3.3 the a priori pdf's and the approximated marginal a posteriori pdf's are given. In the region where the latter are nonzero they are Gaussian pdf's with the estimated values as means and the standard deviations as given above. A convenient way of display for this type of problem is to plot the a priori and a posteriori models with the standard deviations as uncertainty bounds, see figure 3.4. From the numbers and the figures it is clear that hardly any information has been gained on the acoustic impedance. The standard deviation of the thickness on the other hand has been reduced by nearly a factor 8. The a posteriori standard deviations provide uncertainty intervals for the estimated model. In figure 3.4c it is shown that for this specific example the true model indeed lies within these intervals.

The interpretation of a posteriori covariance matrix may be easier when the parameters are reweighted with the a priori covariance matrix. We then have for the transformed covariance matrices:

$$\tilde{\mathbf{C}}_x = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \tilde{\mathbf{C}}_A = \begin{bmatrix} .98 & -.066 \\ -.066 & 0.017 \end{bmatrix}$$

The elements are dimensionless. The a posteriori covariance matrix can now directly be compared with the identity matrix which is advantageous when, for larger problems, the a posteriori covariance matrix is plotted rather than given in explicit numbers, see chapter 5. The off-diagonal elements, the covariances, are most easily studied when the covariance

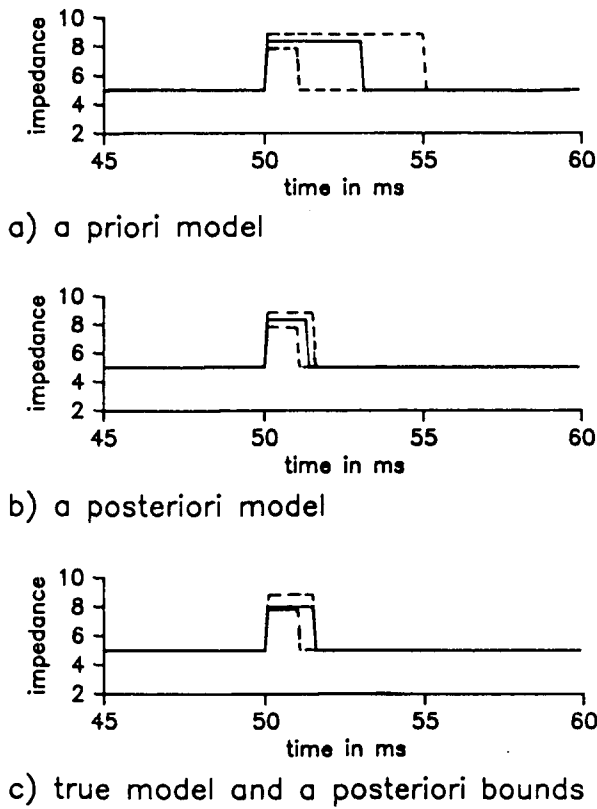


Figure 3.4 A priori (a) and a posteriori (b) model with uncertainty intervals. In (c) the true model is given with the a posteriori uncertainty intervals.

matrices are normalized on their variances. The correlation matrix (see section 1.6) is the result. We get:

$$\begin{array}{l} \text{a priori} \\ \text{correlation} = \\ \text{matrix} \end{array} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \begin{array}{l} \text{a posteriori} \\ \text{correlation} = \\ \text{matrix} \end{array} \begin{bmatrix} 1 & .51 \\ .51 & 1 \end{bmatrix}$$

There is no correlation in the a priori information. The a posteriori correlation coefficient of .51 indicates that, due to the exchange effect, the parameters are not determined independently.

3.3.3 Linearity check

In an analysis procedure as the one described above one would like to know over what range the quadratic approximation (3.40) is valid. Sometimes it can analytically be shown that the forward model is linear enough. A practical numerical procedure is to scan the parameter space along the eigenvectors and to plot the exact and approximated a posteriori pdf's. For the two-parameter example this is illustrated in figure 3.5. Along eigenvector 2 the approximation is excellent. Along eigenvector 1 there are some very small differences.

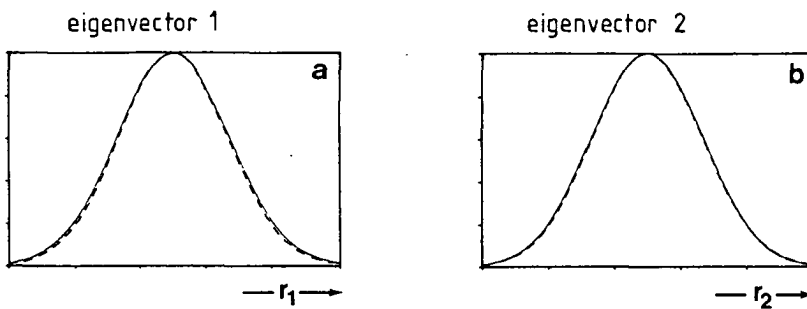


Figure 3.5 Linearity analysis along the eigenvectors.

— true a posteriori pdf
 approximated a posteriori pdf

3.3.4 Standard deviations and marginal pdf's

The standard deviation σ_i of a parameter is defined as the square root of:

$$\sigma_i^2 = \int (x_i - E x_i)^2 p(x_i) dx_i, \quad (3.48)$$

where $p(x_i)$ is the marginal pdf for x_i , see section 1.6. The marginal pdf reflects the 'average' information for x_i . The standard deviation σ_i is therefore an 'average' or 'overall' uncertainty measure. For linear models equation (3.41) for the a posteriori pdf is exact. The marginal pdf of a Gaussian pdf with mean $\bar{x}$ and covariance C_x is itself Gaussian with mean $\bar{x}_i$ and standard deviation σ_i and is therefore exactly known.

For nonlinear models the true marginal pdf may differ from the approximated one as derived from (3.41). In figure 3.6 this is illustrated for the two-parameter problem. The true marginal pdf's are computed from the two-dimensional pdf. The approximation is very good for the time thickness. The true marginal pdf of the acoustic impedance is slightly

shifted to lower values. This can be explained from the contours for lower values of the

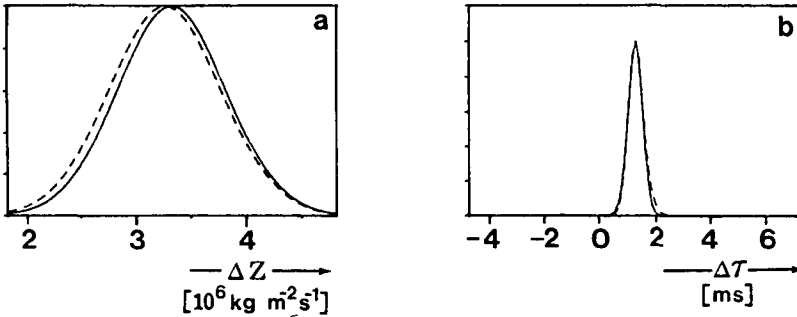


Figure 3.6 True and approximated marginal pdf's for the acoustic impedance (a) and the time thickness (b).

— approximated marginal pdf
 true marginal pdf

two-dimensional pdf as shown in figure 1.5. A point estimate of ΔZ based on the maximum of the marginal pdf would yield a lower value than the MAP estimate given in table 1.1 and would therefore be a better estimate. This is an illustration of the fact that parameters in which one is not interested, so-called nuisance parameters, should ideally be integrated out. In this example, integrating over $\Delta \tau$ yields a better estimate for ΔZ . Unfortunately integrating over nuisance parameters is often not possible analytically and too costly to be done numerically.

3.4 SAMPLING DISTRIBUTIONS

A concept that hardly ever occurs in a Bayesian context but is essential for the frequentist school is that of the sampling distribution. It arises when the point estimate is considered as a random variable that takes different values for different realizations of the data. The sampling distribution describe the variations of the estimate. The corresponding covariance matrix C_{sy} can be approximated by (Bard, 1974):

$$C_{sy} \approx H^{-1} A^T C_n^{-1} A H^{-1}, \quad (3.49)$$

for Gaussian assumptions and sufficiently linear models. The Hessian H and the sensitivity matrix A are evaluated at the estimated model for the actual data set under consideration. Accordingly, a sampling distribution due to variations in the a priori information can also be defined. Its covariance matrix can be approximated by:

$$\mathbf{C}_{sx} \approx \mathbf{H}^{-1} \mathbf{C}_x^{-1} \mathbf{H}^{-1} \quad , \quad (3.50)$$

It has the following interesting interpretation. When the a priori information is varied with covariance $\mathbf{C}_x$, the resulting point estimate will vary with covariance $\mathbf{C}_{sx}$. For the linear Gaussian case Jackson and Matsu'ura (1985) derived the combination of these two expressions in the following way (although they do not mention the concept of sampling distribution). The Bayesian estimator for the linear Gaussian case can be written as the sum of two operators, working on data and 'a priori data' respectively:

$$\hat{\mathbf{x}} = \mathbf{K}\mathbf{y} + \mathbf{L}\mathbf{x}^i \quad , \quad (3.51)$$

with

$$\mathbf{K} = \mathbf{H}^{-1} \mathbf{A}^T \mathbf{C}_n^{-1} \quad (3.52)$$

and

$$\mathbf{L} = \mathbf{H}_x^{-1} \mathbf{C}_x^{-1} \quad . \quad (3.53)$$

$\mathbf{H}$ is the Hessian for the linear case:

$$\mathbf{H} = \mathbf{A}^T \mathbf{C}_n^{-1} \mathbf{A} + \mathbf{C}_x^{-1} \quad (3.54)$$

Jackson and Matsu'ura compute the "a posteriori" covariance, which is in fact the sampling covariance due to data and a priori information, here denoted with $\mathbf{C}_s$. It follows from (3.51):

$$\begin{aligned} \mathbf{C}_s &= \mathbf{K} \mathbf{C}_n^{-1} \mathbf{K}^T + \mathbf{L} \mathbf{C}_x^{-1} \mathbf{L}^T \\ &= \mathbf{H}^{-1} \mathbf{A}^T \mathbf{C}_n^{-1} \mathbf{A} \mathbf{H}^{-1} + \mathbf{H}^{-1} \mathbf{C}_x^{-1} \mathbf{H}^{-1} \quad . \end{aligned} \quad (3.55)$$

Although for linear models the righthand side, the sum of the two sampling covariance matrices, is indeed equal to the a posteriori covariance as given in (3.42), this is not at all obvious. The a posteriori pdf is a fundamentally different concept from the one of sampling distribution. It is nevertheless interesting to study the extreme cases of (3.55). For very low noise levels the first term on the right hand side of (3.55) will dominate the second one, assuming $\mathbf{A}$ has no zero singular values. This is not surprising. For low noise levels the data determines the answer so that varying the a priori information does not alter the solution much. For very accurate a priori information the second term will dominate the first one.

As an illustration the terms as appearing in (3.55) are given for the two-parameter example in a statistically normalized parameter system:

$$\begin{bmatrix} .980 & -.066 \\ -.066 & 0.17 \end{bmatrix} = \begin{bmatrix} .016 & 0. \\ 0. & .012 \end{bmatrix} + \begin{bmatrix} .964 & -.066 \\ -.066 & .005 \end{bmatrix}$$

In accordance with the preceding remarks it turns out that for the acoustic impedance the 'a priori sampling distribution' is dominant while for the thickness the data sampling distribution is dominant. Within the numerical accuracy in this example, the off-diagonal elements of the total sampling covariance matrix C_s are entirely due to those of the sampling distribution of the a priori information. Again this need not be surprising. Providing at random uncorrelated a priori models yields a correlation in the sampling distribution of the estimates because the long stretched likelihood function forces a preferred linear combination of parameter estimates. On the other hand, when keeping the a priori information fixed and changing the data at random, no correlation enters the distribution of the estimates, because the a priori information does not force preferred linear combinations.

In a true Bayesian interpretation sampling distributions are unnatural concepts. Their utilization is not recommended. The a priori and a posteriori covariance matrices provide all information required.

3.5 THE RESOLUTION MATRIX

Another concept that is often introduced in estimation problems is the one of the resolution matrix. Like the sampling distributions it is not a natural concept in a Bayesian analysis and as we shall see its interpretation is not without problems in the nonlinear case. In order to derive it an (approximate) linear relation between the estimated and the true parameter values has to be found. The Gaussian case is analyzed here. The position of the maximum of the a posteriori pdf is denoted by $\mathbf{x}_m$. We can write for the estimated model, using (2.5) and (2.8) in a different notation:

$$\mathbf{H}(\hat{\mathbf{x}} - \mathbf{x}_m) = -\mathbf{J}^T \mathbf{e} , \quad (3.56)$$

where $\mathbf{H}$ and $\mathbf{J}$ are evaluated at $\mathbf{x}_m$. The actual data $\mathbf{d}$ as it occurs in $\mathbf{e}$ can be written as:

$$\mathbf{d} = \mathbf{g}(\mathbf{x}_t) + \mathbf{n}_t , \quad (3.57)$$

where $\mathbf{x}_t$ and $\mathbf{n}_t$ denote the true model and the true noise respectively. If $\mathbf{g}(\mathbf{x})$ can be approximated by the first terms of a Taylor expansion around the maximum $\mathbf{x}_m$, we have for $\mathbf{x}=\mathbf{x}_t$:

$$\mathbf{g}(\mathbf{x}_t) \approx \mathbf{g}(\mathbf{x}_m) + \mathbf{A}(\mathbf{x}_t - \mathbf{x}_m) , \quad (3.58)$$

where, again, $\mathbf{A}$ is evaluated at $\mathbf{x}_m$. We can substitute this in (3.56) and (3.57) in order to

obtain:

$$\mathbf{H}(\hat{\mathbf{x}} - \mathbf{x}_m) = -\mathbf{J}^T \begin{bmatrix} \mathbf{C}_n^{-1/2} \mathbf{A}(\mathbf{x}_i - \mathbf{x}_m) + \mathbf{C}_n^{-1/2} \mathbf{n}_i \\ \mathbf{C}_x^{-1/2} (\mathbf{x}^i - \mathbf{x}_m) \end{bmatrix}. \quad (3.59)$$

The resolution matrix $\mathbf{R}$ is defined through the linear relation between the estimated and the true model:

$$\hat{\mathbf{x}} - \mathbf{x}_m = \mathbf{R}(\mathbf{x}_i - \mathbf{x}_m) + \mathbf{c}, \quad (3.60)$$

with $\mathbf{c}$ a vector of constants. The combination of (3.59) with (3.60) yields, using (2.13):

$$\mathbf{R} = \mathbf{H}^{-1} \mathbf{A}^T \mathbf{C}_n^{-1} \mathbf{A}. \quad (3.61)$$

Again, for nonlinear models this expression is more accurate than the usually given expression for the linear case, see e.g. Jackson (1979):

$$\mathbf{R} = (\mathbf{A}^T \mathbf{C}_n^{-1} \mathbf{A} + \mathbf{C}_x^{-1})^{-1} \mathbf{A}^T \mathbf{C}_n^{-1} \mathbf{A}. \quad (3.62)$$

It is stressed that the expressions for the resolution matrix (3.61) and (3.62) are only valid when approximation (3.58) is valid, i.e. when the true model is close enough to the maximum of the a posteriori pdf. This is of course never known in practice, so we never know whether $\mathbf{R}$ is the true operator that maps the true parameters on the estimated ones. The a posteriori pdf however, gives the probability that the true model is close to the maximum. Linearizing at another model makes no sense. If it has a higher probability of being close to the true model it should replace the maximum as the point estimate.

From the equations (3.42) and (3.62) an interesting relation between the resolution matrix and the a posteriori covariance matrix can be derived for the linear Gaussian case:

$$\mathbf{R} = \mathbf{I} - \mathbf{C}_x \mathbf{C}_x^{-1} \quad (3.63)$$

This relation has also been found by Tarantola (1987). The extreme situations can again provide some insight. When the data fully determines the answer we will have $\mathbf{C}_x \ll \mathbf{C}_x$ and thus $\mathbf{R}=\mathbf{I}$, meaning full resolution from the data. When the a priori information fully determines the answer we will have $\mathbf{C}_x = \mathbf{C}_x$ and thus $\mathbf{R}=\mathbf{O}$, meaning no resolution from the data at all. From equation (3.63) it can be seen that the resolution matrix is typically a measure of the balance between the information supplied by the data and the information supplied by the a priori information.

When the parameters are statistically normalized on the a priori information we have the very simple relation:

$$\tilde{\mathbf{R}} = \mathbf{I} - \tilde{\mathbf{C}}_x. \quad (3.64)$$

The interesting features of the resolution matrix are the deviations from the identity matrix and these turn out to be exactly given by the a posteriori covariance matrix $C_{\hat{x}}$!. The relation with the eigenvalue analysis is immediately clear for the linear normalized case. Using (3.46) we can rewrite (3.64) as:

$$\tilde{\mathbf{R}} = \mathbf{I} - \sum_i \frac{1}{s_i^2 + 1} \mathbf{v}_i \mathbf{v}_i^T \quad (3.65)$$

The low singular values s_i determine the deviations of $\mathbf{R}$ from the ideal shape $\mathbf{I}$. Jackson and Matsu'ura (1985) and Tarantola (1987) suggest the following interpretation of the traces of matrices $\mathbf{I}$, $\mathbf{R}$ and $\mathbf{I}-\mathbf{R}$:

$$\text{trace}(\mathbf{I}) = n = \text{the total number of parameters}$$

$$\text{trace}(\tilde{\mathbf{R}}) = \sum_i \frac{s_i^2}{s_i^2 + 1} = \text{the number of parameters determined by the data}$$

$$\text{trace}(\mathbf{I}-\tilde{\mathbf{R}}) = \sum_i \frac{1}{s_i^2 + 1} = \text{the number of parameters determined by the a priori information}$$

The expressions with the singular values hold only for the linear normalized case.

The resolution matrix for the two-parameter example is:

$$\mathbf{R} = \begin{bmatrix} .02 & 1.65 \cdot 10^7 \text{ kgm}^{-2} \text{ s}^{-1} \\ 2.64 \cdot 10^{-10} \text{ kgm}^2 \text{ s}^2 & .983 \end{bmatrix}$$

The off-diagonal elements are not dimensionless, unless all parameters are of the same type. It is clear that the resolution matrix in the normalized system is much easier to interpret:

$$\tilde{\mathbf{R}} = \begin{bmatrix} .02 & .066 \\ .066 & .983 \end{bmatrix}$$

The acoustic impedance is very poorly resolved from the data. The thickness is very well resolved.

3.6 SUMMARY OF UNCERTAINTY ANALYSIS

In this section a summary is given of the steps that can be taken in an uncertainty or resolution analysis for small and medium sized problems. Four types of quantities can be distinguished, given in the order of simple to complex:

- (1) standard deviations
- (2) covariance matrices
- (3) eigenvalue spectra
- (4) pdf's

Standard deviations are the simplest type of information. They can be interpreted as overall error bounds on the parameters when the a posteriori pdf does not deviate too much from the Gaussian form, and will be comprehensible to a person with little knowledge of statistics. Often they can be displayed in an easy to comprehend way, see e.g. chapter 5. A priori and a posteriori standard deviations can be compared in order to assess how much information has been gained from the data. It has to be kept in mind however that standard deviations form a limited amount of information. Strong dependencies (correlations) between parameters are *not* visualized. Furthermore the problem has to be linear enough in order to allow the computation of the a posteriori covariance matrix from which they are derived.

More detailed information can be derived from covariance and correlation matrices. Like above, a priori and a posteriori covariance matrices can be compared for an assessment of the information brought in by the data. For a comprehensible display, the parameters can best be statistically normalized on the a priori information. The a priori covariance matrix then equals the identity matrix. For the a posteriori correlations, the correlation matrix may be more useful than the covariance matrix. Note that for nonlinear problems only an *approximation* of the covariance matrix can be computed. The problem has to be linear enough for the approximation to be accurate. If this is not the case, the computed matrix may nevertheless be useful, because it describes the curvature of the a posteriori pdf around the maximum.

The third type of information are eigenvalue spectra. The square roots of eigenvalues are most useful, because they are inversely proportional to the lengths of the ellipsoids that are the contours of the a posteriori pdf around the maximum. Again it is most useful to statistically transform the parameters according to the a priori information, especially for the Gaussian problem. The a priori spectrum then has values of one. Where the eigenvalues of the data part of the Hessian matrix are lower than one, the a priori information determines the answer for the corresponding linear combination of parameters. From these plots an impression can be obtained along how many directions the data determines the answer.

The information from the eigenvalue spectra can be augmented by plotting the functions (a priori pdf, likelihood function and a posteriori pdf) along desired directions. This gives some impression of the actual shape of the functions around the estimated model. Unlike the covariance matrix, these plots can be computed without approximations.

It is once more stated that for Gaussian problems the Hessian matrix can be computed more accurately than is done in most studies by taking the second derivatives of the forward model into account. It is also emphasized that the eigenvalue analysis is still useful when one rejects the probabilistic basis of the inversion approach. Instead of pdf's quantities like energy of data mismatch etc. can then be plotted. The procedures as sketched in this chapter are used and illustrated in the sequel of this thesis, see especially chapters 5 and 6.

GENERAL ASPECTS OF DETAILED INVERSION OF POSTSTACK DATA

4.1 INTRODUCTION

Conventional processing of seismic data renders a stacked and possibly migrated data set, representing an image of the subsurface in terms of reflectivity. In practice it is bandlimited and contaminated by noise. An interpreter tries to derive detailed information from it usually concerning some target zone. A visual derivation of the geometries and physical properties like densities and propagation velocities can be very difficult or even impossible especially when thin layers are involved.

In the past wavelet processing techniques have been developed in order to improve the vertical resolution. However even after an optimal wavelet deconvolution the data remain bandlimited and contaminated by noise. The vertical resolution of seismic traces and the resolving power of wavelets have been studied by several authors, see e.g. Widess (1973), Kalweit and Wood (1982), Koefoed (1981) and Berkhout (1974, 1984). In particular the response of a thin layer has been studied extensively and methods for estimating the thickness of thin layers have been proposed, see e.g. De Voogd and Den Rooyen (1983). Robertson and Nogami (1984) propose to use complex trace attributes to verify the presence of thin beds in seismic sections. These methods however cannot yield reliable estimates in more complex situations.

Detailed inversion methods aiming at high accuracy and/or the handling of more complex geometries are usually model based and can be seen as parameter estimation

methods. Before giving a short overview of methods proposed in literature (section 4.3) an overview of some basic considerations in setting up an inversion scheme is given (section 4.2).

4.2 DESIGN CONSIDERATIONS

When setting up a scheme for the detailed inversion of poststack seismic data decisions have to be made concerning a number of items. Seven items may be distinguished:

1. parameterization
2. forward model (theoretical relations)
3. noise properties
4. the wavelet
5. a priori information
6. optimization
7. uncertainty/resolution analysis.

Most of these items are strongly related. Decisions concerning one item influence decisions concerning others. The items are discussed in more detail now:

1. *Parameterization*

First of all a choice concerning the type of physical quantities to be estimated has to be made. One may think of reflectivity, of acoustic impedance, of density and velocity or even parameterizations including elastic or rock parameters. Vertical profiles can be functions of time or depth. Second, a choice concerning the lateral and vertical variations has to be made. As far as the vertical variations are concerned, one may think of a general model, which usually will have parameters on an equidistant grid, or of a geology based model, in which layers will be defined, possibly containing vertical gradients and of which the thicknesses are also parameters. The number of layers may or may not be fixed. As far as lateral variations are concerned one may think of single trace approaches, in which the subsurface is "sampled", or multi trace approaches, in which lateral variations are described by functions.

2. *Forward model, theoretical relations*

The forward model may be simple, e.g. the 1-dimensional convolutional model, with primary reflections only, or it may be more complex, containing multiples, be 2- or 3-dimensional etc. The data can be used in the time or in the frequency domain.

3. *Noise properties*

The noise properties may or may not be considered or assumed to be known. In the latter case noise properties will have to be estimated as well, or integrated out. A choice concerning the distribution of the noise will have to be made. Is it Gaussian, uniform etc.

4. *The wavelet*

The physical forward model will contain the seismic wavelet, which also may or may not be known. In the latter case it will have to be estimated or integrated out. A choice concerning its parameterization will then have to be made.

6. *Optimization aspects*

A resulting optimization problem must be solvable in an efficient and accurate way. Ideally it should be a global optimization scheme, in practice this will be too costly for most problems.

7. *Uncertainty analysis*

An inversion scheme is not complete without an uncertainty or resolution analysis. Choices concerning what type of information is displayed will have to be made. It should of course be possible to compute this information efficiently.

The decisions concerning the 7 items discussed must be such that as much detailed information of the desired type as possible comes out of the estimation scheme. Assumptions made must be reasonable and a priori knowledge put in must be actually available. The scheme must furthermore be practical, i.e. efficient enough.

4.3 SOME METHODS DESCRIBED IN LITERATURE

There is a large amount of literature on the subject of detailed inversion of poststack seismic data. It is beyond the scope of this thesis to describe and compare them in detail. Instead a number of papers or groups of related papers will be mentioned that are relatively well-known. Only the most outstanding features of the approaches will be mentioned.

Bilgeri and Carlini (1981) describe a single trace inversion scheme in which reflection coefficients and lagtimes of reflectors are estimated using a least squares scheme in the frequency domain. The number of reflectors can be updated in the scheme and the wavelet can be estimated simultaneously. No a priori information is used.

Cooke and Schneider (1983) also describe a single trace least squares scheme. It is a time domain approach. The parameters are acoustic impedances, including vertical gradients and lagtimes of layers. A priori information can be put in by fixing the values of parameters. The wavelet can be estimated simultaneously and is parameterized by four frequency points of the amplitude spectrum and a linear time shift. The forward model is not described exactly but includes multiples.

The only multi trace inversion scheme described in literature is the one of Gelfand and Larner (1983). Parameters are either acoustic impedances and lagtimes or densities, velocities and thickness (in depth) of layers. Lateral variations are described by 1st order spline functions, of which the end points are the actual parameters. The number of layers has to be chosen by the user. It is a time domain least squares scheme. A priori information

on the parameters can be put in by hard constraints on their values. The wavelet can be estimated simultaneously and is chosen to be a modulated sinusoid described by three parameters: dominant frequency, duration and energy delay. The optimization seems to be performed by a direct search method and the method is computationally very intensive.

Jerry Mendel and co-workers have published a series of papers on what they call "maximum-likelihood deconvolution". As references the textbook of Mendel (1983) and an excellent overview of previous methods and publications by Goutsias and Mendel (1986) are mentioned. All versions of methods have in common that they are based on (Bayesian) parameter estimation and that they are single trace approaches. Parameters are reflection coefficients on an equidistant time grid. A comb filter separates reflection coefficients with zero and nonzero amplitude. The wavelet is simultaneously estimated and parameterized by a fourth order ARMA (autoregressive, moving average) model. The noise is assumed to be Gaussian and the noise power is also estimated. Because the number of spikes (in direct relation to the comb filter) is also estimated the scheme is computationally intensive despite some sophisticated algorithms. A version with a more elaborate forward model is described by Mendel and Goutsias (1986).

Ursin and Holberg (1985) describe a single trace maximum likelihood method. The parameters are reflection coefficients on an equidistant grid and furthermore the noise power and parameters concerning the covariance or the "colour" of the noise. The wavelet is assumed to be known. Strongly related research has been done by Ozdemir (1985). Brevik and Berg (1986) describe a case study of inversion with the Ursin-Holberg method. Ursin and Zheng (1985) describe a single trace approach in which reflection coefficients and lagtimes are estimated. The nonlinear problem is treated by sequential linearized estimation problems which gives rise to unnecessarily complicated expressions for bias and covariance. They present however the only scheme with an uncertainty analysis. Each combination of reflection coefficient and lagtime is given with a two dimensional confidence interval.

All schemes mentioned so far have Gaussian assumptions or simply postulated least squares formalisms. Taylor, Banks and McCoy (1979) present a l_1 -norm method. The choice of the l_1 -norm is not justified by noise distribution assumptions but by the desire to obtain a sparse result. The parameters are reflection coefficients on an equidistant grid. A stabilization term, mathematically equivalent to the utilization of a priori information is incorporated. The amount of stabilization is determined by trial and error. The wavelet is assumed known in a simple version of the scheme. If it is unknown it can also be estimated with a multi trace l_1 -norm scheme, using an estimate of the reflectivity. In the latter case an iterative scheme where wavelet and reflectivity are estimated alternately is the result.

Levy and Fullagar (1981) also use the l_1 -norm in a single trace scheme with reflection coefficients on an equidistant grid. The l_1 -norm that is minimized however concerns only

the reflection coefficients, *not* the data fit. The data is used only in the form of hard constraints, which from a probabilistic point of view corresponds with the assumption of uniformly distributed noise. In Oldenburg, Scheuer and Levy (1983) the method is extended with the incorporation of constraints and with the estimation of acoustic impedances instead of reflection coefficients. They also describe an AR (auto-regressive) model as a basis for inversion with similar results. This latter method is further refined by Walker and Ulrich (1983). An overview of the methods described by this group of researchers is given by Oldenburg, Levy and Stinson (1986).

No critical comparison of the above mentioned methods will be given here. It is merely concluded that they show mutual differences with respect to the items mentioned in section 4.2. In the next section the approach discussed in the sequel of this thesis will be sketched while addressing these items once more. A number of good elements of the methods discussed in this section will be incorporated.

4.4 OUTLINE OF A PROCEDURE FOR DETAILED INVERSION OF POSTSTACK DATA

The procedure for detailed inversion should be applicable to stacked and possibly migrated data. Such a data set can be regarded as a bandlimited estimate of the zero-offset reflectivity as a function of travelt ime. The procedure can also be applied to data that is downward continued to just above the target zone and imaged i.e. to a zero-offset data set shot just over the target. In the near future however, practical application is envisaged for the first case.

1. Parameterization

The parameterization will be in terms of acoustic impedance (in this situation the product of density and velocity). Compared to reflectivity it is easier to provide a priori information for and, more important, it is easier to interpret and of more direct interest to a geologist. The selection of densities and velocities as parameters would lead to a situation in which the a priori information strongly determines the answer. Furthermore it would make the scheme more complex. It is assumed that the layers are homogeneous in the time direction. The acoustic impedance profile is built up of the impedance values of the layers and the positions in travelt ime of the boundaries between them. In comparison with an equidistant grid this parameterization can yield a much higher accuracy. Layers thinner than the sample rate of the data can be estimated. Moreover, in many cases the numbers of parameters will be smaller than with an equidistant grid parameterization. and this fact itself improves the accuracy. No attempt will be made to also estimate vertical gradients within the layers. Although addition of these parameters is possible, it will complicate the scheme. Such an estimation procedure will be less stable because the number of parameters is higher and

because the response of a gradient is not strong. The latter point implies that not much information concerning gradients is present in the data. Both single and multi trace procedures will be described. The single trace case will serve as an introduction to the more accurate multi trace case. It may however also be of practical value because it is much faster, allowing implementation in interactive software on interpretation workstations.

2. *Forward model*

The simple one-dimensional convolutional model will be used with primary reflections only. Although any model can be used in the Bayesian formulation, multiples are not taken into account because it would make the scheme less efficient and because their contribution will usually be lower than other noise components. It will be shown in chapter 5 that the convolutional model can be implemented in a very efficient algorithm in the time domain. This domain is used for inversion because transformation to the frequency domain of a selected data gate would introduce errors. The data selected for inversion will concern a selected target zone. Such a data gate will usually not be large, from 100 to 500 ms. Unnecessary large data gates and especially a large number of parameters would slow down the optimization. Moreover, with a small data gate the wavelet can be assumed to be constant in time.

3. *Noise properties*

Most of the noise properties will be assumed to be known. Only an approximate estimation of the noise power will be included in an iterative scheme described in chapter 8. The distribution of the noise will be taken to be Gaussian for the reasons described in section 1.12.

4. *The wavelet*

In chapters 5 and 6 inversion procedures are described in which the wavelet is assumed to be known. In chapter 7 it will be described how the wavelet can be estimated using reflectivity information. In chapter 8 it is shown how inversion and wavelet estimation can be combined in integral schemes. The wavelet will be parameterized as a time series in the time domain. Because the wavelet is always bandlimited this parameterization, unlike the ones mentioned in the previous section, causes no restrictions. No phase assumptions are made.

5. *A priori information*

A priori information can be incorporated in the form of Gaussian pdf's, see section 1.10. This can be augmented by hard constraints in the form of bounds on the parameters. A priori information may of course vary from parameter to parameter. Additional hard constraints are formed by the requirement that the thickness of a layer can not be less than zero.

6. *Optimization*

The choices made above allow a fast and accurate optimization scheme. Because the

number of parameters will be moderate (typically less than 100) and because derivatives are available the Newton optimization methods as discussed in chapter 2 can be used. In particular the least squares schemes are ideally suited.

7. *Uncertainty analysis*

The choices made above allow an extensive uncertainty analysis along the lines of chapter 3, in particular the results for the Gaussian case.

A few side remarks have to be made. The number of layers is assumed to be known. In practice an inversion procedure will be started at a well position where this information, together with a priori information on the parameters is indeed available. At some distance from the well the number of layers may change. The detection of an incorrect number of layers as well as the detection of wavelet errors were part of the project in which this research was done. It is however not described here. For a scheme for the detection of an incorrect number of layers and a corresponding resolution analysis, see Kaman et al. (1984).

In utilizing the one-dimensional convolutional model a forward model is adopted that since long forms the basis for deconvolution and inversion techniques on poststack data. Some additional remarks however have to be made. The adoption of the one-dimensional model for the target zone does not imply that the complete subsurface has to have a one-dimensional structure. Consider the application of the inversion technique to a stacked, unmigrated, section. When lateral velocity variations are present in the overburden the reflectors at the target zone are not positioned correctly, see figure 4.1. The inversion procedure can nevertheless be used as it merely solves for interference effects in the time direction at the target zone. When the positioning errors are large of course one would first want to apply migration to the data.

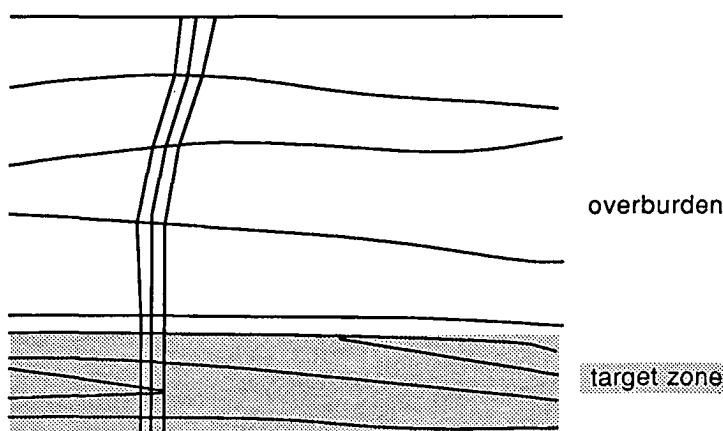


Figure 4.1 Incorrect imaging of unmigrated data for a laterally varying overburden. The normal rays indicate at which CMP positions the reflectors will be imaged.

A two or three-dimensional model will of course be more accurate than the one-dimensional one. In this research the forward model has not been a major topic. In order to arrive at a practical scheme (in terms of computation time) the often used one-dimensional convolutional model is taken as a first step. In later studies the practicality of more sophisticated models can be investigated.

The zero-offset reflectivity that a poststack section represents is a summation over angle dependent reflection coefficients. The forward model with acoustic impedances as used in this thesis however is based on the normal incidence plane wave model. This will nevertheless be fairly accurate because at the target zone the incident angles will usually not be very large and the reflection coefficient of an interface remains constant with a good approximation when the velocity change across the interface is relatively small (Berkhout, 1985), which it is most of the time. A check on the relevant conditions should of course be made. When these conditions are violated a stack should be made with smaller offsets. Another alternative is to estimate reflection coefficients instead, which then represent averages over the angles. The advantages of the acoustic impedances are then lost.

5

SINGLE TRACE INVERSION GIVEN THE WAVELET

5.1 INTRODUCTION

In this chapter the inversion of poststack seismic data will be developed along the lines sketched in chapter 4, assuming the wavelet and the noise properties to be known and using a single trace approach. The technical formulation given in this chapter also forms the basis for the multi trace problem discussed in the next chapter, which is for the major part a rather straightforward extension of the single trace one. In section 5.2 the trace inversion problem is worked out in detail. In section 5.3 optimization aspects and in particular the derivatives of the forward model are discussed. A practical implementation of the inversion scheme requires an efficient calculation of the forward model and its derivatives. A particularly fast algorithm is described in section 5.4. The linearity of the forward model is analytically examined in section 5.5. The analysis provides physical insight in the problem especially with respect to the uncertainty analysis procedure. A closely related sensitivity analysis is also given revealing differences in the expected accuracies with which acoustic impedances and lagtimes can be estimated from the data. Synthetic examples for the single trace inversion problem are given in section 5.6. In section 5.7 it is discussed how a seismic section can be inverted by sequential trace inversion.

5.2 FORMULATION OF THE INVERSE PROBLEM

In this section the technical formulation of the single trace inverse problem is given, following the lines set out in section 4.4. In single trace as well as in multi trace inversion the forward model used is the one-dimensional convolutional model. The subsurface at the target zone is assumed to consist of a stack of locally plane, homogeneous, layers, see figure 5.1. The thickness of the layers is described in terms of two-way traveltimes. The physical property of the layers that is estimated is the acoustic impedance which, for plane

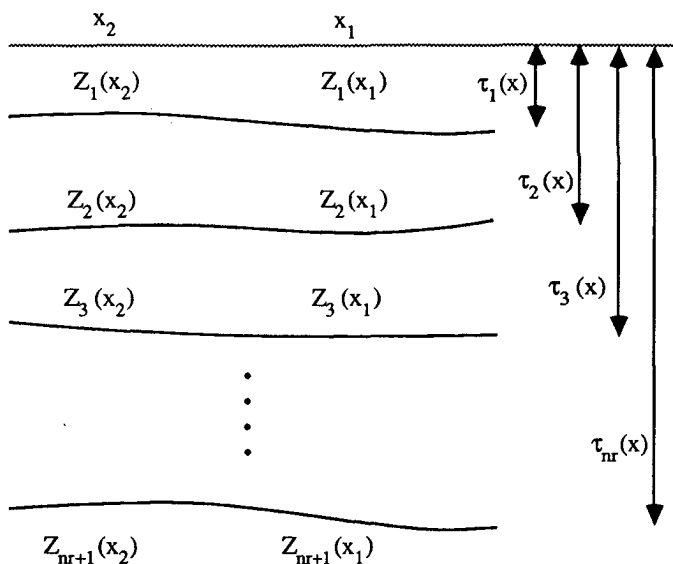


Figure 5.1 The local one-dimensional structure of the target zone.

waves, is the product of the density and the propagation velocity. When transmission effects and multiples are neglected the forward model for the trace $s(t)$ becomes:

$$s(t) = \sum_{i=1}^{nr} r_i w(t - \tau_i) \quad , \quad (5.1)$$

where nr is the number of reflectors (the interfaces separating the layers), $w(t)$ is the wavelet and τ_i are the lagtimes. For normal incidence the reflection coefficients r_i are determined by the acoustic impedances Z_i :

$$r_i = \frac{Z_{i+1} - Z_i}{Z_{i+1} + Z_i} \quad . \quad (5.2)$$

The parameter vector $\mathbf{x}$ is built up by the acoustic impedances Z_i and the lagtimes τ_i :

$$\mathbf{x}^T = (Z_1, \dots, Z_{nr+1}, \tau_1, \dots, \tau_{nr}) \quad (5.3)$$

A priori information on the parameters is given in the form of a Gaussian or normal pdf with mean $\mathbf{x}^i$ and covariance matrix $\mathbf{C}_x$. For the lagtimes there is the additional hard constraint that they cannot cross:

$$\tau_{i+1} - \tau_i \geq 0 \quad (5.4)$$

Furthermore it is possible to set absolute bounds on the parameter values:

$$\mathbf{l} \leq \mathbf{x} \leq \mathbf{u} \quad (5.5)$$

where $\mathbf{l}$ and $\mathbf{u}$ are vectors containing the lower and upper bounds. The noise is taken to be Gaussian with zero mean and covariance $\mathbf{C}_n$. Let the vector $\mathbf{g}(\mathbf{x})$ contain the sampled forward model:

$$g_i = s((i-1)\Delta t + t_s) \quad (5.6)$$

where Δt is the sample rate and t_s the start time of the data gate. When the vector $\mathbf{d}$ contains the sampled data, the MAP estimator is the vector $\mathbf{x}$ that minimizes the l_2 norm (see section 1.11):

$$2F(\mathbf{x}) = (\mathbf{d} - \mathbf{g}(\mathbf{x}))^T \mathbf{C}_n^{-1} (\mathbf{d} - \mathbf{g}(\mathbf{x})) + (\mathbf{x}^i - \mathbf{x})^T \mathbf{C}_x^{-1} (\mathbf{x}^i - \mathbf{x}) \quad (5.7)$$

subject to the constraints:

$$\mathbf{B}\boldsymbol{\tau} \geq 0 \quad (5.8a)$$

$$\mathbf{l} \leq \mathbf{x} \leq \mathbf{u} \quad (5.8b)$$

where (5.8a) is a vector notation for the combination of $nr-1$ constraints of the form (5.4). The optimization problem is therefore a nonlinear least squares problem with linear and bounds constraints, allowing special algorithms to be used, see chapter 2.

Ideally the data should be resampled such that the Nyquist frequency equals the frequency from whereon the amplitude spectrum is negligible. This increases the efficiency of the optimization without any loss of accuracy. In the experiments done in this research the sample rate is invariably 4 ms (a very usual sample rate in seismics) even though the highest frequency with significant amplitude is usually not higher than 80 Hz. This is done for practical reasons. When within the frequency range from DC to the Nyquist frequency significant variations within the amplitude spectrum of the noise are known, they can be incorporated in the noise covariance matrix $\mathbf{C}_n$. The experiments presented in this thesis are done with a stationary white noise assumption so that:

$$\mathbf{C}_n = \sigma_n^2 \mathbf{I} \quad (5.9)$$

with σ_n the standard deviation of the noise samples.

In the single trace experiments the a priori information is chosen such that it contains no correlations. The a priori covariance matrix is then a diagonal matrix:

$$C_x = \text{diag} (\sigma_1^2, \dots, \sigma_N^2) \quad , \quad (5.10)$$

with N the number of parameters. Each σ_i is a measure of the uncertainty in the a priori information of the corresponding parameter i . In the experiments done, the bounds on the parameters as formulated in (5.8b) are derived from the standard deviations. The range from lower to upper bound is 4 standard deviations, symmetrically around the mean. The software built also allows the setting of additional hard constraints that are derived from some reference model, in practice for example the well log model. This option is not used in the experiments described in this chapter.

5.3 OPTIMIZATION AND DERIVATIVES

For the solution of the constrained nonlinear optimization problem as defined in (5.7) and (5.8) a quasi-Newton and a modified Newton scheme are used. In the latter case the Hessian is not computed according to the exact equation (2.10) but with the Gauss-Newton approximation $J^T J$ (equation 2.11). Therefore, for the computation of derivatives only the Jacobian matrix J as defined in equation (2.9) needs to be computed. It is given by:

$$J = - \begin{bmatrix} C_n^{-1/2} A \\ C_x^{-1/2} \end{bmatrix}, \quad (5.11)$$

with the forward matrix A given by:

$$A_{ij} = \frac{\partial g_i}{\partial x_j} = \frac{\partial s((i-1)\Delta t + t_s)}{\partial x_j} \quad . \quad (5.12)$$

The derivatives of $s(t)$ with respect to the acoustic impedances Z_i can be found from (5.1) and (5.2):

$$\begin{aligned} \frac{\partial s(t)}{\partial Z_i} &= \frac{-2 Z_{i+1}}{(Z_{i+1} + Z_i)^2} w(t - \tau_i) & i=1 \\ &= \frac{2 Z_{i-1}}{(Z_i + Z_{i-1})^2} w(t - \tau_{i-1}) - \frac{2 Z_{i+1}}{(Z_{i+1} + Z_i)^2} w(t - \tau_i) & \begin{matrix} i \neq 1 \\ i \neq nr+1 \end{matrix} \\ &= \frac{2 Z_{i-1}}{(Z_i + Z_{i-1})^2} w(t - \tau_{i-1}) & i=nr+1 \end{aligned} \quad (5.13)$$

The derivatives for the lagtimes are:

$$\frac{\partial s(t)}{\partial \tau_i} = -\frac{Z_{i+1} - Z_i}{Z_{i+1} + Z_i} w'(t - \tau_i) \quad , \quad i=1, \dots, nr \quad (5.14)$$

Note that in this expression the time derivative of the wavelet occurs. Second derivatives of the forward model can also easily be derived. They are not implemented in the software because of the time consuming character of such a job and because the optimization performed satisfactorily with first derivatives only.

First and second derivatives are necessary for the constraints as given in equation (5.8a). They follow directly from this expression.

5.4 FAST TRACE AND DERIVATIVE GENERATION

Inversion software should ideally be run interactively. It requires quite some interaction from the user and an interpreter would like to see results on the screen with as little waiting time as possible. Hence the algorithms should be fast. Of all algorithms involved in the inversion of seismic data nonlinear optimization takes by far the most computation time. Most of the time is used for function and derivatives evaluation. In this section the most straightforward and a more sophisticated and fast algorithm to compute these quantities are discussed.

Computation of the function involves evaluation of expressions (5.1) and (5.2). The derivatives are computed using (5.13) and (5.14). In these expressions the wavelet and its derivative appear with continuous arguments, while the wavelet is available as a sampled time series. The most straightforward way to compute a synthetic trace therefore seems to be to use the frequency domain and compute the trace in three steps, assuming the wavelet to be available in the frequency domain:

- (1) Computation of the reflectivity $R(f)$ in the frequency domain:

$$R(f) = \sum_i r_i e^{-j\omega\tau_i} \quad , \quad (5.15)$$

where f denotes the frequency and ω the circular frequency $2\pi f$.

- (2) Multiplication with the wavelet $W(f)$ in the frequency domain in order to obtain the trace $S(f)$:

$$S(f) = W(f)R(f) \quad (5.16)$$

- (3) Inverse Fourier transformation to the time domain:

$$s(t) = F^{-1}\{S(f)\} \quad , \quad (5.17)$$

where F denotes Fourier transformation, for which step in practice the fast Fourier transform (FFT) is used.

Computation of the derivatives can be performed analogously, with the time derivative operator for the wavelet given by $j\omega$ in the frequency domain. This method can be considered to be exact within the numerical accuracies of the algorithms. It turns out however to be rather time consuming. Parameter models with a moderate number of reflectors (10, say) still take in the order of a minute of optimization time on a Gould 32/67 super minicomputer, implying that sections will have to be inverted in batch jobs on such a machine.

A much faster scheme has been developed which is based on an interpolation in the time domain, using a Taylor series approximation. The formulation is given here for a second order expansion, but any higher or lower order can be used as well. A wavelet sample with some arbitrary argument t can be approximated by:

$$w(t) \approx w(t_g) + hw'(t_g) + \frac{1}{2} h^2 w''(t_g), \quad (5.18)$$

with

$$h = t - t_g, \quad (5.19)$$

and t_g denoting a time value on a grid point on which the wavelet and its first and second time derivative *are* available. The accuracy of the approximation depends on the size of h in combination with the frequency content of the wavelet, see below. The computations consist of an initialization step and the actual trace generation. In the initialization step the wavelet is resampled to a fine time grid with a sample rate Δt_f that is an integer factor (most practically a power of 2) smaller than Δt . The derivatives are computed on the same time grid according to:

$$w'(t) = F^{-1}\{j\omega W(f)\}, \quad (5.20a)$$

and

$$w''(t) = F^{-1}\{-\omega^2 W(f)\}. \quad (5.20b)$$

The contribution of one reflector with reflection coefficient r_i to the trace samples $s(k\Delta t)$ can now be computed according to:

$$s(k\Delta t) = r_i w(l\Delta t_f) + r_i h w'(l\Delta t_f) + \frac{1}{2} r_i h^2 w''(l\Delta t_f), \quad (5.21)$$

with

$$l = \text{start value} + k \frac{\Delta t}{\Delta t_f}. \quad (5.22)$$

Note that the factors for the wavelet and its derivatives are constant for the reflector, h being determined by the lagtime τ_i , amongst others. Equation (5.21) is ideally suited for

computers with vector architecture.

The accuracy of this scheme depends on the sample rate Δt_f and the highest order that is used in the computation. The maximum error introduced by the approximation has been evaluated for a 10–70 Hz zero phase wavelet. It is given in table 5.1 in the form of the energy of the error, relative to the error free trace, expressed in dB.

Table 5.1 Errors of fast trace generation algorithm [dB].

Δt_f [ms] \ order	0	1	2
4	-5.6	-15.7	-28.5
2	-11.5	-27.6	-46.5
1	-17.5	-39.6	-64.6
.5	-23.5	-51.7	-83.3
.25	-29.5	-63.8	-107.5

It is clear that the order and the sample rate can be chosen such that the error introduced by using the approximation is negligible compared to other sources of error.

Once the initialization step is done the computation time hardly depends on the sample rate of the fine time grid. An indication of the efficiency of the scheme is given in table 5.2 where the computation times are given for the generation of a trace of 64 samples with 10 reflectors. The timings are given for standard Fortran 77 code, running on a Gould 32/67 super minicomputer (no vector architecture) and a Convex C1-xp mini supercomputer which has vector architecture on the CPU.

Table 5.2 Computation time in ms for the generation of a trace of 64 samples containing 10 reflectors.

hardware \ order	0	1	2
Gould 32/67	6.	9.	12.
Convex C1-xp	.25	.29	.33

There is an expected linear increase of the computation time with the order of the approximation used. A significant amount of time is spent on overhead, consisting of the subroutine call, computation of start and stop pointers etc. The overhead is especially large relative to the actual computation time for the Convex, where the actual vector arithmetic takes less time than the overhead. The computation time on this machine can significantly be reduced by some reorganizations that reduce the overhead.

Derivatives of the forward model containing only first and/or higher order derivatives of the wavelet can be computed in exactly the same way, using the same software.

When only a synthetic data trace is to be computed a zero order approximation can be used. In an optimization algorithm however higher order approximations have to be used

because the algorithm may get stuck on a locally flat part of the function, the flatness being introduced by the zero order approximation. Furthermore the derivatives should ideally be consistent with the function values. The discontinuities in the function introduced by the interpolation scheme are very small and cause no problems in practice. An analogous algorithm can be set up that is continuous but shows inconsistencies between function and derivative values.

The differences in efficiency between the frequency domain algorithm and the interpolation algorithm depends on a number of factors. It can be shown that the latter is always faster. The experience on the Gould computer is that for traces with about 50 samples and a relatively small number of reflectors (less than 10) the latter method is 20 to 30 times faster. This means that single trace inversion can be run interactively on a powerful workstation when the number of reflectors is not too high. Optimization of a trace with 6 or 7 reflectors usually will take less than a second. For a workstation with integrated vector capabilities the fast algorithm will be especially attractive.

5.5 LINEARITY AND SENSITIVITY ANALYSIS

When dealing with practical problems it is advantageous to have an idea about the linearity of the forward model, especially with respect to uncertainty analysis and, very closely related, with respect to the sensitivity of the data to parameter changes. The latter aspect determines how much information the data contains concerning the parameters. Van Riel and Berkhout (1985) give a numerical linearity analysis for lagtimes for a delayed pulse model. Here a simple analytical analysis is given for both acoustic impedances and lagtimes. It gives a rougher indication than can be given for a specific model, but provides more physical insight.

A forward model can be called linear within a certain range around a parameter model $\mathbf{x}_0$ if it can be approximated by the first two terms of a Taylor expansion:

$$\mathbf{g}(\mathbf{x}) \approx \mathbf{g}(\mathbf{x}_0) + \mathbf{A}(\mathbf{x} - \mathbf{x}_0) \quad , \quad (5.23)$$

with

$$\mathbf{A} = \left. \frac{\partial \mathbf{g}}{\partial \mathbf{x}} \right|_{\mathbf{x}=\mathbf{x}_0} \quad . \quad (5.24)$$

The analysis for acoustic impedances and lagtimes is separated. An analysis is given for one acoustic impedance and for one lagtime, keeping all other parameters fixed. It should be kept in mind however that this may give a too optimistic view for certain linear combinations of parameters where a simultaneous change of all parameters give a greater deviation from the linear model than the changing of only one parameter.

With respect to the acoustic impedance it should first be noted that the model is exactly

linear in the reflection coefficients r_i , see equation (5.1). Expansion of equation (5.2) for Z_i in a Taylor series keeping Z_{i+1} fixed gives:

$$r_i(Z_i^0 + \Delta Z) = r_i(Z_i^0) - \frac{\Delta Z}{Z_{i+1} + Z_i} \frac{2 Z_{i+1}}{Z_{i+1} + Z_i} + \left(\frac{\Delta Z}{Z_{i+1} + Z_i} \right)^2 \frac{2 Z_{i+1}}{Z_{i+1} + Z_i} . \quad (5.25)$$

It follows that the quadratic term is smaller than the linear term by a factor $\Delta Z/(Z_{i+1} + Z_i)$. In practice the range under consideration will yield values lower than .1 for this factor, corresponding with variations of approximately 20%. It can be concluded that the problem is quasi-linear in the acoustic impedances.

The analysis for the lagtimes can best be done in the frequency domain. This has no consequences for the analysis. A linear model in the frequency domain is also linear in the time domain. The forward model in the frequency domain is:

$$S(f) = W(f) R(f) \quad (5.26)$$

The lagtimes are contained in the reflectivity $R(f)$ through:

$$R(f) = \sum_{i=1}^{nr} r_i e^{-j\omega\tau_i} . \quad (5.27)$$

The linearity analysis is done by considering one reflector. An expansion in a Taylor series around $\tau = \tau_0$ yields:

$$R(\tau^0 + \Delta\tau) = e^{-j\omega\tau^0} \left\{ 1 - j\omega\Delta\tau + \frac{1}{2} (\omega\Delta\tau)^2 \right\} \quad (5.28)$$

A fairly good linear approximation is obtained under the condition:

$$\omega\Delta\tau < .5 . \quad (5.29)$$

This condition for the lagtime of course depends on the frequency. For a frequency of 40 Hz (5.29) leads to the condition $\Delta\tau < 2$ ms, in accordance with the numerical results of Van Riel and Berkhout (1985) and the results of the previous section. The fact that the response of a thin layer can be approximated by the time derivative of the wavelet is of course a direct result of the linearity of the forward model. For the validity range of this approximation one finds larger values for the time thickness (6 to 8 ms, for practical frequency ranges) due to the fact that the linearization can be done around the middle of the thin layer and the fact that the second order terms cancel, see appendix B.

From condition (5.29) it follows that in practice the detailed inversion of poststack seismic data can *not* be treated as a linear problem by linearizing at the a priori model. The differences in lagtimes between the estimated and the a priori model may very well be larger than 2 ms. The only consequence of this however is that the optimization will have to be done by a nonlinear scheme. More important is whether an uncertainty analysis with

(co)variances can be done. For this to be possible in a practical way approximation (3.40) has to be accurate enough. When the residuals and/or the second derivatives of the forward model are low enough this is equivalent to the demand that the forward model is linear enough in the region where the a posteriori pdf is significant. Condition (5.29) therefore should hold for this region when linearizing at the *estimated* model. In practice this means that when the approximated standard deviations of the lagtimes are less than 2 ms, the approximation will be fairly reliable. A clear classification of inverse problems with respect to linearity is given by Tarantola (1987).

Very closely related to the linearity analysis discussed above is an analysis of the sensitivity of the data due to parameter variations. When only a slight variation of a parameter gives a significant change in the data that parameter can be estimated accurately. In the convolutional model the parameters are contained in the reflectivity only, which is furthermore a sum of contributions of reflectors. When only one of the reflectors, with reflection coefficient r and lagtime τ is considered we have for the reflectivity in the frequency domain:

$$R = r e^{-j\omega\tau} \quad (5.30)$$

From a Taylor series expansion with only the linear terms it follows:

$$R(r_0 + \Delta r, \tau_0 + \Delta \tau) - R(r_0, \tau_0) = \Delta r e^{-j\omega\tau_0} - j\omega\Delta\tau r_0 e^{-j\omega\tau_0} \quad (5.31)$$

For the (monochromatic) energy of this difference it follows:

$$|R - R_0|^2 = r_0^2 \left\{ \left(\frac{\Delta r}{r_0} \right)^2 + (\omega\Delta\tau)^2 \right\} \quad (5.32)$$

Practical ranges are in the order of:

$$\frac{\Delta r}{r_0} = .1 \quad (5.33)$$

and, for $\Delta\tau = 2$ ms and $f = 40$ Hz:

$$\omega\Delta\tau = .5 \quad (5.34)$$

When considering these ranges of parameters it follows that the lagtime variations have a much larger effect on the data than reflection coefficients and therefore also a much larger effect than acoustic impedances. Note however that the sensitivity of the data to a lagtime is proportional to the corresponding reflection coefficient. From this it immediately follows that lagtimes of reflectors with a high reflection coefficient are determined more accurately than the lagtimes corresponding with low reflection coefficients, as is illustrated in the next section.

5.6 RESULTS FOR SINGLE TRACE INVERSION

In this section three examples are given as an illustration of the inversion procedure on a single trace. For one example a full uncertainty analysis is performed. In figure 5.2 the zero-phase wavelet is shown with which all experiments in this and the next chapter are done.

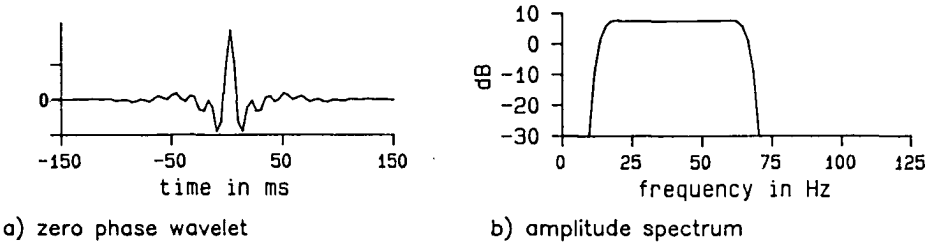


Figure 5.2 The zero-phase wavelet used in the single trace inversion examples.

The first example is for a noise free trace. The true model, the data and the a priori model are shown in figures 5.3.a,b,c respectively. The basic structure is present in the a priori model but the parameter values are significantly wrong, with differences of up to 6 ms between true and a priori lagtime values. It is clear that the detail of the true model can not visually be derived from the data. The estimated model is shown in figure 5.3.e in comparison with the true model. The errors in the lagtimes are virtually zero. There is a small error in all acoustic impedances due to the fact that on the average the a priori values are a little too high. Remember that a constant factor in the acoustic impedances can never be resolved from the data when using the convolutional model. The residual or data mismatch is shown in figure 5.3.d. Its amplitude is of course very low for this example with noise free data.

In figure 5.4 the estimated model is plotted with an a posteriori uncertainty interval around each parameter, formed by the boundaries of one standard deviation above and below the estimated value. The uncertainty in the lagtimes is virtually zero. The uncertainty interval of the acoustic impedances is more significant. This is due to the fact that there is one singular value in the likelihood spectrum that is exactly zero, leading to nonzero elements in the a posteriori covariance matrix, see section 3.3.1, in particular equations (3.46) and (3.47). The corresponding eigenvector points in the direction formed by a constant factor for each impedance. The a priori standard deviation of the acoustic impedances is $2 \cdot 10^6 \text{ kgm}^{-2}\text{s}^{-1}$. The uncertainty therefore is nevertheless significantly reduced.

Before discussing examples with noise the measures with which the estimates will be assessed are given. In the most complete form the errors are given by the differences between the estimated and the true parameters, gathered in a vector $\mathbf{0}$:

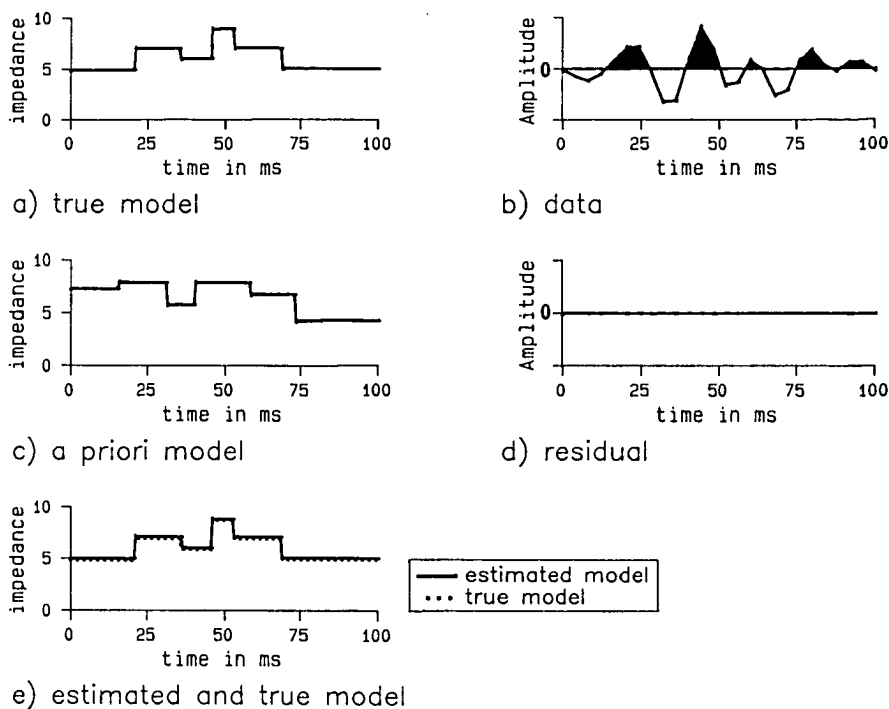


Figure 5.3 Example of single trace inversion with noise free data. The unit of the impedance is $10^6 \text{ kgm}^{-2}\text{s}^{-1}$.

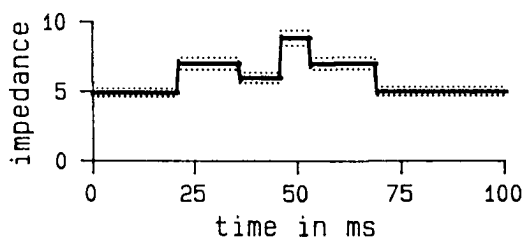


Figure 5.4 The a posteriori uncertainty intervals for the noise free data example.

$$\mathbf{0} = \hat{\mathbf{x}} - \mathbf{x}^{\text{true}} \quad (5.35)$$

A reduction in the amount of information is achieved by separating the vectors into sets of acoustic impedances and sets of lagtimes and computation for each set of the following quantities:

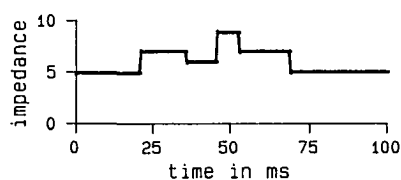
$$\text{error}_1 = \frac{1}{N_p} \sum_i |\hat{x}_i - x_i^{\text{true}}| \quad (5.36)$$

$$\text{error}_2 = \sqrt{\frac{1}{N_p} \sum_i (\hat{x}_i - x_i^{\text{true}})^2}, \quad (5.37)$$

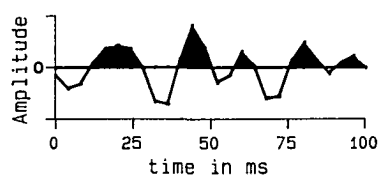
and

$$\text{error}_3 = \max_i |\hat{x}_i - x_i^{\text{true}}|, \quad (5.38)$$

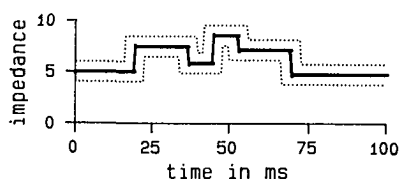
where N_p is the number of acoustic impedances or lagtimes whichever is appropriate. The last measure is equivalent to the so called l_∞ -norm of the vector $\mathbf{x} - \mathbf{x}^{\text{true}}$. The first two quantities resemble the l_1 and the l_2 -norm. The norms however are modified such that the resulting quantity is an averaged measure over the parameters and is meaningful in practice.



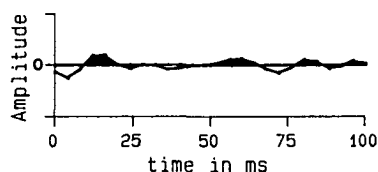
a) true model



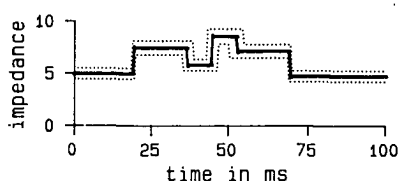
b) data



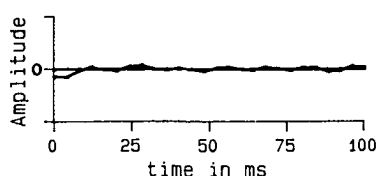
c) a priori model



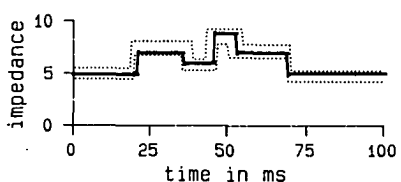
d) noise



e) estimated model



f) residual



g) true model / a post. uncertainties

— true model
... a posteriori uncertainty

Figure 5.5 Example of single trace inversion with signal/noise ratio of 10 dB.

Now a single trace example is discussed which includes noise. The true model is shown in figure 5.5.a. The data and the noise realization are shown in figure 5.5.b,d. The noise is bandlimited (10. – 70 Hz.) Gaussian noise. The signal to noise ratio, which will be expressed in terms of energy throughout this thesis is 10 dB. The a priori model with uncertainty bounds is given in figure 5.5.c. The standard deviations of the Gaussian pdf's are 3 ms for the lagtimes and $10^6 \text{ kgm}^{-2}\text{s}^{-1}$ for the acoustic impedances. The estimated model with a posteriori uncertainty intervals is shown in figure 5.5.e. It is clear that much information has been gained, especially for the first and the last lagtime. Their a posteriori standard deviation is .5 ms. They can be resolved from the data very well because the impedance contrasts for the interfaces are relatively large and because they are well separated from the other reflectors. The cluster of the three reflectors in the middle is less well resolved. The standard deviation is approximately 2 ms for the lagtimes. The interpretation of the uncertainty intervals is that the true parameter values fall within these regions with a probability of 67%. In this synthetic experiment it can be verified that for all but one parameter (the first lagtime) the true model parameters indeed fall within these regions. The residual is shown in figure 5.5.f. Its energy level is -17 dB relative to the data.

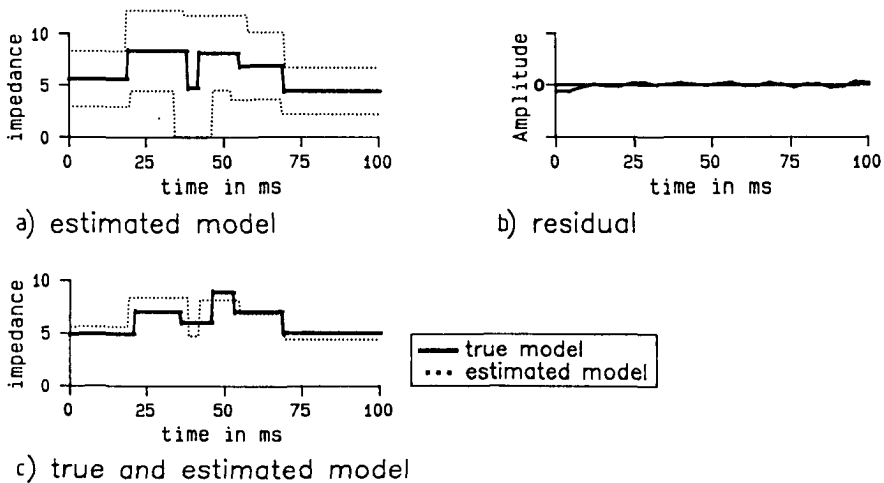


Figure 5.6 Results with wide constraints.

The utilization of a priori information stabilizes the solution as can be seen by comparison of the results with those of figure 5.6 where the only a priori information used is that the reflectors cannot cross and that acoustic impedance values should be larger than 0. The a posteriori uncertainty intervals shown in figure 5.6.a are much larger than in figure 5.5.e. The estimated model is far more wrong, see figure 5.6.e and table 5.3,

below. The residual is lower (-18 dB) because the parameters have more freedom to explain the data. This illustrates the well known fact that a better data fit does not always imply a better estimate. The numerical details of the errors in the estimates and the data mismatch are given in table 5.3. in comparison with those of the a priori model.

Table 5.3 Errors in the estimates and the data mismatch for the single trace examples.

model	impedance [$10^6 \text{ kgm}^{-2} \text{ s}^{-1}$]			lag time [ms]			residual [dB]
	error ₁	error ₂	error ₃	error ₁	error ₂	error ₃	
a priori	.18	.21	.30	1.6	1.7	2.6	n.a.
result for normal constraints	.23	.26	.43	.83	.93	1.40	-17.
result for wide constraints	.81	.90	1.33	2.26	2.53	4.17	-18.

The acoustic impedances of the a priori model are quite accurate and therefore the estimated impedance values are more wrong than the a priori values. For the lagtimes the situation is reversed, the estimated lagtimes are more accurate than the a priori lagtimes. The errors for wide constraints are much larger than those for normal constraints.

For the problem with normal constraints a full uncertainty analysis is performed. The standard deviations are already given in figure 5.5. The next type of information in the line of complexity and detail are the various matrices discussed in chapter 3. In figure 5.7 the layout of the matrices for the single trace example is given. It corresponds with the parameter vector as given in equation (5.3). The problem is statistically normalized on the a

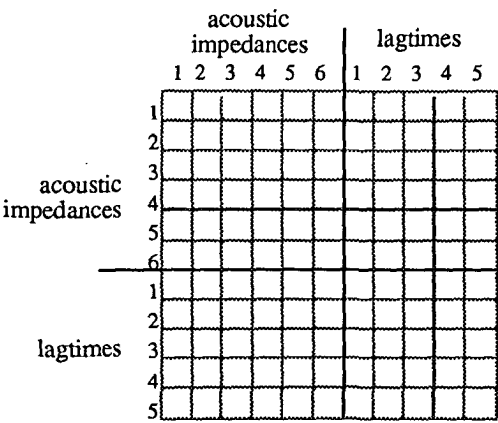
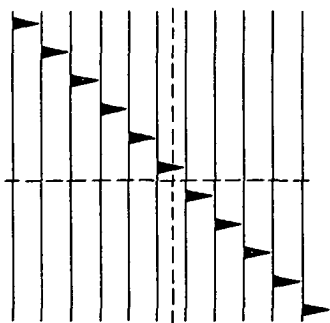
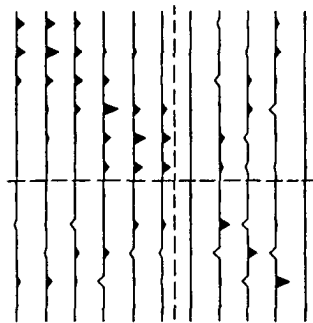


Figure 5.7. Layout of the matrices used for uncertainty analysis in the single trace example.

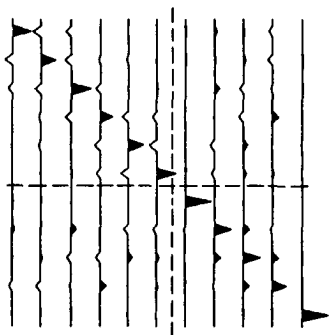
a priori covariance



a posteriori covariance



resolution matrix



a posteriori correlation

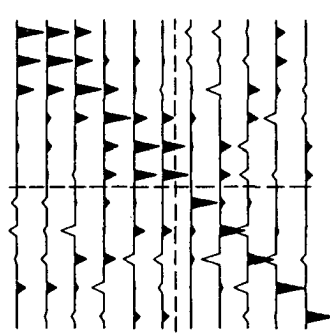


Figure 5.8 Results for the matrices used in uncertainty analysis for the single trace example.

priori information (by dividing each parameter with its a priori standard deviation in this problem) so that the a priori covariance matrix, shown in figure 5.8 equals the identity matrix. The other matrices are displayed on the same scale. The a posteriori covariance matrix computed along the lines sketched in section 3.3.1, using the approximated Hessian, can directly be compared with the a priori covariance matrix. It is clear that the variances of lagtimes 1 and 5 are strongly reduced. The other lagtimes and the acoustic impedances are less well resolved. Nonzero covariances occur especially between adjacent impedances and between the lagtimes and impedances of the cluster of middle layers in the model. The covariances can better be studied in the correlation matrix. Some strong correlations occur in the impedances as well as in the lagtimes. See e.g. the correlations between lagtimes 2 and 3, and those between lagtimes 3 and 4, due to the thin layer effect.

The resolution matrix is the identity matrix minus the a posteriori covariance matrix so that essentially the same type of information can be deduced from it. The lagtimes 1 and 5 are resolved best; they have a unit value on the diagonal and near zero off-diagonal elements. The other parameters are less well resolved, but the values on the diagonal dominate the off-diagonal elements.

The most detailed information can be obtained from an eigenvector analysis, discussed in chapter 3. The singular value spectra are shown in figure 5.9. The a priori values are 1, because of the normalization. The singular value spectrum of the likelihood function contains three values lower than 1. The lowest is exactly zero. There are therefore 3 weak directions along which, close to the estimated model, the curvature of the a priori pdf is stronger than that of the likelihood function and therefore the a priori pdf more strongly determines the a posteriori pdf and thereby the answer to the inverse problem.

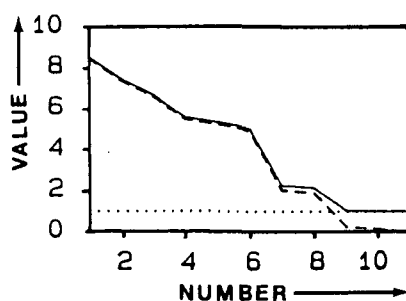


Figure 5.9 Singular value spectra.

..... a priori spectrum
 ----- likelihood spectrum
 ————— a posteriori spectrum

The probability density functions are plotted along the eigenvectors in figure 5.10 in the same type of display as for the two-parameter example, see figure 3.2. The distances plotted are $(-3/\lambda_i, 3/\lambda_i)$ around the estimated model. The one sided range corresponds with 3 a posteriori standard deviations of the canonical parameters. This ensures that the significant part of the a posteriori pdf will fit nicely in the pictures. Along the x-axis the corresponding models are plotted. The estimated model is always in the middle of the x-axis and is plotted with a thicker line. For some directions the changes can easier be evaluated by plotting the centre (estimated) and the two extreme models. These plots appear at the bottom of each picture.

Along eigenvector 11 the likelihood function is perfectly flat. The corresponding singular value is exactly zero. As stated before this direction corresponds with a constant factor in the impedances (see the extreme models in the bottom picture). The a priori information fully determines the answer along this direction. Direction 10 is more complex.

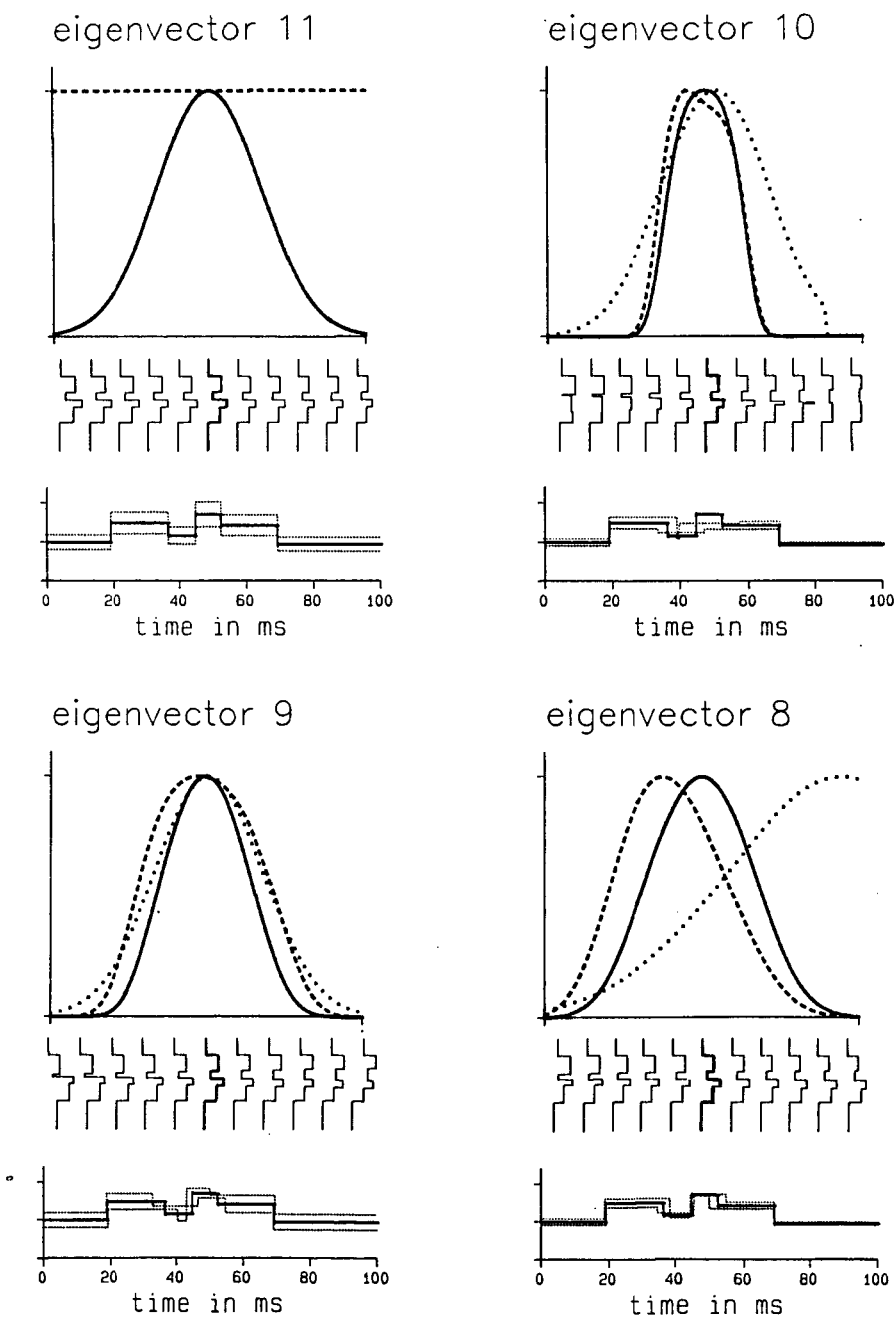
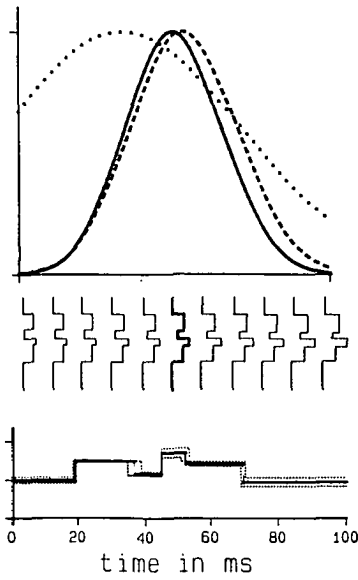


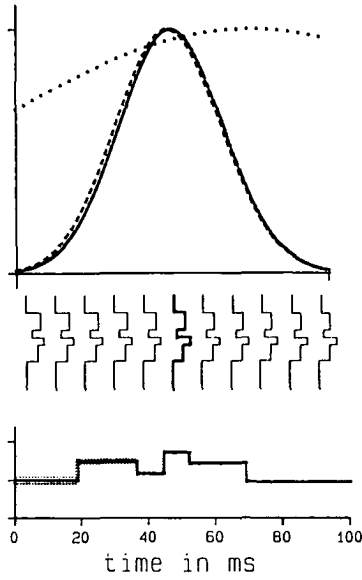
Figure 5.10 Plots of the pdf's along the eigenvectors.

..... a priori pdf
----- likelihood function
—— a posteriori pdf

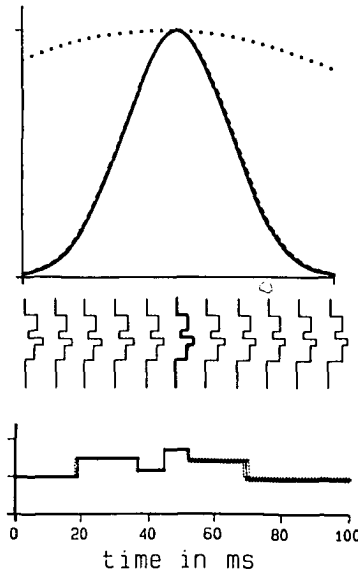
eigenvector 7



eigenvector 6



eigenvector 5



eigenvector 4

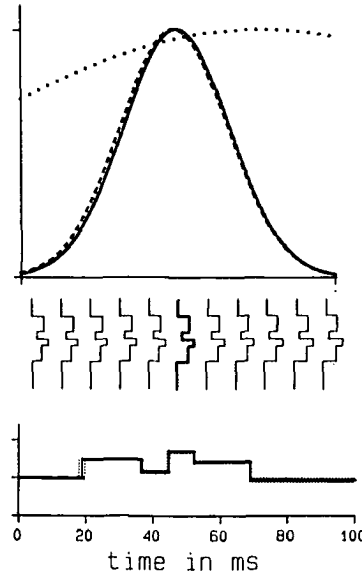


Figure 5.10 continued.

This corresponds to the exchange effects between acoustic impedances and lagtimes of the two thin layers. The gradient of the likelihood function at the estimated model is nonzero but the curvature is very small. Close around the estimated model the curvature of the a priori pdf determines the curvature of the a posteriori pdf. After a short distance (relative to the range of the plot) the likelihood function changes strongly, due to the nonlinearity of the forward model. Further away from the estimated model therefore the likelihood function determines the answer. At the right of the picture the a priori pdf drops to zero due to the hard constraint of non crossing lagtimes 3 and 4.

Along eigenvector 9 the a priori information still more strongly determines the answer to the inverse problem. At eigenvector 8 the situation is reversed (see the singular value spectra). For higher singular values the likelihood function dominates the a priori pdf more and more. The a priori pdf becomes flatter and flatter. From the model variations it can be seen that the length of the distances covered in the parameter space becomes lower for higher singular values, see also equation (3.6).

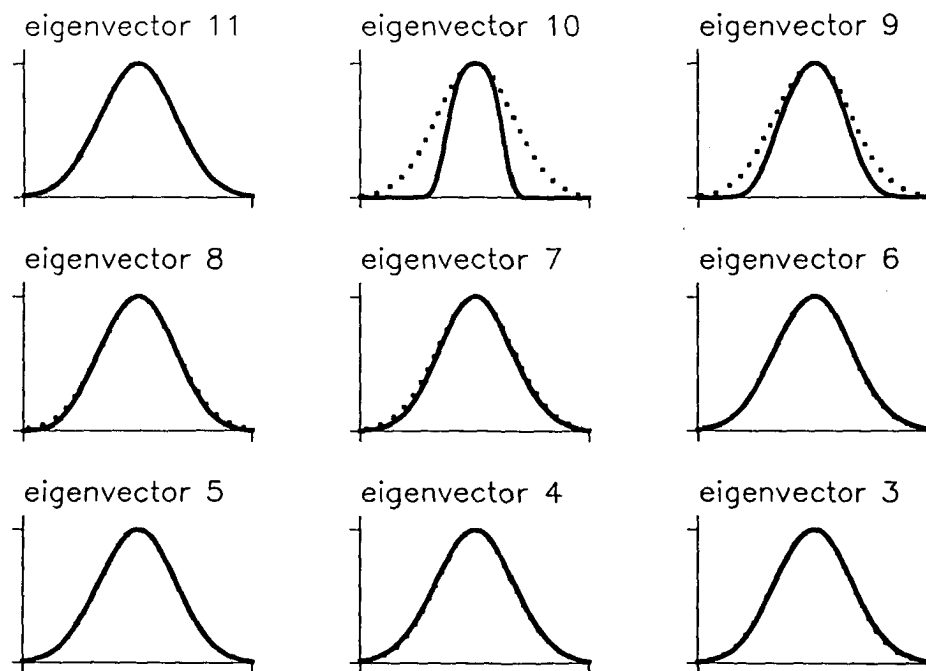


Figure 5.11 Plots of the pdf's along the eigenvectors for linearity analysis.

— true a posteriori pdf
 approximated a posteriori pdf

A linearity analysis can be performed by stepping along the eigenvectors and plotting the true a posteriori pdf and the approximated Gaussian one as given by equation (3.41). The results are given in figure 5.11. When changing the acoustic impedances only (direction 11) the linear approximation is excellent. In combination with the lagtimes however the approximation is less good, see the plots for directions 10 and 9. The true a posteriori pdf is more restricted than the approximated ones. This means that the uncertainty intervals as plotted in figure 5.5 are a little too large. In general it holds that the higher the singular value of the a posteriori pdf the smaller the distances covered in parameter space and therefore the better the linearity of the forward model in the range considered and the better the quadratic approximation of the objective function.

5.7 INVERSION OF A SECTION BY SEQUENTIAL SINGLE TRACE INVERSION

The inversion of a seismic section can of course be done trace by trace. The inversion procedure is the same for each trace and is as described in the previous section. Only the issue of supplying a priori information for each trace needs closer attention.

The whole procedure can be started at a well location, where the number of layers and a priori information on the parameters can be derived from well log data. The first trace is inverted and the result can be used as a priori knowledge for the inversion of the next trace, together with a priori knowledge on trace to trace variations which may stem from general geological knowledge or from other geophysical data on the area. In the most general formulation the inversion result on trace i is the a posteriori pdf on data y_i , denoted by $p(x_i|y_i)$. Trace to trace variations can be described by the difference Δx_i between the parameter vectors of traces i and $i+1$:

$$x_{i+1} = x_i + \Delta x_i \quad . \quad (5.39)$$

The knowledge on Δx_i is described by a pdf $p(\Delta x_i)$. The a priori knowledge on x_{i+1} can now be obtained from the a posteriori knowledge on x_i and the a priori knowledge on Δx_i . In particular, when the latter two are independent we have:

$$p(x_{i+1}) = p(x_i|y_i) * p(\Delta x_i) \quad . \quad (5.40)$$

When both pdf's are Gaussian it follows:

$$p(x_{i+1}) = N(\hat{x}_i + E\Delta x_i, C_{\hat{x}_i} + C_{\Delta x_i}) \quad . \quad (5.41)$$

For a practical scheme this will still be complicated. In the software built, the assumptions are used that the expectation of the lateral variations $E\Delta x_i$ is zero (horizontal, homogeneous layers) and that the a posteriori uncertainty is negligible compared to the trace to trace

variations, which are taken constant for the part of the section that is inverted in one run, so that:

$$p(x_{i+1}) = N(\hat{x}_i, C_{\Delta x}) \quad (5.42)$$

$C_{\Delta x}$ is furthermore a diagonal matrix. The variations may of course be different for each parameter. In order to compensate for the neglect of the a posteriori uncertainty, $C_{\Delta x}$ should be chosen somewhat larger than in the exact formulation (5.41), otherwise, speaking in classical terms, the results may become seriously biased. Additional hard constraints, setting absolute lower and upper bounds on parameter values can be utilized. This may be particularly appropriate for the acoustic impedances for which absolute bounds can be derived from the physical properties of e.g. the type of material of which a layer consists.

An example of the inversion of a section by sequential trace inversion is given in figure 5.12. In figure 5.12a,b the true model and the data are shown. The signal to noise ratio is 10dB. The noise is Gaussian, bandlimited (10 – 70 Hz) and spatially uncorrelated. The

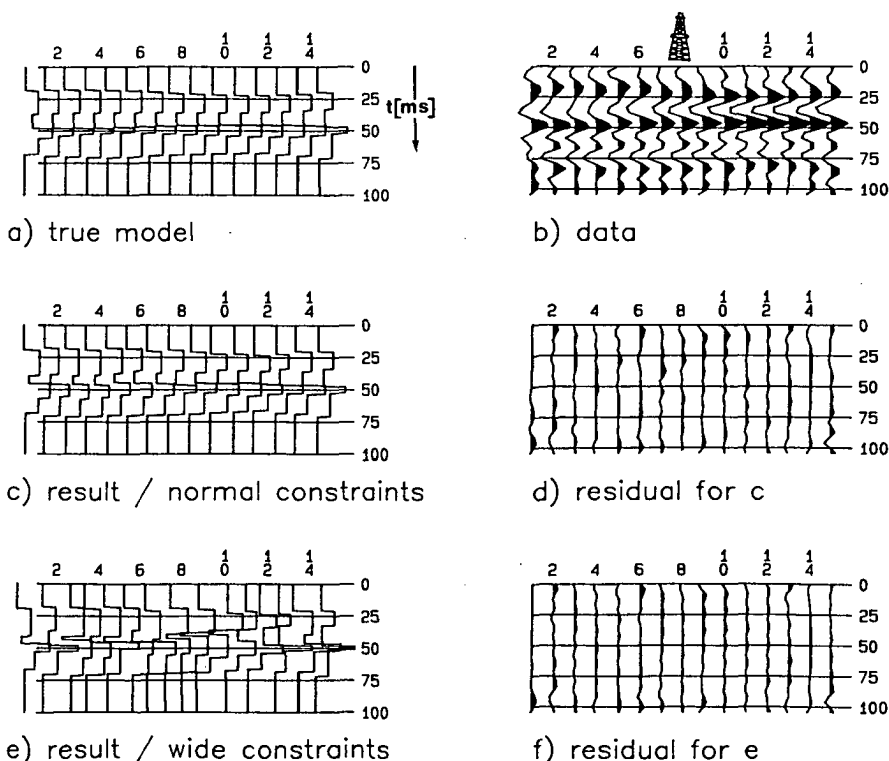


Figure 5.12 Inversion of a synthetic data section by sequential single trace inversion using narrow and wide constraints.

well is located at trace number 8. The section is processed to the left (from trace 8 to 1) and to the right (from trace 8 to 15) from this position. The well log model is taken to be very accurate in this example. The trace to trace standard deviations used are $.1 \text{ } 10^6 \text{ kg}^{-2} \text{ s}^{-1}$ for the impedances and 1. ms for the lagtimes, corresponding to accurate a priori knowledge. The same intervals were used for the well log uncertainties. The estimated models are shown in figure 5.12c and the corresponding residual traces, with an average energy of -13.7 dB relative to the data, in figure 5.12.d. The result is not bad. Especially the first and last lagtime are estimated well. For some numerical details see table 5.4 below. The section is also inverted with wider constraints, standard deviations of $.5 \text{ } 10^6 \text{ kg}^{-2} \text{ s}^{-1}$ for the impedances and 3.5 ms for the lagtimes, so that little a priori information is used. The estimated model is shown in figure 5.12e and the corresponding residuals, with an energy of -15.1 dB relative to the data in figure 5.12f. The residual is lower than for inversion with narrow constraints because the parameters have more freedom to explain the data. The result is far worse than for narrow constraints. There is a lot of jitter in the estimated model, due to the fact that the noise is spatially uncorrelated. Not only the lagtimes are seriously wrong, also the acoustic impedances show a lot of jitter.

Table 5.4 Errors in the estimates and the data mismatch for the inversion of a synthetic section by sequential single trace inversion.

model	impedance [$10^6 \text{ kgm}^{-2} \text{ s}^{-1}$]			lag time [ms]			residual [dB]
	error ₁	error ₂	error ₃	error ₁	error ₂	error ₃	
narrow constraints	.19	.24	.66	1.1	1.5	4.8	-13.7
wide constraints	.71	.92	2.29	1.9	2.5	6.9	-15.1

Note that the numerical results for inversion with wide constraints, although they are clearly worse than for narrow constraints, hide the fact that in practice the result is very poor due to the jitter.

6

MULTI TRACE INVERSION GIVEN THE WAVELET

6.1 INTRODUCTION

In parametric inversion as few parameters as possible should be used. The more parameters the higher the uncertainty of the inversion result and the less stable the estimation procedure. It is clear in this respect that the single trace approach as discussed in chapter 5 is not ideal. Despite a priori information concerning trace to trace variations the parameter estimates may show jitter when the noise level is high and the noise on different traces is uncorrelated. A more accurate inversion result can be obtained by describing the lateral variations of acoustic impedances and lagtimes by so-called spline functions. The parameters describing these functions can then be estimated. The consequence of this approach however is that several traces have to be inverted at a time, which complicates the scheme and increases computation time. In the next sections the parameterization and the practical incorporation of a priori information are discussed. In section 6.6 an example is discussed in which the same synthetic data set as used in chapter 5 is inverted. A full uncertainty analysis is given. Several possibilities of parameterization are compared.

6.2 PARAMETERIZATION

The lateral variation of e.g. the acoustic impedance of a layer can in general be described by a function $Z(h)$, when h denotes lateral distance. In single trace inversion this function is

sampled at equidistant points and the value of Z estimated at each point or trace location. In multi trace inversion a function can be used to describe the lateral variations. The parameters to be estimated are parameters describing the function. A practical way of setting up a function description is to divide the total lateral distance under consideration in smaller parts and to let the variation within each part be described by a so-called spline function. Each part is a block of traces on a section. In this thesis the class of polynomial functions will be used for the spline functions. Each acoustic impedance and lagtime in a block is described by:

$$f(h) = \sum_{j=0}^{n_c} x_j h^j, \quad (6.1)$$

where n_c is the order of the polynomial. A nice feature of this class of functions is that it is linear in the coefficients x_j . Let the vector z^s denote a vector with sampled values of $f(h)$, $z_i^s = f(h_i)$ and let the vector x^s contain the coefficients. The superscript s denotes that the description is for one spline. When $f(h)$ is sampled at the trace positions the vector z^s contains the single trace parameters. The coefficients x_j are called the multi trace parameters. The single trace parameters can be generated from the multi trace parameters by:

$$z^s = E^s x^s, \quad (6.2)$$

with the matrix E^s defined by (6.1):

$$E_{ij}^s = h_i^j, \quad (6.3)$$

where h_i is the lateral distance for the sample z_i . In the total set of multi trace parameters all ns spline coefficient vectors are gathered in the total parameter vector x :

$$x = \begin{bmatrix} x_1^s \\ \vdots \\ x_{ns}^s \end{bmatrix}. \quad (6.4)$$

The total vector of single trace parameters is similarly defined:

$$z = \begin{bmatrix} z_1^s \\ \vdots \\ z_{ns}^s \end{bmatrix}. \quad (6.5)$$

The two are related by:

$$z = Ex, \quad (6.6)$$

with E the block diagonal matrix:

$$E = \begin{bmatrix} E^s & \dots & E^s \end{bmatrix} \quad (6.7)$$

Relation (6.6) generates all single trace parameters from all multi trace parameters. This formulation allows a very elegant and general set up of software, see the next section. In principle software allowing any choice of the order of the spline function could be developed. In practice however numerical problems and storage considerations limit the number of possibilities. In this research three polynomial functions are implemented. The possible orders are, see also figure 6.1:

- 1) $n_c=0$ (constant function)
- 2) $n_c=1$ (straight, dipping line)
- 3) $n_c=3$ (cubic spline).

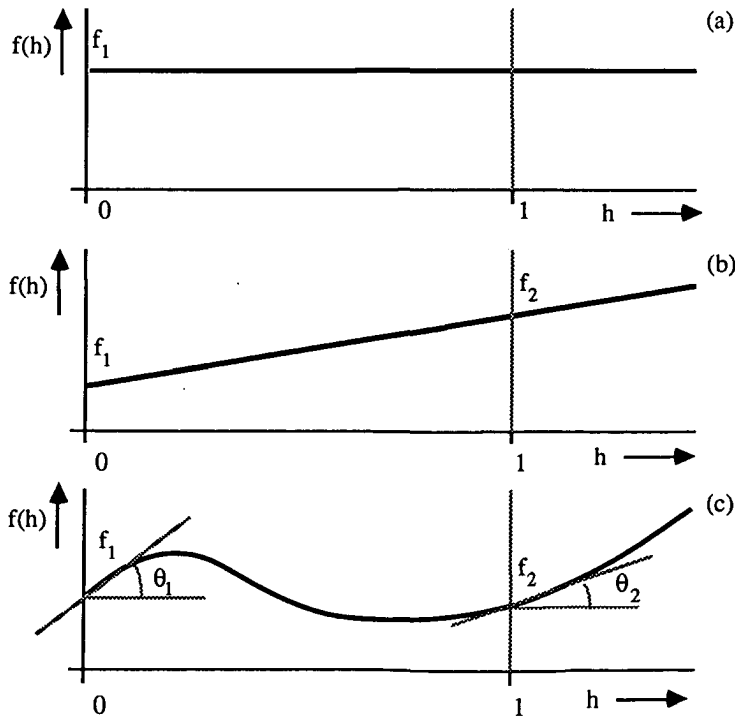


Figure 6.1 The three types of spline functions used in multi trace inversion.

- a) $n_c = 0$, parameter f_1 .
- b) $n_c = 1$, parameters f_1 and f_2 .
- c) $n_c = 3$, cubic spline, parameters $f_1, \theta_1, f_2, \theta_2$.

In principle each lagtime and each acoustic impedance can have its own order. For practical reasons this flexibility is not implemented. The spline functions have the same order. The spline function for the orders 1 and 3 are not straightforwardly implemented in the form (6.1) but by the formulas (see also figure 6.1):

$$f(h) = (1-h)f_1 + hf_2, \quad 0 \leq h \leq 1 \quad (6.8)$$

for $n_c=1$ and

$$f(h) = (2h^3 - 3h^2 + 1)f_1 + (h^3 - 2h^2 + h)\theta_1 + (-2h^3 + 3h^2)f_2 + (h^3 - h^2)\theta_2, \quad 0 \leq h \leq 1 \quad (6.9)$$

for $n_c=3$. For both functions it holds that the lateral distance is transformed so as to have values between 0 and 1. The function values at $h=0$ and $h=1$ are: $f(0) = f_1$ and $f(1) = f_2$. For the third order polynomial as given in (6.9) the parameters θ_1 and θ_2 are the derivatives at the end points of the cubic spline: $f'(0) = \theta_1$, $f'(1) = \theta_2$, see also figure 6.1. The choice for this parameterization leads to better numerical properties and the parameters are now physically meaningful.

6.3 THE ESTIMATOR AND DERIVATIVES

Because multi trace parameters describe the spline function over the lateral distance on which it is defined the consequence of this parameterization is that the whole block of traces has to be inverted integrally (whence: "multi trace inversion"). The basic formula for the estimator is again the least squares function (5.7) but now the data vector $\mathbf{d}$ is built up from several traces:

$$\mathbf{d}^T = (\mathbf{d}_1^T \dots \mathbf{d}_{nt}^T), \quad (6.10)$$

where $\mathbf{d}_i$ is the data trace i and nt is the number of traces. A similar description of course is used for the forward model $\mathbf{g}(\mathbf{x})$. In multi trace inversion it is possible in principle to account for laterally bandlimited noise, which is not possible in single trace inversion. In the software built however the covariance matrix of the noise is $\mathbf{C}_n = \sigma_n^2 \mathbf{I}$, implying white stationary noise, uncorrelated from trace to trace. For the specification of the a priori information in $\mathbf{x}^i$ and $\mathbf{C}_x$, see section 6.5.

An important consequence of the linearity of the polynomial formulation is that linear constraints for single trace parameters remain linear in the multi trace formulation. Let the linear constraints for all traces be gathered in the single trace formulation:

$$\mathbf{l} \leq \mathbf{B}\mathbf{z} \leq \mathbf{u} \quad (6.11)$$

For the multi trace parameters we then have:

$$l \leq BEx \leq u, \quad (6.12)$$

which is also linear. Bounds on single trace parameters however are converted to linear constraints in a multi trace formulation.

The linear relationship between multi trace and single trace parameters is also advantageous for the calculation of derivatives. The Jacobian matrix for the multi trace problem is given by:

$$\begin{aligned} J &= \frac{\partial g}{\partial x} \\ &= \frac{\partial g}{\partial z} \frac{\partial z}{\partial x} \\ &= \begin{bmatrix} J_1 \\ \vdots \\ J_{nt} \end{bmatrix} E, \end{aligned} \quad (6.13)$$

where J_i is the Jacobian matrix of trace i with respect to all single trace parameters. This matrix of course has a special structure because it is nonzero only for the single trace parameters corresponding to trace i . Relation (6.13) should therefore not be implemented straightforwardly, also because of storage considerations. However once the corresponding manipulations are implemented in general form, new (e.g. higher order) spline functions can be added to the software very easily.

6.4 BASIC STEPS IN MULTI TRACE INVERSION

Directly specifying a priori information for multi trace parameters is difficult. Direct interpretation of results for multi trace parameters is also difficult. A practical sequence of steps for multi trace inversion therefore is:

- 1) Formulation of a priori information on a "single trace" basis.
- 2) Transformation of the results of step 1 to multi trace parameters.
- 3) Inversion for the multi trace parameters using the a priori information as derived in step 2.
- 4) Inspection of the inversion results for the multi trace parameters where practical.
- 5) Transformation of the inversion results to single trace parameters and inspection on a single trace basis.

Steps 4 and 5 are not necessarily in that order. The steps are discussed in more detail now.

Step 1. The interpreter specifies an a priori model for the single trace parameters z^i . The specification and computation of the corresponding covariance matrix can be done in a practical way. This is discussed in the next section. For the moment assume that a covariance matrix C_z is available.

Step 2. The a priori information for the multi trace parameters is "estimated" from the a priori information of the single trace parameters using the Gauss Markov estimator in combination with (6.6) as the forward model:

$$\mathbf{x}^i = (\mathbf{E}^T \mathbf{C}_z^{-1} \mathbf{E})^{-1} \mathbf{E}^T \mathbf{C}_z^{-1} \mathbf{z}^i . \quad (6.14)$$

The a priori covariance matrix is accordingly:

$$\mathbf{C}_x = (\mathbf{E}^T \mathbf{C}_z^{-1} \mathbf{E})^{-1} . \quad (6.15)$$

Equation (6.14) corresponds to fitting spline functions through the single trace parameters. Normally with consistent a priori information and spline type selection the fit should be perfect. When the pdf for the single trace parameters is Gaussian, the pdf for the multi trace parameters is also Gaussian.

Step 3. Inversion for the multi trace parameters is done, using the results of (6.14) and (6.15) as a priori information. The estimate is the vector that minimizes:

$$2F(\mathbf{x}) = (\mathbf{d} - \mathbf{g}(\mathbf{x}))^T \mathbf{C}_n^{-1} (\mathbf{d} - \mathbf{g}(\mathbf{x})) + (\mathbf{x}^i - \mathbf{x})^T \mathbf{C}_x^{-1} (\mathbf{x}^i - \mathbf{x}) . \quad (6.16)$$

It is easily seen that this can also be derived when the single trace a priori information is interpreted as independently obtained additional data: the second term on the right hand side in that case would be:

$$(\mathbf{z}^i - \mathbf{E}\mathbf{x})^T \mathbf{C}_z^{-1} (\mathbf{z}^i - \mathbf{E}\mathbf{x}) . \quad (6.17)$$

The only condition for the two viewpoints to be equivalent is that the a priori model $\mathbf{z}^i$ can be described exactly by a spline function $\mathbf{z}^i = \mathbf{E}\mathbf{x}^i$ so that (6.17) can be written as:

$$(\mathbf{x}^i - \mathbf{x})^T \mathbf{E}^T \mathbf{C}_z^{-1} \mathbf{E} (\mathbf{x}^i - \mathbf{x}) , \quad (6.18)$$

which is equivalent to the second term of (6.16), in view of (6.15). The results of the inversion are denoted by $\mathbf{x}$, $\mathbf{C}_x$ for the model and a posteriori covariance matrix respectively.

Step 4. Result and uncertainty analysis is done for the multi trace parameters where practical. The covariance matrix, the resolution matrix and the weak directions may be computed. For the plotting of the models however conversion to single trace parameters as discussed below is necessary in practice.

Step 5. The multi trace results can be converted to single trace results using the relation (6.6). For the estimated parameters we have:

$$\hat{\mathbf{z}} = \mathbf{E}\hat{\mathbf{x}} . \quad (6.19)$$

The a posteriori covariance matrix for the single trace parameters is given by:

$$C_z = EC_x E^T \quad (6.20)$$

The a priori covariance matrix also can be transformed back, analogous to (6.20). Denoting the result with C_{z2} gives:

$$C_{z2} = EC_x E^T \quad (6.21)$$

Substituting (6.15) in (6.21) yields:

$$C_{z2} = E(E^T C_z^{-1} E)^{-1} E^T \quad (6.22)$$

In general C_{z2} will not be equal to C_z . This is not surprising because the single trace parameters are first projected on the multi trace parameter space, a space of smaller dimensions. Degrees of freedom are lost. Transforming the result back will not render the same matrix. The results (6.19) and (6.20) can be used as in a normal inversion procedure. The transformation back to single trace parameters is in principle a sampling of the spline function. The sampling rate can be anything one likes. It can for example also be chosen to be smaller than the trace to trace distance.

6.5 PRACTICAL INCORPORATION OF A PRIORI INFORMATION

For this section it will be convenient to describe the total vector of single trace parameters z as the combination of the vectors z_i containing the single trace parameters of one trace:

$$z^T = (z_1^T \dots z_{nt}^T) \quad (6.23)$$

Compared to (6.5) the vector simply is reordered. The total covariance matrix for z will be denoted by C_z , the partial covariance matrix for the single trace parameters of trace i by C_i and the crosscovariance matrix of traces i and j by C_{ij} . The trace corresponding with the well position has index l .

First the situation around the well is discussed. One type of a priori information available is the parameter model derived from the logs and the related uncertainties which can be specified through the covariance matrix C_l . Usually this will be a diagonal matrix. Furthermore an interpreter may provide a priori models for all traces in the block around the well. This may for example contain trend information. When no significant information concerning lateral changes is available the log model can be used for all traces, corresponding with an expectation of a laterally invariant medium. Mathematically the specification of models can be described by:

$$z_{i+l} = z_i + \Delta z_i \quad (6.24)$$

where Δz_i is the vector containing the variations from trace i to trace $i+l$. Specifying a

priori models for all traces is of course equivalent to specifying a priori values for the variations Δz_i . The model is built up starting at the well.

Of course the interpreter is not 100% sure of the a priori models and hence an uncertainty should be specified for the variations Δz_i . For simplicity these uncertainties are assumed to be equal for all locations in the block. They are denoted by C_Δ which usually will be a diagonal matrix. In the case of a laterally invariant a priori model the uncertainties in C_Δ simply represent the trace to trace variations. Using statistical results, relations like:

$$C_{i+1} = C_i + C_\Delta, \quad (6.25)$$

and

$$C_{i+1,i} = C_i, \quad (6.26)$$

can be derived. Starting from the well position the full covariance matrix C_z can be derived and is given by:

$$C_z = \begin{bmatrix} \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & C_\Delta + 2C_\Delta & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & C_\Delta + C_\Delta & C_\Delta & C_\Delta & \dots & \dots & \dots \\ \dots & \dots & C_\Delta & C_\Delta & C_\Delta & \dots & \dots & \dots \\ \dots & \dots & C_\Delta & C_\Delta & C_\Delta + C_\Delta & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & C_\Delta + 2C_\Delta & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{bmatrix}$$

or, in formula:

$$C_{ij} = C_\Delta \quad \text{for } i \geq \Delta, j \leq \Delta, \\ \text{and } i \leq \Delta, j \geq \Delta.$$

and

$$C_{ij} = C_\Delta + \min(|i-\Delta|, |j-\Delta|) C_\Delta \quad \text{for } i > \Delta, j > \Delta, \\ \text{and } i < \Delta, j < \Delta. \quad (6.27)$$

This constitutes the desired information for step 2 of the previous section. This procedure allows to force the results to be very close to an accurate well log model (C_Δ small) or to constrain the flexibility of the spline functions when little trace to trace variations are expected (C_Δ small).

In this procedure a well log is assumed. Processing from block to block and transferring information in that situation is not extensively studied. A simple option for a block of data not close to the well is to take the adjacent trace of the previous block as a pseudo well log. The a posteriori knowledge of that trace serves as a priori information for the present block.

6.6 RESULTS FOR MULTI TRACE INVERSION

In this section examples for multi trace inversion are given. First an example with cubic splines and a full uncertainty analysis are presented. The true model and the data (both the same as in figure 5.12) are shown in figure 6.2.a,b. The lateral variations of the true model can exactly be described by cubic splines. The a priori model is a flat layer model and is significantly wrong at the well position, trace number 8. The standard deviations of the well log model and the trace to trace uncertainties are given in table 6.1.

The inversion procedure is performed as described in the previous section. The result is given in figure 6.2.d. The residual traces are shown in figure 6.2.e, in comparison with the noise realization on the data in figure 6.2.f. The energy of the residual is -11.7 dB relative

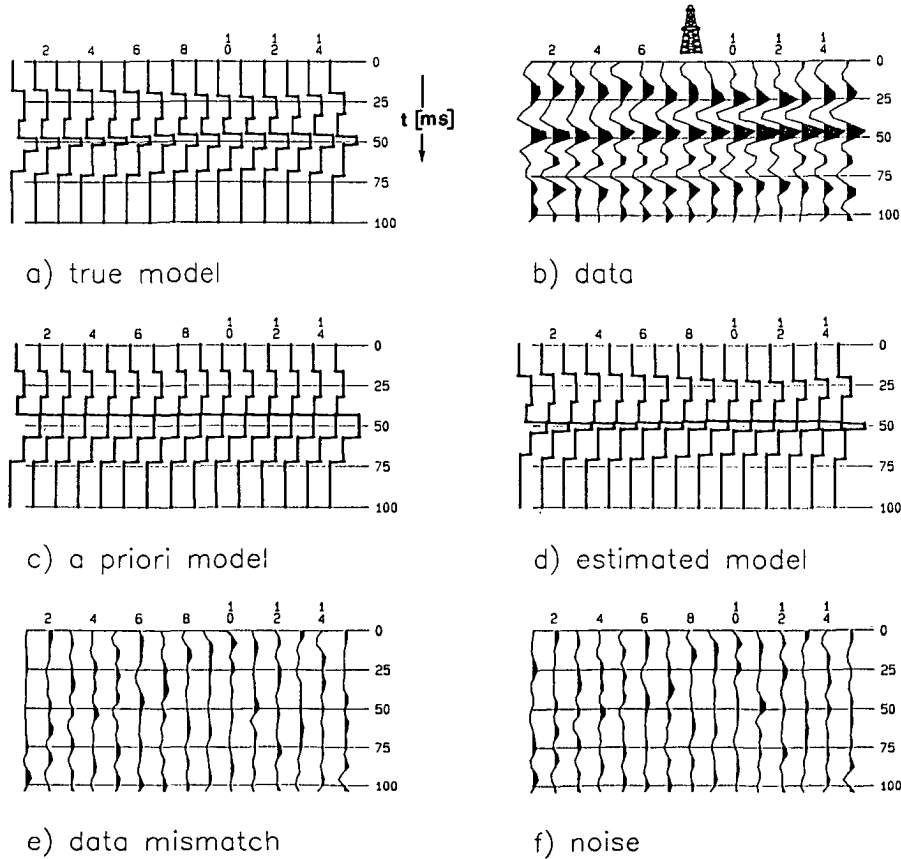


Figure 6.2 Example for multi trace inversion with cubic splines.

Table 6.1 Standard deviations of a priori information in multi trace experiment.

	impedance [$10^6 \text{ kgm}^{-2} \text{ s}^{-1}$]	lag time [ms]
well log model	.5	6.
trace to trace	.2	2.

to the data while the noise to signal ratio is -10 dB. Because there are relatively few parameters compared to the single trace procedure the residual remains higher. A close comparison of the noise and the residuals show many similarities. The procedure, to some extent, separates the signal from the noise.

In figure 6.3 the lagtimes of the a priori, the estimated and the true model are shown to allow a detailed comparison. It is clear that the lagtimes are estimated very well, especially those of the first and last interface. Note that also the thin layer, which is only 4 ms thick at trace positions 4 and 5 is very well determined. The resolution of this type of inversion

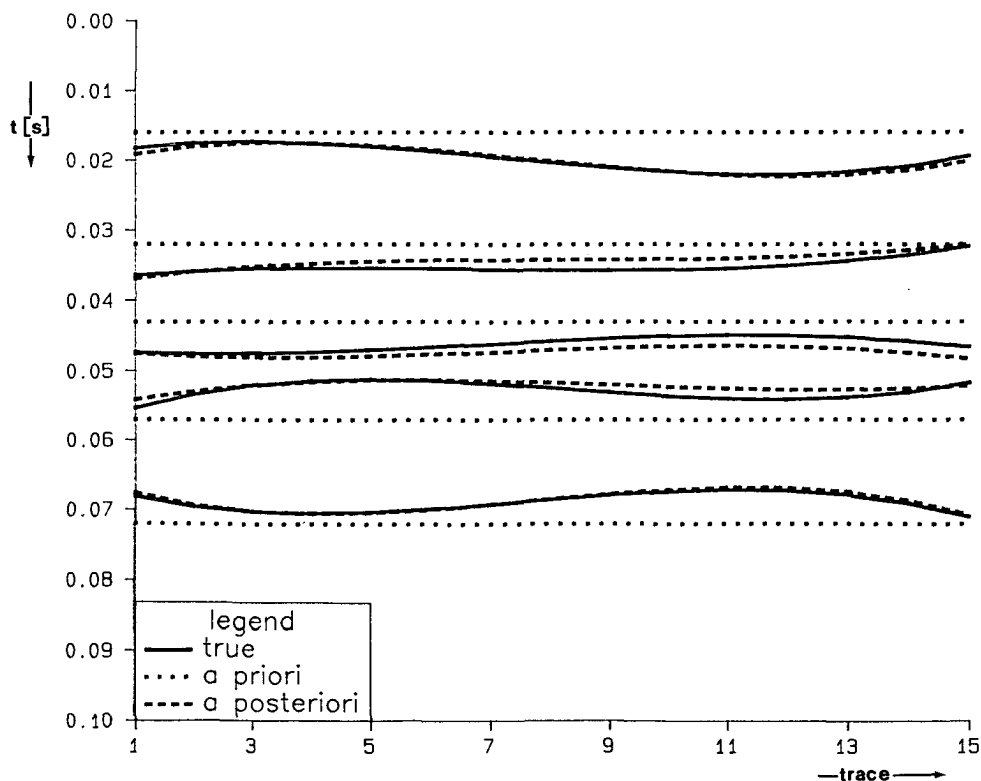


Figure 6.3 True, a priori and a posteriori lagtime values.

Table 6.2 Errors in the estimates and the data mismatch for the multi trace inversion of a synthetic section.

model	impedance [$10^6 \text{ kgm}^{-2}\text{s}^{-1}$]			lag time [ms]			residual [dB]
	error ₁	error ₂	error ₃	error ₁	error ₂	error ₃	
a priori model	.39	.41	.75	3.5	3.7	6.1	n.a
estimated model	.28	.35	.73	.7	.9	1.7	-11.7

procedure is about an order of magnitude higher than that of wavelet deconvolution. The numerical presentation of the size of the errors is given in table 6.2. The values are calculated by using the single trace parameters of all traces.

The lagtimes are estimated with an average accuracy of less than 1 ms. The acoustic impedances are also improved albeit slightly. A more detailed picture is given in figure 6.4,

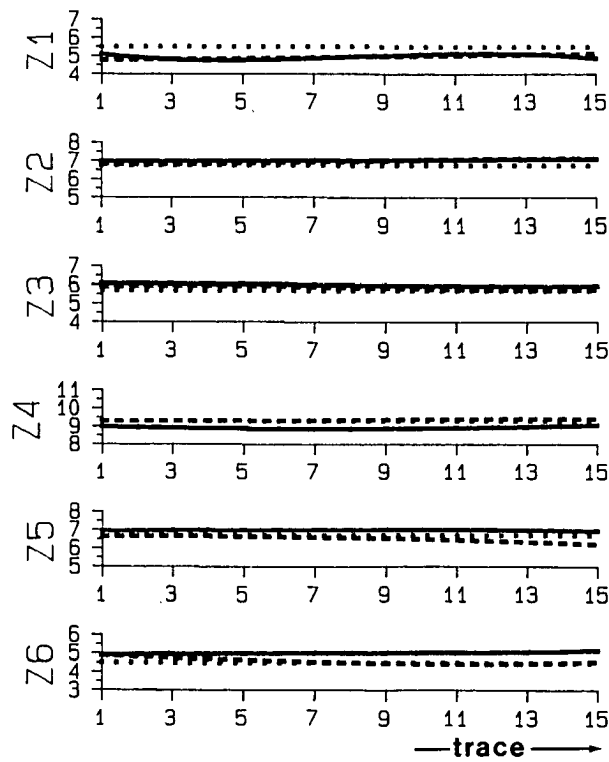


Figure 6.4 True, a priori and a posteriori impedance values for all layers. The units of the impedances are $10^6 \text{ kgm}^{-2}\text{s}^{-1}$.

..... a priori
----- a posteriori
—— true

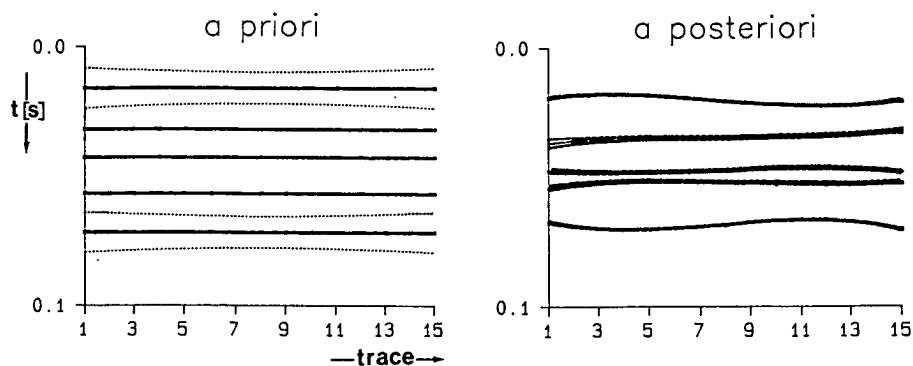


Figure 6.5 A priori and a posteriori uncertainty intervals for the lagtimes. The a priori uncertainty intervals are shown for reflectors 1 and 5 only. The other intervals are of equal size.

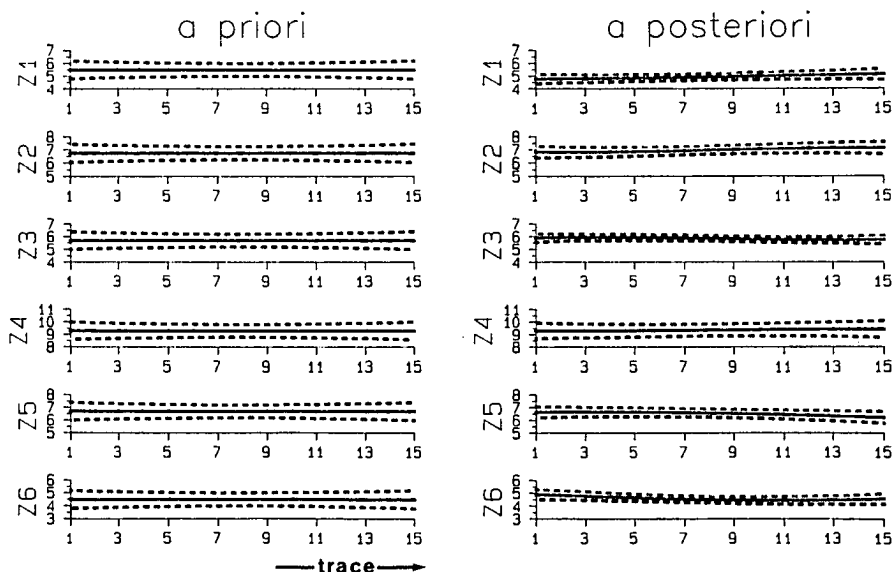


Figure 6.6 A priori and a posteriori uncertainty intervals for the impedances.

where the true, the a priori and the estimated impedances are given for each layer. Some layers are fairly well estimated, e.g. layers 1 and 3. Others, e.g. 4 and 5, have not been improved.

The uncertainty analysis is started with the standard deviations on a single trace parameter basis. The a priori and a posteriori uncertainty intervals for the lagtimes are given in figure 6.5. The uncertainty is drastically reduced. The uncertainty in interfaces 1 and 5 is

approximately .2 ms and for the others it is approximately .6 ms. Note that these uncertainties are about 3 times as small as for single trace inversion as discussed in section 5.6 for similar data. Similar plots are shown in figure 6.6 for the acoustic impedances. The reduction of uncertainty is not as strong as for the lagtimes. For layers 1, 3 and 6 there is clearly a reduction of uncertainty of a factor 2 to 3. For layers 2, 4 and 5 little information has been gained.

In figure 6.7 the structure of the matrices used in the analysis procedure is given. Before statistical normalization all elements have direct physical interpretation. After the statistical normalization however the elements in the 4*4 spline submatrices are linear combinations of the original elements and are not directly interpretable anymore. The elements of one submatrix remain together because there are no a priori correlations between the different

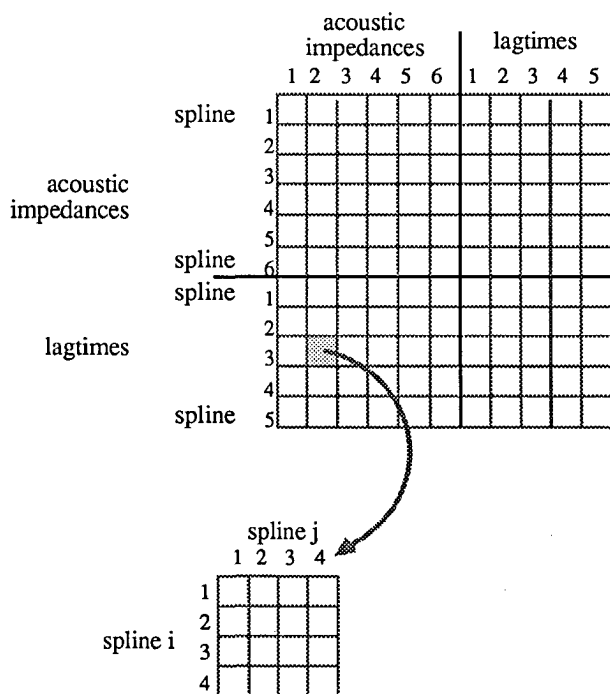


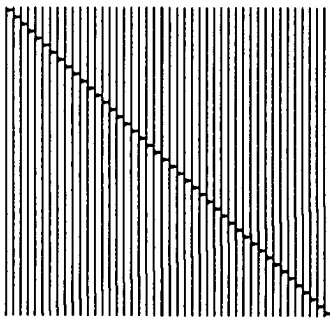
Figure 6.7 Layout of the matrices used for uncertainty analysis in the multi trace example with cubic splines.

spline functions. In figure 6.8 the a priori and a posteriori covariance matrices, the resolution matrix and the a posteriori correlation matrix are given. The problem is statistically normalized so that the a priori covariance matrix equals the identity matrix. From the a posteriori covariance matrix it can be seen that much information has been

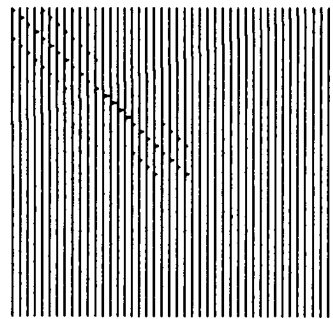
gained on the lagtimes. The acoustic impedance of layer 4 is very poorly resolved, which can also clearly be seen in the resolution matrix. Significant correlations occur between the first and second acoustic impedance, the fifth and the sixth impedance and between the lagtimes and the impedances of the thin layers in the middle of the model.

The singular value spectra for the multi trace example are given in figure 6.9. The full range is given in figure 6.9a. The blowup in figure 6.9.b shows that the 13 lowest singular values of the likelihood spectrum are smaller than 1, so that we have 13 weak directions. Only the smallest singular value (number 44) is exactly zero. The parameter models along its corresponding eigenvector can be obtained from one another by multiplication of all acoustic impedance parameters with the same factor, see also figure 6.10. The next three singular values (numbers 43, 42 and 41) are close to zero. In figure 6.10 the a priori pdf, the likelihood function and the a posteriori pdf are plotted along a number of interesting

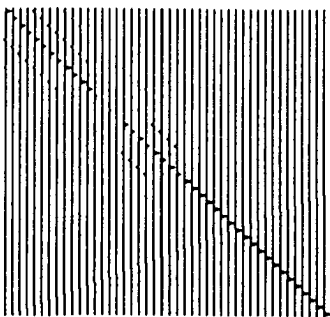
a priori covariance



a posteriori covariance



resolution matrix



a posteriori correlation

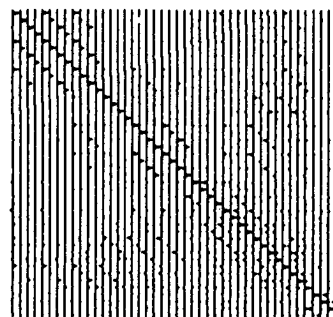


Figure 6.8 Results for the matrices used in uncertainty analysis for the multi trace example.

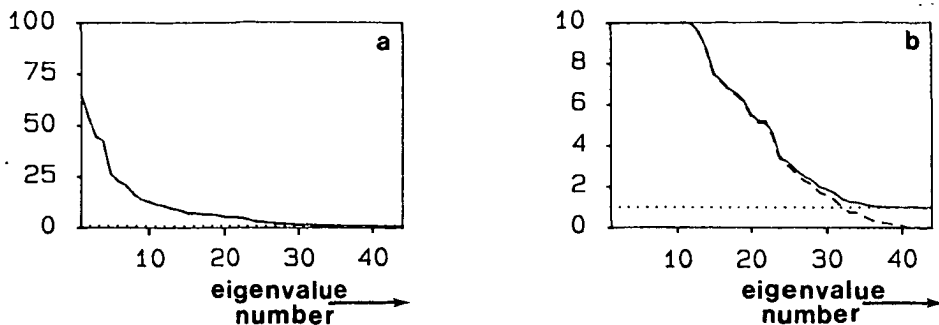


Figure 6.9 Singular value spectra for the multi trace example. a) full range; b) blow up.

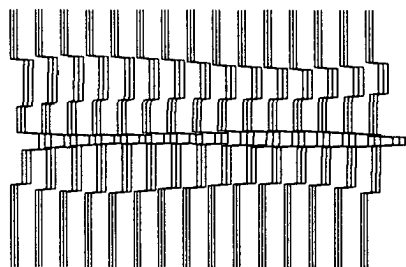
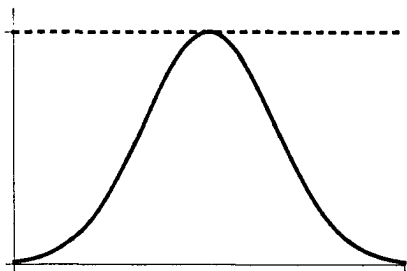
..... *a priori spectrum*
 - - - - - *likelihood spectrum*
 ——— *a posteriori spectrum*

eigenvectors. At the bottom of each plot the estimated and the extreme models, each $3/\lambda_i$ in distance away from the estimated model are shown in overlay. Only the likelihood function along eigenvector 44 is exactly constant. The others show a slight dip. The direction of the eigenvectors 43, 42 and 41 show fluctuations in the acoustic impedances corresponding approximately with laterally varying constant factors for each trace. Eigenvectors 40 – 37 correspond mainly with exchanges between the thickness and the acoustic impedance of the fourth layer. The likelihood function shows nonlinearity effects of the forward model. It is clear that the a priori information determines the answer along these directions. Along the eigenvectors 36 – 32 the likelihood function gets more and more important until at eigenvector 32 the a priori pdf and the likelihood function have more or less equal influence on the a posteriori pdf. Starting from eigenvector 31 the likelihood function starts to dominate the a priori pdf. For high singular values, see eigenvectors 15, 10 and 5 the a priori pdf is virtually flat compared to the likelihood function. It can be seen from the bottom pictures that the distances covered along the eigenvectors become very small.

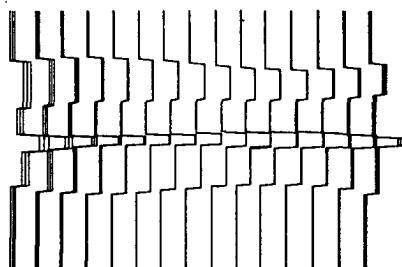
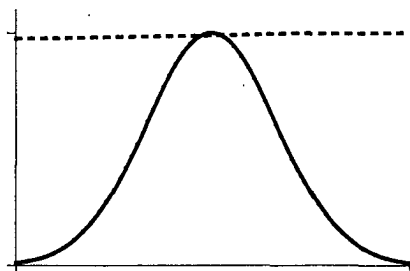
In figure 6.11 the approximated a posteriori pdf $N(\hat{x}, (\hat{J}^T \hat{J})^{-1})$ is plotted in comparison with the true a posteriori pdf. The approximation is excellent for the four lowest singular values corresponding approximately to constant factors for the acoustic impedances. Along eigenvectors 40 and 37 there are slight deviations from the true one. Eigenvector 36 shows the largest deviations. For higher eigenvalues the approximation is excellent. It can be concluded that the computed a posteriori covariance matrix is sufficiently accurate.

The same data set is also inverted with different orders of the spline function. For comparison of the results the true model and the noise realization are once more given in figure 6.12a,b. The data set is the same as given in figure 6.2. The a priori information used is the same as in the cubic spline example, see table 6.1. The estimated models for the

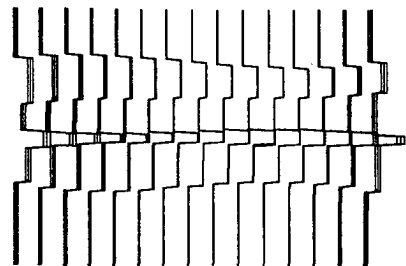
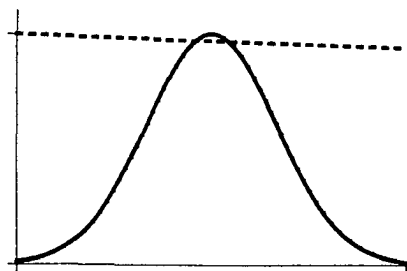
eigenvector 44



eigenvector 43



eigenvector 42



eigenvector 41

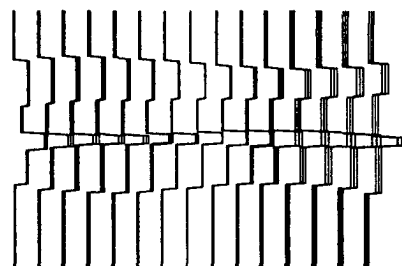
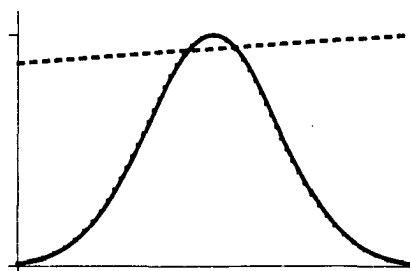
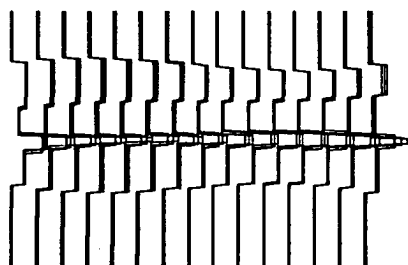
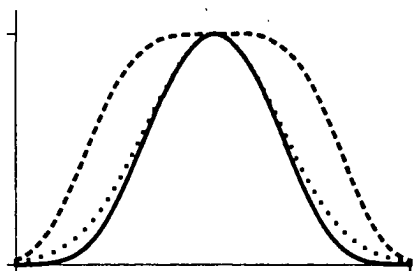
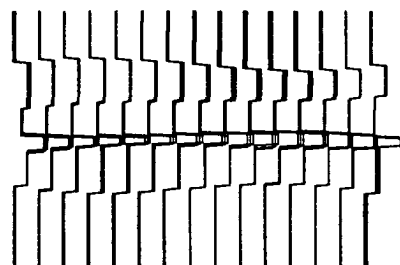
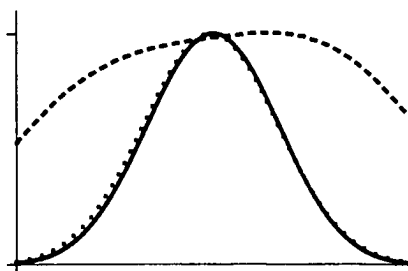


Figure 6.10 Plots of the pdfs along the eigenvectors for the multi trace example.
 a priori pdf; - - - - - likelihood function; ——— a posteriori pdf.

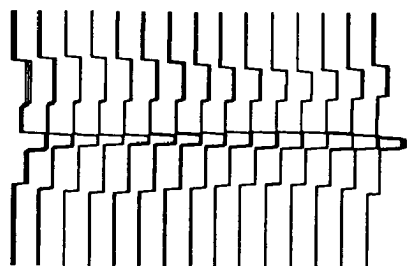
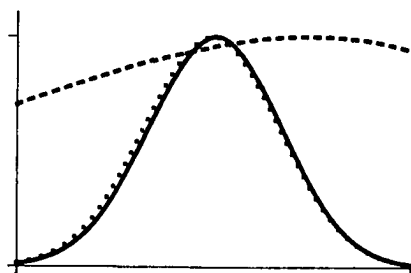
eigenvector 40



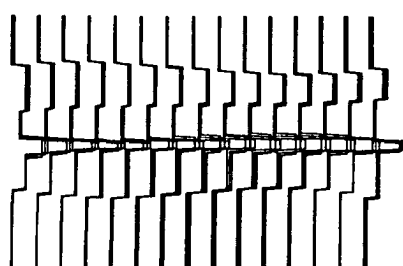
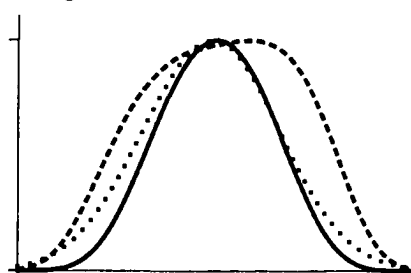
eigenvector 39



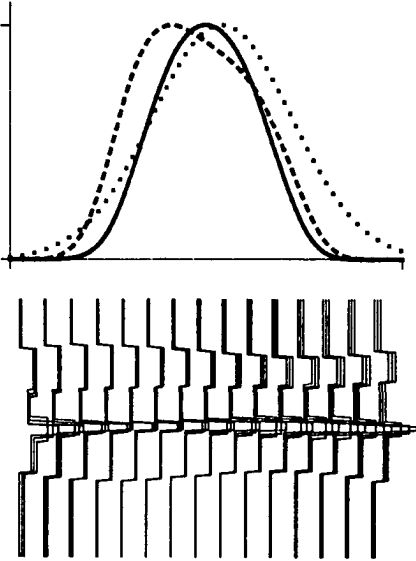
eigenvector 38



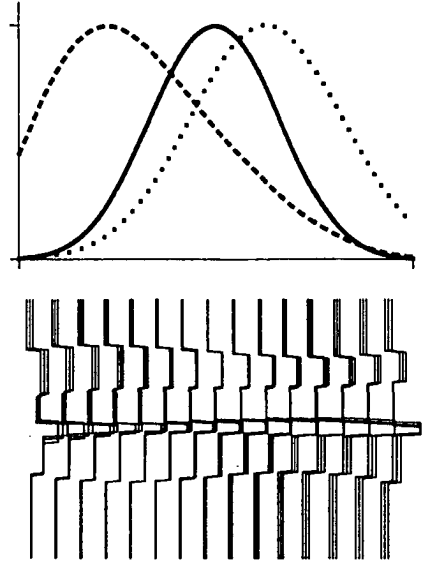
eigenvector 37

*Figure 6.10 continued.*

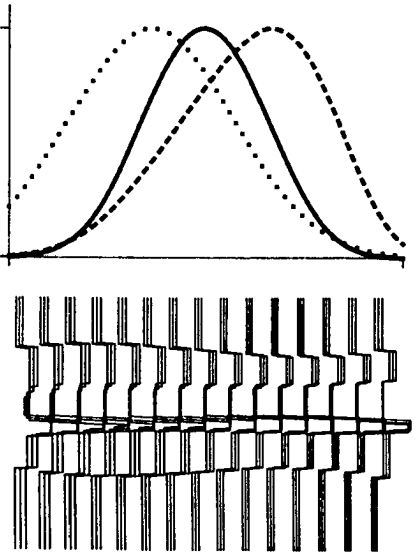
eigenvector 36



eigenvector 35



eigenvector 34



eigenvector 33

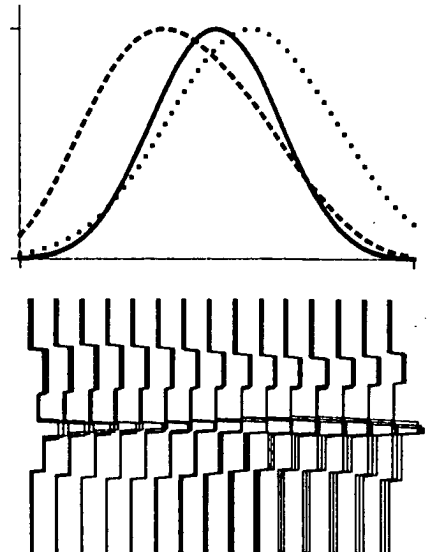
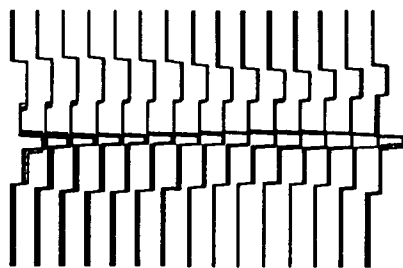
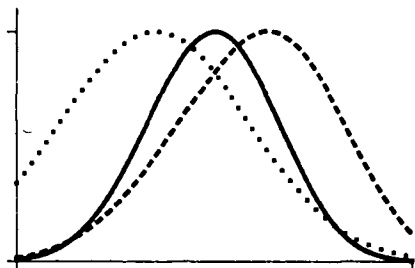
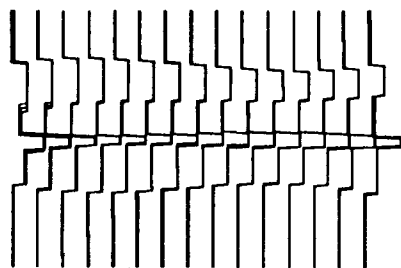
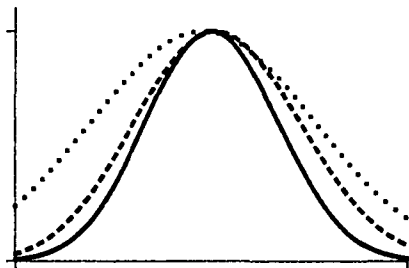


Figure 6.10 continued.

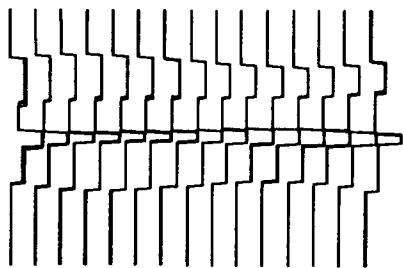
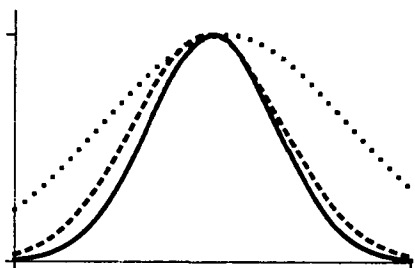
eigenvector 32



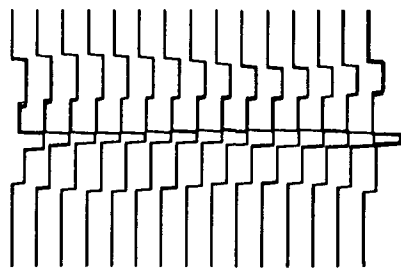
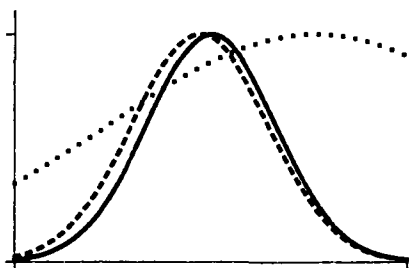
eigenvector 31



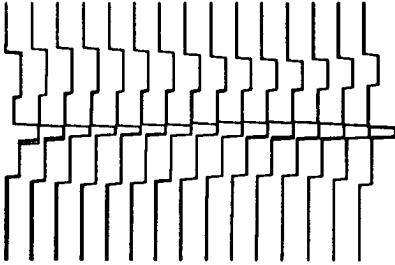
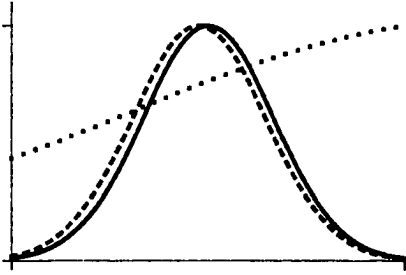
eigenvector 30



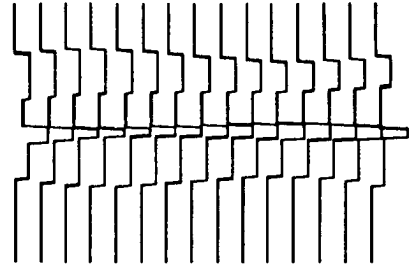
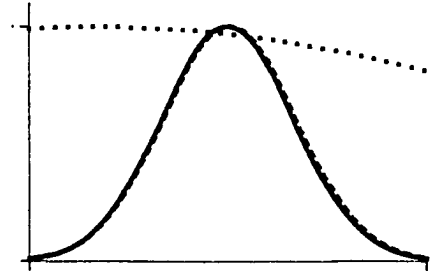
eigenvector 25

*Figure 6.10 continued.*

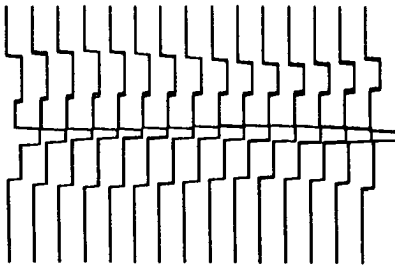
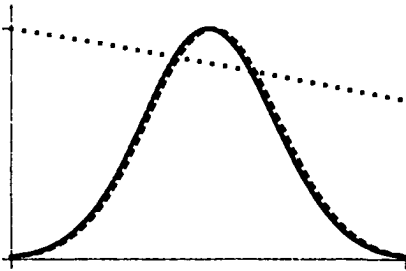
eigenvector 20



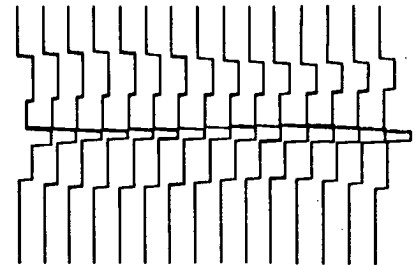
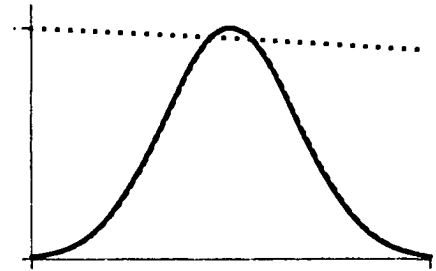
eigenvector 15



eigenvector 10



eigenvector 5

*Figure 6.10 continued.*

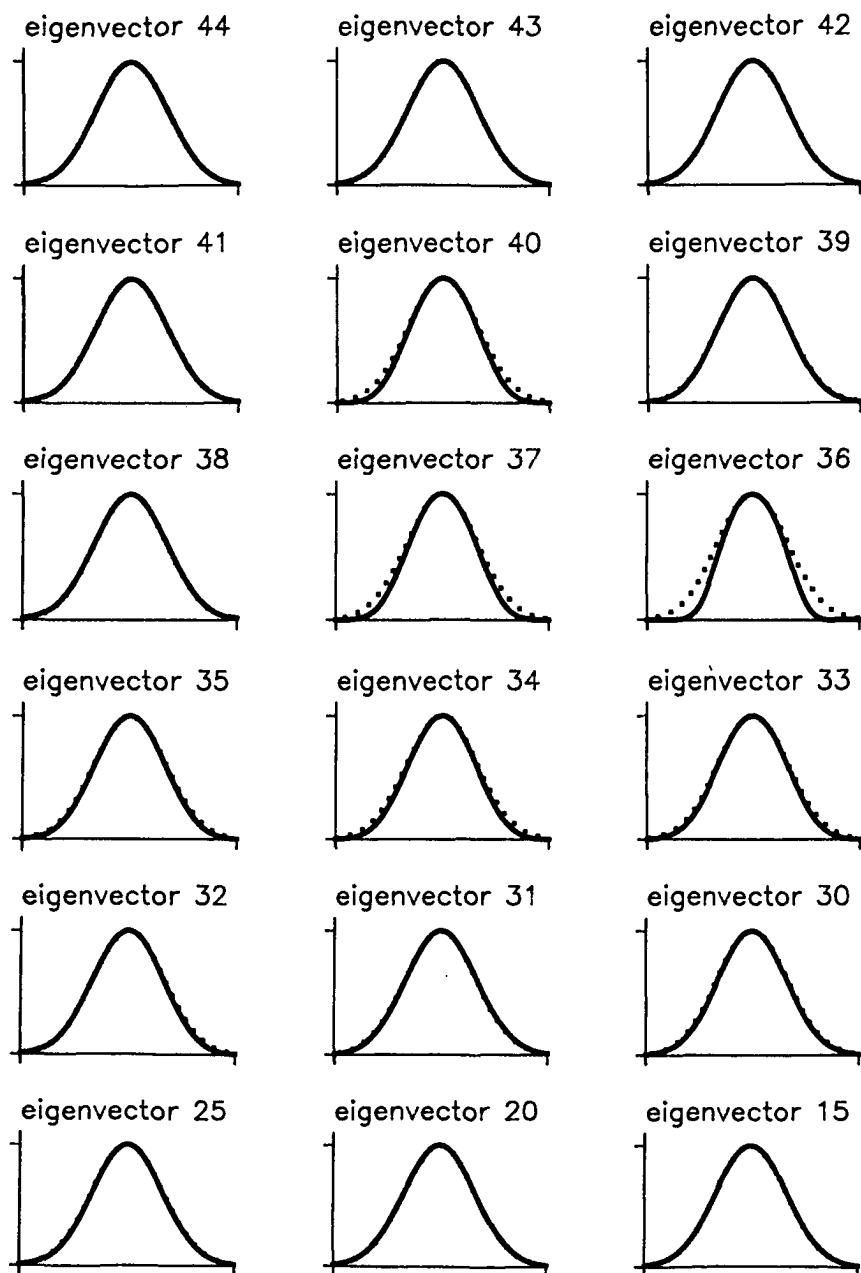
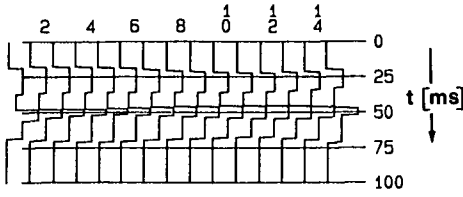
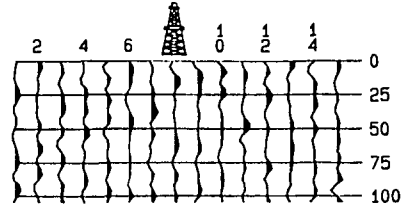


Figure 6.11 Plots of the pdf's along the eigenvectors for linearity analysis.

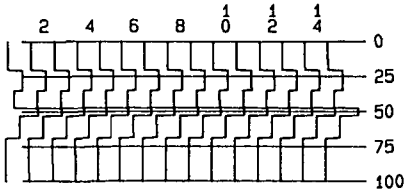
— true a posteriori pdf
 approximated a posteriori pdf



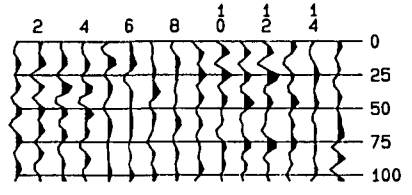
a) true model



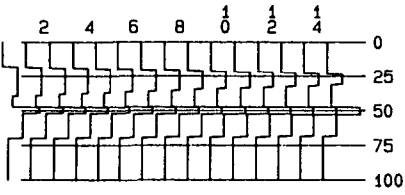
b) noise



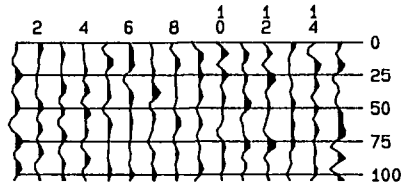
c) result 0-order spline



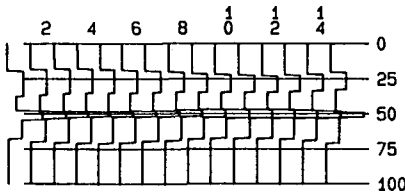
d) residual for 0-order



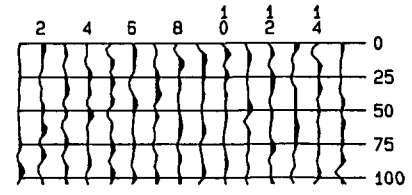
e) result 1st order spline



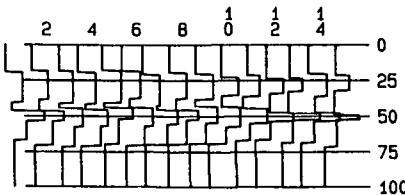
f) residual for 1st order



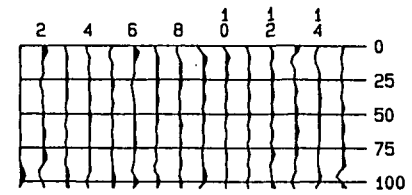
g) result 3rd order spline



h) residual for 3rd order



i) result single trace



j) residual single trace

Figure 6.12 Multi trace and single trace inversion of a data set.

Table 6.3 Errors in the estimates and the data mismatch for different parameterizations.

model	impedance [$10^6 \text{ kg m}^{-2} \text{ s}^{-1}$]			lag time [ms]			residual [dB]
	error ₁	error ₂	error ₃	error ₁	error ₂	error ₃	
a priori model	.39	.41	.75	3.5	3.7	6.1	n.a.
0-order spline	.16	.19	.35	1.2	1.3	2.8	-7.4
1 st order spline	.28	.32	.67	1.2	1.4	4.0	-9.0
cubic spline	.28	.35	.73	.7	.9	1.7	-11.7
single trace	.46	.54	1.1	1.6	2.2	6.4	-14.7

different orders are given in figures 6.12c,e,g and the corresponding residuals in figures 6.12d,f,h. The same data set is also inverted with the single trace approach with trace to trace a priori standard deviations of $.25 \cdot 10^6 \text{ kg}^{-2} \text{ s}^{-1}$ for the impedances and 2.5 ms for the lagtimes. These standard deviations had to be set larger than the narrow constraints of section 5.7 because the well log has much larger errors than in the example discussed there. The numerical details of the experiment are given in table 6.3. Note that the a priori information is only applicable for the multi trace parameterization.

From the residuals as plotted in figure 6.12 and their energy as given in table 6.3 it is clear that the more parameters used, the lower the residual. For the 0-order and the first order parameterization the energy in the residual is higher than the noise energy, which is an indication of a violation of one of the assumptions in practice. In this case the cause is an underparameterization. From the results it will be clear however that an underparameterization not necessarily leads to a bad result, on the contrary, the 0-order spline function gives the best result for the acoustic impedances. This is due to the fact that the true impedance profiles, see figure 6.3 can reasonably be described by a constant function. In classical terms we can state that the 0-order then causes little systematic errors (or bias) while having much lower nonsystematic errors (or variance) due to the smaller number of parameters. For the lagtimes however this parameterization is too coarse and the cubic spline is superior. The first order spline function is slightly better than the cubic spline for the acoustic impedances. A combination of 0-order for the acoustic impedances and cubic splines for the lagtimes would have given still better results. All multi trace parameterizations give better results then the single trace approach.

WAVELET ESTIMATION

7.1 INTRODUCTION

The inversion schemes discussed in chapter 5 and 6 assumed the wavelet at the target zone to be known. In practice of course it never is. Apart from the fact that the source wavelet is seldom measured, the wavelet is distorted when propagating through the overburden. The wavelet therefore has to be estimated. Two major classes of wavelet estimation procedures can be distinguished:

- 1) the "statistical" class that uses seismic data only,
- 2) the "deterministic" class that also uses reflectivity information.

The statistical estimation schemes have to make certain assumptions concerning the reflectivity in order to separate the reflectivity from the wavelet. Often this entails the assumptions of white reflectivity and the minimum phase assumption for the wavelet. In homomorphic deconvolution techniques it is assumed that the reflectivity and the wavelet can be separated in the complex cepstrum of the seismic trace. An overview of statistical wavelet estimation techniques can be found in Lines and Ulrich (1977). More recently Angeleri (1983) proposed a statistical method in which no phase assumption for the wavelet is needed.

Methods that use reflectivity information can be expected to yield better estimates. The statistical methods are therefore not considered in this thesis. Reflectivity information can be derived from a well log, which is assumed to be available. The deterministic methods of course also use statistical estimation techniques. In a number of papers the Wiener filtering technique is used to extract the wavelet, see e.g. Stone (1976). In this approach the seismic

trace is deconvolved for the reflectivity. Lines and Treitel (1985) show the close relationship of this approach to least squares inversion. Danielson and Karlsson (1984) discuss a method with a straightforward division of the trace spectrum by the reflectivity spectrum in the frequency domain. The results show rather long tails in the estimated wavelets. That the length of the wavelet is indeed an important issue is discussed by Hampson and Galbraith (1981). They also mention the strong deteriorating effects of well log errors. In order to arrive at an acceptable result they use an extensive procedure of selection of several data gates, selection of the shortest wavelets, using Berkhout's (1977) length criterion and averaging of those estimates. Oldenburg, Levy and Whittall (1981) present a.o. a time domain method in which a spectral expansion technique described by Parker (1977) is used which shows strong correspondence with the SVD method to be discussed in section 7.6. Only significant eigenvalues are used to construct a wavelet estimate. Additionally they can set smoothness constraints. They stress the inherent nonuniqueness of the problem when reflectivity also has to be estimated and mention the necessity of subjective input of the user or of additional constraints. White (1980) and Walden and White (1984) use a partial coherence matching technique in the frequency domain. It is a multi channel technique in which one channel e.g. corresponds to primaries and internal multiples and another to surface multiples. Each channel has its own wavelet that has to be estimated. An important issue is the proper choice of the spectral smoothing factor, which is closely related to the length of the wavelet. In section 7.7 some additional remarks on this issue will follow.

All of the methods mentioned above with (in a sense) the exception of Hampson and Galbraith are single trace techniques. Each method also parameterizes the wavelet as a time series. A number of inversion schemes which simultaneously estimate the reflectivity or impedance and the wavelet are discussed in chapter 4. Some of those techniques do not parameterize the wavelet directly as a time series but use another parameterization, for example an ARMA model.

In this chapter a parameter estimation based approach will be used to estimate the wavelet. A straightforward parameterization in the time domain is used, which leads to a simple linear forward model. A multi trace procedure is proposed which strongly reduces noise influence when compared to the single trace procedure. A priori information on the wavelet cannot be used as straightforwardly as in acoustic impedance inversion because, due to arbitrary scale factors often introduced in the processing of seismic data the absolute scale of the wavelet is unknown. Nevertheless some stabilization procedures like the SVD cutoff method, ridge regression and parameter selection can be utilized in order to improve the estimator performance. In chapter 8 it is discussed how wavelet estimation can be integrated with impedance inversion.

7.2 THE FORWARD MODEL AND PARAMETERIZATION

The basic physical relation used is the one-dimensional convolutional model:

$$s(t) = r(t) * w(t) , \quad (7.1)$$

with $s(t)$ the synthetic trace, $w(t)$ the wavelet and $r(t)$ the impulse response of the target zone. In the sequel $r(t)$ will be referred to as the reflectivity. The wavelet $w(t)$ contains the effects of propagating through the overburden. The adoption of the convolutional model, implies the assumption that the wavelet is constant in time, which is reasonable for small data gates.

Because $w(t)$ is a bandlimited time series the most straightforward parameterization, not imposing any restrictions on the shape is a sampling in the time domain:

$$w_j = w(j\Delta t_w) , \quad (7.2)$$

where Δt_w is the sample rate, which should be chosen such that

$$\Delta t_w < \frac{1}{2f_{\max}} , \quad (7.3)$$

where $f_{\max}$ is the highest frequency component in the wavelet. In general $f_{\max}$ will not be exactly known so Δt_w should be chosen safely small when errors due to undersampling are to be avoided. Note however that for a given wavelet length a decrease of the sample rate will increase the number of parameters and therefore the uncertainty in the estimated parameters. Hence there is a trade-off in the sample rate for an optimal estimator.

In order to arrive at a discrete formulation of the forward model $r(t)$ will be considered to be a bandlimited version of the true wide band reflectivity $r_1(t)$:

$$s(t) = b(t) * r_1(t) * w(t) , \quad (7.4)$$

where $b(t)$ is a zero-phase bandpass filter. This formulation is exact as long as the condition:

$$B(f) = 1 \quad \text{for } f: W(f) \neq 0. \quad (7.5)$$

is satisfied. This bandlimitation for the reflectivity $r(t)$ allows the convolutional model to be written in discretized form:

$$s(i\Delta t_s) = \sum_{j=j_1}^{j_N} r(i\Delta t_s - j\Delta t_w) w(j\Delta t_w) , \quad (7.6)$$

where Δt_s is the sample rate of the data, j_1 and j_N are the indices of the first and last sample of the wavelet and N is the number of wavelet samples. The sample rate of the data need

not equal that of the wavelet. Equation (7.6) can be written in a vector notation:

$$\mathbf{s} = \mathbf{R}\mathbf{w} \quad , \quad (7.7)$$

where the data vector $\mathbf{s}$ is given by:

$$\mathbf{s}^T = (s(i_1 \Delta t_s) \dots s(i_M \Delta t_s)) \quad , \quad (7.8)$$

with i_1 and i_M the indices of the first and last sample of the data gate and M the number of data points. The convolution matrix $\mathbf{R}$ is given by:

$$R_{ij} = r(i \Delta t_s - j \Delta t_w) \quad . \quad (7.9)$$

Note that this formulation of the forward model allows any selection of rows and columns of $\mathbf{R}$ and of the corresponding elements of $\mathbf{s}$ and $\mathbf{w}$.

The forward model (7.7) can be extended to a multi trace formulation when the wavelet is assumed to be constant for a number of traces:

$$\begin{bmatrix} s_1 \\ \vdots \\ s_{nt} \end{bmatrix} = \begin{bmatrix} R_1 \\ \vdots \\ R_{nt} \end{bmatrix} \mathbf{w} \quad , \quad (7.10)$$

where s_i and R_i are the data and the reflectivity matrix for trace i respectively and nt is the number of traces. This can be written in short:

$$\mathbf{s} = \mathbf{R}\mathbf{w} \quad , \quad (7.11)$$

where from this point onwards $\mathbf{s}$ and $\mathbf{R}$ will contain the data and reflectivity of several traces. With this formulation we have arrived at the well known linear parametric forward model. A similar formulation with the reflectivity as spikes on an equidistant grid has been given by Taylor, Banks and McCoy (1979). In the next sections estimation procedures based on (7.11) are discussed.

7.3 THE GAUSS-MARKOV ESTIMATOR

The data traces $\mathbf{y}$ of course contain errors:

$$\mathbf{y} = \mathbf{s} + \mathbf{n} \quad . \quad (7.12)$$

Like in previous chapters the noise is assumed to be Gaussian with zero mean and covariance matrix C_n .

A priori information on the wavelet can not be specified as straightforwardly as for acoustic impedance or lagtimes because the absolute scale of the wavelet is often unknown, due to the facts that the source wavelet is hardly ever measured and that arbitrary factors may have been used for data scaling in processing. When the a priori pdf is therefore simply taken to be (locally) uniform the a posteriori pdf is also Gaussian with mean and

maximum (which will be used as a point estimate) at the location:

$$\hat{\mathbf{w}} = (\mathbf{R}^T \mathbf{C}_n^{-1} \mathbf{R})^{-1} \mathbf{R}^T \mathbf{C}_n^{-1} \mathbf{y} \quad , \quad (7.13)$$

and covariance matrix:

$$\mathbf{C}_{\hat{\mathbf{w}}} = (\mathbf{R}^T \mathbf{C}_n^{-1} \mathbf{R})^{-1} \quad . \quad (7.14)$$

The estimator (7.13) is known as the Gauss-Markov estimator. The covariance matrix of the noise $\mathbf{C}_n$ contains temporal as well as spatial bandlimitations. When the latter are neglected (as is also done in the inversion procedures described in chapters 5 and 6) the estimator (7.13) can be computed using:

$$\hat{\mathbf{w}} = \left(\sum_i \mathbf{R}_i^T \mathbf{C}_i^{-1} \mathbf{R}_i \right)^{-1} \sum_i \mathbf{R}_i^T \mathbf{C}_i^{-1} \mathbf{y}_i \quad , \quad (7.15)$$

where $\mathbf{y}_i$ and $\mathbf{C}_i$ are the data vector and the noise covariance matrix for trace i respectively. Equation (7.15) allows great savings in storage and computation time compared to the general form (7.13).

Although the mathematical form of the estimator is given by (7.13) or (7.15) there are a number of important design considerations for practical applications:

1) *the number of traces.*

The wavelet should be fairly constant over the traces used in estimation. A number of 10 traces seems very reasonable in practice. From the residuals gradual changes in the wavelet can be detected and the number of traces can accordingly be changed.

2) *the reflectivity.*

The reflectivity should be known over a time span that equals at least the sum of the data and wavelet length. Reflectors before the time gate map the tail of the wavelet into the data gate. For noncausal wavelets reflectivity after the gate also contribute to the data.

3) *the data length.*

The data gate should be as long as possible with the restrictions that the wavelet should be constant within this gate and the reflectivity must be known. Due to these restrictions the data gate will usually be small in practice, 200 to 500 ms, say. In the software built the data gate is the same for each trace but this is by no means a principle restriction.

4) *time origin of the wavelet.*

The time origin of the wavelet usually will only be approximately known. This issue corresponds with the alignment of the well log derived reflectivity with the seismic data. The reflectivity can always be shifted in time such that the wavelet is causal. Such a time shift should however not be unnecessarily large because it increases the number of parameters and hence the uncertainties in the estimates.

5) length of the wavelet.

The length of the wavelet can become an important issue when the estimator is unstable, see the next sections. In multi trace estimation with reflectors before the gate the wavelet can in principle be longer than the time gate. In practice however it is advisable to take the time gate significantly longer, at least 1.5 times as large as the wavelet. The wavelet should be chosen as short as possible for optimal results. The length however is not always accurately known in practice.

6) sample rates.

The sample rate of the data is not very important, as long as it is not undersampled of course. The sample rate of the wavelet should be chosen as large as possible as discussed above. Because the bandlimitations of the wavelet are not accurately known in practice, the sample rate should be small enough so that systematic errors due to undersampling will not be large. A practical procedure can be to roughly estimate the bandlimitations from the data under the assumption that the reflectivity is wide band and to choose the sample rate of both data and wavelet according to this bandlimitation using the Nyquist criterion. In the synthetic examples given below the sample rates of the wavelet and the data are chosen to be 4 ms, although the highest frequency in the data is 75 Hz. The reasons for this are simplicity in software and the fact that in practice one will be inclined to take a safe margin.

7) noise properties.

In the software built the noise is taken to be white and stationary, so that the covariance matrix equals the identity matrix. Note that the noise power need not be known for the estimator (7.13).

A synthetic example is discussed now. The true wavelet used in all examples in this chapter is the minimum phase wavelet shown in figure 7.1. The reflectivity of 10 traces is given in figure 7.2.a. The noise free data and the bandlimited Gaussian noise realization are shown in figure 7.2.b,c. The signal to noise ratio is 10 dB. The noisy data used for estimation is shown in figure 7.2.d. The true reflectivity is used for estimation but shifted 12 ms back in time because its time origin will not be exactly known in practice. The length of the wavelet is chosen to be 160 ms. The estimated wavelet is shown in figure 7.2.e in comparison with the true wavelet. It is resampled to 1 ms for display purposes. The

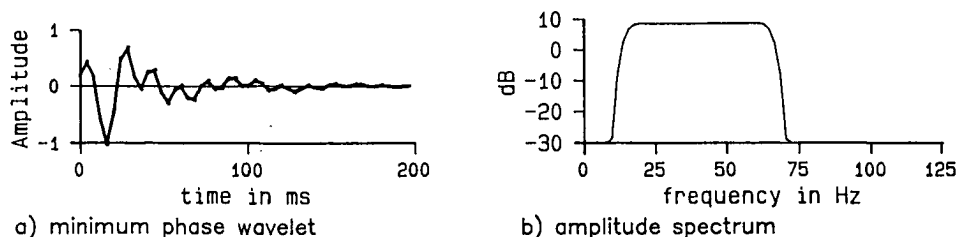


Figure 7.1 Minimum phase wavelet used in the wavelet estimation examples.

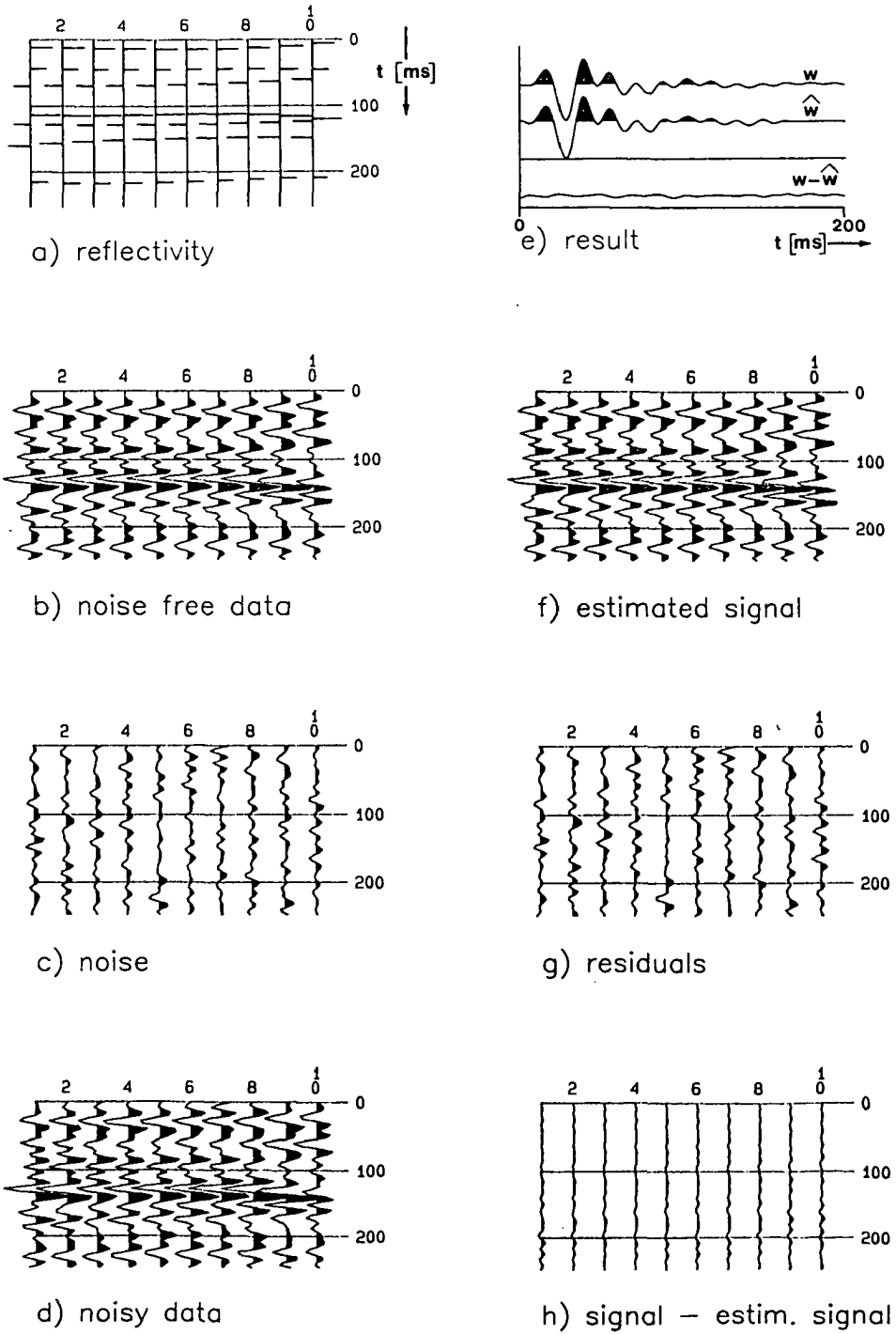


Figure 7.2 Example of wavelet estimation.

estimate is excellent, the errors very small. As a quantitative measure of the errors in the wavelet the normalized error (NE) will be used, defined as:

$$NE = \frac{(\mathbf{w} - \hat{\mathbf{w}})^T (\mathbf{w} - \hat{\mathbf{w}})}{\mathbf{w}^T \mathbf{w}} \quad (7.16)$$

The NE for this example is .011. The synthetic data for the estimated wavelet can be called the estimated signal:

$$\hat{\mathbf{s}} = \mathbf{R} \hat{\mathbf{w}} \quad (7.17)$$

The differences with the data are the residuals:

$$\mathbf{e} = \mathbf{y} - \mathbf{R} \hat{\mathbf{w}} \quad (7.18)$$

The estimated signal and the residuals are given in figure 7.2.f,g respectively. The estimated signal is very close to the noise free data and the residuals very strongly resemble the noise. The differences:

$$\mathbf{s} - \hat{\mathbf{s}} = \mathbf{e} - \mathbf{n} \quad (7.19)$$

are given in figure 7.2.h. The energy of these differences is much lower than the noise power. It follows that the estimation procedure very well separates the signal from the noise. The noise power, if unknown, can therefore also be estimated very well in these circumstances, see also section 7.6 and 7.7. The NE of .011 in the wavelet estimate implies roughly an additional noise power of -20 dB when the estimated wavelet is used for inversion instead of the true wavelet. This error level is therefore sufficiently small in practice.

In order to illustrate the advantages of multi trace wavelet estimation over single trace estimation a single trace result (for trace 5) is given in figure 7.3. Because there is very little information concerning the tail of the wavelet in the data (there are no reflectors before the gate) the estimate shows large errors there. The NE is .19 and is much larger than for multi trace estimation.

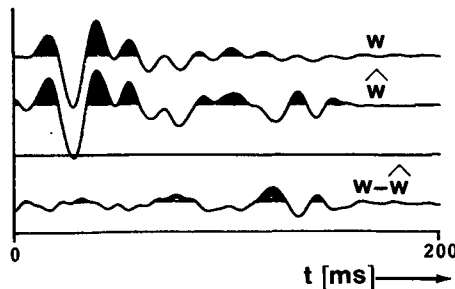


Figure 7.3 Result of wavelet estimation when only one trace is used.

7.4 INCORPORATION OF A PRIORI INFORMATION

When a priori information concerning the wavelet is available it can be incorporated along the lines of Bayesian estimation discussed in chapter 1. This may e.g. be the case in marine seismics where the source wavelet can be measured. Together with knowledge concerning the overburden an a priori wavelet for the target zone could be derived. If this knowledge is insufficient additional parameters concerning the overburden would have to be estimated. A wavelet estimation scheme in which e.g. also the Q-factor of the overburden is estimated has been described by Rosland and Sandvin (1987).

A priori information however will usually consist only of knowledge concerning bandlimitations and the length of the wavelet. With such information the a priori expected values of the wavelet samples will be 0. The a priori information can be used in the parameterization itself, it effects the choice of the sample rate and the number of samples. More detailed knowledge concerning the amplitude spectrum and the decay of the wavelet with time, can be incorporated in the a priori covariance matrix. This will make the choice of the sample rate and the length less critical. It is already mentioned in the previous sections that the absolute scale of the wavelet is usually unknown. The a priori covariance matrix is therefore written in the form:

$$C_w = k_w^2 W, \quad (7.20)$$

where W contains the information mentioned above and k_w^2 is an unknown positive factor.

A similar situation may hold for the noise, about which some information may be available concerning the shape of the amplitude spectrum, but the power may be unknown. The noise covariance can then be written as:

$$C_n = k_n^2 V_n, \quad (7.21)$$

where k_n is unknown. A solution of the inverse problem can now be found by treating k_w and k_n as unknown parameters. The parameter vector is built up according to:

$$x^T = (w^T | k_w | k_n). \quad (7.22)$$

Under Gaussian assumptions and locally uniform a priori pdf's for k_w and k_n the a posteriori pdf is given by:

$$p(x|y) = c \frac{1}{|C_w|} \frac{1}{|C_n|} \exp \left\{ -\frac{1}{2} \left[(y-s)^T C_n^{-1} (y-s) + w^T C_w^{-1} w \right] \right\} \quad (7.23)$$

and, using (7.20) and (7.21):

$$p(x|y) = c \frac{1}{k_w^N |W|} \frac{1}{k_n^M |V_n|} \exp \left\{ -\frac{1}{2} \left[\frac{1}{k_n^2} (y-Rw)^T V_n^{-1} (y-Rw) + \frac{1}{k_w^2} w^T W^{-1} w \right] \right\} , \quad (7.24)$$

where N is the number of parameters and M the number of data samples. Its maximum is found for the parameter values where the derivative of $p(x|y)$ is zero. For k_n it follows:

$$\hat{k}_n^2 = \frac{(y-R\hat{w})^T V_n^{-1} (y-R\hat{w})}{M} , \quad (7.25)$$

where $\hat{w}$ denotes the maximum of $p(x|y)$ for w . Similarly we have for k_w :

$$\hat{k}_w^2 = \frac{\hat{w}^T W^{-1} \hat{w}}{N} . \quad (7.26)$$

From substitution of (7.25) and (7.26) in (7.24) it follows that the wavelet $\hat{w}$ that maximizes $p(x|y)$ is the one that minimizes the function F_1 , given by:

$$2F_1 = \ln \left\{ (y-Rw)^T V_n^{-1} (y-Rw) \right\} + \ln \left\{ w^T W^{-1} w \right\} . \quad (7.27)$$

A similar function as the l_2 norm arises but now the terms are the logarithms of the terms as occurring in (1.91). The main peak of this function will be less sharp. Equation (7.27) does not contain k_w and k_n . With this formula the wavelet can be estimated and the noise power and the wavelet power can then be estimated using (7.25) and (7.26). Similar procedures can be developed when only one of the two factors k_w and k_n are unknown.

Because (7.27) poses a difficult optimization problem it has not been tried yet. Instead the more simple procedures of ridge regression and singular value cutoff method have been tried. They are discussed in the next sections.

7.5 RIDGE REGRESSION

When the a priori and noise covariance matrices are known the maximum of the a posteriori pdf is given by:

$$\hat{w} = (R^T C_n^{-1} R + C_w^{-1})^{-1} R^T C_n^{-1} y . \quad (7.28)$$

Using equations (7.20) and (7.21) this can be rewritten as:

$$\hat{w} = \left(R^T V_n^{-1} R + \frac{k_n^2}{k_w^2} W^{-1} \right)^{-1} R^T V_n^{-1} y . \quad (7.29)$$

It turns out that this estimator depends on the ratio $k=k_n/k_w$. When either k_n or k_w or both are unknown equation (7.29) can simply be tried for different values of k . This procedure

is called ridge regression, see e.g. Hoerl and Kennard (1970a,b) or Marquardt (1970). It is also referred to as damped least squares. The usual formula is given for statistically normalized formulations or for the case where V_n and W equal the identity matrix:

$$\hat{w} = (R^T R + k^2 I)^{-1} R^T y, \quad (7.30)$$

In this form the utilization of the term $k^2 I$ is also called diagonal stabilization. Nonzero values of k stabilize the estimator. The residual energy monotonically increases with increasing k . The choice for k is a subjective one. It can be made by inspection of the so called ridge trace, which is the residual energy for different values of k . The corresponding estimates of course are also studied.

This type of estimation procedure (see also the next section) is discussed in the literature from a classical viewpoint. The covariance matrix of the estimator given is the covariance matrix of the sampling distribution for data variations, but is always shortly referred to as "covariance matrix". For the ridge estimator it is given by:

$$C_{\hat{w}_s} = k_n^2 (R^T V_n^{-1} R + k^2 W^{-1})^{-1} R^T V_n^{-1} R (R^T V_n^{-1} R + k^2 W^{-1})^{-1}. \quad (7.31)$$

This covariance matrix, unlike that of the a posteriori pdf in Bayesian inversion, can *not* be interpreted as an uncertainty measure for the parameters, which is directly clear from (7.31) which can be made arbitrary small by choosing k very large. The variance of the reweighted noise k_n has to be estimated when it is unknown. In classical statistics this is usually done with an unbiased estimator:

$$\hat{k}_n^2 = \frac{1}{M-N} (y - R \hat{w})^T V_n^{-1} (y - R \hat{w}), \quad (7.32)$$

where $\hat{w}$ is the unstabilized estimate, from (7.30) with $k=0$.

An example of wavelet estimation with ridge regression is given using the same true reflectivity and wavelet as in the previous example, see figures 7.1 and 7.2. The signal to noise ratio is taken smaller, only 3 dB. The length of the data gate is reduced to 220 ms. In

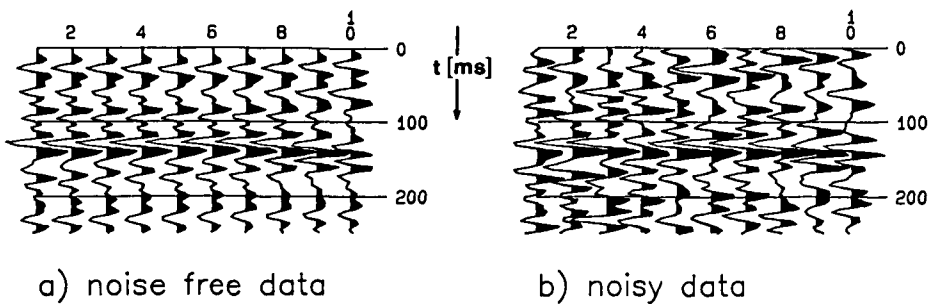


Figure 7.4 Noise free data and the noisy data set that is used for the wavelet estimation examples with stabilization.

figure 7.4a,b the noise free data and the noisy data used in the experiment are shown. The reflectivity used in estimation is again the true reflectivity but shifted 12 ms backwards in time. The length of the wavelet is chosen to be 200 ms, meaning that there is very little information concerning the tail in the data. A set of estimated wavelets for increasing stabilization factor k is shown in figure 7.5. For the first wavelet the stabilization is zero.

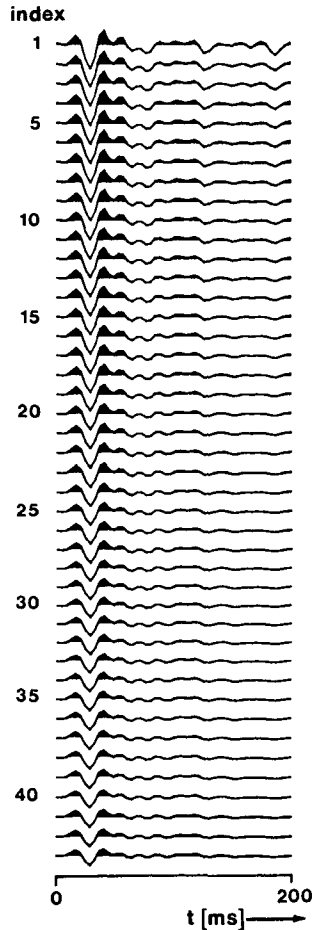
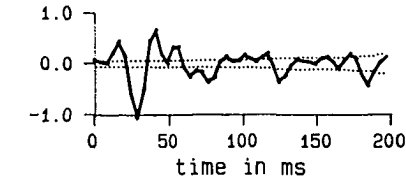
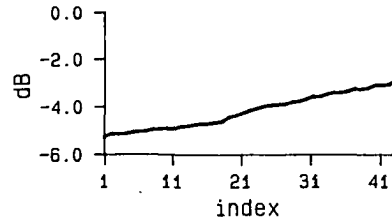


Figure 7.5 Panel of estimated wavelets using ridge regression. The stabilization factor increases with increasing index.

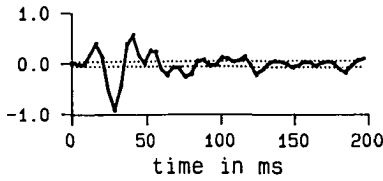
Wavelets 1, 7 and 13 are also plotted in figure 7.6.a,b,c, together with an interval of plus and minus one standard deviation around zero, derived from (7.31) and (7.32). This allows a comparison of the estimated amplitudes with the standard deviation. It is clear that without stabilization there is a slight instability in the tail of the wavelet. For increasing



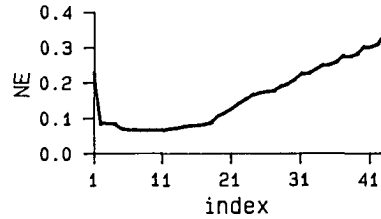
a) wavelet 1



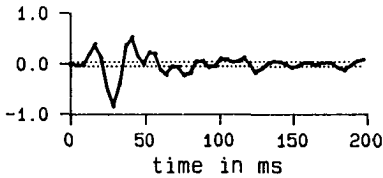
d) residual energy



b) wavelet 7



e) normalised error



c) wavelet 13

Figure 7.6 Results for ridge regression. = standard deviation (in a, b, c).

values of k the amplitude of the wavelet as well as the standard deviations decrease. Note that also the main lobes decrease in amplitude. In figure 7.6.d the ridge trace, the residual energy as a function of the stabilization is plotted relative to the data energy, in dB. It may be difficult to select an estimate using this information only, because the energy simply gradually increases. Fortunately this choice is not very critical, as can be seen from figure 7.6.e where the normalized error is given as a function of the stabilization level. With only a slight stabilization the error is already drastically reduced and remains at approximately the same level until it significantly increases again, when the main lobes of the estimates are getting too low in amplitude. The best possible estimate from this set is shown resampled in figure 7.7 in comparison with the true wavelet. The normalized error is .066.

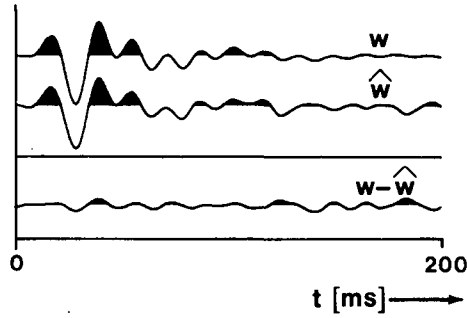


Figure 7.7 Best result for the method of ridge regression.

7.6 STABILIZATION USING THE SVD

Another method of stabilization that is well known in geophysical literature uses the singular value decomposition of the statistically normalized forward matrix:

$$\tilde{\mathbf{R}} = \mathbf{C}_n^{-1/2} \mathbf{R} \mathbf{C}_w^{1/2} . \quad (7.33)$$

A very lucid introduction to the SVD and its properties is given by Lanczos (1961). In this section only the case $M \geq N$ is considered. The SVD of $\mathbf{R}$ can be written as:

$$\tilde{\mathbf{R}} = \mathbf{U} \mathbf{S} \mathbf{V}^T , \quad (7.34)$$

where $\mathbf{V}$ and $\mathbf{U}$ are $(N \times N)$ and $(M \times M)$ orthonormal matrices respectively and $\mathbf{S}$ is a $(M \times N)$ matrix, containing the singular values in decreasing order in the matrix elements S_{ii} . Textbooks that discuss the relations between the SVD and least squares problems and that include software code for its computation are Lawson and Hanson (1974) and Menke (1984). An inverse operator or estimator can be constructed by using only the N_c largest singular values:

$$\hat{\mathbf{w}} = \mathbf{V}_c \mathbf{S}_c^{-1} \mathbf{U}_c^T \mathbf{y} , \quad (7.35)$$

where $\mathbf{S}_c$ is a square diagonal matrix containing only the N_c largest singular values and $\mathbf{V}_c$ and $\mathbf{U}_c$ contain the corresponding columns of $\mathbf{V}$ and $\mathbf{U}$. The properties of this estimator have been extensively discussed in the literature, see e.g. Wiggins (1972), Jackson (1972), Matsu'ura and Hirata (1982) and Van Riel (1982). The covariance matrix of the sampling distribution is given by:

$$\mathbf{C}_{\hat{\mathbf{w}}} = \mathbf{V}_c \mathbf{S}_c^{-2} \mathbf{V}_c^T . \quad (7.36)$$

In the software built, the actual computations use the eigenvalue decomposition of the matrix $\mathbf{R}^T \mathbf{R}$:

$$\mathbf{R}^T \mathbf{R} = \mathbf{V} \mathbf{S}^2 \mathbf{V}^T, \quad (7.37)$$

and the normal least squares formulation. Using the cutoff on the eigenvalue spectrum can easily be shown to give equal results:

$$\begin{aligned} \hat{\mathbf{w}} &= (\mathbf{V}_c \mathbf{S}_c^2 \mathbf{V}_c^T)^{-1} \mathbf{V} \mathbf{S} \mathbf{U}^T \mathbf{y} \\ &= \mathbf{V}_c \mathbf{S}_c^{-2} \mathbf{V}_c^T \mathbf{V} \mathbf{S} \mathbf{U}^T \mathbf{y} \\ &= \mathbf{V}_c \mathbf{S}_c^{-1} \mathbf{U}_c^T \mathbf{y}, \end{aligned} \quad (7.38)$$

in which a number of properties has been used that can be found in the references mentioned above.

Matsu'ura and Hirata (1982) and Van Riel (1982) derive the optimal cutoff point, which turns out to be the point where the singular values become smaller than one. This renders the smallest error in the estimator. Note that this criterion holds only for the statistically normalized system. In both references as well as in Marquardt (1970) the SVD estimator is compared to ridge regression.

Because in our wavelet estimation problem the factors k_n and k_w in the covariance matrices are unknown, the absolute scale of the singular value spectrum is unknown and the above mentioned criterion for the cutoff point cannot be applied. Instead, analogously to the procedure of ridge regression, different choices of the cutoff point N_c can be tried and a subjective choice has to be made using the different wavelet estimates and the residual energy as a function of the cutoff point.

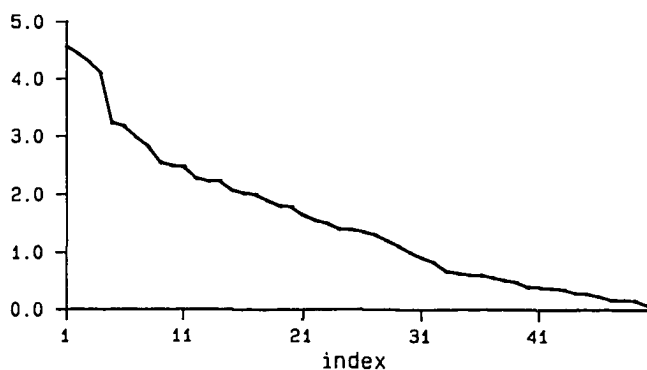


Figure 7.8 Eigenvalue spectrum of $\mathbf{R}^T \mathbf{R}$.

As an example the same data set as in the previous section is used. The eigenvalue spectrum of $\mathbf{R}^T\mathbf{R}$ (the square of the singular value spectrum!) is given in figure 7.8. Note that the absolute scale is unimportant. An estimate of the so called condition number of $\mathbf{R}^T\mathbf{R}$ is given by $\lambda_{\max}/\lambda_{\min}$ and equals approximately 50 for this example. The panel of wavelet estimates is shown in figure 7.9, where the increasing index corresponds with decreasing N_c . For wavelet 1, $N_c = 50$, no cutoff is applied and this wavelet estimate equals wavelet 1 in figure 7.5 and 7.6 of the ridge regression example. It can be seen that SVD cutoff has a different effect than diagonal stabilization. The main lobes only start decreasing for low cutoff values. This can be verified in more detail in figure 7.10,a,b,c, where wavelets 8, 15 and 22 are displayed with the standard deviations plotted similarly as

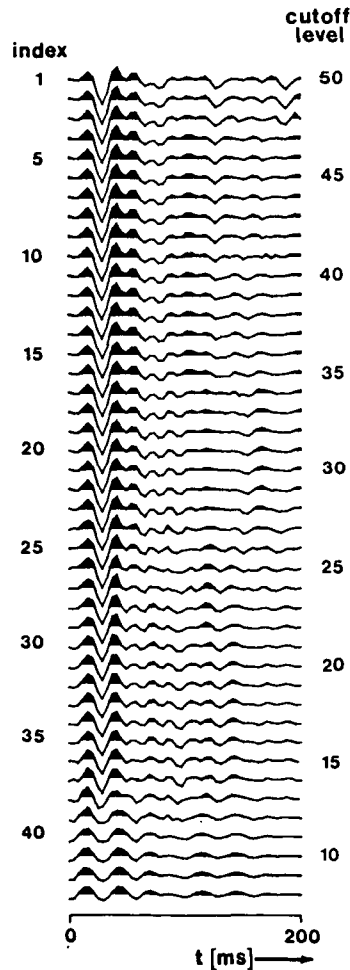


Figure 7.9 Panel of estimated wavelets for the SVD cutoff method.

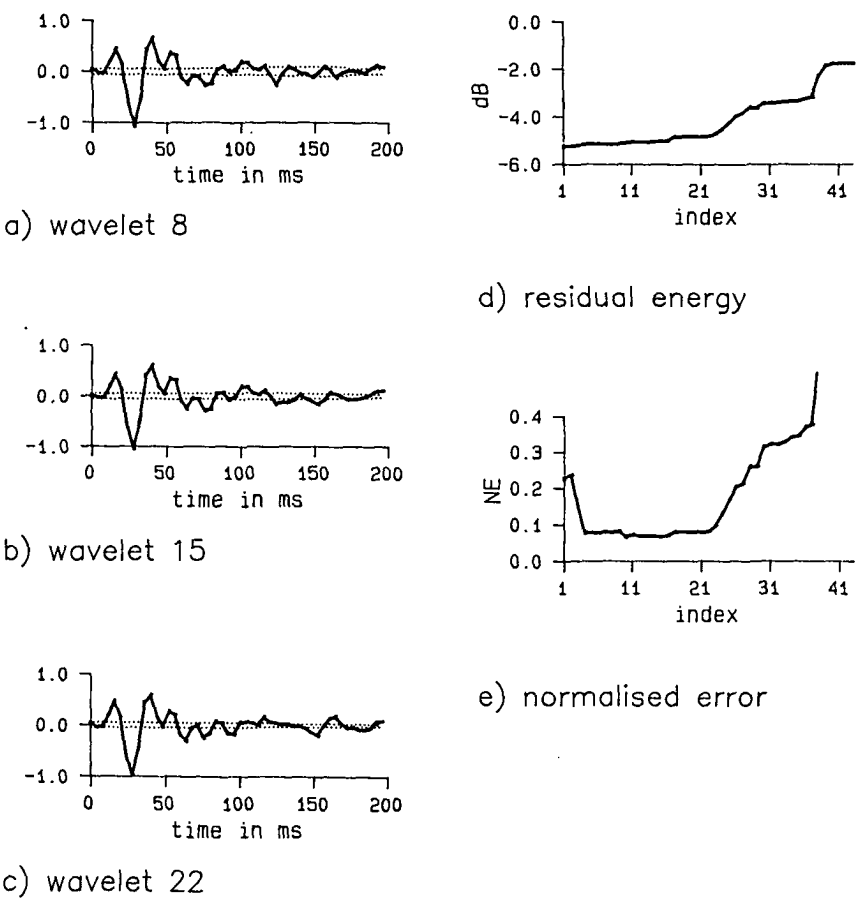


Figure 7.10 Results for SVD cutoff method. = standard deviation (in a, b, c).

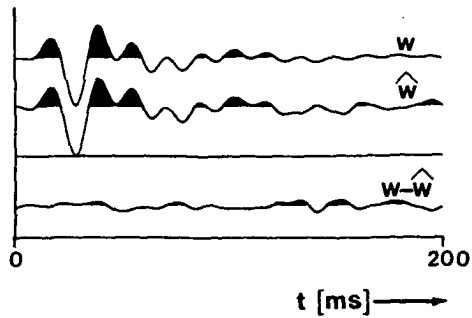


Figure 7.11 Best result for SVD cutoff method.

in figure 7.6. The negative main lobe remains approximately at amplitude -1 . In figure 7.10d the residual energy is plotted as a function of the index (decreasing N_c). Again the residual energy increases slightly and it is difficult to select a wavelet estimate from this picture alone. Fortunately the choice is not very critical as can be seen from figure 7.10e, where the normalized errors are given for each wavelet estimate. The optimal choice would be wavelet 15 ($N_c = 36$) which is displayed in comparison with the true wavelet in figure 7.11. Its NE is .068, slightly higher than for diagonal stabilization.

7.7 PARAMETER SELECTION

The stabilization procedures as discussed in the previous sections are especially beneficial in the example discussed because the simple least squares solution is slightly unstable in the tail of the wavelet. This is caused a.o. by the large number of parameters chosen for the wavelet estimate. There is very little information concerning the tail of the wavelet in the data. A shorter wavelet gives better results as is illustrated in figure 7.12,

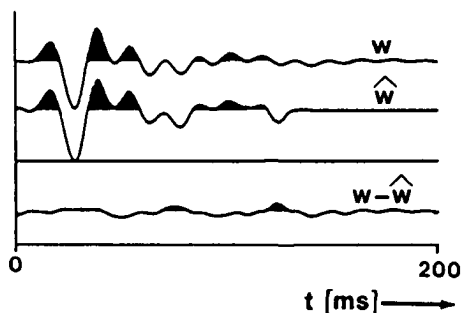


Figure 7.12 Result of unstabilized least squares estimation with a wavelet length of 140 ms.

where the unstabilized least squares estimate for the same data set is given, but now with a wavelet length of 140 ms. The normalized error is .054, lower than the best results of the previous sections. Another subjective procedure for “stabilization” would be to try an unstabilized least squares estimator for different lengths of the wavelet. Inspection of the residual energy and the wavelet estimates should supply the information for the final choice. This type of procedure however can be performed automatically and more efficiently by parameter selection techniques that are well known in applied statistics. In this section such a technique is discussed and illustrated.

The subset selection techniques mentioned above are especially well known for linear least squares estimation, which in applied statistics is often referred to as linear regression analysis. In the literature on this subject parameters are often referred to as variables or regressors and the problem of parameter selection as subset selection. An extensive

discussion of subset selection is beyond the scope of this thesis. The reader is referred to the excellent overview of Hocking (1976). For a more recent further discussion, see Miller (1984). A number of criteria for the selection of the optimal subset have been proposed, see Hocking (1976). All of them are closely related to the residual energy. The choice of the criterion depends on the application of the regression problem. In the wavelet estimation problem under discussion this is mainly prediction. The wavelet estimate is used in the forward model for the inversion of the acoustic impedance profile.

The standard reduced model (including noise) used in inversion can be given as:

$$\mathbf{s} = \mathbf{R}\mathbf{w} + \mathbf{n} \quad , \quad (7.39)$$

where the acoustic impedances and the lagtimes are contained in the matrix $\mathbf{R}$. Like in previous sections and in most literature on regression analysis it is assumed that the covariance matrix of the noise is known except for a constant factor. In the sequel of this section the reweighted system will be considered in which the covariance matrix can be written as:

$$\mathbf{C}_n = \sigma_n^2 \mathbf{I} \quad , \quad (7.40)$$

with σ_n the standard deviation of the independently identically distributed noise samples n_i . When only a wavelet estimate $\hat{\mathbf{w}}$ instead of the true wavelet $\mathbf{w}$ is available for inversion the standard reduced model can be rewritten as:

$$\mathbf{s} = \mathbf{R}\hat{\mathbf{w}} + (\mathbf{R}\mathbf{w} - \mathbf{R}\hat{\mathbf{w}}) + \mathbf{n} \quad . \quad (7.41)$$

The first term on the right hand side is used as the forward model and therefore the bracketed term can be interpreted as an additional noise term, due to the fact that the estimated instead of the true wavelet is used. The energy of this additional noise term plays an important role in the subset selection scheme discussed below.

Let N denote the maximum number of parameters. When only a subset of p parameters is used for estimation, the matrix $\mathbf{R}$ can be partitioned according to:

$$\mathbf{R} = (\mathbf{R}_p | \mathbf{R}_r) \quad , \quad (7.42)$$

where $\mathbf{R}_p$ contains the columns of the parameters that are included in the subset. The wavelet $\mathbf{w}$ can be partitioned accordingly, the wavelet parameters included in the subset are gathered in the vector $\mathbf{w}_p$. The other parameters are in the vector $\mathbf{w}_r$ and their estimated values are set to zero. The least squares estimator for the subset of $\mathbf{w}_p$ is given by:

$$\hat{\mathbf{w}}_p = (\mathbf{R}_p^T \mathbf{R}_p)^{-1} \mathbf{R}_p^T \mathbf{y} \quad . \quad (7.43)$$

The corresponding residual energy, or residual sum of squares, is denoted by RSS_p and is given by:

$$RSS_p = (y - R_p \hat{w}_p)^T (y - R_p \hat{w}_p) . \quad (7.44)$$

The energy of the additional noise term in (7.41) for $\hat{w}_p$, scaled on the variance σ_n^2 of the noise samples is denoted by J_p and is given by:

$$\begin{aligned} J_p &= \frac{(Rw - R\hat{w}_p)^T (Rw - R\hat{w}_p)}{\sigma_n^2} \\ &= \frac{(w - \hat{w}_p)^T R^T R (w - \hat{w}_p)}{\sigma_n^2} . \end{aligned} \quad (7.45)$$

The subset should be chosen such that J_p is as low as possible. In the part of the seismic section where the wavelet estimate is used for inversion the reflectivity will not exactly equal the reflectivity as used in the part where the wavelet is estimated, but it will usually be close to it. For white reflectivity, $R^T R = I$ and J_p is then a scaled version of the normalized error used as evaluation measure so far. The normalized error is also used in this section for the evaluation of results. J_p can of course not be computed in practice because the true wavelet w and the noise variance σ_n^2 are unknown. Mallows (1973) introduced a criterion called C_p that *can* be computed in a practical situation and is given by:

$$C_p = \frac{RSS_p}{s^2} - M + 2p , \quad (7.46)$$

where s^2 is an estimate of σ_n^2 , computed from the residual energy when all N parameters are used:

$$s^2 = \frac{RSS_N}{M-N} . \quad (7.47)$$

The expectation of J_p is given by Mallows (1973) and Broersen (1984):

$$EJ_p = p + \frac{w_r^T B_r^{-1} w_r}{\sigma_n^2} , \quad (7.48)$$

where B_r is the part of $(R^T R)^{-1}$ belonging to w_r , the part of w that is not in the subset. Under the assumption $s^2 = \sigma_n^2$ it can be shown that (Mallows, 1973):

$$EC_p = EJ_p , \quad (7.49)$$

which is the actual reason for using C_p as criterion. It is hoped that by finding the subset with the lowest C_p also the subset with the lowest J_p has been found. In this respect the

following properties, given by Broersen (1984) are important. The variance of J_p is given by:

$$\text{var } J_p = 2p \quad (7.50)$$

The variance of C_p is given approximately by:

$$\text{var } C_p \approx 2r + \frac{4\mathbf{w}_r^T \mathbf{B}_r^{-1} \mathbf{w}_r}{\sigma_n^2} \quad (7.51)$$

and the covariance between C_p and J_p :

$$\text{cov}(C_p, J_p) \approx 0 \quad (7.52)$$

This latter property implies that the subset with the smallest C_p will certainly not always be the subset with the smallest J_p . Important in this respect is the variance of the difference between C_p and J_p :

$$\text{var}(C_p - J_p) \approx 2N + \frac{4\mathbf{w}_r^T \mathbf{B}_r^{-1} \mathbf{w}_r}{\sigma_n^2} \quad (7.53)$$

From this equation it follows that the standard deviation of $C_p - J_p$ equals at least $\sqrt{2N}$. With $N=50$ in the example given below this standard deviation equals at least 10. It is therefore not sensible to assign significance to quantities of a few units on the C_p scale, see also Mallows (1973) and Broersen (1984).

Mallows (1973) advocates the utilization of plots of C_p vs. p as a help in the analysis of the data and discourages the uncritical selection of the subset with smallest C_p as the best subset. He states that "using the 'minimum C_p ' rule to select a subset of terms for least squares fitting cannot be recommended universally", and "the case where the 'minimum C_p ' rule gives bad results are exactly those where a large number of subsets are close competitors for the honour". In the application of wavelet estimation in an iterative scheme for impedance inversion and wavelet estimation (see chapter 8) however it is desirable to have an automatic scheme. The subset with the smallest C_p then at least seems to be a practical possibility.

The search strategy for the subset with lowest C_p is of great practical importance because of the computational effort involved. A straightforward evaluation of all possibilities for the case of 50 parameters is certainly beyond reach. By very efficiently using the mathematical properties of the problem orders of computation time however can be gained. Furnival and Wilson (1974) describe an algorithm that efficiently inspects all possible subsets, i.e. computes the residual energy or a related property. They give no formula for the number of floating point operations as a function of the maximum number

of parameters in the set. Extrapolating their table of operations, that runs up to a number of 35, to a number of 50 parameters yields an approximate number of 10^9 operations, which may very well be feasible on a minicomputer with an attached array processor or for "mini supercomputers" with integrated vector architecture in the CPU. Miller (1984) further discusses a number of computational algorithms.

In this section the efficient stepwise directed search algorithm described by Broersen (1986) is considered. Note that this algorithm does *not* compute all possibilities and may fail to find the subset with the lowest C_p . Broersen (1986) however discusses its excellent properties and performances on several tests. The method requires approximately 3 times more computation time than a single least squares estimate with all parameters included. It inspects approximately $3N^2$ subsets by efficiently using a formula with which the residual sum of squares can be computed without actually computing the parameter estimates. The procedure starts with forward selection, in which the parameters are inserted in the subset one at a time, in the order of greatest reduction of residual energy. When all parameters are in the subset s^2 is computed using formula (7.47) as well as the C_p value for all subsets encountered in forward selection. At this point a stepwise backward elimination procedure is executed, where the parameters which give the greatest reduction in C_p are eliminated from the subset. Parameters that give a greater reduction in the residual energy than $2s^2$ can be shown to always render a lower C_p when removed from the subset. This corresponds to the situation where the square of the estimated value of the parameter is lower than twice its estimated variance. These parameters are defined as "weak parameters" by Broersen. Other parameters are only temporarily removed from the subset and may reenter the subset at a later stage, see the example below. When all parameters are removed the subset with the lowest C_p is selected. For further details the reader is referred to Broersen (1986).

An example of this scheme is given for the same data set as in previous sections. The maximum length wavelet contains 50 samples. In figure 7.13 a part of the output of the computer program is given, which shows the compositions of the subsets in the stepwise backward elimination part that are actually computed. Note that of $N-1$ times as many subsets the C_p values are computed, but not displayed. The corresponding wavelet estimates are shown in figure 7.14. The C_p values as well as the number of parameters are also shown graphically in figure 7.15a. The C_p value for $p=N$ equals N , see the definitions (7.46) and (7.47). In the first part C_p decreases with a value of 2 each time a parameter is removed from the subset, because the weakest parameters hardly give rise to an increase in the residual energy, see definition (7.46). The residual energy, relative to the data and expressed in dB, is plotted in figure 7.15b. When after the decrease C_p increases again some parameters are temporarily removed from the subset and reenter the subset at index 31 where a weak parameter is permanently removed. When too many parameters are

INDEX	SAMPLE STATUS ("*" = IN, "-" = OUT)	Cp	#pars
1	*****	50.00	50
2	*****	48.00	49
3	*****	46.01	48
4	*-*****	44.03	47
5	*-*****	42.10	46
6	*-*****	40.18	45
7	*-*****	38.31	44
8	*-*****	36.49	43
9	*-*****	34.68	42
10	*-*****	32.84	41
11	*-*****	31.02	40
12	*-*****	29.23	39
13	*-*****	27.46	38
14	*-*****	25.74	37
15	*-*****	24.07	36
16	*-*****	22.64	35
17	*-*****	21.14	34
18	*-*****	19.76	33
19	*-*****	18.46	32
20	*-*****	17.12	31
21	*-*****	15.91	30
22	-----	15.23	29
23	-----	14.91	28
24	-----	14.02	27
25	-----	13.72	26
26	-----	13.34	25
27	-----	13.02	24
28	-----	13.29	23
29	-----	13.79	22
30	-----	14.44	21
31	-----	15.70	23
32	-----	14.49	22
33	-----	14.04	21
34	-----	14.26	20
35	-----	15.61	19
36	-----	17.36	18
37	-----	19.21	17
38	-----	21.92	16
39	-----	28.01	15
40	-----	33.72	14
41	-----	44.96	13
42	-----	59.25	12
43	-----	73.51	11
44	-----	90.80	10
45	-----	110.41	9
46	-----	130.99	8
47	-----	159.09	7
48	-----	193.22	6
49	-----	256.11	5
50	-----	330.77	4
51	-----	436.13	3
52	-----	557.21	2
53	-----	727.91	1
54	-----	1142.55	0

FINAL SELECTED SET:

27	-----	13.02	24
----	-------	-------	----

TIME →

Figure 7.13 Sample selection by stepwise backward elimination, part of computer output.
 ("*" = sample in set of wavelet samples, "-" sample out set of wavelet samples.

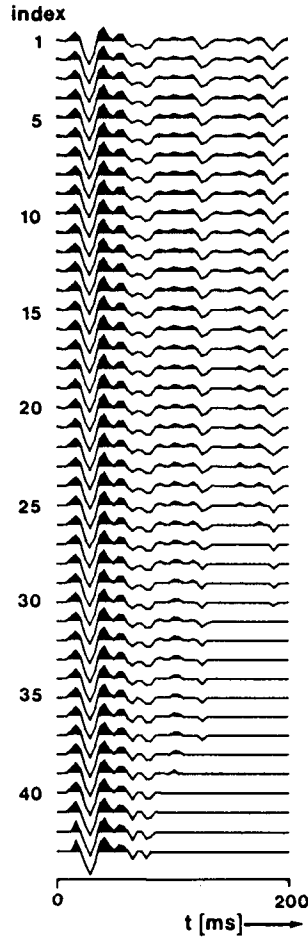


Figure 7.14 Estimated wavelets in stepwise backward elimination, corresponding with figure 7.13.

removed the residual energy starts to increase significantly and therefore also C_p increases sharply. The normalized error of the estimates, given in figure 7.15c decreases gradually and is lowest for indices between 30 and 38. There is a fairly broad range of wavelet estimates with low error. When more samples are removed the error increases sharply, but this is outside the range of the plot. Similar shapes in the error curve has also been found by Walden and White (1984) who varied the spectral smoothing factor in the frequency domain, corresponding with the length of the wavelet in the time domain. Similar plots also occurred in previous sections for ridge regression and SVD cutoff. The subset with lowest C_p has index 27, $C_p = 13.02$. The selected samples are shown in figure 7.15.d in comparison with the true wavelet. It is clear that the scheme succeeds in removing the

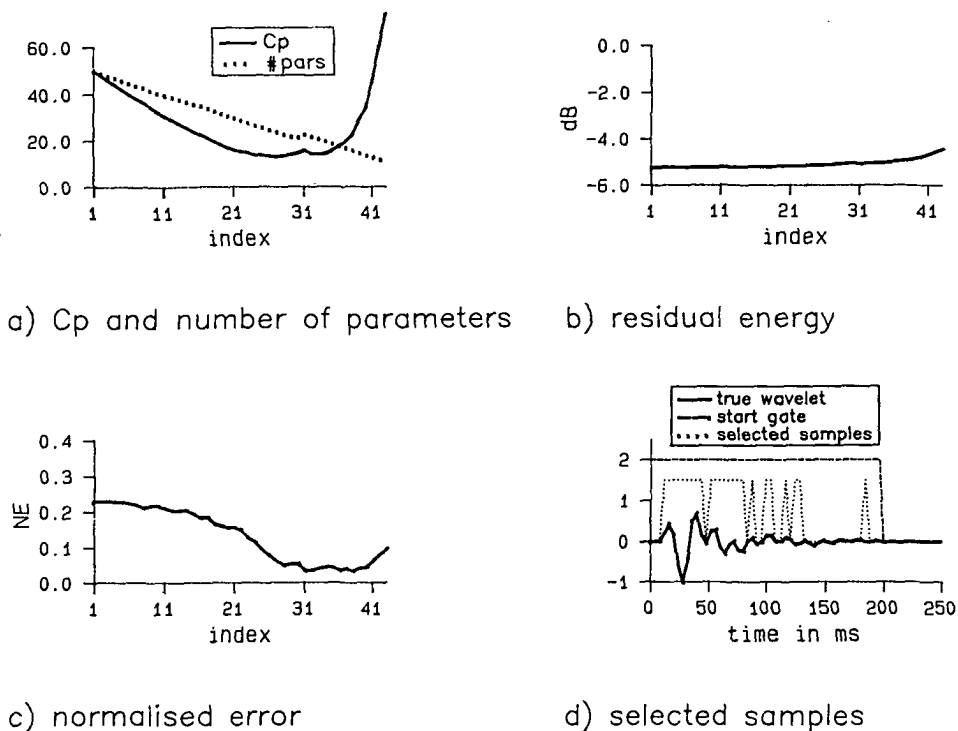


Figure 7.15 Results for the sample selection method. The part of stepwise backward elimination is shown only.

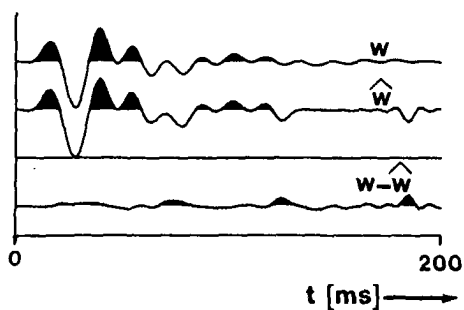


Figure 7.16 Final selected wavelet in the sample selection method for a start gate of 200 ms.

samples with low amplitude. The resampled wavelet estimate with lowest C_p is shown in figure 7.16 in comparison with the true wavelet. The normalized error is .063, slightly better than the best estimates of ridge regression and the SVD cutoff method. Note that this estimate is found automatically while for the others a subjective (and therefore possibly

erroneous) choice is necessary. When user interaction is feasible it is clear that the wavelet with the lowest C_p will not be selected by the user. First of all, from the numbers for C_p in figure 7.14 and the plot in figure 7.15 it is clear that there are a large number of wavelets with C_p close to 13, which with regard to equations (7.52) and (7.53) are all good candidates. Besides those there are also the subsets, with C_p value close to 13 that are not displayed here and of which only the C_p value is computed. All samples but one are removed from the last 60 ms of the wavelet. A user will in practice be inclined to also discard this parameter and rather select one of the wavelets with indices 31 to 35, which would indeed yield a lower error. Such a situation may also be automatically detected. Therefore the scheme is rerun with a start gate of the wavelet of 160 ms. The selected samples are shown in figure 7.17a. The wavelet with lowest C_p of this run is displayed in figure 7.17b in comparison with the true wavelet. Its normalized error is .04, significantly better than the best results of ridge regression and SVD cutoff.

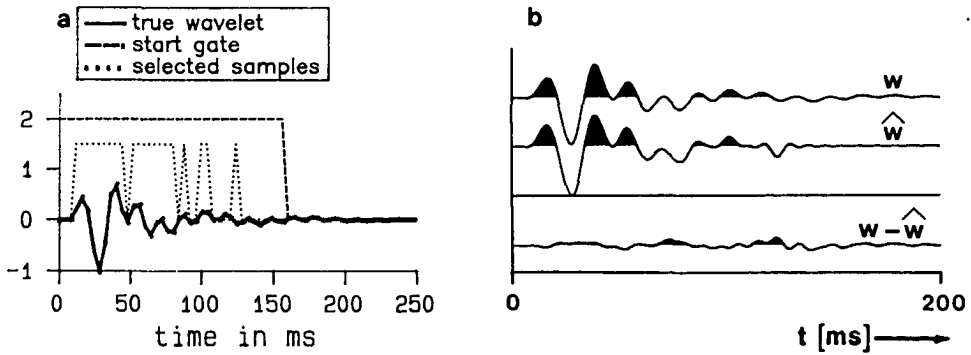


Figure 7.17 Results for the sample selection method for a start gate of 160 ms.

a) selected samples; b) wavelet estimate.

7.8 REFLECTIVITY ERRORS

In all examples presented so far the reflectivity used was correct. The “random” noise added to the data was the only source of error. In this section an example with the same data is given but now incorrect reflectivity is used. It corresponds to the situation where the correct reflectivity for one trace (in practice derived e.g. from a well log) is used for all traces while the true reflectivity contains lateral changes. The true reflectivity and the reflectivity used in estimation (shifted 12 ms backwards in time) are shown in figure 7.18a,b. Trace 5 has the correct reflectivity. The utilization of incorrect reflectivity can be seen as corresponding with an additional noise term. This noise term is displayed in figure 7.18.d and can be compared with the random noise on the data, given in figure 7.18c. The latter has an energy of -3dB, relative to the noise free data. The energy of the additional noise term is -4 dB. The errors for trace 5 are of course zero. Unlike the random noise the

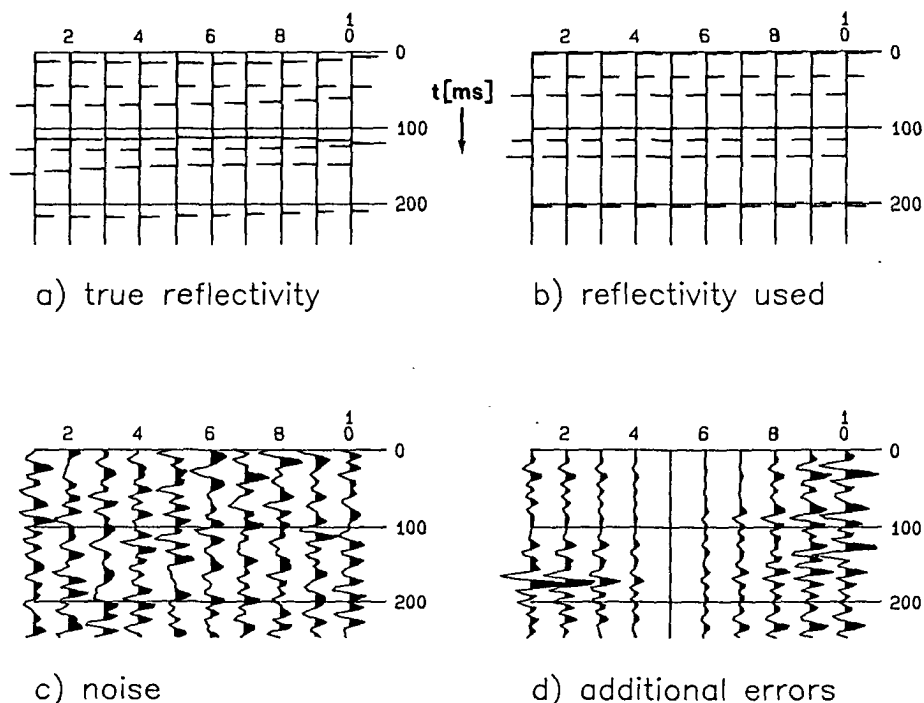


Figure 7.18 Set up of an experiment of wavelet estimation with reflectivity errors.

a) selected samples; b) wavelet estimate.

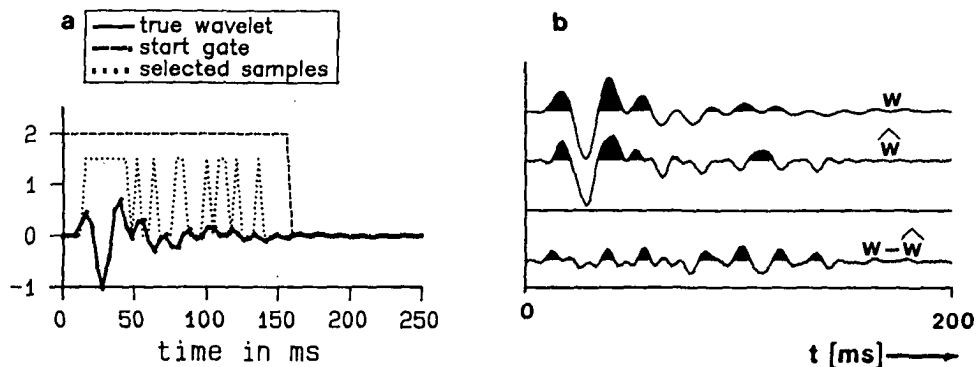


Figure 7.19 Results for the experiment with reflectivity errors.

additional noise is strongly laterally correlated because of the gradual lateral changes in the true reflectivity. This causes large errors in the wavelet estimate as can be seen from the estimated result in figure 7.19b, where it is displayed in comparison with the true wavelet. The normalized error is .22, which is significantly higher than in the examples where only the random noise was present. The laterally uncorrelated noise averages out more strongly

in the estimation procedure. In figure 7.19a the selected samples are given. Although the most significant samples of the wavelet are in the final set the selection shows stronger inconsistencies with the true wavelet than in the previous examples. It can be concluded that incorrect reflectivity of this nature has strong effects on wavelet estimation.

7.9 CONCLUSIONS

Provided that the correct reflectivity is available multi trace wavelet estimation gives very accurate estimates in "normal" circumstances, see the example of section 7.3. It is much more accurate than single trace estimation. The multi trace wavelet estimation problem is usually well posed and will not be very critical in what could be called the forward model design, i.e. the choices of the number and the length of the data traces, the length and sample rate of the wavelet. It is the experience of the author that extra stabilization gives only marginal improvements in this situation, if it gives improvements at all.

For more difficult problems in which a.o. the length of the wavelet is not accurately known a priori (e.g the example discussed in sections 7.5, 7.6 and 7.7) the straightforward application of unstabilized least squares can give poor results. The stabilization procedures of ridge regression and SVD cutoff can improve the results but have the disadvantage of needing a subjective and interactive choice in practice. A parameter selection scheme with Mallows' C_p as criterion and Broersen's efficient stepwise directed search procedure usually gives better results. For optimal results some additional intelligence should be built into the scheme. The procedure is automatic which is a great advantage for the iterative scheme discussed in chapter 8.

8

INTEGRAL ESTIMATION OF ACOUSTIC IMPEDANCE AND THE WAVELET

8.1 INTRODUCTION

If the reflectivity, derived from a well log approximates the reflectivity of a block of traces around the well good enough, the multi trace wavelet estimation scheme described in chapter 7 can be used to estimate the wavelet which in its turn is used to estimate the acoustic impedance profile for a larger part of the section. If the well log derived reflectivity is not good enough, e.g. due to well log errors or lateral variations of the geology, the wavelet estimate may be poor. The example of section 7.8 illustrates this. It is expected that this will be the case for the majority of practical data sets. A solution, described in this chapter is to integrally estimate the wavelet and the acoustic impedance.

The most straightforward way of integral estimation is to combine wavelet and acoustic impedance parameters in one parameter set to be estimated and to use a consistent Bayesian estimation scheme. This approach is described in section 8.2. A sample selection procedure (other than by trial and error) however has not yet been worked out in detail for this approach. An iterative scheme in which sample selection as well as any other stabilization technique for wavelet estimation *can* easily be utilized is described in section 8.3. The disadvantage of this scheme is that it can not strictly be shown to converge to a meaningful result. One can only check in practice whether the scheme converges and if it does, assess the results.

Results on synthetic data are given in section 8.4. The inversion scheme has also been tried on real data. The results are given in section 8.5.

8.2 A CONSISTENT BAYESIAN SCHEME

formulation of the problem

For integral estimation the parameter vector can be built up of three vectors:

$$\mathbf{x}^T = (\mathbf{z}^T / \boldsymbol{\tau}^T / \mathbf{w}^T) , \quad (8.1)$$

where $\mathbf{z}$ denotes the vector containing all parameters concerning acoustic impedances, $\boldsymbol{\tau}$ denotes the vector containing all lagtime parameters and $\mathbf{w}$ contains the wavelet parameters. In principle we are not interested in the wavelet so an ideal solution for the inverse problem when the wavelet is unknown is to treat the wavelet parameters as nuisance parameters and integrate them out. The solution is then the marginal a posteriori pdf for $\mathbf{z}$ and $\boldsymbol{\tau}$:

$$p(\mathbf{z}, \boldsymbol{\tau} | \mathbf{y}) = \int p(\mathbf{z}, \boldsymbol{\tau}, \mathbf{w} | \mathbf{y}) d\mathbf{w} . \quad (8.2)$$

This approach has not been tried yet due to its mathematical complexity. Furthermore it may nevertheless be practical to have an estimate of the wavelet so that it can be used in other parts of the section than where it is estimated. In the sequel of this chapter therefore the situation is considered where all three components of (8.1) are estimated. As a point estimator the maximum of the a posteriori pdf is again used. Under Gaussian assumptions the least squares estimator (1.91) results. In principle a sample selection scheme for the wavelet could also be worked out for this estimation problem, see e.g. Rosenkrantz (1977). In the approach described in this section no parameter selection is used, so that the wavelet length has to be chosen a priori.

physical aspects

From the mathematical form of the forward model it is clear beforehand that there are a number of directions through the parameter space about which no information can be derived from the data:

- 1) A linear combination of parameters corresponding with a constant factor in the acoustic impedances, see also chapters 5 and 6.
- 2) A direction corresponding with an exchange between the amplitude of the wavelet and the reflection coefficients, which follow directly from the acoustic impedances.
- 3) A direction corresponding with an exchange in time origin of the wavelet and the lagtimes. The data may provide some information concerning this direction due to truncation effects at the edges of the data gate but it will usually be very little. When there is no a priori information on the wavelet the a priori information concerning the acoustic impedances will yield the answer along the directions (1) and (2). For the third

direction it will be the a priori information on the lagtimes, again assuming that there is no a priori information on the wavelet. Note that this situation occurs because the inversion is performed on poststack data using the convolutional forward model discussed in chapter 5. On prestack data, with other forward models the situation is different and more knowledge can be derived from the data.

optimization

The optimization problem to be solved is again a nonlinear least squares problem. Due to the special form of the problem however two optimization procedures are shortly discussed:

1) an integral optimization scheme.

In this procedure the problem is treated in a truly integral way. The Jacobian matrix of the problem can simply be partitioned in the Jacobian matrix for the acoustic impedances and the lagtimes as discussed in chapter 5 and 6, and the forward matrix for the wavelet estimation problem as discussed in chapter 7. Second derivatives for the wavelet samples are zero.

2) mixed linear-nonlinear optimization.

In this approach line searches are alternatingly performed for the wavelet parameters and the impedance/lagtime parameters, see figure 8.1. The line search along the wavelet parameters is a linear least squares problem and can be performed in one step. The line search for the impedance/lagtime parameters is exactly the nonlinear optimization problem as discussed in chapter 6. The sequence of points in parameter space will

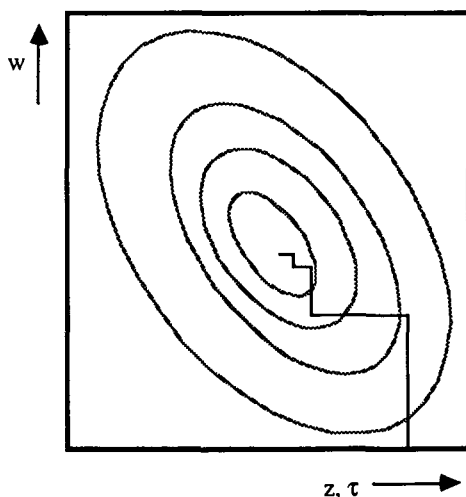


Figure 8.1 Optimization for wavelet parameters w , impedance parameters z and lagtime parameters τ by alternating line searches for w and the combination z, τ .

follow a zigzag course, like steepest descent methods and can therefore may show slow convergence. This procedure is nevertheless used in the experiments discussed below, because it allows a quick implementation through a combination of the software related to the material discussed in chapters 6 and 7. In the experiments discussed below the scheme always converged in less than 20 iterations.

There are a number of practical issues involved in the optimization problem. When the *a priori* wavelet has zero amplitude (as will often be the case in practice) and is used as the start wavelet the derivatives for the acoustic impedance and lagtime parameters are zero. The first line search in the integral optimization scheme will then be along the wavelet parameters only. The knowledge of this fact may be utilized to increase the efficiency of the scheme. When furthermore the procedure starts with a parameter model that has equal acoustic impedance values for all layers, the derivatives for the wavelet parameters are also zero. The start model as a whole then lies exactly on a saddle point! A sophisticated optimization scheme will succeed in moving away from it, but this may very well be in the wrong direction, leading to a local minimum. The second procedure can not work in this situation, because the reflectivity matrix, with which the first wavelet has to be determined is zero. When a source wavelet is available this may be used as a start wavelet (apart from using it in *a priori* knowledge) and the above mentioned problems do not arise.

uncertainty analysis

An uncertainty analysis can be performed along the lines discussed in chapter 3. A problem with respect to the scaling of the Hessian matrix may occur when no *a priori* information is available on the wavelet. The scaling of the reflectivity matrix (which is dimensionless) can then be chosen arbitrarily by the user. A meaningful choice for scaling may be to derive some “*a priori*” uncertainty scale from the noise power or covariance matrix and the reflectivity. A full uncertainty analysis has however not been implemented yet. An approximation for the *a posteriori* covariance matrix of the impedances and the lagtimes is obtained instead by inverting only the part of the Hessian matrix corresponding to those parameters. This corresponds with an uncertainty analysis conditional on the fact that the wavelet is the estimated wavelet. The approximation is exact when there are no *a posteriori* correlations between the wavelet and the impedance/lagtime parameters or when the *a posteriori* covariance matrix of the wavelet parameters is zero. Neither of these conditions however is likely to be fulfilled in practice. It can be shown that the approximation renders an optimistic picture so that one has to be careful in the interpretation of the results.

8.3 AN ITERATIVE SCHEME

A possibility to incorporate the stabilization methods for wavelet estimation as discussed in chapter 7 is to iteratively estimate the wavelet and the acoustic impedance profile, see figure

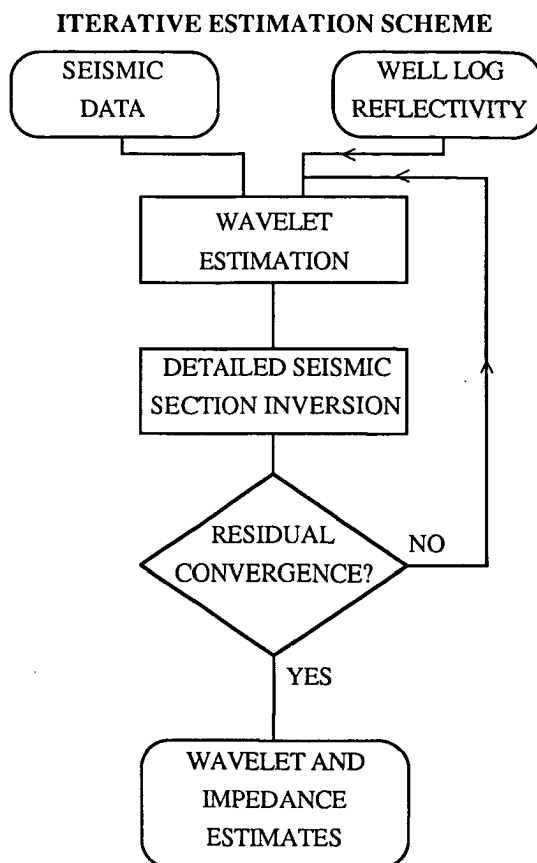


Figure 8.2 Flowchart of the iterative scheme for wavelet and acoustic impedance estimation.

8.2. The procedure is started with wavelet estimation, in which a reflectivity model is used that is derived from well log data. The lagtimes of the major events can roughly be estimated first by using crosscorrelation of the seismic traces. Stabilization can be used for wavelet estimation. An automatic stabilization procedure is of course strongly preferable because the wavelet estimation step has to be done in each iteration. The step of inversion for acoustic impedance may consist of multi trace inversion as discussed in chapter 6 or of sequential single trace inversion, optionally followed by the fitting of spline functions through the results to remove jitter. It is important to realize that the a priori model should be kept the same in each iteration and not be replaced by the previous estimated model because the stabilization effect is then lost.

The iterations are repeated until the scheme converges to a final solution, i.e. until no significant changes occur in consecutive wavelet or impedance estimates and no significant

change occurs in the residual energy after the impedance inversion step. Note that convergence can not be proved to occur. It is simply hoped that each iteration will yield better estimates for the wavelet and the acoustic impedance profile and that the scheme converges to an optimal solution.

The iterative scheme is only directed at point estimation. An uncertainty analysis can not be performed in the conceptually clear way of Bayesian inversion. An uncertainty analysis for the impedance/lagtime parameters has to be performed in a similar way as the approximation discussed in the previous section.

8.4 RESULTS ON SYNTHETIC DATA

In this section some examples are given of the performance of the schemes discussed in the previous sections on synthetic data. The first example concerns a moderately complex model and is inverted with the consistent Bayesian scheme discussed in section 8.2. The relevant plots of this example are shown in figure 8.3. The true impedance profile shown in figure 8.3a contains 4 reflectors in 70 ms. The third layer is rather thin, 8 ms at some locations. The noisy data used for inversion is shown in figure 8.3b. Note that the time scale differs from figure 8.3a. The data is generated by convolving the true reflectivity with the minimum phase wavelet shown in figure 7.1, see also figure 8.3f. Bandlimited noise (10–70 Hz) is added to the noise free data. The signal to noise ratio is 10 dB, measured on the first 100 ms and 7.5 dB when measured over 250 ms, the data gate used for inversion. The a priori model is a flat layered model, derived from an erroneous well log located at trace position 8. The a priori standard deviations of the single trace parameters at the well location are taken as $.5 \cdot 10^6 \text{ kgm}^{-2} \text{ s}^{-1}$ for the acoustic impedances and 4 ms for the lagtimes. The trace to trace uncertainties are taken as $.2 \cdot 10^6 \text{ kgm}^{-2} \text{ s}^{-1}$ for the impedances and 2 ms for the lagtimes. Numerical details on the errors are given in table 8.1 below. The a priori model in this example is shifted backwards in time with a time shift that is likely to be applied in practice to ensure that the wavelet is causal. In the errors for the lagtimes given in table 8.1 the average lagtime shift is removed. For a definition of the three error quantities, see section 5.6. In figure 8.3.e the estimated model is shown. From the numerical details given in table 8.1 it is clear that the estimated model is very accurate in the lagtimes, with an average error of less than .5 ms and a largest error of only 1.2 ms. The improvements for the acoustic impedances are not very large.

The estimated wavelet is shown in figure 8.3.f in comparison with the true wavelet. The wavelets are resampled to 1 ms for display purposes. The wavelet estimate is excellent. The normalized error (defined in chapter 7) is only 0.04. From a close inspection of the figure it can be seen that a large part of the wavelet error is due to the fact that the estimate is wrong by a constant factor. This is caused by the erroneous a priori model. As stated in

section 8.2 the data can not provide information concerning the exchange in wavelet and reflection coefficient amplitudes. When the a priori model is wrong in this direction of parameter space, the true model can not be recovered and the wavelet estimate will be wrong by a constant factor.

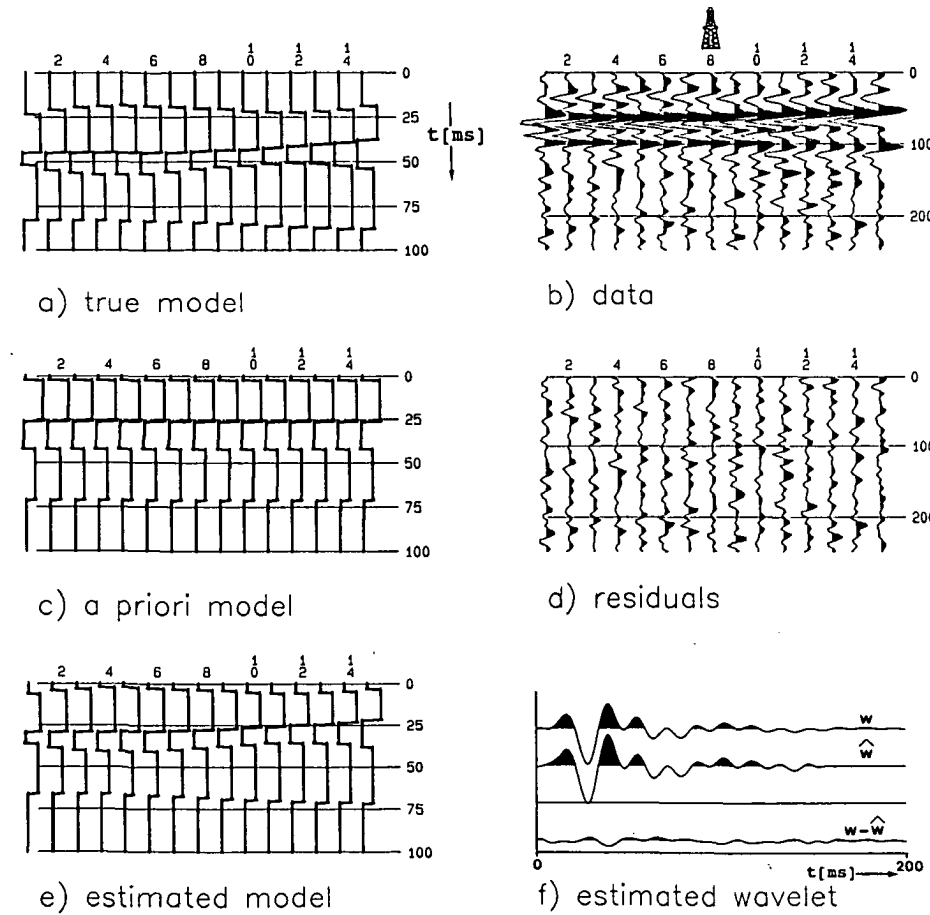


Figure 8.3 First synthetic example of integral estimation.

Table 8.1 Errors in the estimates for the first example of integral estimation.

model	impedance [$10^6 \text{ kgm}^{-2} \text{ s}^{-1}$]			lagtimes [ms]		
	error ₁	error ₂	error ₃	error ₁	error ₂	error ₃
a priori model	.35	.42	1.11	3.0	3.4	6.6
estimated model	.27	.38	1.17	.3	.4	1.2

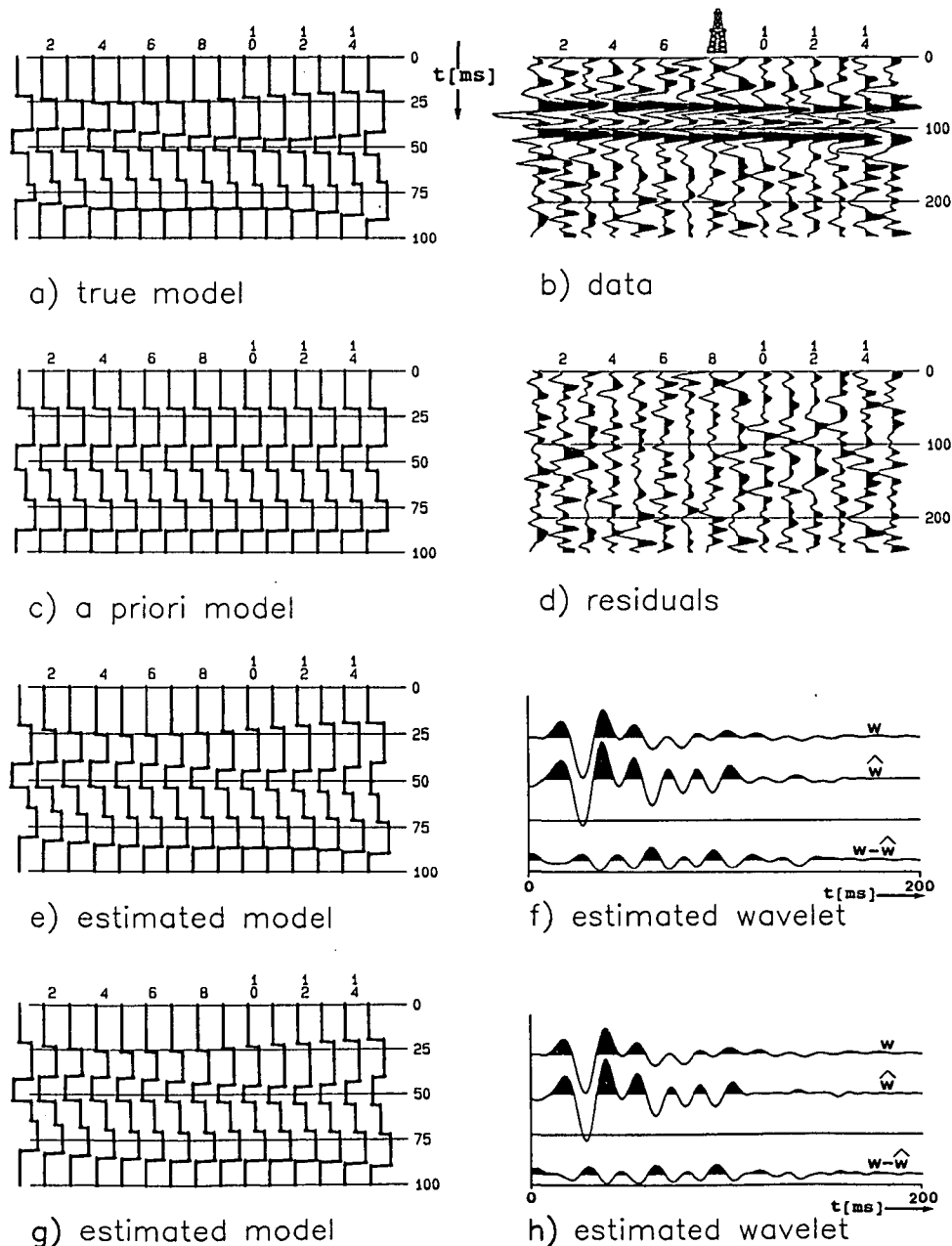


Figure 8.4 Second synthetic example for integral estimation.

e,f) results for the consistent scheme

g,h) results for the iterative scheme with sample selection in wavelet estimation.

The residual traces are shown in figure 8.3.d. The residual energy is -8.8 dB, relative to the data over the full data gate, and is lower than the noise level. From a detailed comparison of the lower parts of the traces it can be seen that the residuals closely resemble the noise on the traces. The estimation procedure to a large extent separates the noise from the signal.

The second example is for a more complex model, shown in figure 8.4.a. There are 5 reflectors in 70 ms. For the fourth reflector there is a low impedance contrast. The noisy data used for inversion is shown in figure 8.4.b. The signal to noise ratio is 6 dB when measured over the first 100 ms and 2.7 dB when measured over the full trace length. The a priori model is again flat layered and derived from an erroneous well log located at trace position 8. The uncertainties in the a priori model are the same as in the previous example. The a priori and estimated model are now displayed after alignment with the true model to allow an easier comparison. The inversion is performed with the consistent Bayesian scheme as well as with the iterative scheme with the sample selection method in wavelet estimation. The length of the wavelet in the consistent scheme as well as the start length of the wavelet in the iterative scheme is 160 ms. The results are shown in figure 8.4.e,f for the consistent scheme and in figure 8.4.g,h for the iterative scheme. Numerical details concerning the errors in the impedance model are given in table 8.2.

Table 8.2 Errors in the estimates for the second example of integral estimation.

model	impedance [$10^6 \text{ kgm}^{-2} \text{ s}^{-1}$]			lagtimes [ms]		
	error ₁	error ₂	error ₃	error ₁	error ₂	error ₃
a priori model	.21	.27	.70	2.4	3.0	7.8
estimated model consistent scheme	.29	.36	.80	1.3	1.5	6.5
estimated model iterative scheme	.29	.36	.86	1.0	1.4	6.9

The errors in the lagtimes are larger than for the simple model as expected. The largest errors in the lagtime parameters are in the fourth interface, with the low reflection coefficient.

The normalized errors in the wavelet estimates are .27 for the consistent scheme and .19 for the iterative scheme. From figures 8.4f,h it can easily be seen that a large portion of the error is due to a scale factor. A very rough correction in the estimated wavelets gives normalized errors of .16 and .1 for the respective procedures. It can be concluded that the two approaches give similar results, the iterative scheme with sample selection may be called slightly better.

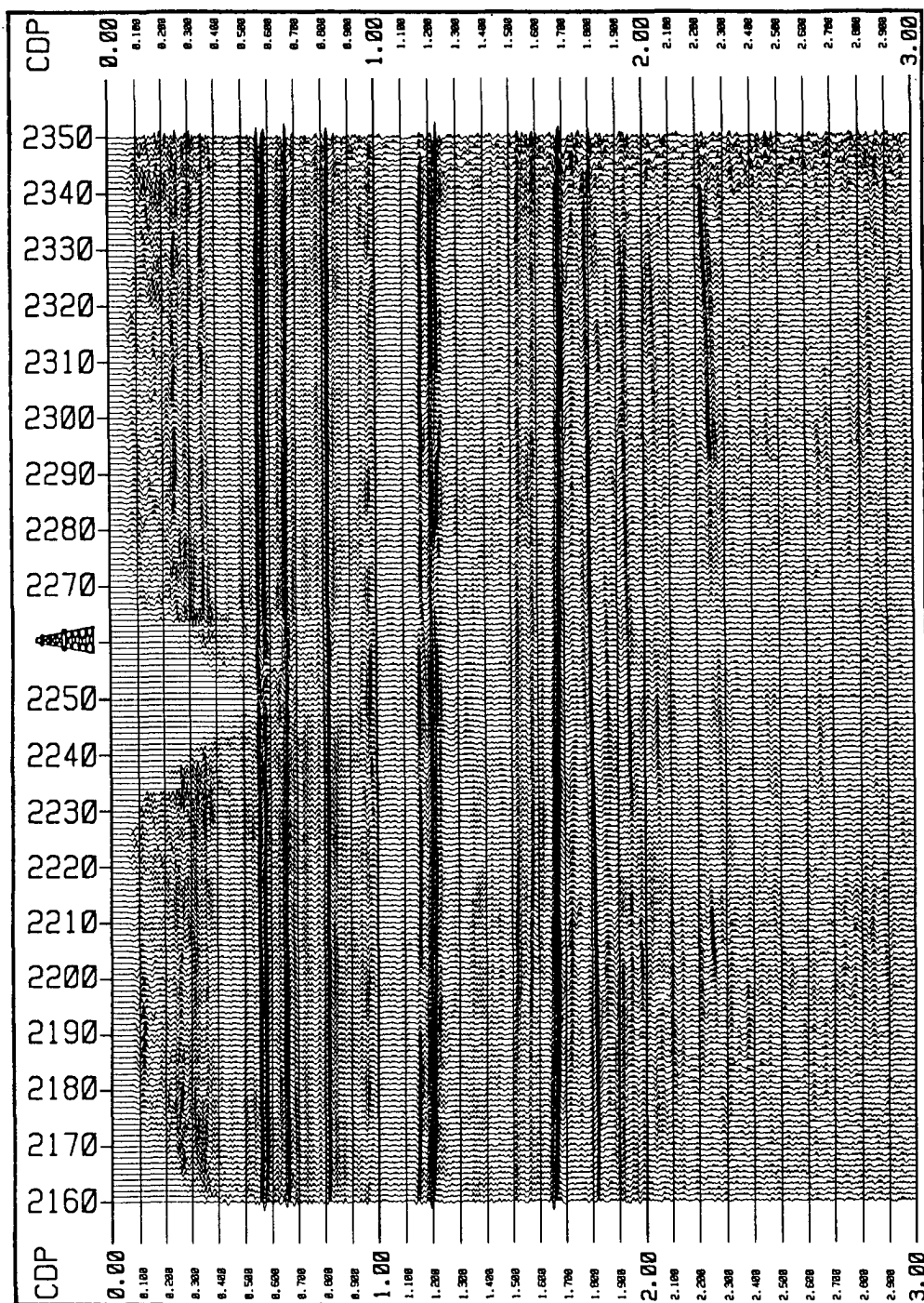


Figure 8.5 Poststack section used in real data example.

The residual traces for the two approaches are also very similar. Therefore only those of the consistent scheme are shown in figure 8.4.d. The residual energy is -3.2 dB relative to the data. Again it can be seen that the residuals very closely resemble the noise on the data.

8.5 A REAL DATA EXAMPLE

In this section some results on real data are shown. An oil company, which wished to remain unknown, provided a poststack data set. A large part of the section is shown in figure 8.5. Conventional processing had already been applied. This means that the data set probably cannot strictly be regarded as a true amplitude representation of the reflectivity because in conventional processing no special care is taken to preserve amplitudes perfectly. One may however expect that multiplicative factors for different traces that result from processing steps are close to each other for neighbouring traces. This makes application of the inversion scheme still possible. The wavelet may have to be reestimated more often than for a true amplitude data set.

Well log data has been provided. The well is located at CDP number 2260. The target zone is around 1.7 s. A blow up of that zone is shown in figure 8.6, for 40 traces around the well. The traces are renumbered for convenience such that the well is located at trace location 60. At the target zone four layers are distinguished in the composite well log provided. The material and depth locations are:

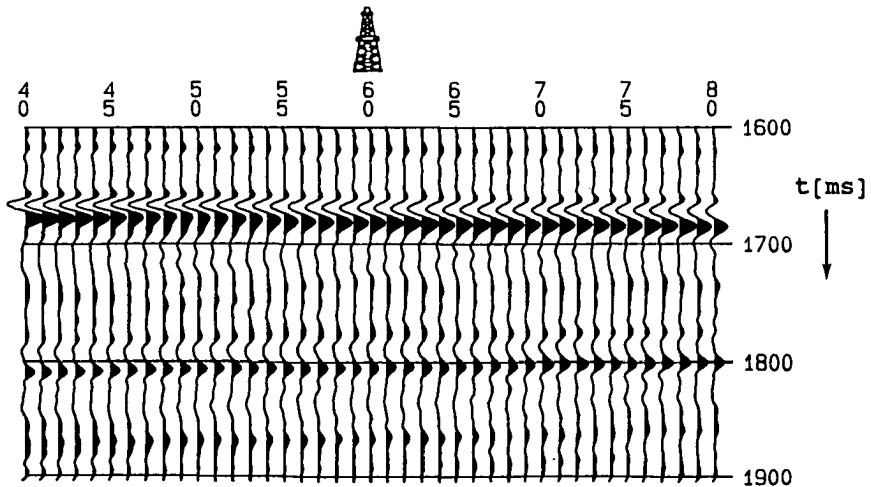


Figure 8.6 Target zone for inversion.

- 1) shale (1874 – 1975.7 m)
- 2) sandstone (1975.7 – 1996.5 m)
- 3) anhydrite (1996.5 – 2037 m)
- 4) salt (2037 – 2269 m)

The geophysicist of the oil company was mainly interested in the sandstone, which may turn into a gas-sand in the left part of the section. There are three reflectors which constitute the event just above 1.7 s. In table 8.3 the density, velocity and the impedance derived from the sonic and density log are given, together with the uncertainties.

Table 8.3 Properties of the layers at the target zone, derived from sonic and density logs.

layer	density		velocity		acoustic impedance		
	value [kg/m ³]	relative error (%)	value [m/s]	relative error (%)	value [10 ⁶ kgm ⁻² s ⁻¹]	error	
						rel.(%)	abs.
shale	2530	5	2875	5	7.27	10	.73
sandstone	2470	5	3277	11	8.09	16	1.29
anhydrite	2850	5	6096	5	17.37	10	1.74
salt	2100	2.5	4482	3	9.41	5.5	.52

A number of remarks have to be made concerning the values in table 8.3. The values for the densities and velocities are determined by a block averaging of the logs. The amplitude of the variations are taken as the size of the (one sided) uncertainty interval. The sonic log of the sandstone showed a gradient in depth. The averaged value is used and the largest deviation taken as the uncertainty interval. The interpretation of the density log posed difficulties for the anhydrite layer, possibly due to the presence of the casing shoe in that layer. The value for the density found corresponds well with the general values given for anhydrite in the literature, see e.g Telford et al. (1976), who use data of Gardner et al. (1974). The uncertainty for the density of the anhydrite is taken so as to cover the range of density values given in the above mentioned references. The impedance is the product of the density and the velocity. The relative errors of the latter two add to the relative error of the acoustic impedance. The computed absolute errors in the last column of table 8.3 are used as the standard deviations in the a priori Gaussian pdf's for the impedances.

The lagtimes of the reflectors are determined from the thickness and the velocities of the layers. The sandstone/anhydrite interface is used as a reference in time. After alignment with the data it is more or less set fixed by taking a very small a priori uncertainty for its lagtime (only at the well location of course). The time thicknesses and their uncertainties for the sandstone and the anhydrite layers are given in table 8.4. For the derivation of the uncertainty in the lagtime the relative uncertainty in the velocity is used. The uncertainty in the depths are neglected.

Table 8.4 Relation of thickness, velocity and two-way travelttime of the sandstone and anhydrite layers.

layer	thickness [m]	velocity [ms]	two-way travelttime		
			value [ms]	relative error	stand. dev. [ms]
sandstone	20.8	3277	12.7	11	1.4
anhydrite	40.5	6096	13.3	5	.7

Note that from this derivation it is clear that the a priori information in the impedances and the lagtimes are not uncorrelated. They are nevertheless treated as if they are. The scheme can be improved on this point.

As an a priori model for all traces the well log model is used, thus taking a laterally invariant a priori model.

For the full a priori covariance matrix to be built up uncertainties concerning trace to trace variations are also necessary, see chapter 6. Concerning the lagtimes little is known a priori. A save value of 1 ms for trace to trace lagtime variations is used, implying that according to the a priori information a layer can easily pinch out in 10 traces. A little more can be said concerning the acoustic impedances. An interpreter provided values for the expected maximal changes of the densities and velocities per 100 meter. They are given in table 8.5 together with the resulting relative changes of the impedance and the resulting standard deviation for the trace to trace variations of the impedances. The trace to trace distance is 25 meter.

Table 8.5 Uncertainties in lateral variations of densities, velocities and impedances.

layer	relative uncertainty density, velocity /100 m	relative uncertainty impedance per trace	acoustic impedance [$\text{kgm}^{-2}\text{s}^{-1}$]	standard deviation [$\text{kgm}^{-2}\text{s}^{-1}$]
shale	3	1.5	$7.27 \cdot 10^6$	$.11 \cdot 10^6$
sandstone	3	1.5	$8.09 \cdot 10^6$	$.12 \cdot 10^6$
anhydrite	2	1.	$17.37 \cdot 10^6$	$.17 \cdot 10^6$
salt	1	.5	$9.41 \cdot 10^6$	$.05 \cdot 10^6$

From the well log data it is derived that the shale is approximately 72 ms thick and the salt layer approximately 104 ms. This means that the response of the three reflectors can be regarded as isolated when the wavelet is not longer than 70 ms. This will probably not be the case, but the energy in the tail of the wavelet is expected to be sufficiently small to neglect the reflectors above the target zone. The data gate used for inversion is chosen from 1650 to 1765 ms. The lower boundary of 1765 ms is chosen such that the reflection of the interface of the salt and the next layer falls outside the data gate.

The estimation of the noise level is not fully integrated in the scheme yet. From a

comparison of neighbouring traces a noise level of -15 dB relative to the data is chosen. The residuals after inversion will have to indicate whether this value is significantly too high or too low.

The length of the wavelet is chosen to be 88 ms, somewhat shorter than the data gate, in order to avoid instability in the wavelet tail. The consistent Bayesian scheme described in section 8.2 is used to estimate both the acoustic impedance and the wavelet from 11 traces symmetrically around the well. The a priori model and the data used are shown in figure 8.7a,b. Two spline types are tried: the linear spline and the cubic spline. The estimated model and the residuals for the linear spline type are shown in figure 8.7c,d. and those for the cubic spline in figure 8.7e,f. The estimated models for the two parameterizations

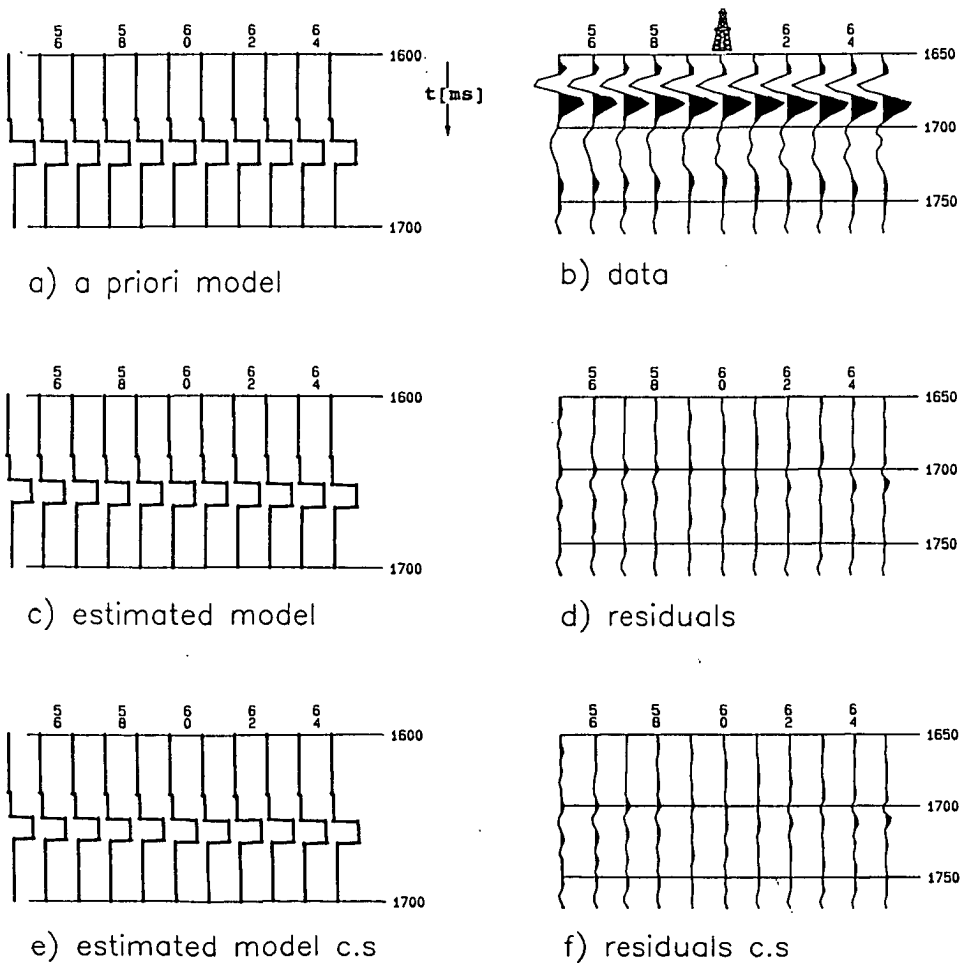


Figure 8.7 Inversion of 11 traces around the well.
c,d) results for linear splines; e,f) results for cubic splines

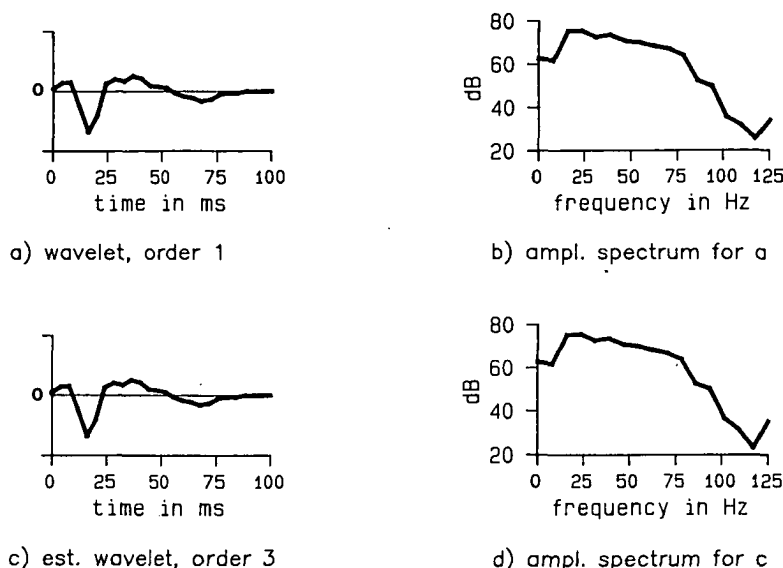


Figure 8.8 Estimated wavelets for 11 traces around the well.

a,b) wavelet for linear splines

c,d) wavelet for cubic splines

closely resemble each other as can be studied in more detail in figures 8.9 – 8.12. The estimated wavelets and their amplitude spectra are shown in figure 8.8. The two wavelets are very similar. Note that there are nonzero dc components, approximately 15 dB down compared to the maximum component in the spectrum. Such a dc component is due to a truncation of the data gate. It can be filtered out if desired.

The data is explained very well as can be seen from the residuals. The energy of the residual traces is -16.9 dB (relative to the data) for the linear spline and -17.1 dB for the cubic spline. For both spline types therefore there is a good data fit. Nonetheless lateral alignment can be seen in the residuals, especially around 1700 ms. It is antisymmetric around the well position. This may be indicative of a gradually varying scale factor for the traces, possibly due to a trace normalization step in the preprocessing. The effect is considered to be of fairly low amplitude so no further attention is paid to it.

Figure 8.9 shows the a priori and the a posteriori lagtime model for the linear spline, together with the uncertainty intervals of one standard deviation of the single trace parameters at each side of the model (see also chapter 6). There is a strong decrease in uncertainty for the second and third reflector. The reduction for the first reflector is also significant but less strong, because the reflection coefficient of the first reflector is much lower than for the second and third one. It can clearly be seen that the reflectors dip to the

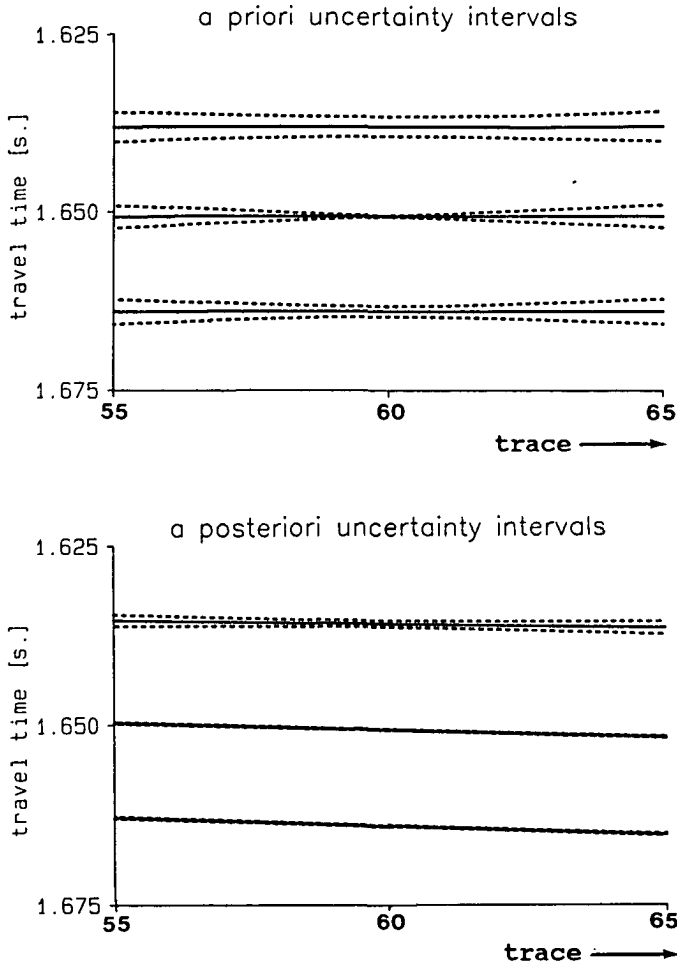


Figure 8.9 A priori and a posteriori models with uncertainty intervals for 11 traces around the well for linear splines.

right, as is expected from the seismic data displayed in figure 8.5. In figure 8.10 similar results are given, now for the cubic splines. The a priori and a posteriori uncertainty intervals are slightly higher than for the linear spline, due to the fact that the cubic spline type contains more parameters. The first reflector in the estimated model shows a slight deviation from a straight line, the other two are perfectly straight. The estimated models of the two parameterizations are very close to each other and within each other's uncertainty intervals.

The same information is given for the impedances, for the linear spline in figure 8.11 and for the cubic spline in figure 8.12. First the linear splines in figure 8.11 are discussed.

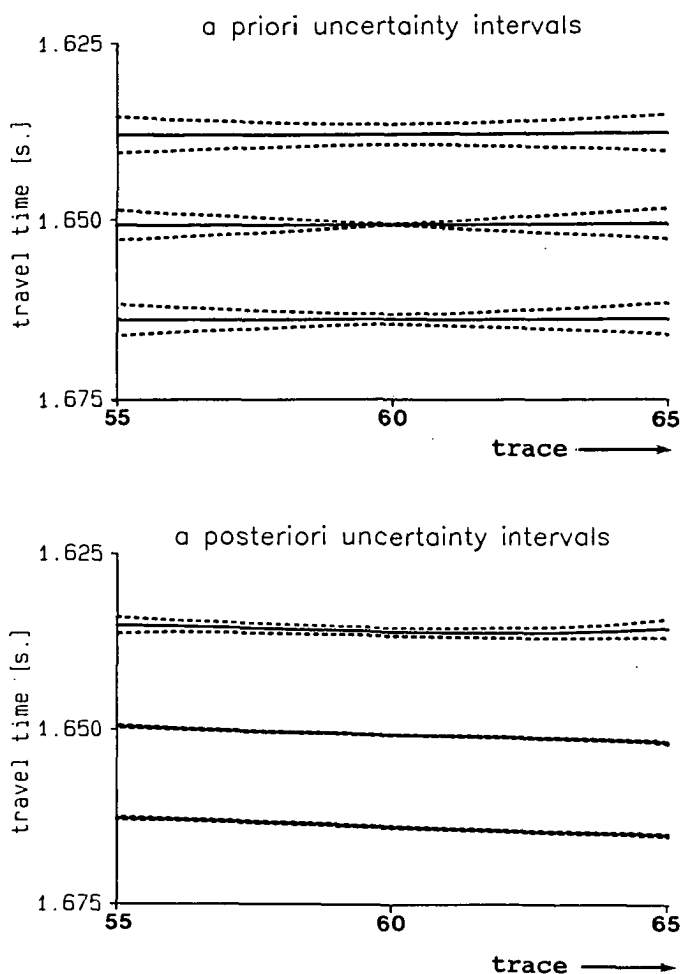


Figure 8.10 As figure 8.9 but for cubic splines.

As in the synthetic experiments the reduction of uncertainty in the acoustic impedances is less strong although significant for all but the salt layer. This result for the salt layer can be explained by the fact that the a priori information for the salt layer is already accurate. Note that the range of the impedance values in the plot for the first three layers is $6 \cdot 10^6 \text{ kgm}^{-2}\text{s}^{-1}$ while it is $3 \cdot 10^6 \text{ kgm}^{-2}\text{s}^{-1}$ for the fourth (the salt) layer. The estimated acoustic impedance profile is fairly constant with lateral distance. Only the acoustic impedance of the anhydrite shows a small increase to the right. The results for the cubic splines are fairly similar to those for the linear splines. The estimated acoustic impedance of the anhydrite shows a small deviation from a linear function.

Figure 8.13 shows the a priori and a posteriori covariance matrices, the resolution

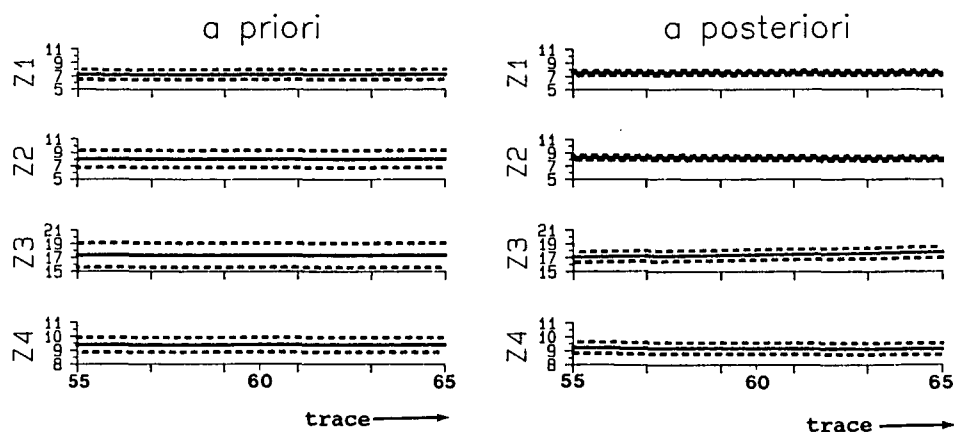


Figure 8.11 *A priori and a posteriori models and uncertainty intervals of acoustic impedances for 11 traces around the well, for linear splines.*

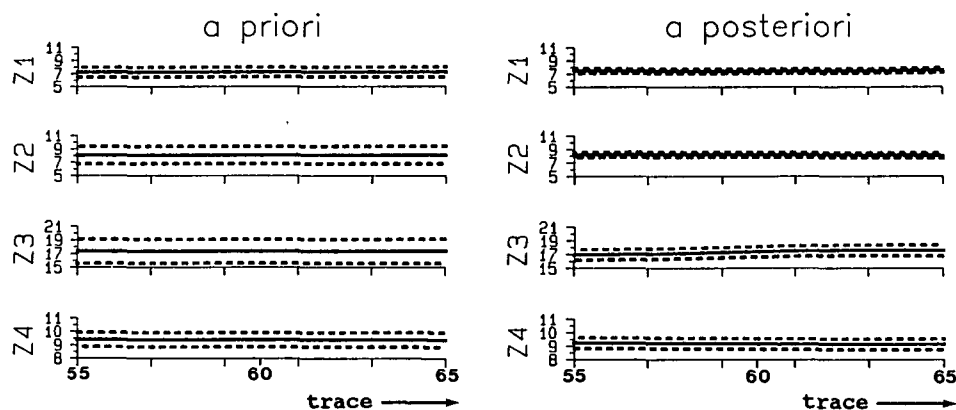
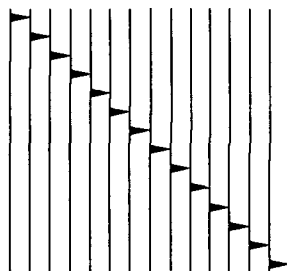


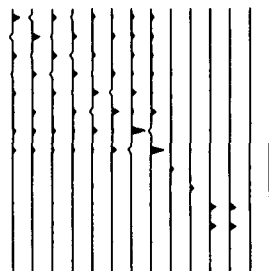
Figure 8.12 *As figure 8.11 but for cubic splines.*

matrix and the a posteriori correlation matrix for the linear spline type in the same type of display as in chapter 6, see figure 6.7. The parameters are transformed such that the a priori covariance matrix equals the identity matrix (see chapter 3). The a posteriori variances are strongly reduced when compared to the a priori variances, except for the acoustic impedance of the fourth layer. The resolution matrix is not too far from the identity matrix. The data thus provides significant information for the parameters. Strong a posteriori covariances, especially well visible in the correlation matrix, can be seen to occur between the acoustic impedance parameters. This is caused by the fact that the acoustic impedances can not be estimated fully independently from the data. There is also a very strong correlation in the lagtime of the 2nd reflector. This is due to the fact that there is very strong correlation in the unweighted a priori information because the middle of the layer is

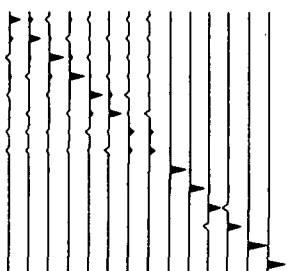
a priori covariance



a posteriori covariance



resolution matrix



a posteriori correlation

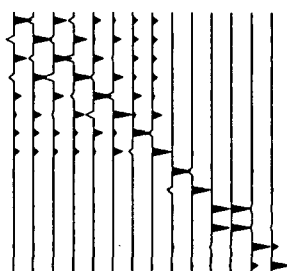


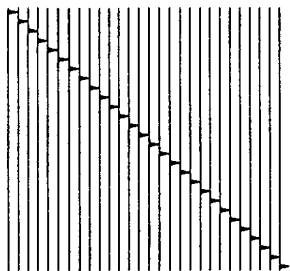
Figure 8.13 Matrices for uncertainty analysis, for linear splines.

more or less set fixed. This correlation is removed from the a priori covariance matrix by the linear transformation. This linear transformation then causes the strong correlation in the a posteriori covariance matrix because the parameter space is strongly “deformed” in the relevant subspace.

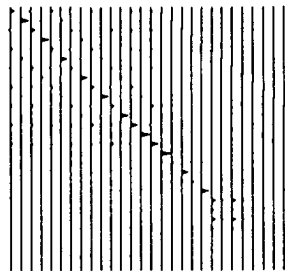
Figure 8.14 shows the same matrices for the cubic splines. This parameterization has 28 parameters, twice as much as the linear spline parameterization. Similar conclusions as for the linear spline can be drawn. The uncertainty reduction is on the average poorer than for the linear spline. The same type of correlations as for the linear spline occurs.

Figure 8.15 shows the singular value spectra for the linear spline type. There are four weak directions, for which the singular value of the likelihood function is lower than the corresponding one of the a priori pdf. The lowest singular value of the likelihood function is exactly zero, corresponding with a constant factor for all impedance parameters. The 13th singular value is also very low, corresponding to an effect close to constant factors for the acoustic impedances, but such that there are different factors for the traces. The plots of the pdf's along the eigendirections are not given because it proved to be very difficult to

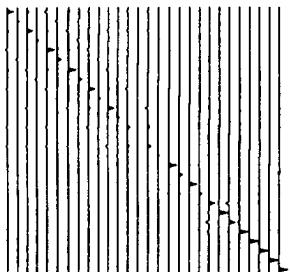
a priori covariance



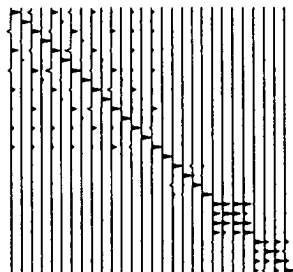
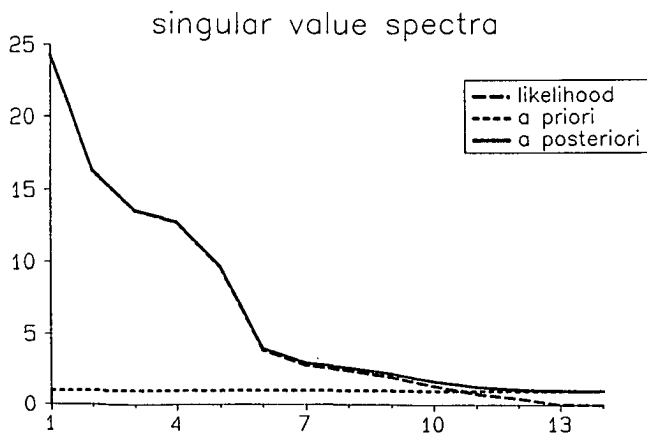
a posteriori covariance



resolution matrix



a posteriori correlation

*Figure 8.14 As figure 8.13 but for cubic splines.**Figure 8.15 Singular value spectra for linear splines.*

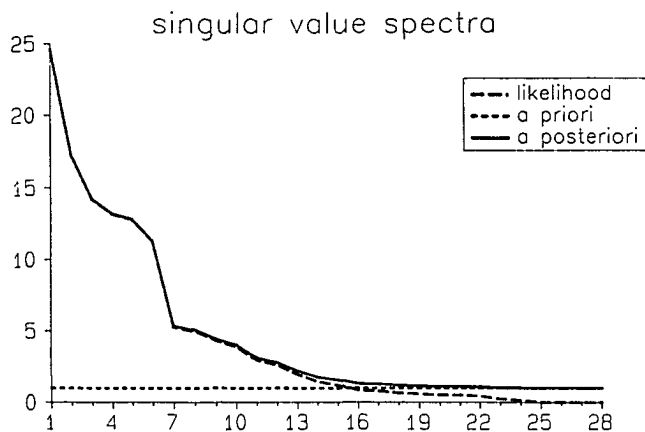


Figure 8.16 Singular value spectra for cubic splines.

visualize the directions through the parameter space caused by the fact that the relevant distances were quite small.

Figure 8.16 shows the singular value spectra for the cubic spline parameterization. Note that for high values the slope of the spectra are similar to those of the linear spline. They probably correspond to changes in the parameters of the lagtimes. There are relatively more weak directions showing that the data can provide significant information for only half the number of parameters.

From the comparison of the results for the two parameterizations it can be concluded that the linear spline is preferable for this data set, because the data fit is almost the same for both parameterizations and the results for the cubic spline are very close to those of the linear spline. The linear spline, with fewer parameters, is then preferable because it is more stable and gives lower uncertainty.

Next, a block of 11 traces to the left of the first block is inverted with the linear spline type, using the wavelet estimate from the first block. The estimated model of trace 55 at the left of the first block is used as a pseudo well log. The a priori model is again flat layered and the trace to trace uncertainties are taken the same as in the first block. The a priori model and the data are shown in figure 8.17a,b for both blocks. The estimated model and the residuals (plotted on the same scale as the data) are shown in figure 8.17c,d. The estimated model shows again the expected dip in the lagtimes. For a closer inspection of the results the reader is referred ahead to figures 8.20 and 8.21. The energy of the residuals in the second block (traces 44 – 54) is –13.5 dB, significantly higher than for the first block. Because the residuals are higher over the full range of the data gate and not strongly localized (which might have indicated a missing reflector in the parameter model) and because the data block is in a part of the section with bad coverage (see figure 8.5) it is

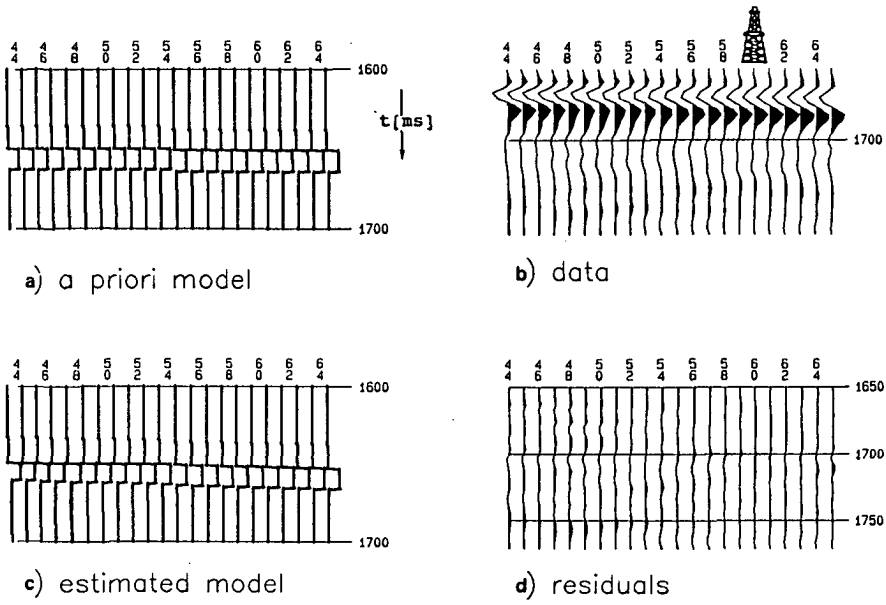


Figure 8.17 Inversion of the second block of traces with the wavelet estimate of the first block.

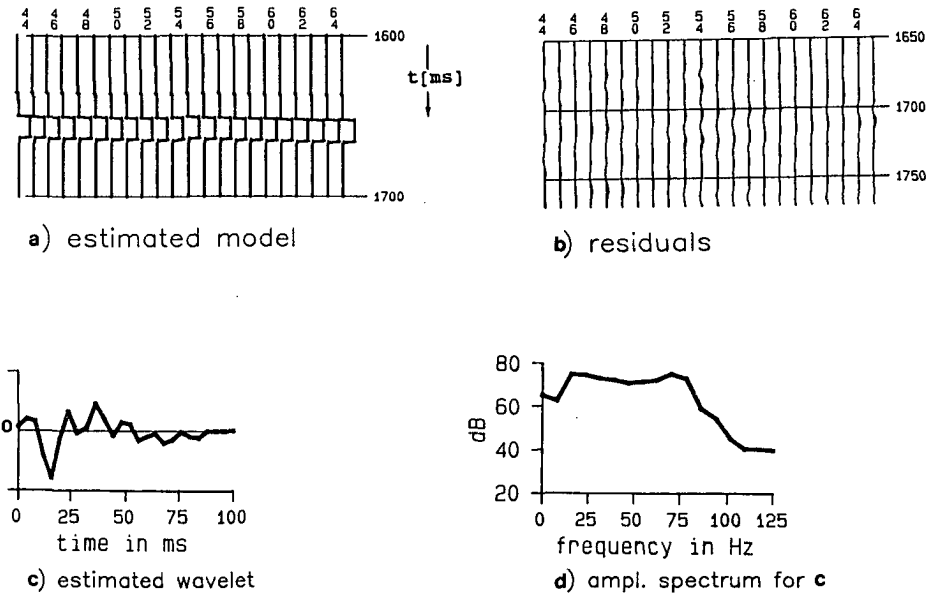


Figure 8.18 Integral estimation of the wavelet and the impedance profile for the second block of traces.

concluded that the wavelet may have to be reestimated. The integral estimation scheme described in section 8.2 therefore is used again to estimate both the wavelet and the acoustic impedance. The results are shown in figure 8.18. The residuals are significantly lower now. The residual energy is -16.4 dB, relative to the data. From the amplitude spectrum of the estimated wavelet it can be seen that this wavelet has relatively more energy at high frequencies (75 Hz) than the wavelet of the first block. The character of the wavelet in time has also clearly changed. The results of this reestimation effectively explain the data.

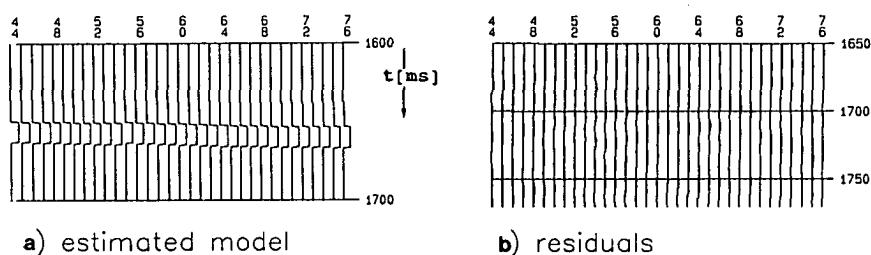


Figure 8.19 Inversion results for 3 blocks of traces.

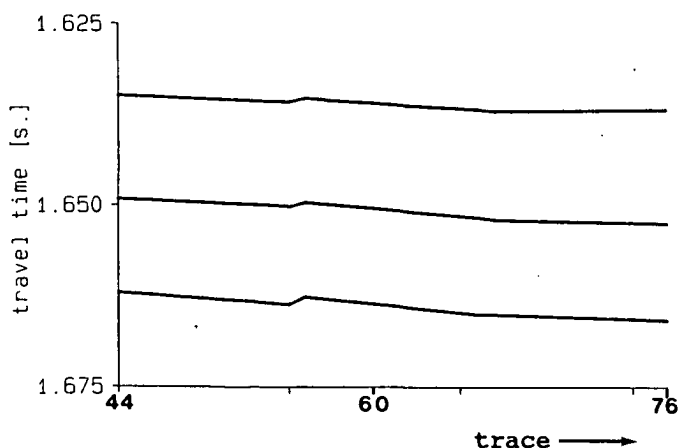


Figure 8.20 Estimated lagtimes for the three blocks of traces.

A third block is estimated to the right of the well log (traces 66–76). The wavelet estimate of the first block is again used. The estimated model and the residuals are shown for all three blocks in figure 8.19a,b. The dip in the lagtimes continues. The residuals of the third block are reasonably low with the estimated wavelet used, the energy is -15.3 dB relative

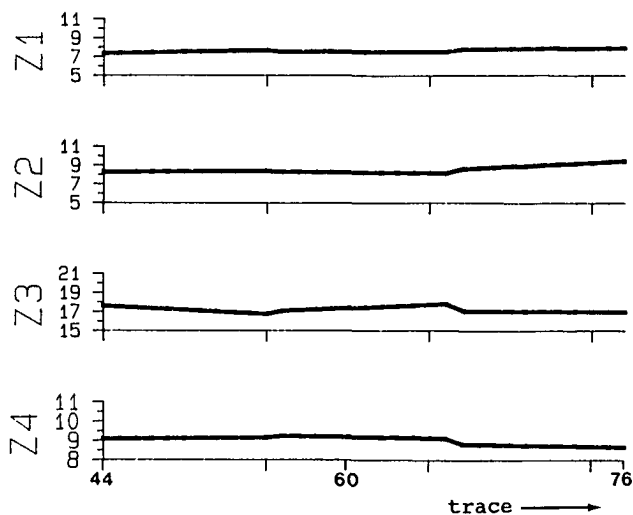


Figure 8.21 Estimated acoustic impedances for the three blocks of traces.

to the data. Although the residual energy could be lower when the wavelet is again reestimated this residual energy level forms no real cause for doing so. Figures 8.20 and 8.21 show the estimated lagtimes and impedances for the three blocks as a function of lateral distance, allowing closer inspection than the wiggle displays. There is a discontinuity between the lagtimes of the middle and the left block. This is due to the fact that the wavelet is reestimated in the left block. Because there was no real tight lagtime constraint the lagtimes turned around the centre of the block to account for the dip in the data. Since a time reference per block is arbitrary when the wavelet is also estimated, the left block may be shifted backwards in time to give greater continuity in the lagtime profile. The acoustic impedances show some variation with lateral distance. It has to be realized however that these changes are not very significant compared to the scale of the a posteriori uncertainties as calculated for the middle block, see figure 8.11. Only the increase of the impedance of the sandstone (second layer) in the right block seems to be significant.

The results of this test on real data are satisfactory in the sense that no “disasters” occurred which would clearly indicate that the scheme does not work on real data. The combination of estimated parameter values and the forward model explain the data very well. An acoustic impedance profile has been derived with a much greater accuracy than is possible for a visual interpretation of the data. On the other hand this test could not be a severe test because there is no other well log available for this data set, with which the inversion results could have been compared. A more detailed discussion on these items is given in the next chapter.

9

EVALUATION OF THE
INVERSION SCHEME**9.1 INTRODUCTION**

For a practical application of the scheme for detailed inversion of poststack data some refinements and extensions are desired. In section 9.2 the basic ingredients of the scheme are summarized and possibilities for refinements indicated. An important element for a practical scheme is testing of the residuals in order to detect the supply of inconsistent information. With this extension the scheme is more complete. In section 9.3 a short comparison is given between what can be called deterministic and statistical inversion techniques. The latter techniques have clear advantages because errors and a priori information are taken into account.

In the evaluation of the inversion scheme proper tests play of course a vital role. The significance of tests and their possible consequences are discussed in section 9.4. It turns out that in the discussion of testing some very fundamental issues come to the foreground. Some notions from the philosophy of science, which have been introduced in geophysics in the past few years, can play an important role. Their applicability to the seismic inverse problem is discussed. The chapter ends with the final conclusions.

9.2 REFINEMENTS AND EXTENSIONS OF THE SCHEME

In this section the nontrivial basic elements of the scheme for detailed inversion of poststack seismic data for acoustic impedances and traveltimes are summarized and discussed. Possible and desirable refinements and extensions are given that are necessary for a complete and robust scheme in practice.

summary and refinements

Basic elements in the scheme are:

1) *Bayesian estimation*

The choice for Bayesian estimation is a fundamental one. It is argued at some length in chapter 1. Some additional remarks are given in section 9.4.

2) *parameterization*

The choice to estimate acoustic impedance is motivated by the considerations that only poststack data is used and that it allows a simple and efficient scheme. It is preferred to reflectivity because it is closer to lithology. It is easier to provide a priori information for and easier to interpret. A disadvantage is that the reflectivity should not depend too strongly on the angle of incidence. If this condition does not hold reflection coefficients are preferable. The choice to describe the acoustic impedance profile as a function of traveltime (for data in time) is motivated by the fact that for a description in depth knowledge about the velocity would have to be built in. The choice of describing it in terms of the impedances and lagtimes of layers that are homogeneous in the vertical direction is motivated by the consideration that it allows a minimum of parameters and still may yield an accurate description of the target zone of the subsurface. A falsification (see section 9.4) of this parameterization may be obtained when in combination with the forward model it fails to explain the data sufficiently. In the multi trace version of the scheme described several spline types are possible. The choice for the spline type and the number of traces in a block is taken by the user of the scheme. It will be motivated by geological considerations and can also be falsified by inspection of the data mismatch.

3) *forward model*

In the scheme discussed in previous chapters the well known convolutional model, with primary reflections only, is used due to its simplicity and elegant possibilities to build an efficient optimization scheme for. It is very important to realize that this model is by no means prerequisite. Any forward model can be put into the framework of parameter estimation as set up in this thesis, allowing e.g. for multiples, absorption, elastic effects etc. The forward model can also be extended to two or three dimensions if desired. The advantage of a more sophisticated forward model is that the data is described more accurately. Going to a more accurate model turns noise into signal, leading to a more

accurate estimation of parameters. The disadvantage is that the optimization strongly decreases in efficiency.

4) *noise properties*

The issue of noise properties is a rather difficult one which will not be discussed at full length here. The following remarks summarize the author's opinion. What is noise? Noise (the term "errors" is preferable) is that part of the data that we either choose not to describe or are incapable of describing in our forward model. Whether the nature of some or all parts of the noise is intrinsically random is an interesting philosophical issue but is not relevant for a practical inverse problem. In the noise properties the uncertainties concerning this part of the data are described. These uncertainties may be called subjective (as is done by subjective Bayesians) but they nevertheless describe aspects of reality (which is implicitly assumed to exist objectively in empirical science) and there may therefore be reasons to change our description of the uncertainties due to the results of an inversion procedure. In this respect the author's position is at present somewhat between the subjective and the logical Bayesian interpretation of probability, the latter interpretation taking the position that probability is part of an objective state of affairs and that for noise on a certain data set only one pdf is the true one. This issue needs a closer philosophical analysis.

In section 1.12 the selection of the type of the pdf is discussed and a number of arguments for the choice of the Gaussian pdf for "normal" circumstances are given. Note that the description of the noise pdf is not very critical. As far as the covariance matrix is concerned for example we have the situation that all synthetic experiments in this thesis are performed with a white noise assumption, while the simulated noise is in fact bandlimited. The results are nevertheless very satisfactory. In practice however one may have serious doubts concerning the choice of the noise power. In the schemes presented in this thesis it has to be given so it has to be estimated from the seismic data using certain assumptions concerning trace to trace variations of signal and noise. A refinement of the inversion scheme would include a simultaneous estimation of the noise power in a consistent Bayesian framework.

Practical application and extensions

With the basic elements of inversion as presented in this thesis the practical inversion of a seismic section would proceed as follows. First the noise power is determined. Then the wavelet and acoustic impedance are estimated integrally at a well location. The estimated wavelet is then used to invert other blocks of data. An uncertainty analysis can be performed wherever desired. The results of the inversion, estimated model and uncertainties are conditional on all premisses used. Probabilistic properties of features of

the residuals (e.g. the energy of the residuals traces) can be determined. When the actual outcome of one of those features is highly improbable one may conclude that at least one the premisses is wrong. When the residual energy for example is higher than a certain threshold, that has to be determined beforehand, one of the following items may need correction:

- 1) the wavelet
- 2) the mathematical form of the forward model
- 3) the choice of the spline type (in multi trace inversion)
- 4) the number of layers in the model
- 5) the noise properties
- 6) the a priori information on the parameters

The determination of which of the items need correction poses a difficult problem. By utilization of several features extracted from the residuals, an automatic selection may be possible. In e.g. a simplified situation where one may doubt between updating either the number of layers or the wavelet the form of the residuals may yield the answer because wavelet errors are expected to give high residuals over the whole time gate, while a missing layer is expected to give more localized high values. A resolution analysis is of course again quite important for this detection procedure, answering the question which size of errors (or inconsistencies) can be determined from the residuals and which can't.

As mentioned before this research was part of a project in which this testing of residuals is investigated by other researchers. Some results can be found in Kaman et al. (1984) and in Kaman et al. (1987). Although the material presented there is approached from a pattern recognition point of view it is of course closely related to statistical hypothesis testing and can probably conceptually be fully integrated in the Bayesian framework as put forward in this thesis.

9.3 DETERMINISTIC VERSUS STATISTICAL TECHNIQUES

In the past few years inversion schemes have been proposed which aim at direct inversion of a one-dimensional seismogram. Often the one-dimensional wave equation is used as the forward model. Ursin and Berteussen (1986) give an overview of some of those methods. They may be called deterministic techniques as they do not account for noise explicitly. Numerical stability is very often (if not always) a problem with those schemes. A recent example is also given by Sarwar and Smith (1987), who succeed, using the one-dimensional wave equation, to derive a perfect acoustic impedance profile with high resolution from noise free data. When either noise is added to the data or an incorrect wavelet is used (which can of course also be considered as having noise) their method becomes unstable. These results need not surprise us. When the forward model has a

mathematical structure such that many parameter models give identical or nearly identical data (as is the case for an acoustic impedance model on a finely sampled time grid) the inverse operator can be expected to become unstable.

In practice we always have to deal with uncertainties, with noise on the data. The solution to the problems mentioned above is thus to take these errors into account, not to aim at a perfect data fit and use a priori information to select the desired model out of the many that explain the data equally well. In a more refined formulation this is exactly what the statistical Bayesian approach does. The mathematical techniques as described in the papers mentioned above should be used to supply a more sophisticated forward model and derivatives to be used in a parameter estimation formulation.

9.4 TESTING OF THE INVERSION PROCEDURE

In considering the application of an inversion scheme as described in this thesis the question of course will arise what the scheme is really worth, how good it is. (This holds for any inversion scheme alike, in or outside geophysics. The discussion given in the sequel of this section therefore is also partly applicable for the seismic inversion techniques of filtering and migration.) Unfortunately the question of how good an inversion scheme really is is difficult to answer. Although it seems a straightforward and simple question it is really fraught with methodological difficulties. It is beyond the competence of the author to give a conclusive philosophical analysis of the problems that arise in answering such a question. A number of remarks however can be made and will be made because of the importance of the issue.

With the laudable aim to improve scientific methodology in geophysics Parasnis (1980) and Ziolkowski (1982) introduce elements of the philosophy of science in geophysical literature. Especially Parasnis is strongly oriented on Popper's philosophy, which has as most characteristic element that scientific effort should be aiming at falsifying theories, rather than trying to confirm them. As a demarcation criterion between scientific and nonscientific theories Popper (see e.g. his publications of 1959, 1962) puts forward the criterion that scientific theories can be disproved or falsified whereas nonscientific theories can't. His philosophy can be seen as a reaction on the school of logical positivism of the Vienna circle. Especially Rudolf Carnap is an exponent of this school and has put great efforts in trying to devise a theory of inductive logic with which scientific theories can be assigned a degree of confirmation. An important element in this effort is probability in a Bayesian interpretation (see Carnap, 1962), which in this thesis is used for parameter estimation only but apparently may have wider applicability. The attempt of logical positivism to erect a philosophy of science on these principles is generally regarded as

highly problematic. Some activity in this direction however is still going on, as is shortly discussed below.

Parasnis is such an enthusiastic follower of Popper's philosophy that he may be called to be "more Popperian than Popper". This is illustrated by comparison of his statement (which is also criticized by Ziolkowski, 1982): "... it is possible to conclusively disprove or falsify a scientific theory" with one of Popper himself*): "No conclusive disprove of a theory can ever be produced". Imre Lakatos (see his essay of 1970) probably would have classified Parasnis as a "dogmatic falsificationist". Though Popper himself recognizes some problems with the falsification of theories the central notion in his philosophy remains that irrespective of the presence of competing theories a theory can be decided to be falsified by observations only. Lakatos (1970) refines the falsificationism of Popper, the most important refinement being that a theory can only be falsified when there is a competing theory that explains all facts that the first theory explains (1), that has excess empirical content (2) and of which some of the excess content is verified (3). In geophysical literature Ziolkowski (1982) also doubts whether falsification without a competing theory is possible. To complete this very brief overview the important work of Thomas Kuhn (1962) should be mentioned. He states that scientific activity can be characterized by periods of "normal science", in which the basic theory is not questioned and in which activity is mainly devoted to puzzle solving within the existing theoretical framework or paradigm. A theory is only replaced by another one in a scientific revolution when the old theory has to face too many anomalies and a new one is put forward. From his analysis much can be learned about scientific activities. Although he is often criticized for describing scientific progress as an irrational process the differences with especially the philosophy of Lakatos are not large as Kuhn himself explains in his "reflections on my critics" (1970). That there is certainly no consensus within the philosophy of science is exemplified by the controversial book "Against method" of Paul Feyerabend (1975), in which Feyerabend strongly argues for an anarchistic science. Progress in science he argues should not be hindered by any rules, restricting the scientist's creativity. If Feyerabend has a principle then it is expressed in his adage "anything goes". Although the scientist thus can not expect any guidelines from Feyerabend's work (see also his paper on the Kuhn-Popper controversy, 1970) it can be recommended as refreshing and stimulating reading.

Before discussing Ziolkowski's application of Popper's theory to theories of deconvolution it should be mentioned that the philosophy of science is a research field in which insights are still changing and possibly always will be. No clear-cut recipe for practicing science is universally applicable. In the sequel of this section the falsification criterion of Popper will be applied to seismic inverse problems. Concepts and

*) Popper (1959), section 9.

considerations of the other philosophies mentioned can also shed some light on this problem.

Ziolkowski (1982) discussed seismic deconvolution theories in the light of Popper's demarcation criterion. He argues that "statistical deconvolution techniques" (that have to assume certain statistical properties of the geology) cannot be refuted and are therefore nonscientific by Popper's criterion. Some comments on this issue can be made. In order to give the discussion a wider scope the statistical deconvolution technique of Ziolkowski's paper will be generalized to inversion techniques. A first question the author would like to raise (without having the hope of being able to conclusively answering it) is whether inversion techniques can be treated as scientific theories. At first sight one would be inclined to say they are certainly not the same. Scientific theories try to explain reality by a set of relatively simple concepts, while inversion techniques estimate free parameters (e.g. the reflectivity of the subsurface) in a theory, using theories. Because one may also reason otherwise (see below) the analysis of possible falsifications of an inversion technique is split up for the two viewpoints:

A) Inversion theories are not regarded as scientific theories.

In this interpretation falsification should not be directed against the inversion techniques themselves but can be directed towards theories (which in a simplified form can also be statistical assumptions as mentioned above) that are premisses of a proper working of the technique. A theory that a certain statistical assumption always holds can be falsified by independent data, in seismics e.g. by well log data. Such falsifications have probably been obtained for most statistical deconvolution techniques. A retreat however can be to relax the falsified theory such that it claims the statistical assumption to be valid only for the data set under consideration (thus, after falsification, leaving open the possibility of application to other data sets). The possibility of falsification then remains for each data set. When the statistical assumptions concern the solution itself however (as in the case of the statistical deconvolution techniques) we find ourselves in the curious position that whether falsification using the independent data actually takes place or not, the whole inversion result becomes useless because more accurate results are available now. So either the inversion result must be taken for granted or it is useless to apply the technique at all.

B) Inversion techniques are regarded as scientific theories.

Although it is not very clear what the explaining power of the "theory" is in this interpretation the theory has an important observational prediction to make: the geology of the subsurface. In this interpretation it is probably clearest to equate the theory with the proposition: "Application of such and such inversion technique to the data always renders a good (or accurate) result". The theory can be falsified by "observation" of the

geology. Strictly spoken this poses another inverse problem and this situation immediately leads to an infinite regress. One could also state that falsification is impossible because the true geology (whatever that may mean) is never known. We can however falsify the first "theory" in favour of the second (if this is a more accurate one). If the first "theory" claims to always work than falsification means definite falsification. If the inverse technique contains situation dependent parameters than it can be falsified for the data set under consideration only, not for future data sets. A similar situation as under interpretation A occurs.

This analysis suggests that when the inverse problem contains situation dependent parameters and when testing is performed on the *results* then either these results are useless (because in the process of testing we already obtain a more accurate answer to the original problem) or one has to take the result for granted. Note that an inverse problem will virtually always contain situation dependent parameters (noise properties for example). A comment on Ziolkowski's article is that with falsification he must mean falsification within practical limitations. The statistical deconvolution results he qualifies as unscientific *are* falsifiable in principle: by drilling wells! In the interpretation of an inversion technique as a technique to determine optimal values for free parameters in a theory the inversion technique is a tool for refutation itself because it may turn out that even for optimal parameter values the theory fails to explain the observed data sufficiently.

Ziolkowski proposed a technique which overcomes the problem of the statistical assumptions in the statistical deconvolution techniques. In this technique additional measurements and a scaling law for wavelets turns the situation of "one equation with two unknowns" into a situation of "three equations with three unknowns". He states that the three equations can be solved for the two wavelets and the reflectivity. This is incorrect in case Ziolkowski means *uniquely* solve. Because the wavelets are always bandlimited the reflectivity can always be chosen arbitrarily outside these bandlimitations. This illustrates that the phrase "three equations with three unknowns" is too simplistic as has also been mentioned by Szaraniec (1985). Secondly from Ziolkowski et al. (1980) it becomes clear that for the actual solution of this technique an inverse problem must again be solved. Inversion techniques seem to be an inevitable fundamental tool to learn from observations. Ziolkowski mentions the possibility of refutation by performing a third experiment. This indeed is a test for the combination of the scaling law and the convolutional model. Note that in practice however such an additional measurement would preferably be used to more accurately estimate the basic form of the wavelet and the reflectivity. The data mismatch of the three measurements may then be used for falsification. A choice of the rejection threshold (in close relation to the noise level) has to be made and introduces an arbitrary element in the falsification procedure (as Lakatos was well aware of).

It thus seems that unfortunately the situation in deconvolution is not that simple and a more refined analysis is necessary to fully clear up the picture concerning inverse problems. From an estimation point of view in any case extra data or additional information is always desirable. The author therefore fully agrees with Ziolkowski's statement that the measurement of the wavelet or an additional measurement (in combination with the scaling law) are to be preferred to the situation with only seismic data where we have to resort to statistical assumptions.

Synthetic versus real data tests

A clear distinction between inversion techniques and empirical theories is that the first can be tested (with regard to some aspects) on synthetic data whereas the second obviously can't. In the seismological community tests on real data are always quickly demanded. A few remarks are made on this issue.

There is a very clear difference between tests on synthetic and tests on real data. With synthetic data the perfect answer is known and an estimation result can perfectly be assessed. Synthetic data tests are therefore *the* means to assess items as estimator accuracy and other statistical properties (if these can not be determined analytically). Furthermore it is possible to assess the effects of the violation of certain (often simplifying) assumptions made in the design of the estimator. Tests on synthetic data are therefore very important and should precede tests on real data. The latter are of course also very important but one has to be careful with the evaluation of results. Because on real data the answer is not available the results can not be compared with anything. Only in the worst case that the results are clearly wrong (which in principle can not even occur in Bayesian estimation because of the a priori information) the scheme will fail the test. In all other cases the results may be more or less seriously wrong without the interpreter of the results noticing it. Conclusions concerning real data tests that are often formulated as "the scheme performed well" should more appropriately be formulated as "it could not be proved that the results are wrong". The only way of properly testing on real data is when independent accurate data (in seismics e.g. from a well log) is available but not used in the inversion scheme and the results are compared to it. Such an additional data set was unfortunately not available for the real data set discussed in chapter 8. But even when such a test fails the inversion scheme need not be refuted because there are often many possible reasons for a failure.

Tests of Bayesian estimation

Now tests of the scheme for detailed inversion of seismic data are discussed in more detail. At the basis of the scheme is of course Bayesian estimation theory. This theory is not an

empirical theory and therefore can not be falsified. On the contrary, it is rather a tool with which theories are falsified! All tests discussed in this section are statistical tests, because uncertainties or noise are always present. The choice between the classical approach to estimation and Bayesian estimation therefore has to be made on conceptual and theoretical grounds. The main arguments for choosing Bayesian estimation have been discussed in chapter 1, the main advantage being conceptual clarity and a higher accuracy. Note that for the problem of inversion for acoustic impedance maximum likelihood estimation could not even be used because the results would be perfectly nonunique. Physically meaningless results like negative acoustic impedances are equally likely as normal values as far as the likelihood function is concerned.

It has been mentioned before that Bayesian probability theory has been used by Carnap to erect a theory of inductive logic in the philosophy of science. Although these attempts are generally regarded as having failed Bayesian estimation has again attracted attention in this field. Rosenkrantz published a book entitled "Inference, method and decision", subtitled "towards a Bayesian philosophy of science" (1977), in which Bayesian probability theory plays a basic role. Note that the Bayesian interpretation of probability allows the probability of a theory. A theory can be called confirmed by data when its posterior probability is increased by it. If T denotes a theory then it can easily be derived that the probability of the theory T , given data y is:

$$P(T|y) = \frac{P(T)}{p(y)} \int p(y|x, T) p(x|T) dx \quad , \quad (9.1)$$

where x are free parameters. Rosenkrantz shows how the "average likelihood" measures the support for a theory. He successfully applies this idea to the well known statistical problem of the order determination in a linear least squares problem (which is also discussed in section 7.7, from a more classical viewpoint). The author suspects however that theories can only meaningfully be compared in this way for equal data sets (with equal interpretation). Hence the failure of this program for a general philosophy of science.

Tests of other parts of the scheme for detailed inversion

Two levels of tests can be distinguished. As discussed above and in section 9.2 tests can be routinely built into the inversion scheme. As an additional result besides the parameters certain features of the residuals, like the residual energy, can be predicted. If the residual energy is too high one may regard this as the result of an inconsistent supply of information. As discussed in section 9.2 several premisses may be wrong, and a decision concerning changes has to be made. Popperian tests are thus routinely used in the scheme! As such the procedure can not be refuted; it is itself a refutation procedure.

Nevertheless one may consider another level in which we have the situation where the scheme is applied to real data, a second well is drilled and the results of the well and of the Bayesian scheme do not correspond within reasonable a posteriori uncertainty intervals as they come out of the seismic data inversion scheme and the well log inversion procedure. (In this procedure noisy data from the sonic and density logs are inverted for an acoustic impedance profile). What is the conclusion? There are several possibilities to look for the cause of the inconsistency. The Bayesian inversion procedure as a whole may be wrong, the user may have supplied wrong information, or the results from the well log data are wrong, or any combination of these. It is concluded that a straightforward falsification of the scheme as proposed in this thesis is not easy.

9.5 FINAL CONCLUSIONS

The scheme for detailed inversion of poststack seismic data as set up in this thesis is capable of achieving a much higher resolution than is possible with wavelet deconvolution. Arguments for the utilization of the scheme (in its full form) are at this point in time formed by the basic concepts of the scheme, the results on synthetic data, and the results on real data, which did not show any real problems but which could not be compared to an independent well log data set. These results can in the future be augmented with the results on real data for which two well log data sets are available. Practical success will also strongly depend on a proper implementation of the algorithms with state of the art user interface and graphics software.

APPENDIX A

BAYES' RULE AND UNCERTAIN DATA

In this appendix it is discussed how Bayes' rule can be used for the case of uncertain data. The result is the same as the generalization of Jeffrey's solution as discussed in section 1.14. A mental construct is used which consists of forcing exactly known data which makes Bayes' rule applicable again. This is in fact what is so often done in practice. An observer, reading e.g. a thermometer writes down a value, e.g. 20 °C, while in fact he doubts whether it should not be 19.9 °C or 20.1 °C. The error or "noise" created in this process is treated as an independent additional source of error.

The problem can be solved using the schematic situation given in figure A.1. The vector y_2 denotes the exactly known (forced) data and n_0 denotes the additional errors introduced by forcing the data y_2 . The relation between the noise n_0 , and the data vectors y and y_2 is:

$$y_2 = y + n_0 \quad . \quad (A.1)$$

The errors are introduced as being independent of y so that:

$$p(y_2|y) = p_n(y_2 - y) \quad , \quad (A.2)$$

where p_n denotes the pdf of n_0 . For the whole construction to make sense, an inverse problem solved for y using y_2 should render as an answer the degree of belief of the observer $p_1(y)$ on:

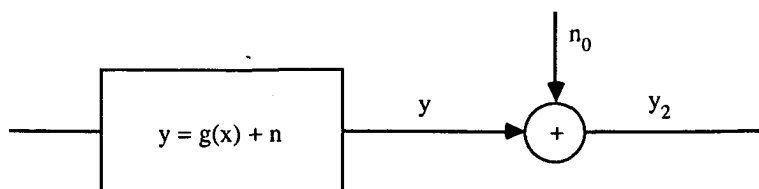


Figure A.1 Schematic of situation with uncertain data y .

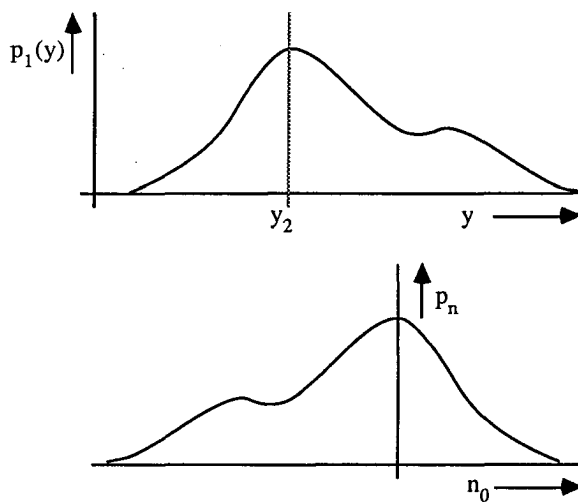


Figure A.2 Relation between the specification of the observation pdf $p_1(y)$ and the corresponding noise pdf p_n .

$$p(y|y_2) = p_1(y) \quad , \quad (\text{A.3})$$

where the pdf's are marginal pdf's, x is disregarded. From Bayes' rule for y and y_2 :

$$p(y_2|y) = \frac{p(y|y_2) p(y_2)}{p(y)} \quad , \quad (\text{A.4})$$

it follows that the noise distribution must be (using (A.2) and (A.3)):

$$p_n(y_2 - y) = \frac{p_1(y) p(y_2)}{p(y)} \quad . \quad (\text{A.5})$$

Choosing $p(y_2)$ equal to $p(y)$ yields:

$$p_n(y_2 - y) = p_1(y) \quad , \quad (\text{A.6})$$

which effectively means that the noise distribution is the reverse of the of $p_1(y)$, see figure

A.2 for an illustration for the univariate case. Note that a value for y_2 can in principle be chosen arbitrarily. A higher chosen value simply shifts the noise pdf up. The thermometer observer could as well specify the data as 30 °C as long as he specifies that the error in observation has a mean of 10 °C (and e.g. a standard deviation of .1 °C). In practice the most convenient choice of data is such that the errors thus introduced have zero mean.

Having set up this mental construction the solution of the inverse problem is found with Bayes' rule using the exact data y_2 :

$$\begin{aligned}
 p(x|y_2) &= \frac{p(y_2|x) p(x)}{p(y_2)} \\
 &= \frac{p(x)}{p(y_2)} \int p(y_2, y|x) dy \\
 &= \frac{p(x)}{p(y_2)} \int p(y_2|y, x) p(y|x) dy \\
 &= \frac{p(x)}{p(y_2)} \int p(y_2|y) p(y|x) dy ,
 \end{aligned} \tag{A.7}$$

where the only condition used is that the observational noise does not depend on x : $p(y_2|y, x) = p(y_2|y)$. Using (A.2) we have:

$$p(x|y_2) = \frac{p(x)}{p(y_2)} \int p_n(y_2 - y) p(y|x) dy . \tag{A.8}$$

The integral is the likelihood function and is equivalent to the solution given in section 1.9 for a situation with two sources of errors. In particular, for independent errors $\mathbf{n}$ we have the convolution of equation (1.79) again. The covariance matrices of $\mathbf{n}$ and $\mathbf{n}_0$ can then be added, which has as a special case the result that for univariate problems the variances of the errors may be added. This last result is a basic "law" taught at first courses in physics.

In order to prove the consistency of this construction with the extension of Jeffrey's approach we can rewrite (A.7) using Bayes' rule as:

$$\begin{aligned}
 p(x|y) &= \int \frac{p(y|y_2) p(y|x) p(x)}{p(y)} dy \\
 &= \int p_1(y) p(x|y) dy .
 \end{aligned} \tag{A.9}$$

This is equivalent to (1.103).

APPENDIX B

THE RESPONSE OF A THIN LAYER

When using the plane wave convolutional model, the response of a thin layer embedded in a homogeneous medium is given by:

$$s(t) = r \left\{ w(t-\tau_1) - w(t-\tau_2) \right\} , \quad (\text{B.1})$$

where r is the reflection coefficient of the two reflectors and τ_1 and τ_2 are the lagtimes. The reflectivity $R(f)$ is given in the frequency domain by:

$$\begin{aligned} R(f) &= r \left\{ e^{-j\omega\tau_1} - e^{-j\omega\tau_2} \right\} \\ &= r e^{-j\omega\tau^0} \left\{ e^{j\omega\Delta\tau} - e^{-j\omega\Delta\tau} \right\} , \end{aligned} \quad (\text{B.2})$$

where τ^0 is the position of the middle of the layer:

$$\tau^0 = \frac{\tau_1 + \tau_2}{2} , \quad (\text{B.3})$$

and $\Delta\tau$ (other than in chapter 1, 2 and 3) is half the thickness:

$$\Delta\tau = \frac{\tau_1 - \tau_2}{2} . \quad (\text{B.4})$$

Using a Taylor expansion for the exponential functions yields:

$$\begin{aligned}
 R &\approx r e^{-j\omega\tau^0} \left\{ \left[1 + j\omega\Delta\tau - \frac{1}{2}(\omega\Delta\tau)^2 - j\frac{(\omega\Delta\tau)^3}{6} \right] \right. \\
 &\quad \left. - \left[1 - j\omega\Delta\tau - \frac{1}{2}(\omega\Delta\tau)^2 + j\frac{(\omega\Delta\tau)^3}{6} \right] \right\} \\
 &= 2j r e^{-j\omega\tau^0} \left\{ \omega\Delta\tau - \frac{(\omega\Delta\tau)^3}{6} \right\} .
 \end{aligned} \tag{B.5}$$

Neglection of the third order term yields:

$$R \approx 2 j \omega \Delta \tau r e^{-j\omega\tau^0} , \tag{B.6}$$

and for the response of the layer in the time domain:

$$\dot{s}(t) \approx 2 r \Delta \tau w'(t - \tau^0) , \tag{B.7}$$

which appears to be the time derivative of the wavelet. The amplitude is proportional to the thickness in time and the reflection coefficient. From equation (B.5) it can be seen that the approximation is fairly good for:

$$\frac{(\omega\Delta\tau)^3}{6} < .3(\omega\Delta\tau) ,$$

or,

$$\Delta\tau < \frac{\sqrt{1.8}}{2\pi f} . \tag{B.8}$$

For $f = 50$ Hz the approximation is valid for thickness $2\Delta\tau < 8.5$ ms.

REFERENCES

- Angeleri, G.P., 1983, A statistical approach to the extraction of the seismic propagating wavelet, *Geophysical Prospecting*, 25, 512-540.
- Bard, Y., 1974, *Nonlinear parameter estimation*, Academic Press.
- Barnett, V., 1982, *Comparative statistical inference*, 2nd edition, John Wiley & Sons.
- Bayes, T., 1763, Essay towards solving a problem in the doctrine of chances, republished in *Biometrika*, 1958, Vol. 45, 293-315.
- Berkhout, A.J., 1974, Related properties of minimum-phase and zero-phase time functions, *Geophysical Prospecting*, 2, 683-709.
- Berkhout, A.J., 1977, Least-squares inverse filtering and wavelet deconvolution, *Geophysics*, 42, 1369-1383.
- Berkhout, A.J., 1984, *Seismic Resolution*, Geophysical Press, Amsterdam.
- Berkhout, A.J., 1985, *Seismic migration, A. theoretical aspects*, 3rd edition, Elsevier.
- Bilgeri, D. and Carlini, A., 1981, Non-linear estimation of reflection coefficients from seismic data, *Geophysical Prospecting*, 29, 672-686.
- Box, G.E.P. and Tiao, G.C., 1973, *Bayesian inference in statistical analysis*, Addison-Wesley.
- Brevik, I. and Berg, E.W., 1986, Maximum likelihood estimation of reflection coefficients from seismic data: a case study, poster presented at 56th annual SEG meeting, Houston.
- Broersen, P.M.T., 1984, Stepwise backward elimination with C_p as selection criterion, Internal Report ST-SV 84-01, Dept. of Applied Physics, Delft University of Technology.
- Broersen, P.M.T., 1986, Subset regression with stepwise directed search, *Journal of the Royal Statistical Society, series C*, vol. 35, no. 2, 168-177.

- Burg, J.P., 1967, Maximum entropy spectral analysis, paper presented at 37th annual SEG meeting, Oklahoma City.
- Carnap R., 1962, Theory of probability.
- Clearbout, J.F., and Muir, F., 1973, Robust modeling 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.
- Cooke, R.M., 1986, Subjective probability and expert opinion, college notes, Delft University of Technology.
- Cox, R.T., 1946, Probability, frequency and reasonable expectation, *American Journal of Physics*, vol. 14, no. 1, 1-13.
- Cox, R.T., 1978, On inference and inquiry, an essay in inductive logic, *in* Levine, R.P. and Tribus, M. (ed), *The maximum entropy formalism*, MIT Press, 119-167.
- Danielson, V. and Karlsson, T.V., 1984, Extraction of signatures from seismic and well data, *First Break*, vol. 2, no 4, 15-21.
- De Finetti, B., 1974, Theory of probability, vol. 1, John Wiley.
- De Voogd, N. and Den Rooijen, H., 1983, Thin-layer response and spectral bandwidth. *Geophysics*, 48, 12-18.
- Feyerabend, P., 1970, Consolations for the specialist, *in*, Lakatos, I. and Musgrave, A. (ed), 1970, *Criticism and the growth of knowledge*, Cambridge University Press.
- Feyerabend, P., 1975, *Against method*, NLB, London.
- Furnival, G.M. and Wilson, W.W., 1974, Regressions by leaps and bounds, *Technometrics*, vol. 16, no 4, 499-511.
- Gardner, G.H.F., Gardner, L.W. and Gregory, A.R., 1974, Formation velocity and density — the diagnostic basics for stratigraphic traps, *Geophysics*, 39, 770-780.
- Gelfand, V. and Lerner, K., 1983, Seismic lithologic modeling, *Leading Edge*, 3, nov., 30-35.
- Gill, P.E., and Murray, W., 1978, Algorithms for the solution of the nonlinear least squares problem, *SIAM Journal of Numerical analysis*, 15, 977-992.
- Gill, P.E., Murray, W., and Wright, M.H., 1981, *Practical Optimization*, Academic Press.

- Gjöstdal, H. and Ursin, B., 1981, Inversion of reflection times in three dimensions, *Geophysics*, 46, 972-983.
- Goutsias, J. and Mendel, J.M., 1986, Maximum-likelihood deconvolution: an optimization theory perspective, *Geophysics*, 51, 1206-1220.
- Hampson, D. and Galbraith, M., 1981, Wavelet extraction by sonic log correlation, paper presented at the 51st SEG meeting.
- Hocking, R.R., 1976, The analysis and selection of variables in linear regression, *Biometrics*, 32, 1-49.
- Hoerl, A.E. and Kennard, R.W., 1970a, Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, vol. 12, no 1, 55-67.
- Hoerl, A.E. and Kennard, R.W., 1970b, Ridge regression: applications to nonorthogonal problems, *Technometrics*, vol. 12, no 1, 69-82.
- Jackson, D.D., 1972, Interpretation of inaccurate, insufficient and inconsistent data, *Geophysical Journal of the Royal Astronomical Society*, 28, 97-109.
- Jackson, D.D., 1973, Marginal solutions to quasi-linear inverse problems in geophysics: the edgohog method, *Geophysical Journal of the Royal Astronomical Society*, 35, 121-136.
- Jackson, D.D., 1976, Most-squares inversion, *Journal of Geophysical Research*, 81, 1027-1030.
- Jackson, D.D., 1979, The use of a priori data to resolve nonuniqueness in linear inversion, *Geophysical Journal of the Royal Astronomical Society*, 57, 137-157.
- Jackson, D.D. and Matsu'ura, M., 1985, A Bayesian approach to nonlinear inversion, *Journal of Geophysical Research*, 90, 581- 591.
- Jaynes, E.T., 1963, Information theory and statistical mechanics, *in*, Ford, K.W. (ed), *Statistical Physics*, vol. 3, W.A. Benjamin Inc., p. 182-218.
- Jaynes, E.T., 1968, Prior probabilities, *IEEE transactions on systems science and cybernetics*, vol. ssc-4, 227-241.
- Jaynes, E.T., 1985, Where do we go from here? *in* Ray Smith, C., and Grandy, W.T. (eds), *Maximum-Entropy and Bayesian methods in inverse problems*, D. Reidel Publishing Company, 21-58.
- Jeffrey, R.C., 1983, *The logic of decision*, 2nd edition, University of Chicago Press, first published 1965.
- Jeffreys, H., 1939, *Theory of probability*, Clarendon Press, (3rd edition 1983).

- Jeffreys, H., 1957, *Scientific inference*, Cambridge University Press.
- Kallweit, R.S. and Wood, L.C., 1982, The limits of resolution of zero-phase wavelets, *Geophysics*, 47, 1035-1046.
- Kaman, E.J., Van Riel, P., and Duijndam, A.J.W., 1984, Detailed inversion of reservoir data by constrained parameter estimation and resolution analysis, paper presented at the 54th annual SEG meeting, Atlanta.
- Kaman, E.J., Protais, J.C., Van Riel P. and Young, I.T., 1987, The application of pattern recognition in detailed model inversion, *in* Aminzadeh, F. (ed), *Handbook of geophysical exploration, section I, seismic exploration*, vol. 20, Pattern recognition and image processing, Geophysical Press, 312-335.
- Kennett, B.L.N., 1978, Some aspects of nonlinearity in inversion, *Geophysical Journal of the Royal Astronomical Society*, 55, 373-391.
- Keynes, J.M., 1929, *A treatise on probability*, MacMillan, London.
- Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P., 1983, Optimization by simulated annealing, *Science*, 220, 671-680.
- Koefoed, O., 1981, Aspects of vertical seismic resolution, *Geophysical Prospecting*, 29, 21-30.
- Kuhn, T., 1962, *The structure of scientific revolutions*, University of Chicago Press.
- Kuhn, T., 1970, Reflections on my critics, *in*, Lakatos, I. and Musgrave, A. (ed), *Criticism and the growth of knowledge*, Cambridge University Press.
- Lakatos, I., 1970, Falsification and the methodology of scientific research programs, *in*, Lakatos, I. and Musgrave, A. (ed), *Criticism and the growth of knowledge*, Cambridge University Press.
- Lanczos, C., 1961, *Linear differential operators*, D. Van Nostrand Co.
- Lawson, C.L. and Hanson, R.J., 1974, *Solving least squares problems*, Prentice-Hall.
- Levy, S. and Fullager, P.K. 1981, Reconstruction of a sparse spike train from a portion of its spectrum, and application to high resolution deconvolution, *Geophysics*, 46, 1235-1243.
- Lindley, D.V., 1972, Bayesian statistics, a review, *SIAM*.
- Lines, L.R. and Treitel, S., 1985, Wavelets, well logs and Wiener filters, *First Break*, vol. 3, no. 8, 9-14.

- Lines, L.R. and Ulrych, T.J., 1977, The old and the new in seismic deconvolution and wavelet estimation, *Geophysical Prospecting*, 25, 512-540.
- Mallows, C.L., 1973, Some comments on Cp, *Technometrics*, 15, 661-675.
- Marquardt, D.W., 1970, Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation, *Technometrics*, vol. 12, no. 3, 591-612.
- Matsu'ura, M. and Hirata, N., 1982, Generalised least-squares solutions to quasi-linear inverse problems with a priori information, *Journal Phys. Earth*, 30, 451-468.
- Mendel, J.M., 1983, *Optimal Seismic deconvolution*, Academic Press.
- Mendel, J.M. and Goutsias, J., 1986, One-dimensional normal-incidence inversion: a solution procedure for band-limited and noisy data, *proceedings of the IEEE*, vol. 74, no. 3, 401-414.
- Menke, W., 1984, *Geophysical data analysis: discrete inverse theory*, Academic Press.
- Miller, A.J., 1984, Selection of subsets of regression variables, *Journal of the Royal Statistical Society A*, 147, 389- 425.
- 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 edition, McGraw-Hill.
- Oldenburg, D.W., Levy, S. and Whittall, K.P., 1981, Wavelet estimation and deconvolution, *Geophysics*, 46, 1528-1542.
- Oldenburg, D.W., Levy, S. and Stinson, J.S., 1986, Inversion of bandlimited reflection seismograms: theory and practice, *proceedings of the IEEE*, vol. 74, no. 3, 487-497.
- Oldenburg, D.W., Scheuer, T. and Levy, S., 1983, Recovery of the acoustic impedance from reflection seismograms, *Geophysics*, 48, 1318-1337.
- Özdemir, H. 1985, Maximum likelihood estimation of seismic reflection coefficients, *Geophysical prospecting*, 33, 828-860.
- Papoulis, A., 1965, *Probability, random variables and stochastic processes*, McGraw-Hill.
- Parasnis, D.S., 1980, A Popperian look at geophysics, *Geophysical Prospecting*, 28, 667-673.
- Parker, R.L., 1977, Understanding inverse theory, *Ann. Rev. Earth Plan. Sci.*, 35-64.
- Peterka, V., 1980, *Bayesian approach to systems identification*, Institute of information theory and automation, Prague.

- Popper, K.R., 1959, The logic of scientific discovery, Hutchinson, first published in 1934 as "Logik der Forschung".
- Popper, K.R., 1962, Conjectures and refutations: the growth of scientific knowledge, Basic Books.
- Pous, J., Marcuello, A. and Queralt, P., 1987, Resistivity inversion with a priori information, Geophysical Prospecting, 35, 590-603.
- Ray Smith, C. and Grandy, W.T. (ed), 1985, Maximum entropy and Bayesian methods in inverse problems, D. Reidel Publishing Company, Dordrecht.
- Rietsch, E., 1977, The maximum entropy approach to inverse problems, Journal of Geophysics, 42, 489-506.
- Robertson, J.D. and Nogami, H.H., 1984, Complex trace analysis of thin beds, Geophysics, 49, 344-352.
- Rosenkrantz, R.D., 1977, Inference, method and decision, D.Reidel Publishing Company, Dordrecht.
- Rosland, B.O. and Sandvin, O.A., 1987, Estimation of pulse and Q-factor at well locations using global minimization, paper presented at the 49th EAEG meeting, Belgrade.
- 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.
- Sarwar, A.K.M., and Smith, D.L., 1987, Wave scattering deconvolution by seismic inversion, Geophysical Prospecting, 35, 491-501.
- 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.
- Stone, D.G., 1976, Robust wavelet estimation by structural deconvolution, paper presented at 46th SEG meeting, Houston.
- Szaraniec, E., 1985, More comments on "Further thoughts on Popperian geophysics - the example of deconvolution" by A. Ziolkowski, Geophysical Prospecting, 33, 895-899.

- Tarantola, A., 1984, Linearized inversion of seismic reflection data, *Geophysical Prospecting*, 32, 998-1015.
- 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, no. 2, 219-232.
- Taylor, H.L., Banks, S.C. and McCoy, J.F., 1979, Deconvolution with the l_1 norm, *Geophysics*, 44, 39-52.
- Telford, W.M., Geldart, L.P., Sheriff, R.E. and Keys, D.A., 1976, *Applied geophysics*, Cambridge University Press.
- Ursin, B. and Berteussen, K.A., 1986, Comparison of some inverse methods for wave propagation in layered media, *Proceedings of the IEEE*, 74, 389-400.
- Ursin, B. and Holberg, O., 1985, Maximum likelihood estimation of seismic impulse responses, *Geophysical Prospecting*, 33, 233- 251.
- Ursin, B. and Zheng, Y., 1985, Identification of seismic reflections using singular value decomposition, *Geophysical Prospecting*, 33, 773-799.
- Van der Made, P.M., Van Riel, P., and Berkhout, A.J., 1984, Velocity and subsurface geometry inversion by parameter estimation in complex inhomogeneous media, paper presented at the 54th annual SEG meeting, Atlanta.
- Van der Schoot, A., Larson, D.E., Romijn, R. and Berkhout, A.J., 1987, Pre-stack migration by shot record inversion and common- depth-point stacking: a case study, paper presented at the 49th EAEG meeting, Belgrade.
- Van Riel, P., 1982, Seismic trace inversion, Msc. Thesis, Delft University of Technology.
- Van Riel, P. and Berkhout, A.J., 1985, Resolution in seismic trace inversion by parameter estimation, *Geophysics*, 50, 1440- 1455.
- Von Mises, R., 1936, *Wahrscheinlichkeit, Statistik und Wahrheit*, Springer Verlag.
- Von Mises, R., 1957, *Probability, Statistics and Truth*, 2nd English edition, George Allen & Unwin, London.

- Walden, A.F. and White, R.E., 1984, On errors of fit and accuracy in matching synthetic seismograms and seismic traces, *Geophysical Prospecting*, 32, 871-891.
- Walker, C. and Ulrich, T.J., 1983, Autoregressive recovery of the acoustic impedance, *Geophysics*, 48, 1338-1350.
- White, R.E., 1980, Partial coherence matching of synthetic seismograms with seismic traces, *Geophysical Prospecting*, 28, 333-358.
- Widess, M.B., 1973, How thin is a thin bed?, *Geophysics*, 38, 1176-1180.
- Wiggins, R.A., 1972, The general linear inverse problem: implication of surface waves and free oscillations for earth structure, *Reviews of Geophysics and Space physics*, vol. 10, no 1, 251-285.
- Wiggins, R., Lerner, K., and Wisecup, D., 1976, Residual statics analysis as a general linear inverse problem, *Geophysics*, 41, 922-938.
- Williams, P.M., 1980, Bayesian conditionalisation and the principle of minimum information, *British Journal of the Philosophy of Science*, 31, 131-144.
- Ziolkowski, A.M., 1982, Further thoughts on Popperian geophysics - the example of deconvolution, *Geophysical Prospecting*, 30, 155-165.
- Ziolkowski, A.M., Lerwill, W.E., March, D.W. and Peardon, L.G., 1980, Wavelet deconvolution using a source scaling law, *Geophysical Prospecting*, 28, 872-901.

SUMMARY

In this thesis a technique is described for the detailed inversion of poststack seismic data. It is based on Bayesian parameter estimation and aims at a resolution much higher than is achievable with wavelet deconvolution.

In Bayesian estimation the answer to an inverse problem is given by the so-called a posteriori probability density function, which represents the degree of belief on the parameter values after the data have been obtained. It is proportional to the likelihood function which reflects information on the parameters obtained from the data and the a priori probability density function, representing a degree of belief on the parameters before the data are obtained. Because inspection of the complete a posteriori probability density function is impossible even for a moderate number of parameters its maximum is taken as a so-called point estimator of the parameters (MAP estimation). The mathematical form of the estimator depends on the type of the probability density functions chosen for noise and a priori information. For the Gaussian distributions used in this thesis the well known least squares estimator results.

For nonlinear forward models, describing the theoretical relations between the parameters and the data, the MAP estimate has to be determined with optimization techniques. For the problems in this thesis Newton methods in general and special nonlinear least squares methods in particular are best suited.

The results of the MAP estimator should be augmented with an uncertainty analysis. Such a procedure can provide the a posteriori uncertainties in the parameter values and a comparison of the relative contributions of data and a priori information to the solution of the inverse problem. Several approaches can be followed, depending on the amount of information desired. Standard deviations can be used as overall uncertainty bounds. From covariance matrices additional information concerning the correlations between the parameters can be obtained. The relevant density functions can also be scanned along essential or other desired directions through the parameter space.

Bayesian estimation is applied to the detailed inversion of poststack data. The acoustic impedances and the thickness in traveltime of geological layers in a certain target zone are estimated, using the one-dimensional convolutional model. This can be done very efficiently on a single trace basis. A seismic section is inverted trace by trace, starting at a well location where the number of layers and a priori information is available. General geological knowledge can also be incorporated as a priori information. The utilization of a priori information renders unique and stable results.

A rather straightforward extension of single trace inversion is inversion on a multi trace basis. The lateral variations of acoustic impedances and lagtimes can be described by various spline functions. The parameters of these functions are estimated, inverting a block of traces at a time. The scheme is less efficient but the results are superior.

In practice the wavelet will have to be estimated. A time domain least squares technique is discussed which assumes the reflectivity to be known. Stabilization methods like ridge regression, the SVD cutoff method and parameter selection can improve the results. The estimates are sensitive to reflectivity errors.

Inversion for impedance and the wavelet can be integrated in a consistent Bayesian scheme as well as in an iterative scheme. Results on synthetic data are excellent. Tests on real data are performed and give satisfactory results.

In the last chapter of this thesis refinements and extensions of the scheme are given. A discussion is given on some fundamental issues concerning testing of inversion schemes.

SAMENVATTING

Gedetailleerde Bayesiaanse inversie van seismische data

In dit proefschrift wordt een techniek voor gedetailleerde inversie van poststack seismische data beschreven. De techniek is gebaseerd op Bayesiaanse parameter schatting en beoogt een resolutie te halen die veel hoger is dan die van wavelet deconvolutie.

In Bayesiaanse parameterschattingstechnieken wordt de oplossing van een inversie-probleem gegeven door de zogenaamde a posteriori verdelingsdichtheid. Deze is recht evenredig met de aannemelijkheidsfunctie, welke de mate van geloof betreffende de parameter waarden weergeeft nadat de meetresultaten beschikbaar zijn, en de a priori verdelingsdichtheid, welke de mate van geloof weergeeft voordat de meetresultaten beschikbaar zijn. Omdat inspectie van de complete a posteriori verdelingsdichtheidfunctie zelfs voor een bescheiden aantal parameters onmogelijk is wordt het maximum van deze verdeling als een puntschatter genomen (MAP schatting). De mathematische vorm van deze schatter hangt af van het type verdelingsdichtheidfunctie die voor de ruis en de a priori informatie worden gekozen. Voor Gaussische verdelingen, zoals gebruikt in dit proefschrift, resulteert dit in de bekende kleinste kwadraten schatter.

Voor niet-lineaire voorwaartse modellen, welke de theoretische relaties tussen de parameters en de meetresultaten beschrijven, moet de MAP schatting berekend worden met optimalisatie technieken. Voor de problemen in dit proefschrift zijn Newton methoden in het algemeen en speciale methoden voor niet-lineaire kleinste kwadraten methoden het best geschikt.

De resultaten van de MAP schatter dienen aangevuld te worden met een onzekerheids-analyse. Zo'n procedure kan de a posteriori onzekerheden in de parameter waarden verschaffen alsmede een vergelijking opleveren van de verschillende bijdragen van de data en de a priori kennis tot de oplossing van het inversieprobleem.

Bayesiaanse parameterschatting wordt toegepast op de gedetailleerde inversie van poststack data. De akoestische impedantie en de dikte van geologische lagen in een bepaalde doelzone worden geschat, waarbij het één-dimensionale konvolutie model gebruikt wordt. Dit kan zeer efficiënt gebeuren op basis van enkelvoudige sporen. Een seismische sectie kan spoor na spoor geïnverteerd worden, beginnend bij een boorput lokatie, waar het aantal lagen en a priori informatie beschikbaar is. Algemene geologische kennis kan ook als a priori informatie gebruikt worden. Het gebruik van a priori informatie levert unieke en stabiele resultaten op.

Een directe uitbreiding van enkelvoudige spoor-inversie is inversie op basis van meervoudige sporen. De laterale variaties van akoestische impedanties en looptijden kunnen beschreven worden door verscheidene "spline" functies. De parameters van deze spline functies worden geschat, waarbij een aantal sporen simultaan wordt geïnverteerd. Dit schema is minder efficiënt maar de resultaten zijn superieur.

In de praktijk zal het wavelet moeten worden geschat. Een kleinste kwadraten methode in het tijddomein wordt besproken, waarbij aangenomen wordt dat de reflektiviteit bekend is. Stabilisatie methoden zoals "ridge regression", de SVD afkapmethode en parameter-selektie kunnen de resultaten verbeteren. De schattingen zijn gevoelig voor fouten in de reflektiviteit.

Inversie voor impedantie en het wavelet kunnen geïntegreerd worden in een consistente Bayesiaanse procedure alsmede in een iteratieve procedure. De resultaten op synthetische data zijn uitstekend. Tests op echte data zijn uitgevoerd en geven bevredigende resultaten.

In het laatste hoofdstuk worden verfijningen en uitbreidingen van de procedure besproken. Een verhandeling wordt gegeven over enkele fundamentele punten betreffende het testen van inversie procedures.

CURRICULUM VITAE

Name : Duijndam, Adrianus, Joseph, Willibrordus
 Born : March 1, 1956
 Nationality : Dutch

EDUCATION

1962 – 1968 primary school
 1968 – 1974 secondary school
 1974 – 1981 Delft University of Technology
 (Applied Physics)
 1980: B.S
 medical ultrasound
 thesis: Two-dimensional focussing with a linear array
 1981: M.Sc.
 medical ultrasound
 thesis: Wide-band simulation of wavefields and acoustic
 design of an electronic sectorscanner
 supervised by Prof. A.J. Berkhout

EMPLOYMENT

1978 – 1981 Research assistant with the group of experimental cardiology,
 Erasmus University Rotterdam
 1981 – 1983 Research scientist with the Dutch Meteorological Institute,
 group of instrument development
 1983 – 1985 Research scientist with institute of applied geoscience, TNO,
 group of seismics. Spent in the research project "Princeps" with
 the group of seismics and acoustics, Delft University
 1985 – present Research scientist with Delft Geophysical