

INTERNATIONAL GEOPHYSICS SERIES • VOLUME 45

Geophysical Data Analysis: Discrete Inverse Theory

Revised Edition

William Menke



GEOPHYSICAL DATA ANALYSIS:
DISCRETE INVERSE THEORY

This is Volume 45 in
INTERNATIONAL GEOPHYSICS SERIES
A series of monographs and textbooks
Edited by RENATA DMOWSKA and JAMES R. HOLTON

A complete list of the books in this series appears at the end of this volume.

GEOPHYSICAL DATA ANALYSIS: DISCRETE INVERSE THEORY

Revised Edition

William Menke

Lamont-Doherty Geological Observatory and
Department of Geological Sciences
Columbia University
Palisades, New York

Formerly of
College of Oceanography
Oregon State University



ACADEMIC PRESS, INC.
Harcourt Brace Jovanovich, Publishers
San Diego New York Berkeley Boston
London Sydney Tokyo Toronto

Copyright © 1989, 1984 by Academic Press, Inc.

All Rights Reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher.

Academic Press, Inc.
San Diego, California 92101

United Kingdom Edition published by
Academic Press Limited
24-28 Oval Road, London NW1 7DX

Library of Congress Cataloging-in-Publication Data

Menke, William

Geophysical data analysis : discrete inverse theory / William Menke. -- Rev. ed.

p. cm.

Bibliography: p.

Includes index.

ISBN 0-12-490921-3 (alk. paper)

1. Geophysics--Measurement. 2. Oceanography--Measurement.

3. Inverse problems (Differential equations)--Numerical solutions.

1. Title. II. Series.

QC802.A1M46 1989

551--dc20

89-31224

CIP

Printed in the United States of America

89 90 91 92 9 8 7 6 5 4 3 2 1

CONTENTS

PREFACE xi

INTRODUCTION 1

1 DESCRIBING INVERSE PROBLEMS

1.1	Formulating Inverse Problems	7
1.2	The Linear Inverse Problem	9
1.3	Examples of Formulating Inverse Problems	10
1.4	Solutions to Inverse Problems	17

2 SOME COMMENTS ON PROBABILITY THEORY

2.1	Noise and Random Variables	21
2.2	Correlated Data	24
2.3	Functions of Random Variables	27
2.4	Gaussian Distributions	29
2.5	Testing the Assumption of Gaussian Statistics	31
2.6	Confidence Intervals	33

3 SOLUTION OF THE LINEAR, GAUSSIAN INVERSE PROBLEM, VIEWPOINT 1: THE LENGTH METHOD

3.1	The Lengths of Estimates	35
3.2	Measures of Length	36
3.3	Least Squares for a Straight Line	39

3.4	The Least Squares Solution of the Linear Inverse Problem	40
3.5	Some Examples	42
3.6	The Existence of the Least Squares Solution	45
3.7	The Purely Underdetermined Problem	48
3.8	Mixed-Determined Problems	50
3.9	Weighted Measures of Length as a Type of A Priori Information	52
3.10	Other Types of A Priori Information	55
3.11	The Variance of the Model Parameter Estimates	58
3.12	Variance and Prediction Error of the Least Squares Solution	58
4	SOLUTION OF THE LINEAR, GAUSSIAN INVERSE PROBLEM, VIEWPOINT 2: GENERALIZED INVERSES	
4.1	Solutions versus Operators	61
4.2	The Data Resolution Matrix	62
4.3	The Model Resolution Matrix	64
4.4	The Unit Covariance Matrix	65
4.5	Resolution and Covariance of Some Generalized Inverses	66
4.6	Measures of Goodness of Resolution and Covariance	67
4.7	Generalized Inverses with Good Resolution and Covariance	68
4.8	Sidelobes and the Backus-Gilbert Spread Function	71
4.9	The Backus-Gilbert Generalized Inverse for the Underdetermined Problem	73
4.10	Including the Covariance Size	75
4.11	The Trade-off of Resolution and Variance	76
5	SOLUTION OF THE LINEAR, GAUSSIAN INVERSE PROBLEM, VIEWPOINT 3: MAXIMUM LIKELIHOOD METHODS	
5.1	The Mean of a Group of Measurements	79
5.2	Maximum Likelihood Solution of the Linear Inverse Problem	82
5.3	A Priori Distributions	83
5.4	Maximum Likelihood for an Exact Theory	87
5.5	Inexact Theories	89

5.6	The Simple Gaussian Case with a Linear Theory	91
5.7	The General Linear, Gaussian Case	92
5.8	Equivalence of the Three Viewpoints	95
5.9	The F Test of Error Improvement Significance	96
5.10	Derivation of the Formulas of Section 5.7	97

6 NONUNIQUENESS AND LOCALIZED AVERAGES

6.1	Null Vectors and Nonuniqueness	101
6.2	Null Vectors of a Simple Inverse Problem	102
6.3	Localized Averages of Model Parameters	103
6.4	Relationship to the Resolution Matrix	104
6.5	Averages versus Estimates	105
6.6	Nonunique Averaging Vectors and A Priori Information	106

7 APPLICATIONS OF VECTOR SPACES

7.1	Model and Data Spaces	109
7.2	Householder Transformations	111
7.3	Designing Householder Transformations	115
7.4	Transformations That Do Not Preserve Length	117
7.5	The Solution of the Mixed-Determined Problem	118
7.6	Singular-Value Decomposition and the Natural Generalized Inverse	119
7.7	Derivation of the Singular-Value Decomposition	124
7.8	Simplifying Linear Equality and Inequality Constraints	125
7.9	Inequality Constraints	126

8 LINEAR INVERSE PROBLEMS AND NON-GAUSSIAN DISTRIBUTIONS

8.1	L_1 Norms and Exponential Distributions	133
8.2	Maximum Likelihood Estimate of the Mean of an Exponential Distribution	135
8.3	The General Linear Problem	137
8.4	Solving L_1 Norm Problems	138
8.5	The L_∞ Norm	141

9 NONLINEAR INVERSE PROBLEMS

9.1	Parameterizations	143
9.2	Linearizing Parameterizations	147
9.3	The Nonlinear Inverse Problem with Gaussian Data	147
9.4	Special Cases	153
9.5	Convergence and Nonuniqueness of Nonlinear L_2 Problems	153
9.6	Non-Gaussian Distributions	156
9.7	Maximum Entropy Methods	160

10 FACTOR ANALYSIS

10.1	The Factor Analysis Problem	161
10.2	Normalization and Physicality Constraints	165
10.3	Q -Mode and R -Mode Factor Analysis	167
10.4	Empirical Orthogonal Function Analysis	167

11 CONTINUOUS INVERSE THEORY AND TOMOGRAPHY

11.1	The Backus–Gilbert Inverse Problem	171
11.2	Resolution and Variance Trade-off	173
11.3	Approximating Continuous Inverse Problems as Discrete Problems	174
11.4	Tomography and Continuous Inverse Theory	176
11.5	Tomography and the Radon Transform	177
11.6	The Fourier Slice Theorem	178
11.7	Backprojection	179

12 SAMPLE INVERSE PROBLEMS

12.1	An Image Enhancement Problem	183
12.2	Digital Filter Design	187
12.3	Adjustment of Crossover Errors	190
12.4	An Acoustic Tomography Problem	194
12.5	Temperature Distribution in an Igneous Intrusion	198
12.6	L_1 , L_2 , and L_∞ Fitting of a Straight Line	202
12.7	Finding the Mean of a Set of Unit Vectors	207
12.8	Gaussian Curve Fitting	210
12.9	Earthquake Location	213
12.10	Vibrational Problems	217

13 NUMERICAL ALGORITHMS

13.1	Solving Even-Determined Problems	222
13.2	Inverting a Square Matrix	229
13.3	Solving Underdetermined and Overdetermined Problems	231
13.4	L_2 Problems with Inequality Constraints	240
13.5	Finding the Eigenvalues and Eigenvectors of a Real Symmetric Matrix	251
13.6	The Singular-Value Decomposition of a Matrix	254
13.7	The Simplex Method and the Linear Programming Problem	256

14 APPLICATIONS OF INVERSE THEORY TO GEOPHYSICS

14.1	Earthquake Location and the Determination of the Velocity Structure of the Earth from Travel Time Data	261
14.2	Velocity Structure from Free Oscillations and Seismic Surface Waves	265
14.3	Seismic Attenuation	267
14.4	Signal Correlation	267
14.5	Tectonic Plate Motions	268
14.6	Gravity and Geomagnetism	269
14.7	Electromagnetic Induction and the Magnetotelluric Method	270
14.8	Ocean Circulation	271

APPENDIX A: Implementing Constraints with Lagrange
Multipliers 273

APPENDIX B: L_2 Inverse Theory with Complex
Quantities 275

REFERENCES 277

INDEX 281

INTERNATIONAL GEOPHYSICS SERIES 288

This page intentionally left blank

PREFACE

Every researcher in the applied sciences who has analyzed data has practiced inverse theory. Inverse theory is simply the set of methods used to extract useful inferences about the world from physical measurements. The fitting of a straight line to data involves a simple application of inverse theory. Tomography, popularized by the physician's CAT scanner, uses it on a more sophisticated level.

The study of inverse theory, however, is more than the cataloging of methods of data analysis. It is an attempt to organize these techniques, to bring out their underlying similarities and pin down their differences, and to deal with the fundamental question of the limits of information that can be gleaned from any given data set.

Physical properties fall into two general classes: those that can be described by discrete parameters (e.g., the mass of the earth or the position of the atoms in a protein molecule) and those that must be described by continuous functions (e.g., temperature over the face of the earth or electric field intensity in a capacitor). Inverse theory employs different mathematical techniques for these two classes of parameters: the theory of matrix equations for discrete parameters and the theory of integral equations for continuous functions.

Being introductory in nature, this book deals only with "discrete

inverse theory,” that is, the part of the theory concerned with parameters that either are truly discrete or can be adequately approximated as discrete. By adhering to these limitations, inverse theory can be presented on a level that is accessible to most first-year graduate students and many college seniors in the applied sciences. The only mathematics that is presumed is a working knowledge of the calculus and linear algebra and some familiarity with general concepts from probability theory and statistics.

The treatment of inverse theory in this book is divided into four parts. Chapters 1 and 2 provide a general background, explaining what inverse problems are and what constitutes their solution as well as reviewing some of the basic concepts from probability theory that will be applied throughout the text. Chapters 3–7 discuss the solution of the canonical inverse problem: the linear problem with Gaussian statistics. This is the best understood of all inverse problems; and it is here that the fundamental notions of uncertainty, uniqueness, and resolution can be most clearly developed. Chapters 8–11 extend the discussion to problems that are non-Gaussian and nonlinear. Chapters 12–14 provide examples of the use of inverse theory and a discussion of the numerical algorithms that must be employed to solve inverse problems on a computer.

Many people helped me write this book. I am very grateful to my students at Columbia University and at Oregon State University for the helpful comments they gave me during the courses I have taught on inverse theory. Critical readings of the manuscript were undertaken by Leroy Dorman, L. Neil Frazer, and Walt Pilant; I thank them for their advice and encouragement. I also thank my copyeditor, Ellen Drake, draftsperson, Susan Binder, and typist, Lyn Shaterian, for the very professional attention they gave to their respective work. Finally, I thank the many hundreds of scientists and mathematicians whose ideas I drew upon in writing this book.

INTRODUCTION

Inverse theory is an organized set of mathematical techniques for reducing data to obtain useful information about the physical world on the basis of inferences drawn from observations. Inverse theory, as we shall consider it in this book, is limited to observations and questions that can be represented numerically. The observations of the world will consist of a tabulation of measurements, or “data.” The questions we want to answer will be stated in terms of the numerical values (and statistics) of specific (but not necessarily directly measurable) properties of the world. These properties will be called “model parameters” for reasons that will become apparent. We shall assume that there is some specific method (usually a mathematical theory or model) for relating the model parameters to the data.

The question, What causes the motion of the planets?, for example, is not one to which inverse theory can be applied. Even though it is perfectly scientific and historically important, its answer is not numerical in nature. On the other hand, inverse theory can be applied to the question, Assuming that Newtonian mechanics applies, determine the number and orbits of the planets on the basis of the observed orbit of Halley’s comet. The number of planets and their orbital ephemerides

are numerical in nature. Another important difference between these two problems is that the first asks us to determine the reason for the orbital motions, and the second presupposes the reason and asks us only to determine certain details. Inverse theory rarely supplies the kind of insight demanded by the first question; it always demands that the physical model be specified beforehand.

The term “inverse theory” is used in contrast to “forward theory,” which is defined as the process of predicting the results of measurements (predicting data) on the basis of some general principle or model and a set of specific conditions relevant to the problem at hand. Inverse theory, roughly speaking, addresses the reverse problem: starting with data and a general principle or model, it determines estimates of the model parameters. In the above example, predicting the orbit of Halley’s comet from the presumably well-known orbital ephemerides of the planets is a problem for forward theory.

Another comparison of forward and inverse problems is provided by the phenomenon of temperature variation as a function of depth beneath the earth’s surface. Let us assume that the temperature increases linearly with depth in the earth; that is, temperature T is related to depth z by the rule $T(z) = az + b$, where a and b are numerical constants. If one knows that $a = 0.1$ and $b = 25$, then one can solve the forward problem simply by evaluating the formula for any desired depth. The inverse problem would be to determine a and b on the basis of a suite of temperature measurements made at different depths in, say, a bore hole. One may recognize that this is the problem of fitting a straight line to data, which is a substantially harder problem than the forward problem of evaluating a first-degree polynomial. This brings out a property of most inverse problems: that they are substantially harder to solve than their corresponding forward problems.

Forward problem:

model parameters → model → prediction of data

Inverse problem:

data → model → estimates of model parameters

Note that the role of inverse theory is to provide information about unknown numerical parameters that go into the model, not to provide the model itself. Nevertheless, inverse theory can often provide a means for assessing the correctness of a given model or of discriminating between several possible models.

The model parameters one encounters in inverse theory vary from discrete numerical quantities to continuous functions of one or more variables. The intercept and slope of the straight line mentioned above are examples of discrete parameters. Temperature, which varies continuously with position, is an example of a continuous function. This book deals only with discrete inverse theory, in which the model parameters are represented as a set of a finite number of numerical values. This limitation does not, in practice, exclude the study of continuous functions, since they can usually be adequately approximated by a finite number of discrete parameters. Temperature, for example, might be represented by its value at a finite number of closely spaced points or by a set of splines with a finite number of coefficients. This approach does, however, limit the rigor with which continuous functions can be studied. Parameterizations of continuous functions are always both approximate and, to some degree, arbitrary properties, which cast a certain amount of imprecision into the theory. Nevertheless, discrete inverse theory is a good starting place for the study of inverse theory in general, since it relies mainly on the theory of vectors and matrices rather than on the somewhat more complicated theory of continuous functions and operators. Furthermore, careful application of discrete inverse theory can often yield considerable insight, even when applied to problems involving continuous parameters.

Although the main purpose of inverse theory is to provide estimates of model parameters, the theory has a considerably larger scope. Even in cases in which the model parameters are the only desired results, there is a plethora of related information that can be extracted to help determine the “goodness” of the solution to the inverse problem. The actual values of the model parameters are indeed irrelevant in cases when we are mainly interested in using inverse theory as a tool in experimental design or in summarizing the data. Some of the questions inverse theory can help answer are the following.

- (a) What are the underlying similarities among inverse problems?
- (b) How are estimates of model parameters made?
- (c) How much of the error in the measurements shows up as error in the estimates of the model parameters?
- (d) Given a particular experimental design, can a certain set of model parameters really be determined?

These questions emphasize that there are many different kinds of answers to inverse problems and many different criteria by which the

goodness of those answers can be judged. Much of the subject of inverse theory is concerned with recognizing when certain criteria are more applicable than others, as well as detecting and avoiding (if possible) the various pitfalls that can arise.

Inverse problems arise in many branches of the physical sciences. An incomplete list might include such entries as

- (a) medical tomography,
- (b) image enhancement,
- (c) curve fitting,
- (d) earthquake location,
- (e) factor analysis,
- (f) determination of earth structure from geophysical data,
- (g) satellite navigation,
- (h) mapping of celestial radio sources with interferometry, and
- (i) analysis of molecular structure by x-ray diffraction.

Inverse theory was developed by scientists and mathematicians having various backgrounds and goals. Thus, although the resulting versions of the theory possess strong and fundamental similarities, they have tended to look, superficially, very different. One of the goals of this book is to present the various aspects of discrete inverse theory in such a way that both the individual viewpoints and the “big picture” can be clearly understood.

There are perhaps three major viewpoints from which inverse theory can be approached. The first and oldest sprang from probability theory—a natural starting place for such “noisy” quantities as observations of the real world. In this version of inverse theory the data and model parameters are treated as random variables, and a great deal of emphasis is placed on determining the probability distributions that they follow. This viewpoint leads very naturally to the analysis of error and to tests of the significance of answers.

The second viewpoint developed from that part of the physical sciences that retains a deterministic stance and avoids the explicit use of probability theory. This approach has tended to deal only with estimates of model parameters (and perhaps with their error bars) rather than with probability distributions per se. Yet what one means by an estimate is often nothing more than the expected value of a probability distribution; the difference is only one of emphasis.

The third viewpoint arose from a consideration of model parameters that are inherently continuous functions. Whereas the other two

viewpoints handled this problem by approximating continuous functions with a finite number of discrete parameters, the third developed methods for handling continuous function explicitly. Although continuous inverse theory is not within the scope of this book, many of the concepts originally developed for it have application to discrete inverse theory, especially when it is used with discretized continuous functions.

This book is written at a level that might correspond to a first graduate course in inverse theory for quantitative applied scientists. Although inverse theory is a mathematical subject, an attempt has been made to keep the mathematical treatment self-contained. With a few exceptions, only a working knowledge of the calculus and matrix algebra is presumed. Nevertheless, the treatment is in no sense simplified. Realistic examples, drawn from the scientific literature, are used to illustrate the various techniques. Since in practice the solutions to most inverse problems require substantial computational effort, attention is given to the kinds of algorithms that might be used to implement the solutions on a modern digital computer.

This page intentionally left blank

1

DESCRIBING INVERSE PROBLEMS

1.1 Formulating Inverse Problems

The starting place in most inverse problems is a description of the data. Since in most inverse problems the data are simply a table of numerical values, a vector provides a convenient means of their representation. If N measurements are performed in a particular experiment, for instance, one might consider these numbers as the elements of a vector \mathbf{d} of length N . Similarly, the model parameters can be represented as the elements of a vector \mathbf{m} , which, is of length M .

$$\begin{aligned} \text{data: } \mathbf{d} &= [d_1, d_2, d_3, d_4, \dots, d_N]^T \\ \text{model parameters: } \mathbf{m} &= [m_1, m_2, m_3, m_4, \dots, m_M]^T \end{aligned} \quad (1.1)$$

Here T signifies transpose.

The basic statement of an inverse problem is that the model parameters and the data are in some way related. This relationship is called the *model*. Usually the model takes the form of one or more formulas that the data and model parameters are expected to follow.

If for instance, one were attempting to determine the density of an object by measuring its mass and volume, there would be two data—mass and volume (say, d_1 and d_2 , respectively)—and one unknown model parameter, density (say, m_1). The model would be the statement that density times volume equals mass, which can be written compactly by the vector equation $d_2 m_1 = d_1$.

In more realistic situations the data and model parameters are related in more complicated ways. Most generally, the data and model parameters might be related by one or more implicit equations such as

$$\begin{aligned} f_1(\mathbf{d}, \mathbf{m}) &= 0 \\ f_2(\mathbf{d}, \mathbf{m}) &= 0 \\ &\cdot \\ &\cdot \\ &\cdot \\ f_L(\mathbf{d}, \mathbf{m}) &= 0 \end{aligned} \tag{1.2}$$

where L is the number of equations. In the above examples concerning the measuring of density, $L = 1$ and $d_2 m_1 = d_1$ would constitute the one equation of the form $f_1(\mathbf{d}, \mathbf{m}) = 0$. These implicit equations, which can be compactly written as the vector equation $\mathbf{f}(\mathbf{d}, \mathbf{m}) = 0$, summarize what is known about how the measured data and the unknown model parameters are related. The purpose of inverse theory is to solve, or “invert,” these equations for the model parameters, or whatever kinds of answers might be possible or desirable in any given situation.

No claims are made either that the equations $\mathbf{f}(\mathbf{d}, \mathbf{m}) = 0$ contain enough information to specify the model parameters uniquely or that they are even consistent. One of the purposes of inverse theory is to answer these kinds of questions and to provide means of dealing with the problems that they imply. In general, $\mathbf{f}(\mathbf{d}, \mathbf{m}) = 0$ can consist of arbitrarily complicated (nonlinear) functions of the data and model parameters. In many problems, however, the equation takes on one of several simple forms. It is convenient to give names to some of these special cases, since they commonly arise in practical problems; and we shall give them special consideration in later chapters.

1.1.1 IMPLICIT LINEAR FORM

The function \mathbf{f} is linear in both data and model parameters and can therefore be written as the matrix equation

$$\mathbf{f}(\mathbf{d}, \mathbf{m}) = 0 = \mathbf{F} \begin{bmatrix} \mathbf{d} \\ \mathbf{m} \end{bmatrix} \quad (1.3)$$

where \mathbf{F} is an $L \times (M + N)$ matrix.

1.1.2 EXPLICIT FORM

In many instances it is possible to separate the data from the model parameters and thus to form $L = N$ equations that are linear in the data (but still nonlinear in the model parameters through a vector function \mathbf{g}).

$$\mathbf{f}(\mathbf{d}, \mathbf{m}) = 0 = \mathbf{d} - \mathbf{g}(\mathbf{m}) \quad (1.4)$$

1.1.3 EXPLICIT LINEAR FORM

In the explicit linear form the function \mathbf{g} is also linear, leading to the $N \times M$ matrix equation (where $L = N$)

$$\mathbf{f}(\mathbf{d}, \mathbf{m}) = 0 = \mathbf{d} - \mathbf{G}\mathbf{m} \quad (1.5)$$

Using this form is equivalent to saying that the matrix \mathbf{F} in Section 1.1.1 is:

$$\mathbf{F} = [-\mathbf{I} \ \mathbf{G}] \quad (1.6)$$

1.2 The Linear Inverse Problem

The simplest and best-understood inverse problems are those that can be represented with the explicit linear equation $\mathbf{G}\mathbf{m} = \mathbf{d}$. This equation, therefore, forms the foundation of the study of discrete inverse theory. As will be shown below, many important inverse problems that arise in the physical sciences involve precisely this equation. Others, while involving more complicated equations, can often be solved through linear approximations.

The matrix \mathbf{G} is called the data kernel, in analogy to the theory of integral equations, in which the analogs of the data and model parameters are two continuous functions $d(x)$ and $m(x)$, where x is some independent variable. Continuous inverse theory lies between these two extremes, with discrete data but a continuous model function:

Discrete inverse theory:

$$d_i = \sum_{j=1}^M G_{ij} m_j \quad (1.7a)$$

Continuous inverse theory:

$$d_i = \int G_i(x) m(x) dx \quad (1.7b)$$

Integral equation theory:

$$d(y) = \int G(y,x) m(x) dx \quad (1.7c)$$

The main difference between discrete inverse theory, continuous inverse theory, and integral equation theory is whether the model m and data d are treated as continuous functions or discrete parameters. The data d_i in inverse theory are necessarily discrete, since inverse theory is concerned with deducing information from observational data, which always has a discrete nature. Both continuous inverse problems and integral equations can be converted to discrete inverse problems by approximating the integral as a summation using the trapezoidal rule or some other quadrature formula.

1.3 Examples of Formulating Inverse Problems

1.3.1 EXAMPLE 1: FITTING A STRAIGHT LINE

Suppose that N temperature measurements T_i are made at depths z_i in the earth. The data are then a vector \mathbf{d} of N measurements of temperature, where $\mathbf{d} = [T_1, T_2, T_3, \dots, T_N]^T$. The depths z_i are not, strictly speaking, data. Instead, they provide some auxiliary information that describes the geometry of the experiment. This distinction will be further clarified below.

Suppose that we assume a model in which temperature is a linear function of depth: $T = a + bz$. The intercept a and slope b then form

the two model parameters of the problem, $\mathbf{m} = [a, b]^T$. According to the model, each temperature observation must satisfy $T = a + bz$:

$$\begin{aligned} T_1 &= a + bz_1 \\ T_2 &= a + bz_2 \\ &\quad \cdot \\ &\quad \cdot \\ &\quad \cdot \\ T_N &= a + bz_N \end{aligned} \tag{1.8}$$

These equations can be arranged as the matrix equation $\mathbf{G}\mathbf{m} = \mathbf{d}$:

$$\begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ \vdots \\ T_N \end{bmatrix} = \begin{bmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ \vdots & \vdots \\ 1 & z_N \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \tag{1.9}$$

1.3.2 EXAMPLE 2: FITTING A PARABOLA

If the model in Example 1 is changed to assume a quadratic variation of temperature with depth of the form $T = a + bz + cz^2$, then a new model parameter is added to the problem, $\mathbf{m} = [a, b, c]^T$. The number of model parameters is now $M = 3$. The data are supposed to satisfy

$$\begin{aligned} T_1 &= a + bz_1 + cz_1^2 \\ T_2 &= a + bz_2 + cz_2^2 \\ &\quad \cdot \\ &\quad \cdot \\ &\quad \cdot \\ T_N &= a + bz_N + cz_N^2 \end{aligned} \tag{1.10}$$

These equations can be arranged into the matrix equation

$$\begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ \vdots \\ T_N \end{bmatrix} = \begin{bmatrix} 1 & z_1 & z_1^2 \\ 1 & z_2 & z_2^2 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & z_N & z_N^2 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} \quad (1.11)$$

This matrix equation has the explicit linear form $\mathbf{Gm} = \mathbf{d}$. Note that, although the equation is linear in the data and model parameters, it is not linear in the auxiliary variable z .

The equation has a very similar form to the equation of the previous example, which brings out one of the underlying reasons for employing matrix notation: it can often emphasize similarities between superficially different problems.

1.3.3 EXAMPLE 3: ACOUSTIC TOMOGRAPHY

Suppose that a wall is assembled from a rectangular array of bricks (Fig. 1.1) and that each brick is composed of a different type of clay. If the acoustic velocities of the different clays differ, one might attempt to

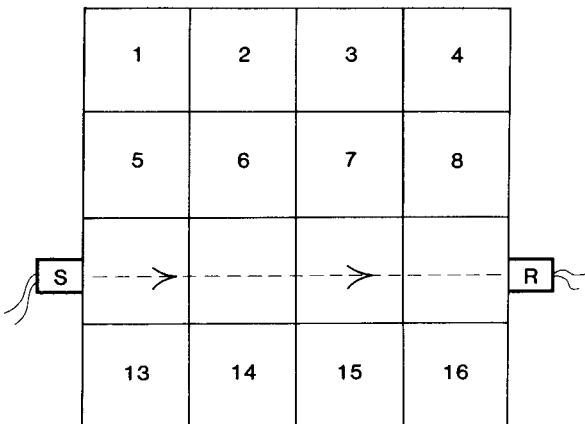


Fig. 1.1. The travel time of acoustic waves (dashed lines) through the rows and columns of a square array of bricks is measured with the acoustic source S and receiver R placed on the edges of the square. The inverse problem is to infer the acoustic properties of the bricks (which are assumed to be homogeneous).

distinguish the different kinds of bricks by measuring the travel time of sound across the various rows and columns of bricks in the wall. The data in this problem are $N = 8$ measurements of travel times, $\mathbf{d} = [T_1, T_2, T_3, \dots, T_8]^T$. The model assumes that each brick is composed of a uniform material and that the travel time of sound across each brick is proportional to the width and height of the brick. The proportionality factor is the brick's *slowness* s_i , thus giving $M = 16$ model parameters $\mathbf{m} = [s_1, s_2, s_3, \dots, s_{16}]^T$, where the ordering is according to the numbering scheme of the figure as

$$\begin{aligned} \text{row 1: } & T_1 = hs_1 + hs_2 + hs_3 + hs_4 \\ \text{row 2: } & T_2 = hs_5 + hs_6 + hs_7 + hs_8 \\ & \vdots \quad \quad \quad \vdots \\ & \vdots \quad \quad \quad \vdots \\ & \vdots \quad \quad \quad \vdots \\ \text{column 4: } & T_8 = hs_4 + hs_8 + hs_{12} + hs_{16} \end{aligned} \tag{1.12}$$

and the matrix equation is

$$\begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_8 \end{bmatrix} = h \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_{16} \end{bmatrix} \tag{1.13}$$

Here the bricks are assumed to be of width *and* height h .

1.3.4 EXAMPLE 4: X-RAY IMAGING

Tomography is the process of forming images of the interior of an object from measurements made along rays passed through that object (“tomo” comes from the Greek word for “slice”). The Computerized Axial Tomography (CAT) scanner is an x-ray imaging device that has revolutionized the diagnosis of brain tumors and certain other medical conditions. The scanner solves an inverse problem for the x-ray opacity

of body tissues using measurements of the amount of radiation absorbed from many criss-crossing beams of x rays (Fig. 1.2).

The basic physical model underlying this device is the idea that the intensity of x rays diminishes with the distance traveled, at a rate proportional to the intensity of the beam, and an absorption coefficient that depends on the type of tissue:

$$dI/ds = -c(x,y)I \quad (1.14)$$

Here I is the intensity of the beam, s the distance along the beam, and $c(x,y)$ the absorption coefficient, which varies with position. If the x-ray source has intensity I_0 , then the intensity at the i th detector is

$$I_i = I_0 \exp\left(-\int_{\text{beam } i} c(x,y) ds\right) \quad (1.15a)$$

$$\ln I_0 - \ln I_i = \int_{\text{beam } i} c(x,y) ds \quad (1.15b)$$

Note that Eq. (1.15a) is a nonlinear function of the unknown absorption coefficient $c(x,y)$ and that the absorption coefficient varies continuously along the beam. This problem is a nonlinear problem in continuous inverse theory. However, it can be linearized by taking the logarithm of both sides or, for small net absorption, by approximating the exponential with the first two terms in its Taylor series expansion.

We convert this problem to a discrete inverse problem of the form

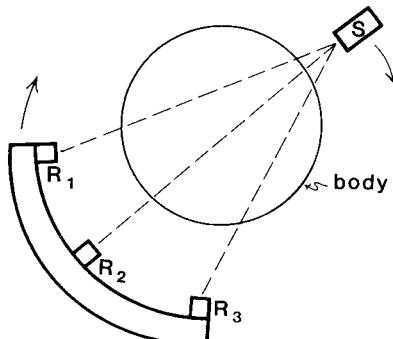


Fig. 1.2. An idealized x-ray tomography device measures x-ray absorption along lines (dashed) passing through body. After a set of measurements is made, the source S and receivers R_1, R_2, R_3, \dots are rotated and the measurements are repeated, so data along many crisscrossing lines are collected. The inverse problem is to determine the x-ray opacity as a function of position in the body.

Gm = d. The first step is to assume that the net absorption of x rays is small; this allows the exponential to be approximated by the first two terms in its Taylor series, $\exp(-x) \approx 1 - x$. The second step is to assume that the continuously varying absorption coefficient can be adequately represented by a grid of many small square boxes, each of which has a constant absorption coefficient. With these boxes numbered $1 - M$, the model parameters are then the vector $\mathbf{m} = [c_1, c_2, c_3, \dots, c_M]^T$. The integral can then be written as the sum

$$\Delta I_i = \frac{I_0 - I_i}{I_0} = \sum_{j=1}^M \Delta s_{ij} c_j \quad (1.16)$$

Here the data ΔI_i represent the differences between the x-ray intensities at the source and at the detector, and Δs_{ij} is the distance the i th beam travels in the j th box.

The inverse problem can then be summarized by the matrix equation

$$\begin{bmatrix} \Delta I_1 \\ \Delta I_2 \\ \vdots \\ \Delta I_N \end{bmatrix} = \begin{bmatrix} \Delta s_{11} & \Delta s_{12} & \cdots & \Delta s_{1M} \\ \Delta s_{21} & \Delta s_{22} & \cdots & \Delta s_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \Delta s_{N1} & \Delta s_{N2} & \cdots & \Delta s_{NM} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_M \end{bmatrix} \quad (1.17)$$

Since each beam passes through only a few of the many boxes, many Δs_{ij} 's are zero, and the matrix is very sparse.

1.3.5 EXAMPLE 5: GAUSSIAN CURVE FITTING

Not every inverse problem can be adequately represented by the discrete linear equation $\mathbf{Gm} = \mathbf{d}$. Consider, for example, an experiment in which the intensity of the x rays scattered from a crystal lattice is measured as a function of Bragg angle. Near one of the diffraction peaks the x-ray intensity $I(\theta)$ varies as

$$I(\theta) = \frac{A}{(2\pi)^{1/2}\sigma} \exp\left[-\frac{(\theta - \Theta)^2}{2\sigma^2}\right] \quad (1.18)$$

Here θ is the Bragg angle, Θ the Bragg angle of the peak's maximum (Fig. 1.3), A the amplitude of the peak, and σ a measure of its width.

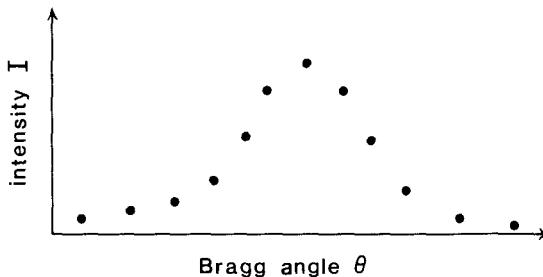


Fig. 1.3. A hypothetical x-ray diffraction experiment. Measurements of x-ray intensity I against Bragg angle θ form a peak that is assumed to have a Gaussian shape. The inverse problem is to determine the peak area, width, and position.

In a typical experiment $I(\theta)$ is measured at many different θ 's, and the model parameters A , Θ , and σ are to be determined. The data and model are therefore related by the *nonlinear* explicit equation $\mathbf{d} = \mathbf{g}(\mathbf{m})$. Furthermore, since the argument of the exponential is large in the vicinity of the peak, the problem cannot be linearized.

1.3.6 EXAMPLE 6: FACTOR ANALYSIS

Another example of a nonlinear inverse problem is that of determining the composition of chemical end members on the basis of the chemistry of a suite of mixtures of the end members. Consider a simplified “ocean” (Fig. 1.4) in which sediments are composed of mixtures of several chemically distinct rocks eroded from the continents. One expects the fraction of chemical j in the i th sediment sample S_{ij} to be related to the amount of end-member rock in sediment

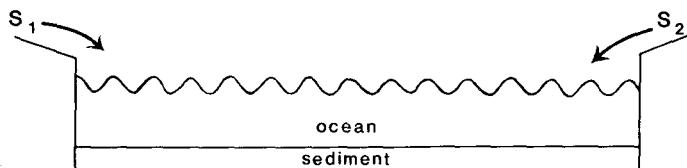


Fig. 1.4. Sediment on the floor of this idealized ocean is a simple mixture of rocks eroded from several sources S_i . The inverse problem is to determine the number and composition of the sources.

sample i (C_{ik}) and to the amount of the j th chemical in the end-member rock (F_{kj}) as

$$\begin{bmatrix} \text{sample} \\ \text{composition} \end{bmatrix} = \sum_{\substack{\text{end} \\ \text{members}}} \begin{bmatrix} \text{amount of} \\ \text{end member} \end{bmatrix} \begin{bmatrix} \text{end member} \\ \text{composition} \end{bmatrix}$$

$$S_{ij} = \sum_{k=1}^p C_{ik} F_{kj} \quad \text{or} \quad \mathbf{S} = \mathbf{C} \mathbf{F} \quad (1.19)$$

In a typical experiment, the number of end members p , the end-member composition \mathbf{F} , and the amount of end members in the samples \mathbf{C} are all unknown model parameters. Since the data \mathbf{S} are on one side of the equations, this problem is also of the explicit nonlinear type. Note that basically the problem is to factor a matrix \mathbf{S} into two other matrices \mathbf{C} and \mathbf{F} . This factoring problem is a well-studied part of the theory of matrices, and methods are available to solve it. As will be discussed in Chapter 10, this problem (which is often called *factor analysis*) is very closely related to the algebraic eigenvalue problem of linear algebra.

1.4 Solutions to Inverse Problems

We shall use the terms “solution” and “answer” to indicate broadly whatever information we are able to determine about the problem under consideration. As we shall see, there are many different points of view regarding what constitutes a solution to an inverse problem. Of course, one generally wants to know the numerical values of the model parameters (we call this kind of answer an estimate of the model parameters). Unfortunately, only very infrequently can an inverse problem be solved in such a way as to yield this kind of exact information. More typically, the practitioner of inverse theory is forced to make various compromises between the kind of information he or she actually wants and the kind of information that can in fact be obtained from any given dataset. These compromises lead to other kinds of “answers” that are more abstract than simple estimates of the model parameters. Part of the practice of inverse theory is the identifying of what features of a solution are most valuable and making the compromises that emphasize these features. Some of the possible forms an “answer” to an inverse problem might take are described below.

1.4.1 ESTIMATES OF MODEL PARAMETERS

The simplest kind of solution to an inverse problem is an estimate \mathbf{m}^{est} of the model parameters. An estimate is simply a set of numerical values for the model parameters, $\mathbf{m}^{\text{est}} = [1.4, 2.9, \dots, 1.0]^T$ for example. Estimates are generally the most useful kind of solution to an inverse problem. Nevertheless, in many situations they can be very misleading. For instance, estimates in themselves give no insight into the quality of the solution. Depending on the structure of the particular problem, measurement errors might be averaged out (in which case the estimates might be meaningful) or amplified (in which case the estimates might be nonsense). In other problems, many solutions might exist. To single out arbitrarily only one of these solutions and call it \mathbf{m}^{est} gives the false impression that a unique solution has been obtained.

1.4.2 BOUNDING VALUES

One remedy to the problem of defining the quality of an estimate is to state additionally some bounds that define its certainty. These bounds can be either absolute or probabilistic. Absolute bounds imply that the true value of the model parameter lies between two stated values, for example, $1.3 \leq m_1 \leq 1.5$. Probabilistic bounds imply that the estimate is likely to be between the bounds, with some given degree of certainty. For instance, $m_1^{\text{est}} = 1.4 \pm 0.1$ might mean that there is a 95% probability that m_1^{true} lies between 1.3 and 1.5.

When they exist, bounding values can often provide the supplementary information needed to interpret properly the solution to an inverse problem. There are, however, many instances in which bounding values do not exist.

1.4.3 PROBABILITY DISTRIBUTIONS

A generalization of the stating of bounding values is the stating of the complete probability distribution for model parameters. The usefulness of this technique depends in part on how complicated the distribution is. If it has only one peak (Fig. 1.5a), then stating the distribution provides little more information than stating an estimate (based on the position of the peak's center) with error bounds based on the peak's shape. On the other hand, if the distribution is very complicated (Fig. 1.5c), it is basically uninterpretable (except in the sense that

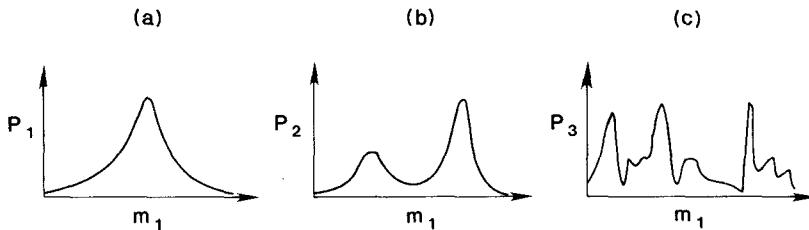


Fig. 1.5. Three probability distributions for a model parameter m_1 . The first is so simple that its properties can be summarized by the central position and width of the peak. The second implies that there are two probable ranges for the model parameter. The third is so complicated that it provides no easily interpretable information about the model parameter.

it implies that the model parameter cannot be well estimated). Only in those exceptional instances in which the distribution has some intermediate complexity (Fig. 1.5b) does it really provide information toward the solution of an inverse problem. In practice, the most interesting distributions are exceedingly difficult to compute, so this technique is of rather limited usefulness.

1.4.4 WEIGHTED AVERAGES OF MODEL PARAMETERS

In many instances it is possible to identify combinations or averages of the model parameters that are in some sense better determined than the model parameters themselves. For instance, given $\mathbf{m} = [m_1, m_2]^T$ it may turn out that $\langle \mathbf{m} \rangle = 0.2m_1 + 0.8m_2$ is better determined than either m_1 or m_2 . Unfortunately, one might not have the slightest interest in such an average, be it well determined or not, because it may not have physical significance.

Averages *can* be of considerable interest when the model parameters represent a discretized version of some continuous function. If the weights are large only for a few physically adjacent parameters, then the average is said to be localized. The meaning of the average in such a case is that, although the data cannot resolve the model parameters at a particular point, they can resolve the average of the model parameters in the neighborhood of that point.

In the following chapters we shall derive methods for determining each of these different kinds of solutions to inverse problems. We note here, however, that there is a great deal of underlying similarity between these types of “answers.” In fact, it will turn out that the same numerical “answer” will be interpretable as any of several classes of solutions.

2

SOME COMMENTS ON PROBABILITY THEORY

2.1 Noise and Random Variables

In the preceding chapter we represented the results of an experiment as a vector \mathbf{d} whose elements were individual measurements. Sometimes, however, a single number is insufficient to represent a single observation. Measurements are known to contain noise, so that if an observation were to be performed several times, each measurement would be different (Fig. 2.1). To characterize the data completely, information about the range and shape of this scatter must also be provided.

The concept of a *random variable* is used to describe this property. Each random variable has definite and precise properties, governing the range and shape of the scatter of values one observes. These properties cannot be measured directly, however; one can only make individual measurements, or *realizations*, of the random variable and try to estimate its true properties from these data.

The true properties of the random variable d are specified by a distribution $P(d)$. This function gives the probability that a particular

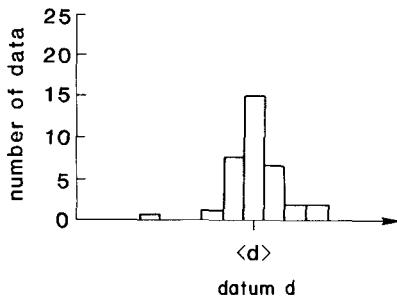


Fig. 2.1. Histogram showing data from 40 repetitions of an experiment. Noise causes the data to scatter about their mean value $\langle d \rangle$.

realization of the random variable will have a value in the neighborhood of d —the probability that the measurement is between d and $d + \partial d$ is $P(d) \partial d$ (Fig. 2.2). We have used the partial derivative sign ∂ only for the sake of clarity, to avoid adding another d to the notation. Since each measurement must have some value, the probability that d lies somewhere between $-\infty$ and $+\infty$ is complete certainty (usually given the value of 100% or unity, which is written as

$$\int_{-\infty}^{+\infty} P(d) \partial d = 1 \quad (2.1)$$

The distribution completely describes the random variable. Unfortunately, it is a continuous function that may be quite complicated. It is helpful, therefore, to derive a few numbers from the distribution that try to summarize its major properties. One such kind of number tries to indicate the typical numerical value of a measurement. The most likely measurement is the one with the highest probability (Fig. 2.3). If

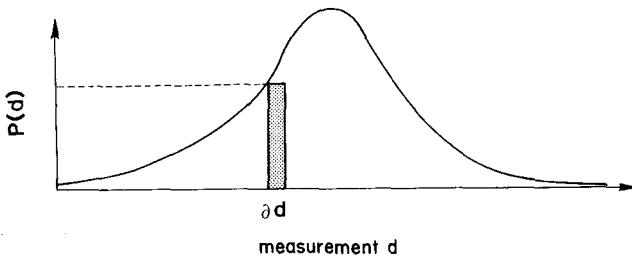


Fig. 2.2. The shaded area $P(d) \partial d$ of the probability distribution gives the probability that the datum will fall between d and $d + \partial d$.

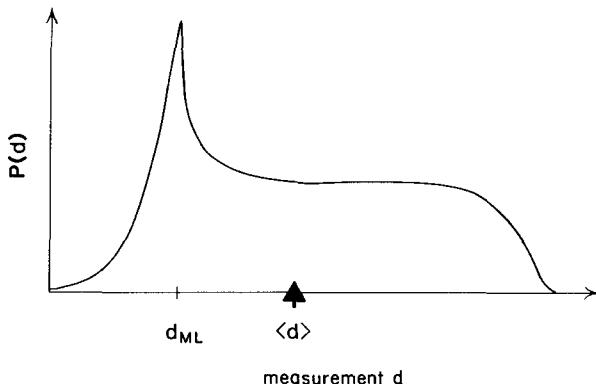


Fig. 2.3. The maximum likelihood point d_{ML} of the probability distribution $P(d)$ for data d gives the most probable value of the data. In general, this value can be different from the mean datum $\langle d \rangle$, which is at the “balancing point” of the distribution.

the distribution is skewed, the peak, or *maximum likelihood point*, may not be a good indication of the typical measurement, since a wide range of other values also have high probability. In such instances the mean, or expected measurement, $E(d)$ is a better characterization of a typical measurement. This number is the “balancing point” of the distribution and is given by

$$E(d) = \int_{-\infty}^{+\infty} dP(d) \partial d \quad (2.2)$$

Another property of a distribution is its overall width. Wide distributions imply very noisy data, and narrow ones imply relatively noise-free data. One way of measuring the width of a distribution is to multiply it by a function that is zero near the center (peak) of the distribution, and that grows on either side of the peak (Fig. 2.4). If the distribution is narrow, then the resulting function will be everywhere small; if the distribution is wide, then the result will be large.

A quantitative measure of the width of the peak is the area under the resulting function. If one chooses the parabola $(d - \langle d \rangle)^2$ as the function, where $\langle d \rangle = E(d)$ is the expected value of the random variable, then this measure is called the *variance* σ^2 of the distribution and is written as

$$\sigma^2 = \int_{-\infty}^{+\infty} (d - \langle d \rangle)^2 P(d) \partial d \quad (2.3)$$

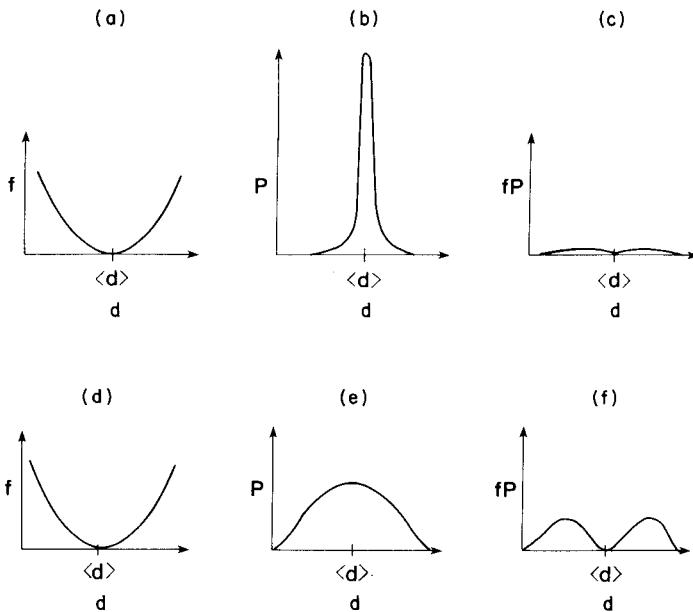


Fig. 2.4. (a and d) A parabola of the form $f = [d - \langle d \rangle]^2$ is used to measure the width of two probability distributions P (b and e) which have the same mean but different widths. The product fP is everywhere small for the narrow distribution (c) but has two large peaks for the wider distribution (f). The area under fP is a measure of the width of the distribution, and is called the variance.

The square root of the variance σ is a measure of the width of the distribution.

2.2 Correlated Data

Experiments usually involve the collection of more than one datum. We therefore need to quantify the probability that a set of random variables will take on a given value. The joint distribution $P(\mathbf{d})$ is the probability that the first datum will be in the neighborhood of d_1 , that the second will be in the neighborhood of d_2 , etc. If the data are independent—that is, if there are no patterns in the occurrence of the values between two random variables—then this joint distribution is just the product of the individual distributions (Fig. 2.5).

$$P(\mathbf{d}) = P(d_1)P(d_2) \cdots P(d_N) \quad (2.4)$$

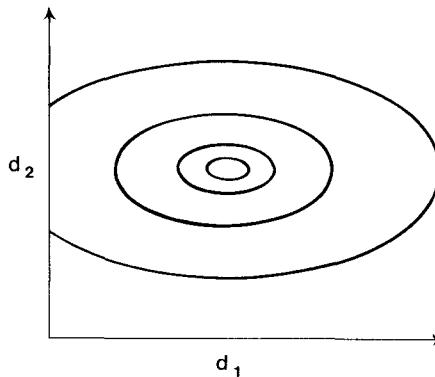


Fig. 2.5. The probability distribution $P(d_1, d_2)$ is contoured as a function of d_1 and d_2 . These data are uncorrelated, since especially large values of d_2 are no more or less likely if d_1 is large or small.

In some experiments, on the other hand, measurements *are* correlated. High values of one datum tend to occur consistently with either high or low values of another datum (Fig. 2.6). The joint distribution for these two data must be constructed to take this correlation into account. Given a joint distribution, one can test for correlation by selecting a function that divides the (d_1, d_2) plane into four quadrants

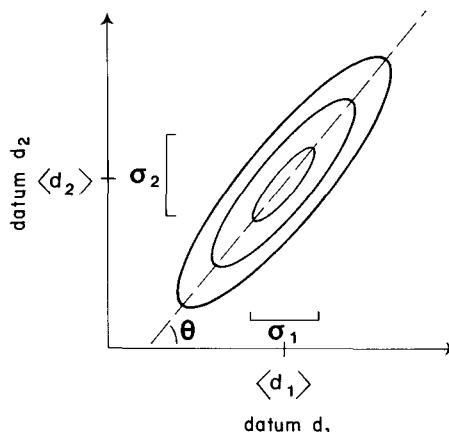


Fig. 2.6. The probability distribution $P(d_1, d_2)$ is contoured as a function of d_1 and d_2 . These data are correlated, since large values of d_2 are especially probable if d_1 is large. The distribution has mean values $\langle d_1 \rangle$ and $\langle d_2 \rangle$ and widths in the coordinate directions given by σ_1 and σ_2 . The angle Θ is a measure of the correlation and is related to the covariance.

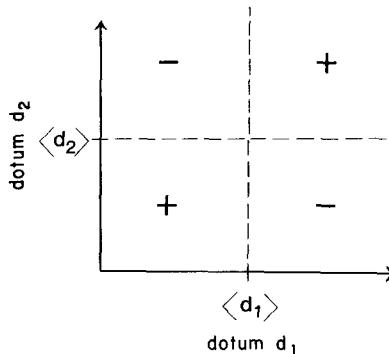


Fig. 2.7. The function $(d_1 - \langle d_1 \rangle)(d_2 - \langle d_2 \rangle)$ divides the (d_1, d_2) plane into four quadrants of alternating sign.

of alternating sign, centered on the center of the distribution (Fig. 2.7). If one multiplies the distribution by this function, and then sums up the area, the result will be zero for uncorrelated distributions, since they tend to lie equally in all four quadrants. Correlated distributions will have either positive or negative area, since they tend to be concentrated in two opposite quadrants (Fig. 2.8). If $[d_1 - \langle d_1 \rangle][d_2 - \langle d_2 \rangle]$ is used as the function, the resulting measure of correlation is called the covariance.

$$\text{cov}(d_1, d_2) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} [d_1 - \langle d_1 \rangle][d_2 - \langle d_2 \rangle] P(\mathbf{d}) \partial d_1 \cdots \partial d_N \quad (2.5)$$

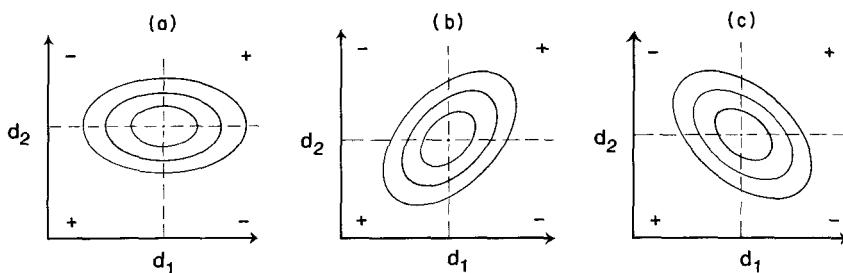


Fig. 2.8. Contour plots of $P(d_1, d_2)$ when the data are (a) uncorrelated, (b) positively correlated, (c) negatively correlated. The dashed lines indicate the four quadrants of alternating sign used to determine correlation (see Fig. 2.7).

Note that the covariance of a datum with itself is just the variance. The covariance, therefore, characterizes the basic shape of joint distribution.

When there are many data given by the vector \mathbf{d} , it is convenient to define a vector of expected values and a matrix of covariances as

$$\begin{aligned}\langle \mathbf{d} \rangle_i &= \int_{-\infty}^{+\infty} \partial d_1 \int_{-\infty}^{+\infty} \partial d_2 \cdots \int_{-\infty}^{+\infty} \partial d_N d_i P(\mathbf{d}) \\ [\text{cov } \mathbf{d}]_{ij} &= \int_{-\infty}^{+\infty} \partial d_1 \int_{-\infty}^{+\infty} \partial d_2 \cdots \int_{-\infty}^{+\infty} \partial d_N [d_i - \langle d \rangle_i][d_j - \langle d \rangle_j] P(\mathbf{d})\end{aligned}\quad (2.6)$$

The diagonal elements of the covariance matrix are a measure of the width of the distribution of the data, and the off-diagonal elements indicate the degree to which pairs of data are correlated.

2.3 Functions of Random Variables

The basic premise of inverse theory is that the data and model parameters are related. Any method that solves the inverse problem —that estimates a model parameter on the basis of data— will, therefore, tend to map errors from the data to the estimated model parameters. Thus the *estimates* of the model parameters are themselves random variables, which are described by a distribution $P(\mathbf{m}^{\text{est}})$. Whether or not the *true* model parameters are random variables depends on the problem. It is appropriate to consider them deterministic quantities in some problems, and random variables in others. Estimates of the parameters, however, are always random variables.

If the distribution of the data is known, then the distribution for any function of the data, including estimated model parameters, can be found. Consider two uncorrelated data that are known to have white distributions on the interval $[0, 1]$, that is, they can take on any value between 0 and 1 with equal probability (Fig. 2.9a). Suppose that some model parameter is estimated to be the sum of these two data, $m_1 = d_1 + d_2$. The probability of this sum taking any particular value is determined by identifying curves of equal m_1 in the (d_1, d_2) plane and integrating the distribution along these curves (Fig. 2.9b). Since in this case $P(\mathbf{d})$ is constant and the curves are straight lines, $P(m_1)$ is just a triangular distribution on the interval $[0, 2]$ (Fig. 2.9c).

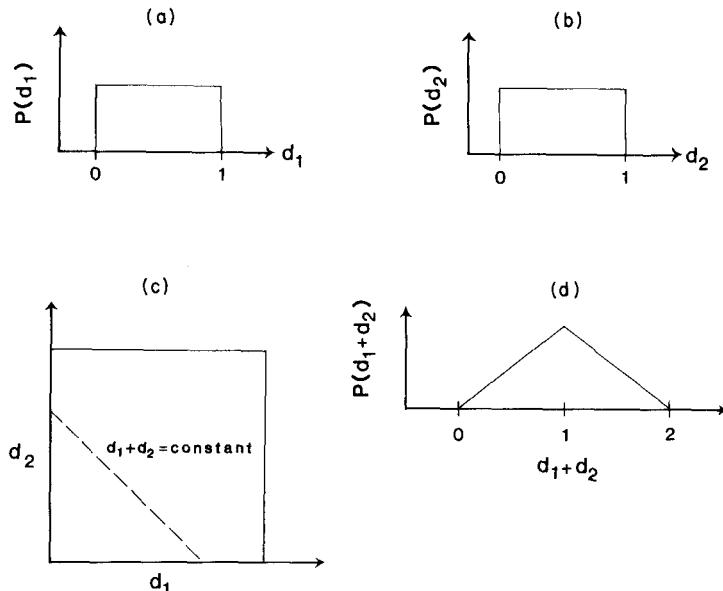


Fig. 2.9. (a and b) The uncorrelated data d_1 and d_2 both have white distributions on the interval $[0, 1]$. (c) Their joint distribution consists of a rectangular area of uniform probability. The probability distribution for the sum $d_1 + d_2$ is the joint distribution, integrated along lines of constant $d_1 + d_2$ (dashed). (d) The distribution for $d_1 + d_2$ is triangular.

Performing this line integral in the most general case can be mathematically quite difficult. Fortunately, in the case of the linear function $\mathbf{m} = \mathbf{M}\mathbf{d} + \mathbf{v}$, where \mathbf{M} and \mathbf{v} are an arbitrary matrix and vector, respectively, it is possible to make some statements about the properties of the resultant distribution without explicitly performing the integration. In particular, the mean and covariance of the resultant distribution can be shown, respectively, to be

$$\langle \mathbf{m} \rangle = \mathbf{M}\langle \mathbf{d} \rangle + \mathbf{v} \quad \text{and} \quad [\text{cov } \mathbf{m}] = \mathbf{M}[\text{cov } \mathbf{d}]\mathbf{M}^T \quad (2.7)$$

As an example, consider a model parameter m_1 , which is the mean of a set of data

$$m_1 = 1/N \sum_{i=1}^N d_i = (1/N)[1, 1, 1, \dots, 1]\mathbf{d} \quad (2.8)$$

That is, $\mathbf{M} = [1, 1, 1, \dots, 1]/N$ and $\mathbf{v} = 0$. Suppose that the data are uncorrelated and all have the same mean $\langle d \rangle$ and variance σ_d^2 . Then

we see that $\langle m_1 \rangle = \mathbf{M}(\mathbf{d}) + \mathbf{v} = \langle d \rangle$ and $\text{var}(m_1) = \mathbf{M}[\text{cov } \mathbf{d}] \mathbf{M}^T = \sigma_d^2/N$. The model parameter m_1 , therefore, has a distribution $P(m_1)$ with mean $\langle m_1 \rangle = \langle d \rangle$ and a variance $\sigma_m^2 = \sigma_d^2/N$. The square root of the variance, which is a measure of the width of the peak in $P(m_1)$ and therefore a measure of the likelihood that any particular experiment will yield an m_1 close to the true mean, is proportional to $N^{-1/2}$. The accuracy of determining the mean of a group of data, therefore, decreases very slowly as the number of observations increases.

2.4 Gaussian Distributions

The distribution for a particular random variable can be an arbitrarily complicated function, but in many instances data possess the rather simple Gaussian distribution

$$P(d) = \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[-\frac{(d - \langle d \rangle)^2}{2\sigma^2}\right] \quad (2.9)$$

This distribution has mean $\langle d \rangle$ and variance σ^2 (Fig. 2.10). The Gaussian distribution is so common because it is the limiting distribution for the sum of random variables. The *central limit theorem* shows (with certain limitations) that regardless of the distribution of a set of independent random variables, the distribution of their sum tends to a Gaussian distribution as the number of summed variables increases. As long as the noise in the data comes from several sources of comparable size, it will tend to follow a Gaussian distribution. This

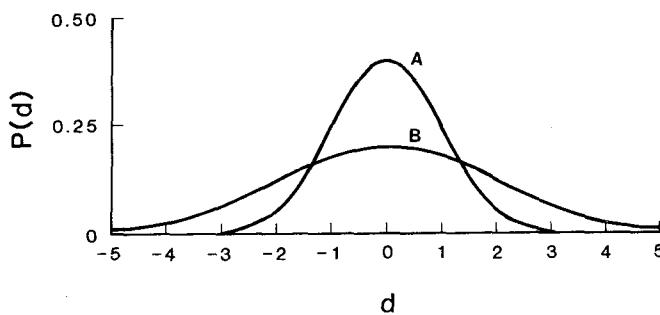


Fig. 2.10. Gaussian distribution with zero mean and $\sigma = 1$ for curve A, and $\sigma = 2$ for curve B.

behavior is exemplified by the sum of the two white distributions in Section 2.3. The distribution of their sum is more nearly Gaussian than the individual distributions (it being triangular instead of rectangular).

The joint distribution for two independent Gaussian variables is just the product of two univariate distributions. When the data are correlated (say, with mean $\langle \mathbf{d} \rangle$ and covariance $[\text{cov } \mathbf{d}]$), the distribution is more complicated, since it must express the degree of correlation. The appropriate generalization turns out to be

$$P(\mathbf{d}) = \frac{|[\text{cov } \mathbf{d}]|^{-1/2}}{(2\pi)^{N/2}} \exp\left(-\frac{1}{2}[\mathbf{d} - \langle \mathbf{d} \rangle]^T [\text{cov } \mathbf{d}]^{-1} [\mathbf{d} - \langle \mathbf{d} \rangle]\right) \quad (2.10)$$

This distribution is chosen because it has the correct mean and variance when the data are uncorrelated and has covariance $[\text{cov } \mathbf{d}]$ when the data are correlated. It can be shown that all linear functions of Gaussian random variables are also Gaussian random variables with a distribution of this form.

The idea that the model and data are related by an explicit relationship $\mathbf{g}(\mathbf{m}) = \mathbf{d}$ can be reinterpreted in light of this probabilistic description of the data. We can no longer assert that this relationship can hold for the data themselves, since they are random variables. Instead, we assert that this relationship holds for the mean data: $\mathbf{g}(\mathbf{m}) = \langle \mathbf{d} \rangle$. The distribution for the data can then be written as

$$P(\mathbf{d}) = \frac{|[\text{cov } \mathbf{d}]|^{-1/2}}{(2\pi)^{N/2}} \exp\left[-\frac{1}{2}[\mathbf{d} - \mathbf{g}(\mathbf{m})]^T [\text{cov } \mathbf{d}]^{-1} [\mathbf{d} - \mathbf{g}(\mathbf{m})]\right] \quad (2.11)$$

The model parameters now have the interpretation of a set of unknown quantities that define the shape of the distribution for the data. One approach to inverse theory (which will be pursued in Chapter 5) is to try to use the data to determine the distribution, and thus the values of the model parameters.

For the Gaussian distribution [Eq. (2.11)] to be sensible, $\mathbf{g}(\mathbf{m})$ must not be a function of any random variables. This is why we differentiated between data and auxiliary variables in Chapter 1; the latter must be known exactly. If the auxiliary variables are themselves uncertain, then they must be treated as data and the inverse problem becomes an implicit one with a much more complicated distribution than the above problem exhibits.

As an example of constructing the distribution for a set of data, consider an experiment in which the temperature d_i in some small

volume of space is measured N times. If the temperature is assumed not to be a function of time and space, the experiment can be viewed as the measurement of N realizations of the same random variable or as the measurement of one realization of N distinct random variables that all have the same distribution. We adopt the second viewpoint.

If the data are independent Gaussian random variables with mean $\langle \mathbf{d} \rangle$ and variance σ_d^2 , then we can represent the assumption that all the data have the same mean by an equation of the form $\mathbf{Gm} = \mathbf{d}$:

$$\begin{bmatrix} 1 \\ 1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix} [m_1] = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ \vdots \\ d_N \end{bmatrix} \quad (2.12)$$

where m_1 is a single model parameter. We can then compute explicit formulas for the expressions in $P(\mathbf{d})$ as

$$\begin{aligned} |[\text{cov } \mathbf{d}]^{-1}|^{1/2} &= (\sigma_d^{-2N})^{1/2} = \sigma_d^{-N} \\ [\mathbf{d} - \mathbf{Gm}]^T [\text{cov } \mathbf{d}]^{-1} [\mathbf{d} - \mathbf{Gm}] &= \sigma_d^{-2} \sum_{i=1}^N (d_i - m_1)^2 \end{aligned} \quad (2.13)$$

The joint distribution is therefore

$$P(\mathbf{d}) = \frac{\sigma_d^{-N}}{(2\pi)^{N/2}} \exp \left[-\frac{1}{2} \sigma_d^{-2} \sum_{i=1}^N (d_i - m_1)^2 \right] \quad (2.14)$$

2.5 Testing the Assumption of Gaussian Statistics

In the following chapters we shall derive methods of solving inverse problems that are applicable whenever the data exhibit Gaussian statistics. In many instances the assumption that the data follow this distribution is a reasonable one; nevertheless, it is important to have some means of testing this assumption.

First, consider a set of v random variables x_i , each possessing a Gaussian distribution with zero mean and unit variance. Suppose we

construct a new random variable

$$\chi^2 = \sum_{i=1}^v x_i^2 \quad (2.15)$$

This random variable is said to have the χ^2 distribution with v degrees of freedom. This distribution can be shown to be unimodal with mean v and variance $2v$ and to have the functional form

$$P(\chi^2, v) = \frac{[\chi^2]^{(v-2)/2} \exp(-\chi^2/2)}{2^{v/2} \Gamma(v/2)} \quad (2.16)$$

where Γ is the gamma function. We shall make use of this distribution in the discussion to follow.

We begin by supposing that we have some method of solving the inverse problem for the estimated model parameters. Assuming further that the model is explicit, we can compute the variation of the data about its estimated mean—a quantity we refer to as the error $e = d - g(m^{est})$. Does this error follow an uncorrelated Gaussian distribution with uniform variance?

To test the hypothesis that it does, we first make a histogram of the errors e_i , in which the histogram intervals have been chosen so that there are about the same number of errors e_i in each interval. This histogram is then normalized to unit area, and the area A_i of each of the, say, p intervals is noted. We then compare these areas with the areas A'_i given by a Gaussian distribution with the same mean and variance as the e_i . The overall difference between these areas can be quantified by using

$$X^2 = \sum_{i=1}^p \frac{(A'_i - A_i)^2}{A'_i} \quad (2.17)$$

If the data followed a Gaussian distribution exactly, then X^2 should be close to zero (it will not be zero since there are always random fluctuations). We therefore need to inquire whether the X^2 measured for any particular data set is sufficiently far from zero that it is improbable that the data follow the Gaussian distribution. This is done by computing the theoretical distribution of X^2 and seeing whether X_{obs}^2 is probable. The usual rule for deciding that the data do not follow the assumed distribution is that values greater than or equal to X_{obs}^2 occur less than 5% of the time (if many realizations of the entire experiment were performed).

The quantity X^2 can be shown to follow approximately a χ^2 distribution, regardless of the type of distribution involved. This method can therefore be used to test whether the data follow any given distribution. The number of degrees of freedom is given by p minus the number of constraints placed on the observations. One constraint is that the total area $\sum A_i$ is unity. Two more constraints come from the fact that we assumed a Gaussian distribution and then estimated the mean and variance from the e_i . The Gaussian case, therefore, has $v = p - 3$. This test is known as the χ^2 test. The χ^2 distribution is tabulated in most texts on statistics.

2.6 Confidence Intervals

The confidence of a particular observation is the probability that one realization of the random variable falls within a specified distance of the true mean. Confidence is therefore related to the distribution of area in $P(d)$. If most of the area is concentrated near the mean, then the interval for, say, 95% confidence will be very small; otherwise, the confidence interval will be large. The width of the confidence interval is related to the variance. Distributions with large variances will also tend to have large confidence intervals. Nevertheless, the relationship is not direct, since variance is a measure of width, not area. The relationship is easy to quantify for the simplest univariate distributions. For instance, Gaussian distributions have 68% confidence intervals 1σ wide and 95% confidence intervals 2σ wide. Other types of simple distributions have similar relationships. If one knows that a particular Gaussian random variable has $\sigma = 1$, then if a realization of that variable has the value 50, one can state that there is a 95% chance that the mean of the random variable lies between 48 and 52 (one might symbolize this by $\langle d \rangle = 50 \pm 2$).

The concept of confidence intervals is more difficult to work with when one is dealing with several correlated data. One must define some volume in the space of data and compute the probability that the true means of the data are within the volume. One must also specify the shape of that volume. The more complicated the distribution, the more difficult it is to choose an appropriate shape and calculate the probability within it.

This page intentionally left blank

3

SOLUTION OF THE LINEAR, GAUSSIAN INVERSE PROBLEM, VIEWPOINT 1: THE LENGTH METHOD

3.1 The Lengths of Estimates

The simplest of methods for solving the linear inverse problem $\mathbf{Gm} = \mathbf{d}$ is based on measures of the size, or length, of the estimated model parameters \mathbf{m}^{est} and of the predicted data $\mathbf{d}^{\text{pre}} = \mathbf{Gm}^{\text{est}}$.

To see that measures of length can be relevant to the solution of inverse problems, consider the simple problem of fitting a straight line to data (Fig. 3.1). This problem is often solved by the so called method of least squares. In this method one tries to pick the model parameters (intercept and slope) so that the predicted data are as close as possible to the observed data. For each observation one defines a prediction error, or misfit, $e_i = d_i^{\text{obs}} - d_i^{\text{pre}}$. The best fit-line is then the one with

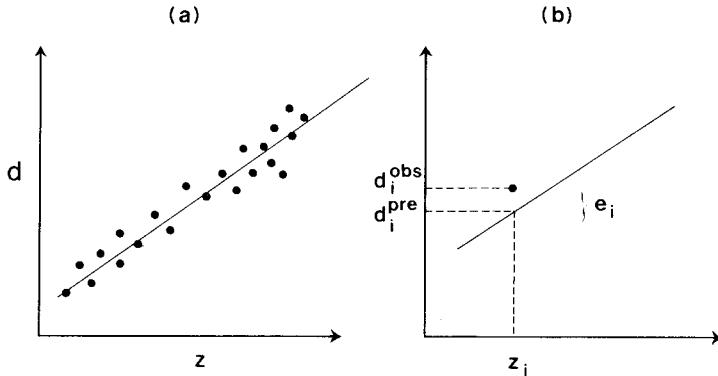


Fig. 3.1. (a) Least squares fitting of a straight line to (z, d) pairs. (b) The error e_i for each observation is the difference between the observed and predicted datum: $e_i = d_i^{\text{obs}} - d_i^{\text{pre}}$.

model parameters that lead to the smallest overall error E , defined as

$$E = \sum_{i=1}^N e_i^2 \quad (3.1)$$

The total error E (the sum of the squares of the individual errors) is exactly the squared Euclidean length of the vector \mathbf{e} , or $E = \mathbf{e}^T \mathbf{e}$.

The method of least squares estimates the solution of an inverse problem by finding the model parameters that minimize a particular measure of the length of the estimated data \mathbf{d}^{est} , namely, its Euclidean distance from the observations. As will be detailed below, it is the simplest of the methods that use measures of length as the guiding principle in solving an inverse problem.

3.2 Measures of Length

Note that although the Euclidean length is one way of quantifying the size or length of a vector, it is by no means the only possible measure. For instance, one could equally well quantify length by summing the absolute values of the elements of the vector.

The term *norm* is used to refer to some measure of length or size and is indicated by a set of double vertical bars: $\|\mathbf{e}\|$ is the norm of the vector

e. The most commonly employed norms are those based on the sum of some power of the elements of a vector and are given the name L_n , where n is the power:

$$L_1 \text{ norm: } \|\mathbf{e}\|_1 = \left[\sum_i |e_i|^1 \right] \quad (3.2a)$$

$$L_2 \text{ norm: } \|\mathbf{e}\|_2 = \left[\sum_i |e_i|^2 \right]^{1/2} \quad (3.2b)$$

.

.

.

$$L_n \text{ norm: } \|\mathbf{e}\|_n = \left[\sum_i |e_i|^n \right]^{1/n} \quad (3.2c)$$

Successively higher norms give the largest element of \mathbf{e} successively larger weight. The limiting case of $n \rightarrow \infty$ gives nonzero weight to only the largest element; therefore, it is equivalent to the selection of the vector element with largest absolute value as the measure of length, and is written as

$$L_\infty \text{ norm: } \|\mathbf{e}\|_\infty = \max_i |e_i| \quad (3.2d)$$

The method of least squares uses the L_2 norm to quantify length. It is appropriate to inquire why this, and not some other choice of norm, is used. The answer involves the way in which one chooses to weight data that fall far from the average trend (Fig. 3.2). If the data are very accurate, then the fact that one prediction falls far from its observed value is important. A high-order norm is used, since it weights the larger errors preferentially. On the other hand, if the data are expected to scatter widely about the trend, then no significance can be placed upon a few large prediction errors. A low-order norm is used, since it gives more equal weight to errors of different size.

As will be discussed in more detail later, the L_2 norm implies that the data obey Gaussian statistics. Gaussians are rather short-tailed distributions, so it is appropriate to place considerable weight on any data that have a large prediction error.

The likelihood of an observed datum falling far from the trend depends on the shape of the distribution for that datum. Long-tailed

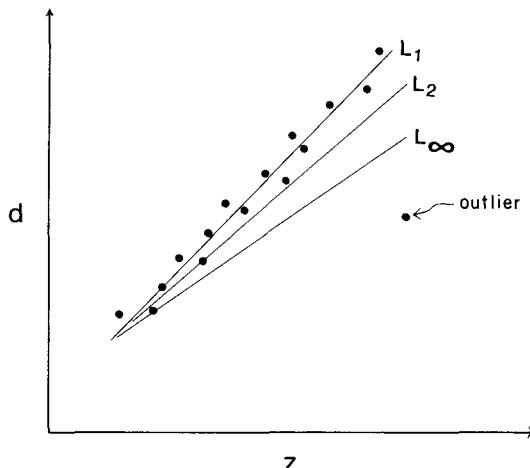


Fig. 3.2. Straight line fits to (z, d) pairs where the error is measured under the L_1 , L_2 and L_∞ norms. The L_1 norm gives least weight to the one outlier.

distributions imply many scattered (improbable) points. Short-tailed distributions imply very few scattered points (Fig. 3.3). The choice of a norm, therefore, implies an assertion that the data obey a particular type of statistics.

Even though many measurements have approximately Gaussian statistics, most data sets generally have a few spurious points that are wildly improbable. The occurrence of these points demonstrates that the assumption of Gaussian statistics is in error, especially in the tails of the distribution. If one applies least squares to this kind of problem, the estimates of the model parameters can be completely erroneous. Least squares weights large errors so heavily that even one “bad” data

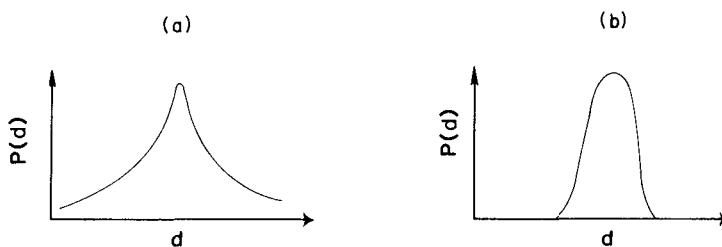


Fig. 3.3. (a) Long-tailed distribution. (b) Short-tailed distribution.

point can completely throw off the result. In these situations methods based on the L_1 norm give more reliable estimates (Methods that can tolerate a few bad data are said to be *robust*.)

Matrix norms can be defined in a manner similar to vector norms [see Eq. (3.3d) below]. Vector and matrix norms obey the following relationships:

Vector norms:

$$\|\mathbf{x}\| > 0 \quad \text{as long as } \mathbf{x} \neq 0 \quad (3.3a)$$

$$\|a\mathbf{x}\| = |a| \|\mathbf{x}\| \quad (3.3b)$$

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| \quad (3.3c)$$

Matrix norms:

$$\|\mathbf{A}\|_2' = \left(\sum_{i=1}^N \sum_{j=1}^N A_{ij}^2 \right)^{1/2} \quad (3.3d)$$

$$\|c\mathbf{A}\| = |c| \|\mathbf{A}\| \quad (3.3e)$$

$$\|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{A}\| \|\mathbf{x}\| \quad (3.3f)$$

$$\|\mathbf{AB}\| \leq \|\mathbf{A}\| \|\mathbf{B}\| \quad (3.3g)$$

$$\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\| \quad (3.3h)$$

Equations (3.3c) and (3.3h) are called triangle inequalities because of their similarity to Pythagoras's law for right triangles.

3.3 Least Squares for a Straight Line

The elementary problem of fitting a straight line to data illustrates the basic procedures applied in this technique. The model is the assertion that the data can be described by the linear equation $d_i = m_1 + m_2 z_i$. Note that there are two model parameters, $M = 2$, and that typically there are many more than two data, $N > M$. Since a line is defined by precisely two points, it is clearly impossible to choose a straight line that passes through every one of the data, except in the instance that they all lie precisely on the same straight line. In practice, when measurements are influenced by noise colinearity rarely occurs.

As we shall discuss in more detail below, the fact that the equation $d_i = m_1 + m_2 z_i$ cannot be satisfied for every i means that the inverse problem is *overdetermined*, that is, it has no exact solution. One therefore seeks values of the model parameters that solve $d_i = m_1 + m_2 z_i$ approximately, where the goodness of the approximation is defined by the error

$$E = \mathbf{e}^T \mathbf{e} = \sum_i (d_i - m_1 - m_2 z_i)^2.$$

This problem is then the elementary calculus problem of locating the minimum of the function $E(m_1, m_2)$ and is solved by setting the derivatives of E to zero and solving the resulting equations.

$$\begin{aligned} \frac{\partial E}{\partial m_1} &= \frac{\partial}{\partial m_1} \sum_{i=1}^N [d_i - m_1 - m_2 z_i]^2 \\ &= 2N m_1 + 2m_2 \sum z_i - 2 \sum d_i \\ \frac{\partial E}{\partial m_2} &= \frac{\partial}{\partial m_2} \sum_{i=1}^N [d_i - m_1 - m_2 z_i]^2 \\ &= 2m_1 \sum z_i + 2m_2 \sum z_i^2 - 2 \sum (z_i d_i) \end{aligned} \tag{3.4}$$

These two equations are then solved simultaneously for m_1 and m_2 , yielding the classic formulas for the least squares fitting of a line.

3.4 The Least Squares Solution of the Linear Inverse Problem

Least squares can be extended to the general linear inverse problem in a very straightforward manner. Again, one computes the derivative of the error E with respect to one of the model parameters, say, m_q , and sets the result to zero as

$$\begin{aligned} E &= \mathbf{e}^T \mathbf{e} = (\mathbf{d} - \mathbf{G}\mathbf{m})^T (\mathbf{d} - \mathbf{G}\mathbf{m}) \\ &= \sum_i^N \left[d_i - \sum_j^M G_{ij} m_j \right] \left[d_i - \sum_k^M G_{ik} m_k \right] \end{aligned} \tag{3.5}$$

Note that the indices on the sums within the parentheses are different dummy variables, to prevent confusion. Multiplying out the terms and

reversing the order of the summations lead to

$$E = \sum_j^M \sum_k^M m_j m_k \sum_i^M G_{ij} G_{ik} - 2 \sum_j^M m_j \sum_i^N G_{ij} d_i + \sum_i^N d_i d_i \quad (3.6)$$

The derivatives $\partial E / \partial m_q$ are now computed. Performing this differentiation term by term gives

$$\begin{aligned} \frac{\partial}{\partial m_q} \left[\sum_j^M \sum_k^M m_j m_k \sum_i^N G_{ij} G_{ik} \right] &= \sum_j^M \sum_k^M [\delta_{jq} m_k + m_j \delta_{kq}] \sum_i^N G_{ij} G_{ik} \\ &= 2 \sum_k^M m_k \sum_i^N G_{iq} G_{ik} \end{aligned} \quad (3.7)$$

for the first term. Note that derivatives of the form $\partial m_i / \partial m_j$ are just the Kronecker delta δ_{ij} . Since both m_i and m_j are independent variables, their derivative is zero unless $i = j$.

The second term gives

$$-2 \frac{\partial}{\partial m_q} \left[\sum_j^M m_j \sum_i^N G_{ij} d_i \right] = -2 \sum_j^M \delta_{jq} \sum_i^N G_{ij} d_i = -2 \sum_i^N G_{iq} d_i \quad (3.8)$$

Since the third term does not contain any m 's, it is zero as

$$\frac{\partial}{\partial m_q} \left[\sum_i^N d_i d_i \right] = 0 \quad (3.9)$$

Combining the three terms gives

$$\partial E / \partial m_q = 0 = 2 \sum_k^M m_k \sum_i^N G_{iq} G_{ik} - 2 \sum_i^N G_{iq} d_i \quad (3.10)$$

Writing this equation in matrix notation yields

$$\mathbf{G}^T \mathbf{Gm} - \mathbf{G}^T \mathbf{d} = 0 \quad (3.11)$$

Note that the quantity $\mathbf{G}^T \mathbf{G}$ is a square $M \times M$ matrix and that it multiplies a vector \mathbf{m} of length M . The quantity $\mathbf{G}^T \mathbf{d}$ is also a vector of length M . This equation is therefore a square matrix equation for the unknown model parameters. Presuming that $[\mathbf{G}^T \mathbf{G}]^{-1}$ exists (an important question that we shall return to later), we have the following solution:

$$\mathbf{m}^{\text{est}} = [\mathbf{G}^T \mathbf{G}]^{-1} \mathbf{G}^T \mathbf{d} \quad (3.12)$$

which is the least squares solution to the inverse problem $\mathbf{Gm} = \mathbf{d}$.

3.5 Some Examples

3.5.1 THE STRAIGHT LINE PROBLEM

In the straight line problem the model is $d_i = m_1 + m_2 z_i$, so the equation $\mathbf{Gm} = \mathbf{d}$ has the form

$$\begin{bmatrix} 1 & z_1 \\ 1 & z_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & z_N \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \cdot \\ \cdot \\ d_N \end{bmatrix} \quad (3.13)$$

forming the matrix products

$$\mathbf{G}^T \mathbf{G} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_N \end{bmatrix} \begin{bmatrix} 1 & z_1 \\ 1 & z_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & z_N \end{bmatrix} = \begin{bmatrix} N & \sum z_i \\ \sum z_i & \sum z_i^2 \end{bmatrix} \quad (3.14)$$

and

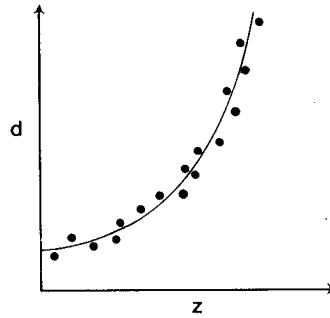
$$\mathbf{G}^T \mathbf{d} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_N \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ \cdot \\ \cdot \\ d_N \end{bmatrix} = \begin{bmatrix} \sum d_i \\ \sum z_i d_i \end{bmatrix} \quad (3.15)$$

This gives the least squares solution

$$\mathbf{m} = [\mathbf{G}^T \mathbf{G}]^{-1} \mathbf{G}^T \mathbf{d} = \begin{bmatrix} N & \sum z_i \\ \sum z_i & \sum z_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum d_i \\ \sum z_i d_i \end{bmatrix} \quad (3.16)$$

3.5.2 FITTING A PARABOLA

The problem of fitting a parabola is a trivial generalization of fitting a straight line (Fig. 3.4). Now the model is $d_i = m_1 + m_2 z_i + m_3 z_i^2$, so

Fig. 3.4. Least squares fitting of a parabola to (z, d) data.

the equation $\mathbf{Gm} = \mathbf{d}$ has the form

$$\begin{bmatrix} 1 & z_1 & z_1^2 \\ 1 & z_2 & z_2^2 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & z_N & z_N^2 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ m_3 \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ \vdots \\ d_N \end{bmatrix} \quad (3.17)$$

forming the matrix products $\mathbf{G}^T \mathbf{G}$.

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_N \\ z_1^2 & z_2^2 & \cdots & z_N^2 \end{bmatrix} \begin{bmatrix} 1 & z_1 & z_1^2 \\ 1 & z_2 & z_2^2 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & z_N & z_N^2 \end{bmatrix} = \begin{bmatrix} N & \sum z_i & \sum z_i^2 \\ \sum z_i & \sum z_i^2 & \sum z_i^3 \\ \sum z_i^2 & \sum z_i^3 & \sum z_i^4 \end{bmatrix} \quad (3.18)$$

and

$$\mathbf{G}^T \mathbf{d} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_N \\ z_1^2 & z_2^2 & \cdots & z_N^2 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix} = \begin{bmatrix} \sum d_i \\ \sum z_i d_i \\ \vdots \\ \sum z_i^2 d_i \end{bmatrix} \quad (3.19)$$

giving the least squares solution

$$\mathbf{m} = [\mathbf{G}^T \mathbf{G}]^{-1} \mathbf{G}^T \mathbf{d} = \begin{bmatrix} N & \sum z_i & \sum z_i^2 \\ \sum z_i & \sum z_i^2 & \sum z_i^3 \\ \sum z_i^2 & \sum z_i^3 & \sum z_i^4 \end{bmatrix}^{-1} \begin{bmatrix} \sum d_i \\ \sum z_i d_i \\ \sum z_i^2 d_i \end{bmatrix} \quad (3.20)$$

3.5.3 FITTING A PLANE SURFACE

To fit a plane surface, two auxiliary variables, say, x and y , are needed. The model (Fig. 3.5) is

$$d_i = m_1 + m_2 x_i + m_3 y_i$$

so the equation $\mathbf{Gm} = \mathbf{d}$ has the form

$$\begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & x_N & y_N \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ m_3 \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \cdot \\ \cdot \\ \cdot \\ d_N \end{bmatrix} \quad (3.21)$$

forming the matrix products $\mathbf{G}^T \mathbf{G}$

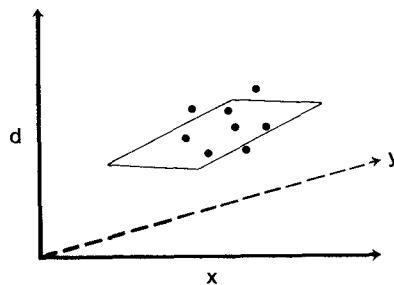


Fig. 3.5. Least squares fitting of a plane to (x, y, d) data.

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_N \\ y_1 & y_2 & \cdots & y_N \end{bmatrix} \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 1 & x_N & y_N \end{bmatrix} = \begin{bmatrix} N & \sum x_i & \sum y_i \\ \sum x_i & \sum x_i^2 & \sum x_i y_i \\ \sum y_i & \sum x_i y_i & \sum y_i^2 \end{bmatrix} \quad (3.22)$$

and

$$\mathbf{G}^T \mathbf{d} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_N \\ y_1 & y_2 & \cdots & y_N \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix} = \begin{bmatrix} \sum d_i \\ \sum x_i d_i \\ \vdots \\ \sum y_i d_i \end{bmatrix} \quad (3.23)$$

giving the least squares solution

$$\mathbf{m} = [\mathbf{G}^T \mathbf{G}]^{-1} \mathbf{G}^T \mathbf{d} = \begin{bmatrix} N & \sum x_i & \sum y_i \\ \sum x_i & \sum x_i^2 & \sum x_i y_i \\ \sum y_i & \sum x_i y_i & \sum y_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum d_i \\ \sum x_i d_i \\ \sum y_i d_i \end{bmatrix} \quad (3.24)$$

3.6 The Existence of the Least Squares Solution

The least squares solution arose from consideration of an inverse problem that had no exact solution. Since there was no exact solution, we chose to do the next best thing: to estimate the solution by those values of the model parameters that gave the best approximate solution (where “best” meant minimizing the L_2 prediction error). By writing a single formula $\mathbf{m}^{\text{est}} = [\mathbf{G}^T \mathbf{G}]^{-1} \mathbf{G}^T \mathbf{d}$, we implicitly assumed that there was only one such “best” solution. As we shall prove later, least squares fails if the number of solutions that give the same minimum prediction error is greater than one.

To see that least squares fails for problems with nonunique solutions, consider the straight line problem with only one data point (Fig. 3.6). It is clear that this problem is nonunique; many possible lines can

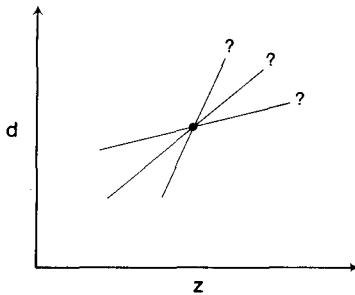


Fig. 3.6. An infinity of different lines can pass through a single point. The prediction error for each is $E = 0$.

pass through the point, and each has zero prediction error. The solution then contains the expression

$$[\mathbf{G}^T \mathbf{G}]^{-1} = \begin{bmatrix} N & \sum_{i=1}^N z_i \\ \sum_{i=1}^N z_i & \sum_{i=1}^N z_i^2 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & z_1 \\ z_1 & z_1^2 \end{bmatrix}^{-1} \quad (3.25)$$

The inverse of a matrix is proportional to the reciprocal of the determinant of the matrix, so that

$$[\mathbf{G}^T \mathbf{G}]^{-1} \propto 1/(z_1^2 - z_1^2).$$

This expression clearly is singular. The formula for the least squares solution fails.

The question of whether the equation $\mathbf{Gm} = \mathbf{d}$ provides enough information to specify uniquely the model parameters serves as a basis for classifying inverse problems. A classification system based on this criterion is discussed in the following sections (3.6.1 — 3.6.3).

3.6.1 UNDERDETERMINED PROBLEMS

When the equation $\mathbf{Gm} = \mathbf{d}$ does not provide enough information to determine uniquely all the model parameters, the problem is said to be *underdetermined*. As we saw in the example above, this can happen if there are several solutions that have zero prediction error. From elementary linear algebra we know that underdetermined problems occur when there are more unknowns than data, that is, when $M > N$.

We must note, however, that there is no special reason why the prediction error must be zero for an underdetermined problem. Frequently the data uniquely determine some of the model parameters but not others. For example, consider the acoustic experiment in Figure 3.7. Since no measurements are made of the acoustic slowness in the second brick, it is clear that this model parameter is completely unconstrained by the data. In contrast, the acoustic slowness of the first brick is *overdetermined*, since in the presence of measurement noise no choice of s_1 can satisfy the data exactly. The equation describing this experiment is

$$h \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix} \quad (3.26)$$

where s_i is the slowness in the i th brick, h the brick width, and the d_i the measurements of travel time. If one were to attempt to solve this problem with least squares, one would find that the term $[G^T G]^{-1}$ is singular. Even though $M < N$, the problem is still underdetermined since the data kernel has a very poor structure. Although this is a rather trivial case in which only some of the model parameters are underdetermined, in realistic experiments the problem arises in more subtle forms.

We shall refer to underdetermined problems that have nonzero prediction error as *mixed-determined problems*, to distinguish them from *purely underdetermined problems* that have zero prediction error.

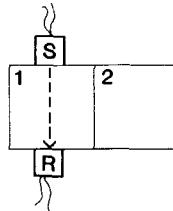


Fig. 3.7. An acoustic travel time experiment with source S and receiver R in a medium consisting of two discrete bricks. Because of poor experiment geometry, the acoustic waves (dashed line) pass only through brick 1. The slowness of brick 2 is completely underdetermined.

3.6.2 EVEN-DETERMINED PROBLEMS

In even-determined problems there is exactly enough information to determine the model parameters. There is only one solution, and it has zero prediction error.

3.6.3 OVERDETERMINED PROBLEMS

When there is too much information contained in the equation $\mathbf{Gm} = \mathbf{d}$ for it to possess an exact solution, we speak of it as being *overdetermined*. This is the case in which we can employ least squares to select a “best” approximate solution. Overdetermined problems typically have more data than unknowns, that is, $N > M$, although for the reasons discussed above it is possible to have problems that are to some degree overdetermined even when $N < M$ and to have problems that are to some degree underdetermined even when $N > M$.

To deal successfully with the full range of inverse problems, we shall need to be able to characterize whether an inverse problem is under- or overdetermined (or some combination of the two). We shall develop quantitative methods for making this characterization in Chapter 7. For the moment we assume that it is possible to characterize the problem intuitively on the basis of the kind of experiment the problem represents.

3.7 The Purely Underdetermined Problem

Suppose that an inverse problem $\mathbf{Gm} = \mathbf{d}$ has been identified as one that is purely underdetermined. For simplicity, assume that there are fewer equations than unknown model parameters, that is, $N < M$, and that there are no inconsistencies in these equations. It is therefore possible to find more than one solution for which the prediction error E is zero. (In fact, we shall show that underdetermined linear inverse problems have an infinite number of such solutions.) Although the data provide information about the model parameters, they do not provide enough to determine them uniquely.

To obtain a solution \mathbf{m}^{est} to the inverse problem, we must have some means of singling out precisely one of the infinite number of solutions with zero prediction error E . To do this, we must add to the problem some information not contained in the equation $\mathbf{Gm} = \mathbf{d}$. This extra information is called *a priori* information [Ref. 10]. *A priori* information can take many forms, but in each case it quantifies expectations about the character of the solution that are not based on the actual data.

For instance, in the case of fitting a straight line through a single data point, one might have the expectation that the line also passes through the origin. This a priori information now provides enough information to solve the inverse problem uniquely, since two points (one datum, one a priori) determine a line.

Another example of a priori information concerns expectations that the model parameters possess a given sign, or lie in a given range. For instance, suppose the model parameters represent density at different points in the earth. Even without making any measurements, one can state with certainty that the density is everywhere positive, since density is an inherently positive quantity. Furthermore, since the interior of the earth can reasonably be assumed to be rock, its density must have values in some range known to characterize rock, say, between 1 and 100 gm/cm³. If one can use this a priori information when solving the inverse problem, it may greatly reduce the range of possible solutions—or even cause the solution to be unique.

There is something unsatisfying about having to add a priori information to an inverse problem to single out a solution. Where does this information come from, and how certain is it? There are no firm answers to these questions. In certain instances one might be able to identify reasonable a priori assumptions; in other instances, one might not. Clearly, the importance of the a priori information depends greatly on the *use* one plans for the estimated model parameters. If one simply wants one example of a solution to the problem, the choice of a priori information is unimportant. However, if one wants to develop arguments that depend on the uniqueness of the estimates, the validity of the a priori assumptions is of paramount importance. These problems are the price one must pay for estimating the model parameters of a nonunique inverse problem. As will be shown in Chapter 6, there are other kinds of “answers” to inverse problems that do not depend on a priori information (localized averages, for example). However, these “answers” invariably are not as easily interpretable as estimates of model parameters.

The first kind of a priori assumption we shall consider is the expectation that the solution to the inverse problem is “simple,” where the notion of simplicity is quantified by some measure of the length of the solution. One such measure is simply the Euclidean length of the solution, $L = \mathbf{m}^T \mathbf{m} = \sum m_i^2$. A solution is therefore defined to be simple if it is small when measured under the L_2 norm. Admittedly, this measure is perhaps not a particularly realistic measure of simplicity. It can be useful occasionally, and we shall describe shortly how it

can be generalized to more realistic measures. One instance in which solution length may be realistic is when the model parameters describe the velocity of various points in a moving fluid. The length L is then a measure of the kinetic energy of the fluid. In certain instances it may be appropriate to find that velocity field in the fluid that has the smallest possible kinetic energy of those solutions satisfying the data.

We pose the following problem: Find the \mathbf{m}^{est} that minimizes $L = \mathbf{m}^T \mathbf{m} = \sum m_i^2$ subject to the constraint that $\mathbf{e} = \mathbf{d} - \mathbf{Gm} = 0$. This problem can easily be solved by the method of Lagrange multipliers (see Appendix A.1). We minimize the function as

$$\Phi(\mathbf{m}) = L + \sum_{i=1}^N \lambda_i e_i = \sum_{i=1}^M m_i^2 + \sum_{i=1}^N \lambda_i \left[d_i - \sum_{j=1}^M G_{ij} m_j \right] \quad (3.27)$$

with respect to m_q , where λ_i are the Lagrange multipliers. Taking the derivatives yields

$$\frac{\partial \Phi}{\partial m_q} = \sum_{i=1}^M 2 \frac{\partial m_i}{\partial m_q} m_i - \sum_{i=1}^N \lambda_i \sum_{j=1}^M G_{ij} \frac{\partial m_j}{\partial m_q} = 2m_q - \sum_{i=1}^N \lambda_i G_{iq} \quad (3.28)$$

Setting this result to zero and rewriting it in matrix notation yields the equation $2\mathbf{m} = \mathbf{G}^T \boldsymbol{\lambda}$, which must be solved along with the constraint equation $\mathbf{Gm} = \mathbf{d}$. Plugging the first equation into the second gives $\mathbf{d} = \mathbf{Gm} = \mathbf{G}[\mathbf{G}^T \boldsymbol{\lambda}/2]$. We note that the matrix \mathbf{GG}^T is a square $N \times N$ matrix. If its inverse exists, we can then solve this equation for the Lagrange multipliers, $\boldsymbol{\lambda} = 2[\mathbf{GG}^T]^{-1}\mathbf{d}$. Then inserting this expression into the first equation yields the solution

$$\mathbf{m}^{\text{est}} = \mathbf{G}^T [\mathbf{GG}^T]^{-1} \mathbf{d} \quad (3.29)$$

We shall discuss the conditions under which this solution exists later. As we shall see, one condition is that the equation $\mathbf{Gm} = \mathbf{d}$ be purely underdetermined—that it contain no inconsistencies.

3.8 Mixed–Determined Problems

Most inverse problems that arise in practice are neither completely overdetermined nor completely underdetermined. For instance, in the x-ray tomography problem there may be one box through which several rays pass (Fig. 3.8a). The x-ray opacity of this box is clearly overdetermined. On the other hand, there may be boxes that have been missed entirely (Fig. 3.8b). These boxes are completely undeter-

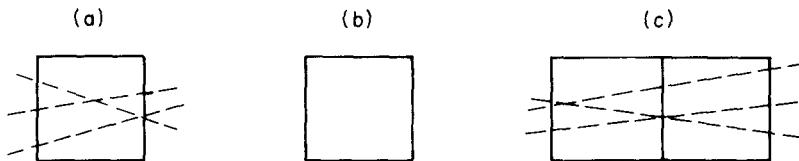


Fig. 3.8. (a) The x-ray opacity of the box is overdetermined, since measurements of x-ray intensity are made along three different paths (dashed lines). (b) The opacity is underdetermined since no measurements have been made. (c) The average opacity of these two boxes is overdetermined, but since each path has an equal length in either brick, the individual opacities are underdetermined.

mined. There may also be boxes that cannot be individually resolved because every ray that passes through one also passes through an equal distance of the other (Fig. 3.8c). These boxes are also underdetermined, since only their mean opacity is determined.

Ideally, we would like to sort the unknown model parameters into two groups; those that are overdetermined and those that are underdetermined. Actually, to do this we need to form a new set of model parameters that are linear combinations of the old. For example, in the two-box problem above, the average opacity $m'_1 = (m_1 + m_2)/2$ is completely overdetermined, whereas the difference in opacity $m'_2 = (m_1 - m_2)/2$ is completely underdetermined. We want to perform this partitioning from an arbitrary equation $Gm = d \rightarrow G'm' = d'$, where m' is partitioned into an upper part m^o that is overdetermined and a lower part m^u that is underdetermined:

$$\begin{bmatrix} G^o & \mathbf{0} \\ \mathbf{0} & G^u \end{bmatrix} \begin{bmatrix} m^o \\ m^u \end{bmatrix} = \begin{bmatrix} d^o \\ d^u \end{bmatrix} \quad (3.30)$$

If this can be achieved, we could determine the overdetermined model parameters by solving the upper equations in the least squares sense and determine the underdetermined model parameters by finding those that have minimum L_2 solution length. In addition, we would have found a solution that added as little a priori information to the inverse problem as possible.

This partitioning process can be accomplished through singular-value decomposition of the data kernel, a process that we shall discuss in Chapter 7. Since it is a relatively time-consuming process, we first examine an approximate process that works if the inverse problem is not too underdetermined.

Instead of partitioning \mathbf{m} , suppose that we determine a solution that minimizes some combination Φ of the prediction error and the solution length for the unpartitioned model parameters:

$$\Phi(\mathbf{m}) = E + \epsilon^2 L = \mathbf{e}^T \mathbf{e} + \epsilon^2 \mathbf{m}^T \mathbf{m} \quad (3.31)$$

where the weighting factor ϵ^2 determines the relative importance given to the prediction error and solution length. If ϵ is made large enough, this procedure will clearly minimize the underdetermined part of the solution. Unfortunately, it also tends to minimize the overdetermined part of the solution. As a result, the solution will not minimize the prediction error E and will not be a very good estimate of the true model parameters. If ϵ is set to zero, the prediction error will be minimized, but no a priori information will be provided to single out the underdetermined model parameters. It may be possible, however, to find some compromise value for ϵ that will approximately minimize E while approximately minimizing the length of the underdetermined part of the solution. There is no simple method of determining what this compromise ϵ should be (without solving the partitioned problem); it must be determined by trial and error. By minimizing $\Phi(\mathbf{m})$ with respect to the model parameters in a manner exactly analogous to the least squares derivation, we obtain

$$\mathbf{m}^{\text{est}} = [\mathbf{G}^T \mathbf{G} + \epsilon^2 \mathbf{I}]^{-1} \mathbf{G}^T \mathbf{d} \quad (3.32)$$

This estimate of the model parameters is called the *damped least squares* solution. The concept of error has been generalized to include not only prediction error but *solution error* (solution length). The underdeterminacy of the inverse problem is said to have been damped.

3.9 Weighted Measures of Length as a Type of A Priori Information

There are many instances in which $L = \mathbf{m}^T \mathbf{m}$ is not a very good measure of solution simplicity. For instance, suppose that one were solving an inverse problem for density fluctuations in the ocean. One may not want to find a solution that is smallest in the sense of closest to zero but one that is smallest in the sense that it is closest to some other value, such as the average density of sea water. The obvious generalization of L is then

$$L = (\mathbf{m} - \langle \mathbf{m} \rangle)^T (\mathbf{m} - \langle \mathbf{m} \rangle) \quad (3.33)$$

where $\langle \mathbf{m} \rangle$ is the a priori value of the model parameters.

Sometimes the whole idea of length as a measure of simplicity is inappropriate. For instance, one may feel that a solution is simple if it is smooth, or if it is in some sense flat. These measures may be particularly appropriate when the model parameters represent a discretized continuous function such as density or x-ray opacity. One may have the expectation that these parameters vary only slowly with position. Fortunately, properties such as flatness can be easily quantified by measures that are generalizations of length. For example, the flatness of a continuous function of space can be quantified by the norm of its first derivative. For discrete model parameters, one can use the difference between physically adjacent model parameters as approximations of a derivative. The flatness \mathbf{I} of a vector \mathbf{m} is then

$$\mathbf{I} = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & \\ & & & & -1 & 1 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ \vdots \\ m_M \end{bmatrix} = \mathbf{D}\mathbf{m} \quad (3.34)$$

where \mathbf{D} is the flatness matrix. Other methods of simplicity can also be represented by a matrix multiplying the model parameters. For instance, solution roughness can be quantified by the second derivative. The matrix multiplying the model parameters would then have rows containing $[\dots 1 \ -2 \ 1 \ \dots]$. The overall roughness or flatness of the solution is then just the length

$$L = \|\mathbf{I}\| = \|\mathbf{D}\mathbf{m}\| = \mathbf{m}^T \mathbf{D}^T \mathbf{D}\mathbf{m} = \mathbf{m}^T \mathbf{W}_m \mathbf{m} \quad (3.35)$$

The matrix $\mathbf{W}_m = \mathbf{D}^T \mathbf{D}$ can be interpreted as a weighting factor that enters into the calculation of the length of the vector \mathbf{m} . Note, however, that $\|\mathbf{m}\|_{\text{weighted}}^2 = \mathbf{m}^T \mathbf{W}_m \mathbf{m}$ is *not* a proper norm, since it violates the positivity condition given in Eq. (3.3a), that is, $\|\mathbf{m}\|_{\text{weighted}}^2 = 0$ for some nonzero vectors (such as the constant vector). This behavior usually poses no insurmountable problems, but it can cause solutions based on minimizing this norm to be nonunique.

The measure of solution simplicity can therefore be generalized to

$$L = [\mathbf{m} - \langle \mathbf{m} \rangle]^T \mathbf{W}_m [\mathbf{m} - \langle \mathbf{m} \rangle] \quad (3.36)$$

By suitably choosing the a priori model vector $\langle \mathbf{m} \rangle$ and the weighting matrix \mathbf{W}_m , we can quantify a wide variety of measures of simplicity.

Weighted measures of the prediction error can also be useful. Frequently some observations are made with more accuracy than others. In this case one would like the prediction error e_i of the more accurate observations to have a greater weight in the quantification of the overall error E than the inaccurate observations. To accomplish this weighting, we define a generalized prediction error

$$E = \mathbf{e}^T \mathbf{W}_e \mathbf{e}$$

where the matrix \mathbf{W}_e defines the relative contribution of each individual error to the total prediction error. Normally we would choose this matrix to be diagonal. For example, if $N = 5$ and the third observation is known to be twice as accurately determined as the others, one might use

$$\text{diag}(\mathbf{W}_e) = [1, 1, 2, 1, 1]^T$$

The inverse problem solutions stated above can then be modified to take into account these new measures of prediction error and solution simplicity. The derivations are substantially the same as for the unweighted cases but the algebra is more lengthy.

3.9.1 WEIGHTED LEAST SQUARES

If the equation $\mathbf{Gm} = \mathbf{d}$ is completely overdetermined, then one can estimate the model parameters by minimizing the generalized prediction error $E = \mathbf{e}^T \mathbf{W}_e \mathbf{e}$. This procedure leads to the solution

$$\mathbf{m}^{\text{est}} = [\mathbf{G}^T \mathbf{W}_e \mathbf{G}]^{-1} \mathbf{G}^T \mathbf{W}_e \mathbf{d} \quad (3.37)$$

3.9.2 WEIGHTED MINIMUM LENGTH

If the equation $\mathbf{Gm} = \mathbf{d}$ is completely underdetermined, then one can estimate the model parameters by choosing the solution that is simplest, where simplicity is defined by the generalized length $L = [\mathbf{m} - \langle \mathbf{m} \rangle]^T \mathbf{W}_m [\mathbf{m} - \langle \mathbf{m} \rangle]^T$. This procedure leads to the solution

$$\mathbf{m}^{\text{est}} = \langle \mathbf{m} \rangle + \mathbf{W}_m \mathbf{G}^T [\mathbf{G} \mathbf{W}_m \mathbf{G}^T]^{-1} [\mathbf{d} - \mathbf{G} \langle \mathbf{m} \rangle] \quad (3.38)$$

3.9.3 WEIGHTED DAMPED LEAST SQUARES

If the equation $\mathbf{Gm} = \mathbf{d}$ is slightly underdetermined, it can often be solved by minimizing a combination of prediction error and solution length, $E + \epsilon^2 L$ [Refs. 8,9,11]. The parameter ϵ is chosen by trial and

error to yield a solution that has a reasonably small prediction error. The estimate of the solution is then

$$\mathbf{m}^{\text{est}} = \langle \mathbf{m} \rangle + [\mathbf{G}^T \mathbf{W}_e \mathbf{G} + \epsilon^2 \mathbf{W}_m]^{-1} \mathbf{G}^T \mathbf{W}_e [\mathbf{d} - \mathbf{G} \langle \mathbf{m} \rangle] \quad (3.39)$$

which is equivalent to

$$\mathbf{m}^{\text{est}} = \langle \mathbf{m} \rangle + \mathbf{W}_m^{-1} \mathbf{G}^T [\mathbf{G} \mathbf{W}_m^{-1} \mathbf{G}^T + \epsilon^2 \mathbf{W}_e^{-1}]^{-1} [\mathbf{d} - \mathbf{G} \langle \mathbf{m} \rangle] \quad (3.40)$$

(see Section 5.9 for a proof.) In both instances one must take care to ascertain whether the inverses actually exist. Depending on the choice of the weighting matrices, sufficient a priori information may or may not have been added to the problem to damp the underdeterminacy.

3.10 Other Types of A Priori Information

One commonly encountered type of a priori information is the knowledge that some function of the model parameters equals a constant. Linear equality constraints of the form $\mathbf{F}\mathbf{m} = \mathbf{h}$ are particularly easy to implement. For example, one such linear constraint requires that the mean of the model parameters must equal some value h_1 :

$$\mathbf{F}\mathbf{m} = \frac{1}{M} [1 \quad 1 \quad 1 \quad \cdots \quad 1] \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_M \end{bmatrix} = [h_1] = \mathbf{h} \quad (3.41)$$

Another such constraint requires that a particular model parameter equal a given value

$$\mathbf{F}\mathbf{m} = [0 \quad \cdots \quad 0 \quad 1 \quad 0 \cdots \quad 0] \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_M \end{bmatrix} = [h_1] = \mathbf{h} \quad (3.42)$$

One problem that frequently arises is to solve a inverse problem $\mathbf{Gm} = \mathbf{d}$ in the least squares sense with the a priori constraint that linear relationships between the model parameters of the form $\mathbf{Fm} = \mathbf{h}$ are satisfied exactly. One way to implement this constraint is to include the constraint equations as rows in $\mathbf{Gm} = \mathbf{d}$ and adjust the weighting matrix \mathbf{W}_e so that these equations are given infinitely more weight than the other equations [Ref. 14]. (In practice one gives them large but finite weight.) The prediction error of the constraints, therefore, is forced to zero at the expense of increasing the prediction error of the other equations.

Another method of implementing the constraints is through the use of Lagrange multipliers. One minimizes $E = \mathbf{e}^T \mathbf{e}$ with the constraint that $\mathbf{Fm} - \mathbf{h} = 0$ by forming the function

$$\Phi(\mathbf{m}) = \sum_{i=1}^N \left[\sum_{j=1}^M G_{ij} m_j - d_i \right]^2 + 2 \sum_{i=1}^p \lambda_i \left[\sum_{j=1}^M F_{ij} m_j - h_i \right] \quad (3.43)$$

(where there are p constraints and $2\lambda_i$ are the Lagrange multipliers) and setting its derivatives with respect to the model parameters to zero as

$$\frac{\partial \Phi(\mathbf{m})}{\partial m_q} = 2 \sum_{i=1}^M m_i \sum_{j=1}^N G_{jq} G_{ji} - 2 \sum_{i=1}^N G_{iq} d_i + 2 \sum_{i=1}^p \lambda_i F_{iq} = 0 \quad (3.44)$$

This equation must be solved simultaneously with the constraint equations $\mathbf{Fm} = \mathbf{h}$ to yield the estimated solution. These equations, in matrix form, are

$$\begin{bmatrix} \mathbf{G}^T \mathbf{G} & \mathbf{F}^T \\ \mathbf{F} & \mathbf{O} \end{bmatrix} \begin{bmatrix} \mathbf{m} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{G}^T \mathbf{d} \\ \mathbf{h} \end{bmatrix} \quad (3.45)$$

Although these equations can be manipulated to yield an explicit formula for \mathbf{m}^{est} , it is often more convenient to solve directly this $M + p$ system of equations for M estimates of model parameters and p Lagrange multipliers by premultiplying by the inverse of the square matrix.

3.10.1 EXAMPLE: CONSTRAINED FITTING OF A STRAIGHT LINE

Consider the problem of fitting the straight line $d_i = m_1 + m_2 z_i$ to data, where one has a priori information that the line must pass through the point (z', d') (Fig. 3.9). There are two model parameters:

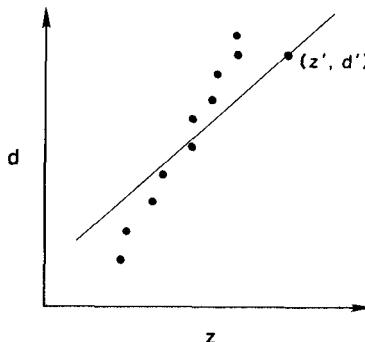


Fig. 3.9. Least squares fitting of a straight line to (z, d) data, where the line is constrained to pass through the point (z', d') .

intercept m_1 and slope m_2 . The $p = 1$ constraint is that $d' = m_1 + m_2 z'$, or

$$\mathbf{Fm} = [1 \quad z'] \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} = [d'] = \mathbf{h} \quad (3.46)$$

Using the $\mathbf{G}^T\mathbf{G}$ and $\mathbf{G}^T\mathbf{d}$ computed in Section 3.5.1, the solution is

$$\begin{bmatrix} m_1^{\text{est}} \\ m_2^{\text{est}} \\ \lambda_1 \end{bmatrix} = \begin{bmatrix} N & \sum z_i & 1 \\ \sum z_i & \sum z_i^2 & z' \\ 1 & z' & 0 \end{bmatrix}^{-1} \begin{bmatrix} \sum d_i \\ \sum z_i d_i \\ d' \end{bmatrix} \quad (3.47)$$

Another kind of a priori constraint is the *linear inequality constraint*, which we can write as $\mathbf{Fm} \geq \mathbf{h}$, (the inequality being interpreted component by component). Note that this form can also include \leq inequalities by multiplying the inequality relation by -1 . This kind of a priori constraint has application to problems in which the model parameters are inherently positive quantities, $m_i > 0$, and to other cases when the solution is known to possess some kind of bounds. One could therefore propose a new kind of constrained least squares solution of overdetermined problems, one that minimizes the error subject to the given inequality constraints. A priori inequality constraints also have application to underdetermined problems. One can find the smallest solution that solves both $\mathbf{Gm} = \mathbf{d}$ and $\mathbf{Fm} \geq \mathbf{h}$. These problems can be solved in a straightforward fashion, which will be discussed in Chapter 7.

3.11 The Variance of the Model Parameter Estimates

The data invariably contain noise that cause errors in the estimates of the model parameters. We can calculate how this measurement error “maps into” errors in \mathbf{m}^{est} by noting that all of the formulas derived above for estimates of the model parameters are linear functions of the data, of the form $\mathbf{m}^{\text{est}} = \mathbf{M}\mathbf{d} + \mathbf{v}$, where \mathbf{M} is some matrix and \mathbf{v} some vector. Therefore, if we assume that the data have a distribution characterized by some covariance matrix $[\text{cov } \mathbf{d}]$, the estimates of the model parameters have a distribution characterized by a covariance matrix $[\text{cov } \mathbf{m}] = \mathbf{M}[\text{cov } \mathbf{d}]\mathbf{M}^T$. The covariance of the solution can therefore be calculated in a straightforward fashion. If the data are uncorrelated and all of equal variance σ_d^2 , then very simple formulas are obtained for the covariance of some of the more simple inverse problem solutions.

The simple least squares solution $\mathbf{m}^{\text{est}} = [\mathbf{G}^T\mathbf{G}]^{-1}\mathbf{G}^T\mathbf{d}$ has covariance

$$[\text{cov } \mathbf{m}] = [[\mathbf{G}^T\mathbf{G}]^{-1}\mathbf{G}^T]\sigma_d^2\mathbf{I}[[\mathbf{G}^T\mathbf{G}]^{-1}\mathbf{G}^T]^T = \sigma_d^2[\mathbf{G}^T\mathbf{G}]^{-1} \quad (3.48)$$

and the simple minimum length solution $\mathbf{m}^{\text{est}} = \mathbf{G}^T[\mathbf{G}\mathbf{G}^T]^{-1}\mathbf{d}$ has covariance

$$[\text{cov } \mathbf{m}] = [\mathbf{G}^T[\mathbf{G}\mathbf{G}^T]^{-1}]\sigma_d^2\mathbf{I}[\mathbf{G}^T[\mathbf{G}\mathbf{G}^T]^{-1}]^T = \sigma_d^2\mathbf{G}^T[\mathbf{G}\mathbf{G}^T]^{-2}\mathbf{G} \quad (3.49)$$

3.12 Variance and Prediction Error of the Least Squares Solution

If the prediction error $E(\mathbf{m}) = \mathbf{e}^T\mathbf{e}$ of an overdetermined problem has a very sharp minimum in the vicinity of the estimated solution \mathbf{m}^{est} , we would expect that the solution is well determined in the sense that it has small variance. Small errors in determining the shape of $E(\mathbf{m})$ due to random fluctuations in the data lead to only small errors in \mathbf{m}^{est} (Fig. 3.10a). Conversely, if $E(\mathbf{m})$ has a broad minimum, we expect that \mathbf{m}^{est} has a large variance (Fig. 3.10b). Since the curvature of a function is a measure of the sharpness of its minimum, we expect that the variance of the solution is related to the curvature of $E(\mathbf{m})$ at

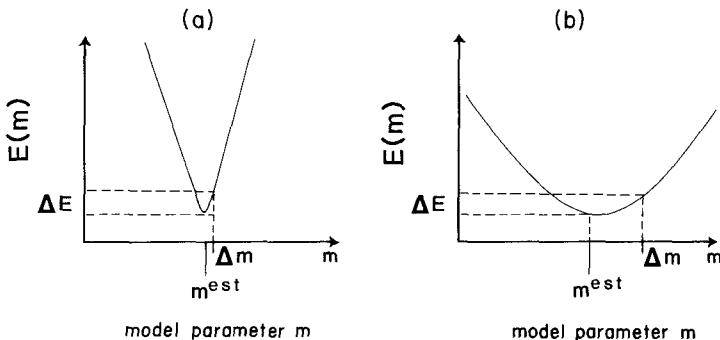


Fig. 3.10. (a) The best estimate m^{est} of model parameter m occurs at the minimum of $E(m)$. If the minimum is relatively narrow, then random fluctuations in $E(m)$ lead to only small errors Δm in m^{est} . (b) If the minimum is wide, then large errors in m can occur.

its minimum. The curvature of the prediction error can be measured by its second derivative, as we can see by computing how small changes in the model parameters change the prediction error. Expanding the prediction error in a Taylor series about its minimum and keeping up to second order terms gives

$$\Delta E = E(\mathbf{m}) - E(\mathbf{m}^{\text{est}}) = [\mathbf{m} - \mathbf{m}^{\text{est}}]^T \left[\frac{1}{2} \frac{\partial^2 E}{\partial \mathbf{m}^2} \right]_{\mathbf{m}=\mathbf{m}^{\text{est}}} [\mathbf{m} - \mathbf{m}^{\text{est}}] \quad (3.50)$$

Note that the first-order term is zero, since the expansion is made at a minimum. The second derivative can also be computed directly from the expression

$$E(\mathbf{m}) = \mathbf{e}^T \mathbf{e} = [\mathbf{d} - \mathbf{G}\mathbf{m}]^T [\mathbf{d} - \mathbf{G}\mathbf{m}]$$

which gives

$$\left[\frac{1}{2} \frac{\partial^2 E}{\partial \mathbf{m}^2} \right]_{\mathbf{m}=\mathbf{m}^{\text{est}}} = \frac{1}{2} \frac{\partial^2}{\partial \mathbf{m}^2} \left[\mathbf{d} - \mathbf{G}\mathbf{m} \right]^2 = \frac{\partial}{\partial \mathbf{m}} \left[-\mathbf{G}^T [\mathbf{d} - \mathbf{G}\mathbf{m}] \right] = \mathbf{G}^T \mathbf{G} \quad (3.51)$$

The covariance of the least squares solution (assuming uncorrelated data all with equal variance σ_d^2) is therefore

$$[\text{cov } \mathbf{m}] = \sigma_d^2 [\mathbf{G}^T \mathbf{G}]^{-1} = \sigma_d^2 \left[\frac{1}{2} \frac{\partial^2 E}{\partial \mathbf{m}^2} \right]_{\mathbf{m}=\mathbf{m}^{\text{est}}}^{-1} \quad (3.52)$$

The prediction error $E = \mathbf{e}^T \mathbf{e}$ is the sum of squares of Gaussian data minus a constant. It is, therefore, a random variable with a χ^2 distribution with $N - M$ degrees of freedom, which has mean $(N - M)\sigma_d^2$ and variance $2(N - M)\sigma_d^4$. (The degrees of freedom are reduced by M since the model can force M linear combinations of the e_i to zero.) We can use the standard deviation of E , $\sigma_E = [2(N - M)]^{1/2}\sigma_d^2$ in the expression for variance as

$$[\text{cov } \mathbf{m}] = \sigma_d^2 [\mathbf{G}^T \mathbf{G}]^{-1} = \frac{\sigma_E}{[2(N - M)]^{1/2}} \left[\frac{1}{2} \frac{\partial^2 E}{\partial \mathbf{m}^2} \right]_{\mathbf{m}=\mathbf{m}^{\text{est}}}^{-1} \quad (3.53)$$

The covariance $[\text{cov } \mathbf{m}]$ can be interpreted as being controlled either by the variance of the data times a measure of how error in the data is mapped into error in the model parameters, or by the standard deviation of the total prediction error times a measure of the curvature of the prediction error at its minimum.

The methods of solving inverse problems that have been discussed in this chapter emphasize the data and model parameters themselves. The method of least squares estimates the model parameters with smallest prediction length. The method of minimum length estimates the simplest model parameters. The ideas of data and model parameters are very concrete and straightforward, and the methods based on them are simple and easily understood. Nevertheless, this viewpoint tends to obscure an important aspect of inverse problems: that the nature of the problems depends more on the *relationship* between the data and model parameters than on the data or model parameters themselves. It should, for instance, be possible to tell a well-designed experiment from a poor one without knowing what the numerical values of the data or model parameters are, or even the range in which they fall. In the next chapter we will begin to explore this kind of problem.

4

SOLUTION OF THE LINEAR, GAUSSIAN INVERSE PROBLEM, VIEWPOINT 2: GENERALIZED INVERSES

4.1 Solutions versus Operators

In the previous chapter we derived methods of solving the linear inverse problem $\mathbf{Gm} = \mathbf{d}$ that were based on examining two properties of its solution: prediction error and solution simplicity (or length). Most of these solutions had a form that was linear in the data, $\mathbf{m}^{\text{est}} = \mathbf{Md} + \mathbf{v}$, where \mathbf{M} is some matrix and \mathbf{v} some vector, both of which are independent of the data \mathbf{d} . This equation indicates that the estimate of the model parameters is controlled by some matrix \mathbf{M} operating on the data (that is, multiplying the data). We therefore shift our emphasis from the estimates \mathbf{m}^{est} to the operator matrix \mathbf{M} , with the expectation that by studying it we can learn more about the properties of inverse problems. Since the matrix \mathbf{M} solves, or “inverts,” the inverse problem $\mathbf{Gm} = \mathbf{d}$, it is often called the *generalized*

inverse and given the symbol \mathbf{G}^{-g} . The exact form of the generalized inverse depends on the problem at hand. The generalized inverse of the overdetermined least squares problem is $\mathbf{G}^{-g} = [\mathbf{G}^T \mathbf{G}]^{-1} \mathbf{G}^T$, and for the minimum length underdetermined solution it is $\mathbf{G}^{-g} = \mathbf{G}^T [\mathbf{G} \mathbf{G}^T]^{-1}$.

Note that in some ways the generalized inverse is analogous to the ordinary matrix inverse. The solution to the square (even-determined) matrix equation $\mathbf{A}\mathbf{x} = \mathbf{y}$ is $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$, and the solution to the inverse problem $\mathbf{G}\mathbf{m} = \mathbf{d}$ is $\mathbf{m}^{est} = \mathbf{G}^{-g}\mathbf{d}$ (plus some vector, possibly). The analogy is very limited, however. The generalized inverse is not a matrix inverse in the usual sense. It is not square, and neither $\mathbf{G}^{-g}\mathbf{G}$ nor $\mathbf{G}\mathbf{G}^{-g}$ need equal an identity matrix.

4.2 The Data Resolution Matrix

Suppose we have found a generalized inverse that in some sense solves the inverse problem $\mathbf{G}\mathbf{m} = \mathbf{d}$, yielding an estimate of the model parameters $\mathbf{m}^{est} = \mathbf{G}^{-g}\mathbf{d}$ (for the sake of simplicity we assume that there is no additive vector). We can then retrospectively ask how well this estimate of the model parameters fits the data. By plugging our estimate into the equation $\mathbf{G}\mathbf{m} = \mathbf{d}$ we conclude

$$\mathbf{d}^{pre} = \mathbf{G}\mathbf{m}^{est} = \mathbf{G}[\mathbf{G}^{-g}\mathbf{d}^{obs}] = [\mathbf{G}\mathbf{G}^{-g}]\mathbf{d}^{obs} = \mathbf{N}\mathbf{d}^{obs} \quad (4.1)$$

Here the superscripts *obs* and *pre* mean observed and predicted, respectively. The $N \times N$ square matrix $\mathbf{N} = \mathbf{G}\mathbf{G}^{-g}$ is called the *data resolution matrix*. This matrix describes how well the predictions match the data. If $\mathbf{N} = \mathbf{I}$, then $\mathbf{d}^{pre} = \mathbf{d}^{obs}$ and the prediction error is zero. On the other hand, if the data resolution matrix is not an identity matrix, the prediction error is nonzero.

If the elements of the data vector \mathbf{d} possess a natural ordering, then the data resolution matrix has a simple interpretation. Consider, for example, the problem of fitting a straight line to (z, d) points, where the data have been ordered according to the value of the auxiliary variable z . If \mathbf{N} is not an identity matrix but is close to an identity matrix (in the sense that its largest elements are near its main diagonal), then the configuration of the matrix signifies that averages of neighboring data can be predicted, whereas individual data cannot. Consider the i th row of \mathbf{N} . If this row contained all zeros except for a

one in the i th column, then d_i would be predicted exactly. On the other hand, suppose that the row contained the elements

$$[\cdots 0 \ 0 \ 0 \ 0.1 \ 0.8 \ 0.1 \ 0 \ 0 \ 0 \cdots] \quad (4.2)$$

where the 0.8 is in the i th column. Then the i th datum is given by

$$d_i^{\text{pre}} = \sum_{j=1}^N N_{ij} d_j^{\text{obs}} = 0.1d_{i-1}^{\text{obs}} + 0.8d_i^{\text{obs}} + 0.1d_{i+1}^{\text{obs}} \quad (4.3)$$

The predicted value is a weighted average of three neighboring observed data. If the true data vary slowly with the auxiliary variable, then such an average might produce an estimate reasonably close to the observed value.

The rows of the data resolution matrix N describe how well neighboring data can be independently predicted, or *resolved*. If the data have a natural ordering, then a graph of the elements of the rows of N against column indices illuminates the sharpness of the resolution (Fig. 4.1). If the graphs have a single sharp maximum centered about the main diagonal, then the data are well resolved. If the graphs are very broad, then the data are poorly resolved. Even in cases where there is no natural ordering of the data, the resolution matrix still shows how much weight each observation has in influencing the

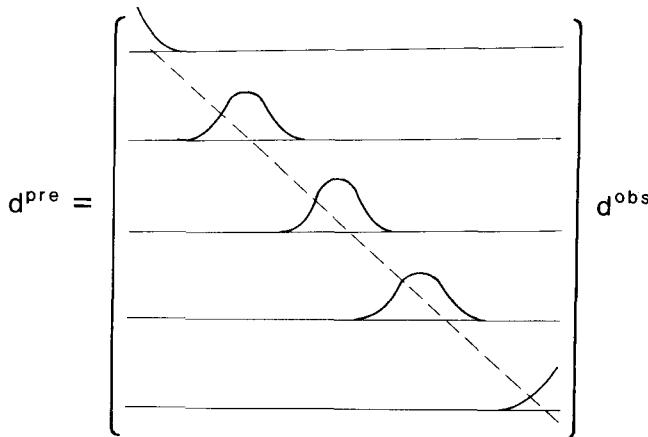


Fig. 4.1. Plots of selected rows of the data resolution matrix N indicate how well the data can be predicted. Narrow peaks occurring near the main diagonal of the matrix (dashed line) indicate that the resolution is good.

predicted value. There is then no special significance to whether large off-diagonal elements fall near to or far from the main diagonal.

Because the diagonal elements of the data resolution matrix indicate how much weight a datum has in its own prediction, these diagonal elements are often singled out and called the *importance* \mathbf{n} of the data [Ref. 15]:

$$\mathbf{n} = \text{diag}(\mathbf{N}) \quad (4.4)$$

The data resolution matrix is not a function of the data but only of the data kernel \mathbf{G} (which embodies the model and experimental geometry) and any a priori information applied to the problem. It can therefore be computed and studied without actually performing the experiment and can be a useful tool in experimental design.

4.3 The Model Resolution Matrix

The data resolution matrix characterizes whether the data can be independently predicted, or resolved. The same question can be asked about the model parameters. To explore this question we imagine that there is a true, but unknown set of model parameters \mathbf{m}^{true} that solve $\mathbf{G}\mathbf{m}^{\text{true}} = \mathbf{d}^{\text{obs}}$. We then inquire how closely a particular estimate of the model parameters \mathbf{m}^{est} is to this true solution. Plugging the expression for the observed data $\mathbf{G}\mathbf{m}^{\text{true}} = \mathbf{d}^{\text{obs}}$ into the expression for the estimated model $\mathbf{m}^{\text{est}} = \mathbf{G}^{-g}\mathbf{d}^{\text{obs}}$ gives

$$\mathbf{m}^{\text{est}} = \mathbf{G}^{-g}\mathbf{d}^{\text{obs}} = \mathbf{G}^{-g}[\mathbf{G}\mathbf{m}^{\text{true}}] = [\mathbf{G}^{-g}\mathbf{G}]\mathbf{m}^{\text{true}} = \mathbf{R}\mathbf{m}^{\text{true}} \quad (4.5)$$

[Ref. 20] Here \mathbf{R} is the $\mathbf{M} \times \mathbf{M}$ *model resolution matrix*. If $\mathbf{R} = \mathbf{I}$, then each model parameter is uniquely determined. If \mathbf{R} is not an identity matrix, then the estimates of the model parameters are really weighted averages of the true model parameters. If the model parameters have a natural ordering (as they would if they represented a discretized version of a continuous function), then plots of the rows of the resolution matrix can be useful in determining to what scale features in the model can actually be resolved (Fig. 4.2). Like the data resolution matrix, the model resolution is a function of only the data kernel and the a priori information added to the problem. It is therefore independent of the actual values of the data and can therefore be another important tool in experimental design.

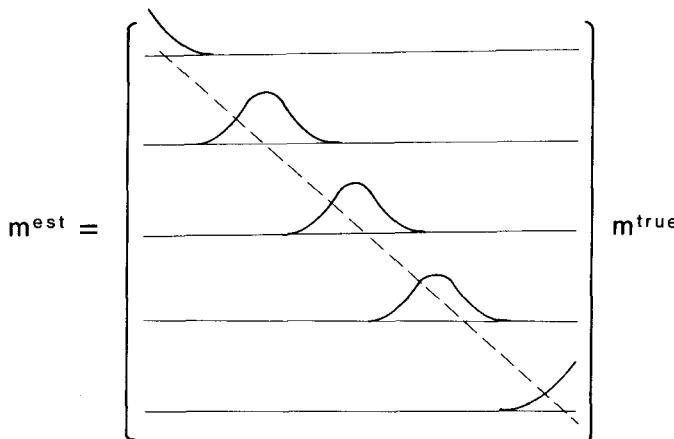


Fig. 4.2. Plots of selected rows of the model resolution matrix R indicate how well the true model parameters can be resolved. Narrow peaks occurring near the main diagonal of the matrix (dashed line) indicate that the model is well resolved.

4.4 The Unit Covariance Matrix

The covariance of the model parameters depends on the covariance of the data and the way in which error is mapped from data to model parameters. This mapping is a function of only the data kernel and the generalized inverse, not of the data itself. It is therefore useful to define a *unit covariance matrix* that characterizes the degree of error amplification that occurs in the mapping. If the data are assumed to be uncorrelated and all have equal variance σ^2 , the unit covariance matrix is given by

$$[\text{cov}_u \mathbf{m}] = \sigma^{-2} \mathbf{G}^{-g} [\text{cov} \mathbf{d}] \mathbf{G}^{-gT} = \mathbf{G}^{-g} \mathbf{G}^{-gT} \quad (4.6)$$

Even if the data are correlated, one can often find some normalization of the data covariance matrix, so that one can define a *unit data covariance matrix* $[\text{cov}_u \mathbf{d}]$, related to the model covariance matrix by

$$[\text{cov}_u \mathbf{m}] = \mathbf{G}^{-g} [\text{cov}_u \mathbf{d}] \mathbf{G}^{-gT} \quad (4.7)$$

Because the unit covariance matrix, like the data and model resolution matrices, is independent of the actual values and variances of the data, it is a useful tool in experimental design.

As an example, reconsider the problem of fitting a straight line to

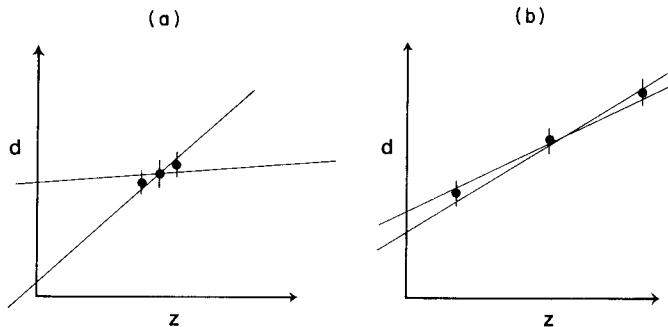


Fig. 4.3. (a) The method of least squares is used to fit a straight line to uncorrelated data with uniform variance. The vertical bars are a measure of the error. Since the data are not well separated in z , the variance of the slope and intercept is large (as indicated by the two different lines). (b) Same as (a) but with the data well separated in z . Although the variance of the data is the same as in (a) the variance of the intercept and slope of the straight line is much smaller.

(z, d) data. The unit covariance matrix for intercept m_1 and slope m_2 is given by

$$[\text{cov}_u \mathbf{m}] = \frac{1}{N \sum z_i^2 - (\sum z_i)^2} \begin{bmatrix} N & -\sum z_i \\ -\sum z_i & \sum z_i^2 \end{bmatrix} \quad (4.8)$$

Note that the estimates of intercept and slope are uncorrelated only when the data are centered about $z = 0$. The overall size of the variance is controlled by the denominator of the fraction. If all the z values are nearly equal, then the denominator of the fraction is small and the variance of the intercept and slope is large (Fig. 4.3a). On the other hand, if the z values have a large spread, the denominator is large and the variance is small (Fig. 4.3b).

4.5 Resolution and Covariance of Some Generalized Inverses

The data and model resolution and unit covariance matrices describe many interesting properties of the solutions to inverse problems. We therefore calculate these quantities for some of the simpler generalized inverses (with $[\text{cov}_u \mathbf{d}] = \mathbf{I}$).

4.5.1 LEAST SQUARES

$$\begin{aligned}
 G^{-g} &= [G^T G]^{-1} G^T \\
 N &= GG^{-g} = G[G^T G]^{-1} G^T \\
 R &= G^{-g} G = [G^T G]^{-1} G^T G = I \\
 [\text{cov}_u m] &= G^{-g} G^{-gT} = [G^T G]^{-1} G^T G [G^T G]^{-1} = [G^T G]^{-1}
 \end{aligned} \tag{4.9}$$

4.5.2 MINIMUM LENGTH

$$\begin{aligned}
 G^{-g} &= G^T [GG^T]^{-1} \\
 N &= GG^{-g} = GG^T [GG^T]^{-1} = I \\
 R &= G^{-g} G = G^T [GG^T]^{-1} G \\
 [\text{cov}_u m] &= G^{-g} G^{-gT} = G^T [GG^T]^{-1} [GG^T]^{-1} G^T = G^T [GG^T]^{-2} G^T
 \end{aligned} \tag{4.10}$$

Note that there is a great deal of symmetry between the least squares and minimum length solutions. Least squares solves the completely overdetermined problem and has perfect model resolution; minimum length solves the completely underdetermined problem and has perfect data resolution. As we shall see later, generalized inverses that solve the intermediate mixed-determined problems will have data and model resolution matrices that are intermediate between these two extremes.

4.6 Measures of Goodness of Resolution and Covariance

Just as we were able to quantify the goodness of the model parameters by measuring their overall prediction error and simplicity, we shall develop techniques that quantify the goodness of data and model resolution matrices and unit covariance matrices. Because the resolution is best when the resolution matrices are identity matrices, one possible measure of resolution is based on the size, or *spread* of the off-diagonal elements.

$$\begin{aligned}\text{spread}(\mathbf{N}) &= \|\mathbf{N} - \mathbf{I}\|_2^2 = \sum_{i=1}^N \sum_{j=1}^N [N_{ij} - I_{ij}]^2 \\ \text{spread}(\mathbf{R}) &= \|\mathbf{R} - \mathbf{I}\|_2^2 = \sum_{i=1}^M \sum_{j=1}^M [R_{ij} - I_{ij}]^2\end{aligned}\quad (4.11)$$

These measures of the goodness of the resolution spread are based on the L_2 norm of the difference between the resolution matrix and an identity matrix. They are sometimes called the *Dirichlet spread functions*. When $\mathbf{R} = \mathbf{I}$, $\text{spread}(\mathbf{R}) = 0$.

Since the unit standard deviation of the model parameters is a measure of the amount of error amplification mapped from data to model parameters, this quantity can be used to estimate the size of the unit covariance matrix as

$$\text{size}([\text{cov}_u \mathbf{m}]) = \|[\text{var}_u \mathbf{m}]^{1/2}\|_2^2 = \sum_{i=1}^M [\text{cov}_u \mathbf{m}]_{ii} \quad (4.12)$$

where the square root is interpreted component by component. Note that this measure of covariance size does not take into account the size of the off-diagonal elements in the unit covariance matrix.

4.7 Generalized Inverses with Good Resolution and Covariance

Having found a way to measure quantitatively the goodness of the resolution and covariance of a generalized inverse, we now consider whether it is possible to use these measures as guiding principles for deriving generalized inverses. This procedure is analogous to that of chapter 3, which involves first defining measures of solution prediction error and simplicity and then using those measures to derive the least squares and minimum length estimates of the model parameters.

4.7.1 OVERTERMINED CASE

We first consider a purely overdetermined problem of the form $\mathbf{Gm} = \mathbf{d}$. We postulate that this problem has a solution of the form $\mathbf{m}^{\text{est}} = \mathbf{G}^{-g} \mathbf{d}$ and try to determine \mathbf{G}^{-g} by minimizing some combination of the above measures of goodness. Since we previously noted that

the overdetermined least squares solution had perfect model resolution, we shall try to determine \mathbf{G}^{-g} by minimizing only the spread of the data resolution. We begin by examining the spread of the k th row of \mathbf{N} , say, J_k :

$$J_k = \sum_{i=1}^N (N_{ki} - I_{ki})^2 = \sum_{i=1}^N N_{ki}^2 - 2 \sum_{i=1}^N N_{ki} I_{ki} + \sum_{i=1}^N I_{ki}^2 \quad (4.13)$$

Since each of the J_k 's is positive, we can minimize the total spread(\mathbf{N}) = $\sum J_k$ by minimizing each individual J_k . We therefore insert the definition of the data resolution matrix $\mathbf{N} = \mathbf{G}\mathbf{G}^{-g}$ into the formula for J_k and minimize it with respect to the elements of the generalized inverse matrix:

$$\partial J_k / \partial G_{qr}^{-g} = 0 \quad (4.14)$$

We shall perform the differentiation separately for each of the three terms of J_k . The first term is given by

$$\begin{aligned} & \frac{\partial}{\partial G_{qr}^{-g}} \left[\sum_{i=1}^N \left[\sum_{j=1}^M G_{kj} G_{ji}^{-g} \right] \left[\sum_{p=1}^M G_{kp} G_{pi}^{-g} \right] \right] \\ &= \frac{\partial}{\partial G_{qr}^{-g}} \left[\sum_{i=1}^N \sum_{j=1}^M \sum_{p=1}^M G_{ji}^{-g} G_{pi}^{-g} G_{kj} G_{kp} \right] \\ &= 2 \sum_{i=1}^N \sum_{j=1}^M \sum_{p=1}^M \delta_{jq} \delta_{ir} G_{pi}^{-g} G_{kj} G_{kp} \\ &= 2 \sum_{p=1}^M G_{pr}^{-g} G_{kq} G_{kp} \end{aligned} \quad (4.15)$$

The second term is given by

$$-2 \frac{\partial}{\partial G_{qr}^{-g}} \sum_{i=1}^N \sum_{j=1}^M G_{kj} G_{ji}^{-g} \delta_{ki} = \sum_{i=1}^N \sum_{j=1}^M G_{kj} \delta_{jq} \delta_{kr} \delta_{ki} = -2 G_{kq} \delta_{kr} \quad (4.16)$$

The third term is zero, since its is not a function of the generalized inverse. Writing the complete equation in matrix form yields

$$\mathbf{G}^T \mathbf{G} \mathbf{G}^{-g} = \mathbf{G}^T \quad (4.17)$$

Since $\mathbf{G}^T\mathbf{G}$ is square, we can premultiply by its inverse to solve for the generalized inverse, $\mathbf{G}^{-g} = [\mathbf{G}^T\mathbf{G}]^{-1}\mathbf{G}^T$, which is precisely the same as the formula for the least squares generalized inverse. The least squares generalized inverse can be interpreted either as the inverse that minimizes the L_2 norm of the prediction error or as the inverse that minimizes the Dirichlet spread of the data resolution.

4.7.2 UNDERDETERMINED CASE

The data can be satisfied exactly in a purely underdetermined problem. The data resolution matrix is, therefore, precisely an identity matrix and its spread is zero. We might therefore try to derive a generalized inverse for this problem by minimizing the spread of the model resolution matrix with respect to the elements of the generalized inverse. It is perhaps not particularly surprising that the generalized inverse obtained by this method is exactly the minimum length generalized inverse $\mathbf{G}^{-g} = \mathbf{G}^T[\mathbf{G}\mathbf{G}^T]^{-1}$. The minimum length solution can be interpreted either as the inverse that minimizes the L_2 norm of the solution length or as the inverse that minimizes the Dirichlet spread of the model resolution. This is another aspect of the symmetrical relationship between the least squares and minimum length solutions.

4.7.3 THE GENERAL CASE WITH DIRICHLET SPREAD FUNCTIONS

We seek the generalized inverse \mathbf{G}^{-g} that minimizes the weighted sum of Dirichlet measures of resolution spread and covariance size.

Minimize: $\alpha_1 \text{spread}(\mathbf{N}) + \alpha_2 \text{spread}(\mathbf{R}) + \alpha_3 \text{size}([\text{cov}_u \mathbf{m}]) \quad (4.18)$

where the α 's are arbitrary weighting factors. This problem is done in exactly the same fashion as the one in Section 4.7.1, except that there is now three times as much algebra. The result is an equation for the generalized inverse:

$$\alpha_1[\mathbf{G}^T\mathbf{G}]\mathbf{G}^{-g} + \mathbf{G}^{-g}[\alpha_2\mathbf{G}\mathbf{G}^T + \alpha_3[\text{cov}_u \mathbf{d}]] = [\alpha_1 + \alpha_2]\mathbf{G}^T \quad (4.19)$$

This equation has no explicit solution for \mathbf{G}^{-g} in terms of an algebraic function of the various matrices. Explicit solutions can be written, however, for a variety of special choices of the weighting factors. The least squares solution is recovered if $\alpha_1 = 1$ and $\alpha_2 = \alpha_3 = 0$; and the

minimum length solution is recovered if $\alpha_1 = 0$, $\alpha_2 = 1$ and $\alpha_3 = 0$. Of more interest is the case in which $\alpha_1 = 1$, $\alpha_2 = 0$, α_3 equals some constant (say, ϵ^2) and $[\text{cov}_u \mathbf{d}] = \mathbf{I}$. The generalized inverse is then given by

$$\mathbf{G}^{-g} = [\mathbf{G}^T \mathbf{G} + \epsilon^2 \mathbf{I}]^{-1} \mathbf{G}^T \quad (4.20)$$

This formula is precisely the damped least squares inverse, which we derived in the previous chapter by minimizing a combination of prediction error and solution length. The damped least squares solution can also be interpreted as the inverse that minimizes a weighted combination of data resolution spread and covariance size.

Note that it is quite possible for these generalized inverses to possess resolution matrices containing *negative* off-diagonal elements. For interpreting the rows of the matrices as localized averages, this is an unfortunate property. Physically, an average would make more sense if it contained only positive weighting factors. In principle, it is possible to include nonnegativity as a constraint when choosing the generalized inverse by minimizing the spread functions. However, in practice this constraint is never implemented, because it makes the calculation of the generalized inverse very difficult.

4.8 Sidelobes and the Backus–Gilbert Spread Function

When there is a natural ordering of data or model parameters, the Dirichlet spread function is not a particularly appropriate measure of the goodness of resolution, because the off-diagonal elements of the resolution matrix are all weighted equally, regardless of whether they are close or far from the main diagonal. If there is a natural ordering, we would much prefer that any large elements be near the main diagonal (Fig. 4.4). The rows of the resolution matrix are then localized averaging functions.

If one uses the Dirichlet spread function to compute a generalized inverse, it will often have *sidelobes*, that is, large amplitude regions in the resolution matrices far from the main diagonal. We would prefer to find a generalized inverse without sidelobes, even at the expense of widening the band of nonzero elements near the main diagonal, since

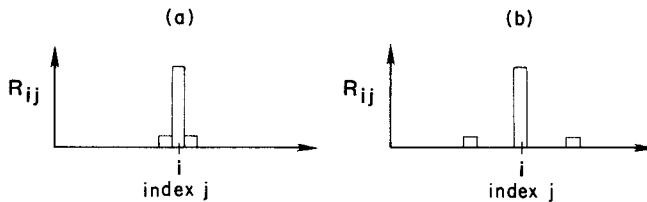


Fig. 4.4. (a and b) Resolution matrices have the same spread, when measured by the Dirichlet spread function. Nevertheless, if the model parameters possess a natural ordering, then (a) is better resolved. The Backus–Gilbert spread function is designed to measure (a) as having a smaller spread than (b).

a solution with such a resolution matrix is then interpretable as a localized average of physically adjacent model parameters.

We therefore add a weighting factor $w(i, j)$ to the measure of spread that weights the (i, j) element of \mathbf{R} according to its physical distance from the diagonal element. This weighting preferentially selects resolution matrices that are “spiky,” or “deltalike.” If the natural ordering were a simple linear one, then the choice $w(i, j) = (i - j)^2$ would be reasonable. If the ordering is multidimensional, a more complicated weighting factor is needed. It is usually convenient to choose the spread function so that the diagonal elements have no weight, i.e., $w(i, i) = 0$, and so that $w(i, j)$ is always nonnegative and symmetric in i and j . The new spread function, often called the Backus–Gilbert spread function, is then given by

$$\text{spread}(\mathbf{R}) = \sum_{i=1}^M \sum_{j=1}^M w(i, j)[R_{ij} - I_{ij}]^2 = \sum_{i=1}^M \sum_{j=1}^M w(i, j)R_{ij}^2 \quad (4.21)$$

[Refs. 1,2].

A similar expression holds for the spread of the data resolution. One can now use this measure of spread to derive new generalized inverses. Their sidelobes will be smaller than those based on the Dirichlet spread functions. On the other hand, they are sometimes worse when judged by other criteria. As we shall see, the Backus–Gilbert generalized inverse for the completely underdetermined problem does not exactly satisfy the data, even though the analogous minimum-length generalized inverse does. These facts demonstrate that there are unavoidable trade-offs inherent in finding solutions to inverse problems.

4.9 The Backus–Gilbert Generalized Inverse for the Underdetermined Problem

This problem is analogous to deriving the minimum length solution by minimizing the Dirichlet spread of model resolution. Since it is very easy to satisfy the data when the problem is underdetermined (so that the data resolution has small spread), we shall find a generalized inverse that minimizes the spread of the model resolution alone.

We seek the generalized inverse \mathbf{G}^{-g} that minimizes the Backus–Gilbert spread of model resolution. Since the diagonal elements of the model resolution matrix are given no weight, we also require that the resulting model resolution matrix satisfy the equation

$$\sum_j^M R_{ij} = 1_i \quad (4.22)$$

This constraint ensures that the diagonal of the resolution matrix is finite and that the rows are unit averaging functions acting on the true model parameters. Writing the spread of one row of the resolution matrix as J_k and inserting the expression for the resolution matrix, we have

$$\begin{aligned} J_k &= \sum_{l=1}^M w(l, k) R_{kl} R_{kl} \\ &= \sum_{l=1}^M w(l, k) \left[\sum_{i=1}^N G_{ki}^{-g} G_{il} \right] \left[\sum_{j=1}^N G_{kj}^{-g} G_{jl} \right] \\ &= \sum_{i=1}^N \sum_{j=1}^N G_{ki}^{-g} G_{kj}^{-g} \left[\sum_{l=1}^M w(l, k) G_{il} G_{jl} \right] \\ &= \sum_{i=1}^N \sum_{j=1}^N G_{ki}^{-g} G_{kj}^{-g} [S_{ij}]_k \end{aligned} \quad (4.23)$$

where the quantity $[S_{ij}]_k$ is defined as

$$[S_{ij}]_k = \sum_{l=1}^M w(l, k) G_{il} G_{jl} \quad (4.24)$$

The left-hand side of the constraint equation $\sum_j R_{ij} = 1_i$ can also be written in terms of the generalized inverse

$$\sum_{k=1}^M R_{ik} = \sum_{k=1}^M \left[\sum_{j=1}^N G_{ij}^{-g} G_{jk} \right] = \sum_{j=1}^N G_{ij}^{-g} \sum_{k=1}^M G_{jk} = \sum_{j=1}^N G_{ij}^{-g} u_j \quad (4.25)$$

Here the quantity u_j is defined as

$$u_j = \sum_{k=1}^M G_{jk} \quad (4.26)$$

The problem of minimizing J_k with respect to the elements of the generalized inverse (under the given constraints) can be solved through the use of Lagrange multipliers. We first define a Lagrange function Φ such that

$$\Phi = \sum_{i=1}^N \sum_{j=1}^N G_{ki}^{-g} G_{kj}^{-g} [S_{ij}]_k - 2\lambda \sum_{j=1}^N G_{kj}^{-g} u_j \quad (4.27)$$

where -2λ is the Lagrange multiplier. We then differentiate Φ with respect to the elements of the generalized inverse and set the result equal to zero as

$$\frac{\partial \Phi}{\partial G_{kp}^{-g}} = 2 \sum_{i=1}^N [S_{pi}]_k G_{ki}^{-g} - 2\lambda u_p = 0 \quad (4.28)$$

(Note that one can solve for each row of the generalized inverse separately, so that it is only necessary to take derivatives with respect to the elements in the k th row.) The above equation must be solved along with the original constraint equation. Treating the k th row of G^{-g} as a vector and the quantity $[S_{ij}]_k$ as a matrix in indices i and j , we can write these equations as the matrix equation

$$\begin{bmatrix} [S_{ij}]_k & \mathbf{u} \\ \mathbf{u}^T & 0 \end{bmatrix} \begin{bmatrix} [G_{kp}^{-g}] \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix} \quad (4.29)$$

This is a square $(N + 1) \times (N + 1)$ system of linear equations that must be solved for the N elements of the k th row of the generalized inverse and for the one Lagrange multiplier λ .

The matrix equation can be solved explicitly using a variant of the “bordering method” of linear algebra, which is used to construct the inverse of a matrix by partitioning it into submatrices with simple properties. Suppose that the inverse of the matrix in Eq. (4.29) exists and that we partition it into an $N \times N$ square matrix \mathbf{A} , vector \mathbf{b} , and scalar c . By assumption, premultiplication by the inverse yields the identity matrix

$$\begin{bmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{b}^T & c \end{bmatrix} \begin{bmatrix} [S_{ij}]_k & \mathbf{u} \\ \mathbf{u}^T & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{O} & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{A}[S_{ij}]_k + \mathbf{b}\mathbf{u}^T & \mathbf{A}\mathbf{u} \\ \mathbf{b}^T[S_{ij}]_k + c\mathbf{u}^T & \mathbf{b}^T\mathbf{u} \end{bmatrix} \quad (4.30)$$

The unknown submatrices \mathbf{A} , \mathbf{b} , and c can now be determined by equating the submatrices

$$\begin{aligned} \mathbf{A}[S_{ij}]_k + \mathbf{b}\mathbf{u}^T &= \mathbf{I} \quad \text{so that} \quad \mathbf{A} = [S_{ij}]_k^{-1}[\mathbf{I} - \mathbf{b}\mathbf{u}^T] \\ \mathbf{A}\mathbf{u} &= \mathbf{O} \quad \text{so that} \quad [S_{ij}]_k^{-1}\mathbf{u} = \mathbf{b}\mathbf{u}^T[S_{ij}]_k^{-1}\mathbf{u} \quad \text{and} \quad \mathbf{b} = \frac{[S_{ij}]_k^{-1}\mathbf{u}}{\mathbf{u}^T[S_{ij}]_k^{-1}\mathbf{u}} \\ \mathbf{b}^T[S_{ij}]_k + c\mathbf{u}^T &= 0 \quad \text{so that} \quad c = \frac{-1}{\mathbf{u}^T[S_{ij}]_k^{-1}\mathbf{u}} \end{aligned} \quad (4.31)$$

Once the submatrices \mathbf{A} , \mathbf{b} , and c are known, it is easy to solve the equation for the generalized inverse. The result is written as

$$G_{kl}^{-g} = \frac{\sum_{i=1}^N [S_{il}]_k^{-1}u_i}{\sum_{i=1}^N \sum_{j=1}^N u_i u_j [S_{ij}]_k^{-1}} \quad (4.32)$$

4.10 Including the Covariance Size

The measure of goodness that was used to determine the Backus–Gilbert inverse can be modified to include a measure of the covariance size of the model parameters [Ref. 3]. We shall use the same measure as we did when considering the Dirichlet spread functions, so that goodness is measured by

$$\begin{aligned} \alpha \text{spread}(\mathbf{R}) + (1 - \alpha) \text{size}([\text{cov}_u \mathbf{m}]) \\ = \alpha \sum_{i=1}^M \sum_{j=1}^M w(i, j) R_{ij}^2 + (1 - \alpha) \sum_{i=1}^M [\text{cov}_u \mathbf{m}]_{ii} \end{aligned} \quad (4.33)$$

where $0 \leq \alpha \leq 1$ is a weighting factor that determines the relative contribution of model resolution and covariance to the measure of the goodness of the generalized inverse. The goodness J'_k of the k th row is

then

$$\begin{aligned}
 J'_k &= \alpha \sum_{l=1}^M w(k, l) R_{kl}^2 + (1 - \alpha) [\text{cov}_u \mathbf{m}]_{kk} \\
 &= \alpha \sum_{i=1}^N \sum_{j=1}^N G_{ki}^{-g} G_{kj}^{-g} [S_{ij}]_k + (1 - \alpha) \sum_{i=1}^N \sum_{j=1}^N G_{ki}^{-g} G_{kj}^{-g} [\text{cov}_u \mathbf{d}]_{ij} \\
 &= \sum_{i=1}^N \sum_{j=1}^N G_{ki}^{-g} G_{kj}^{-g} [S'_{ij}]_k
 \end{aligned} \tag{4.34}$$

where the quantity $[S'_{ij}]_k$ is defined by the equation

$$[S'_{ij}]_k = \alpha [S_{ij}]_k + (1 - \alpha) [\text{cov}_u \mathbf{d}]_{ij} \tag{4.35}$$

Since the function J'_k has exactly the same form as J_k had in the previous section, the generalized inverse is just the previous result with $[S_{ij}]_k$ replaced by $[S'_{ij}]_k$:

$$G_{kl}^{-g} = \frac{\sum_{i=1}^N [S'_{il}]_k^{-1} u_i}{\sum_{i=1}^N \sum_{j=1}^N u_i u_j [S'_{ij}]_k^{-1}} \tag{4.36}$$

4.11 The Trade-off of Resolution and Variance

Suppose that one is attempting to determine a set of model parameters that represents a discretized version of a continuous function, such as x-ray opacity in the medical tomography problem (Fig. 4.5). If the discretization is made very fine, then the x rays will not sample every box; the problem will be underdetermined. If we try to determine the opacity of each box individually, then estimates of opacity will tend to have rather large variance. Few boxes will have several x rays passing through them, so that little averaging out of the errors will take place. On the other hand, the boxes are very small—and very small features can be detected (the resolution is very good). The large variance can be reduced by increasing the box size (or alternatively, averaging several neighboring boxes). Each of these larger regions will then contain several x rays, and noise will tend to be averaged out. But because the regions are now larger, small features can no longer be detected and the resolution of the x-ray opacity has become poorer.

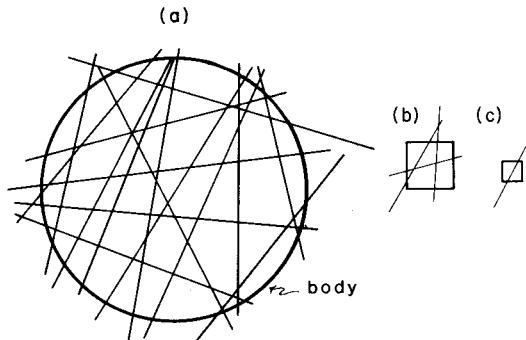


Fig. 4.5. (a) X-ray paths (straight lines) through body in the tomography problem. (b) Coarse parameterizations of the opacity have good variance, since several rays pass through each box. But since they are coarse, they only poorly resolve the structure of the body. (c) Fine parameterizations have poorer variance but better resolution.

This scenario illustrates an important trade-off between model resolution spread and variance size. One can be decreased only at the expense of increasing the other. We can study this trade-off by choosing a generalized inverse that minimizes a weighted sum of resolution spread and covariance size:

$$\alpha \text{ spread}(\mathbf{R}) + (1 - \alpha) \text{ size}([\text{cov}_u \mathbf{m}]) \quad (4.37)$$

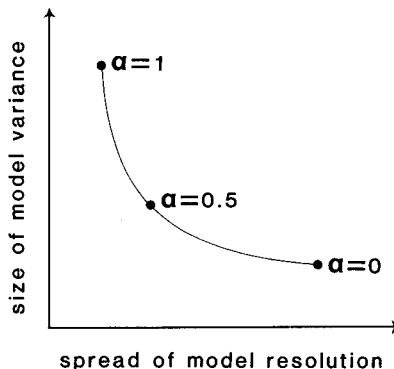


Fig. 4.6. Trade-off curve of resolution and variance for a given discretization of a continuous function. The larger the α , the more weight resolution is given (relative to variance) when forming the generalized inverse. The details of the trade-off curve depend on the parameterization. The resolution can be no better than the smallest element in the parameterization and no worse than the sum of all the elements.

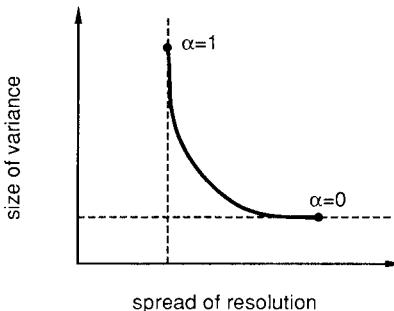


Fig. 4.7. Trade-off curve of resolution and variance has two asymptotes in the case when the model parameter is a continuous function.

If the weighting parameter α is set near 1, then the model resolution matrix of the generalized inverse will have small spread, but the model parameters will have large variance. If α is set close to 0, then the model parameters will have a relatively small variance, but the resolution will have a large spread. By varying α on the interval $[0, 1]$, one can define a *trade-off curve* (Fig. 4.6). Such curves can be helpful in choosing a generalized inverse that has an optimum trade-off in model resolution and variance (judged by criteria appropriate to the problem at hand).

Trade-off curves play an important role in continuous inverse theory, where the discretization is (so to speak) infinitely fine, and all problems are underdetermined. It is known that in this continuous limit the curves are monotonic and possess asymptotes in resolution and variance (Fig. 4.7). The process of approximating a continuous function by a finite set of discrete parameters somewhat complicates this picture. The resolution and variance, and indeed the solution itself, are dependent on the parameterization, so it is difficult to make any definitive statement regarding the properties of the trade-off curves. Nevertheless, if the discretization is sufficiently fine, the discrete trade-off curves are usually close to ones obtained with the use of continuous inverse theory. Therefore, discretizations should always be made as fine as computational considerations permit.

5

SOLUTION OF THE LINEAR, GAUSSIAN INVERSE PROBLEM, VIEWPOINT 3: MAXIMUM LIKELIHOOD METHODS

5.1 The Mean of a Group of Measurements

Suppose that an experiment is performed N times and that each time a single datum d_i is collected. Suppose further that these data are all noisy measurements of the same model parameter m_1 . In the view of probability theory, N realizations of random variables, all of which have the same distribution, have been measured. If these random variables are Gaussian, their joint distribution can be characterized in terms of a variance σ^2 and a mean m_1 (see Section 2.4) as

$$P(\mathbf{d}) = \sigma^{-N} (2\pi)^{-N/2} \exp \left[-\frac{1}{2} \sigma^{-2} \sum_{i=1}^N (d_i - m_1)^2 \right] \quad (5.1)$$

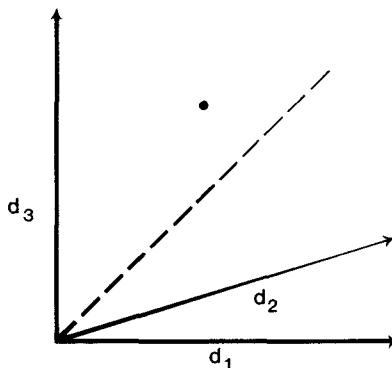


Fig. 5.1. The data are represented by a single point in a space whose dimension equals the number of observations (in this case, 3). These data are realizations of random variables with the same mean and variance. Nevertheless, they do not necessarily fall on the line $d_1 = d_2 = d_3$ (dashed).

The data \mathbf{d}^{obs} can be represented graphically as a point in the N -dimensional space whose coordinate axes are d_1, d_2, \dots, d_N . One such graph is shown in Fig. 5.1. The probability distribution for the data can also be graphed (Fig. 5.2). Note that the distribution is centered about the line $d_1 = d_2 = \dots = d_N$, since all the d_i 's are supposed to have the same mean, and that it is spherically symmetric, since all the d_i 's have the same variance.

Suppose that we guess a value for the unknown data mean and variance, thus fixing the center and diameter of the distribution. We can then calculate the numerical value of the distribution at the data, $P(\mathbf{d}^{\text{obs}})$. If the guessed values of mean and variance are close to being correct, then $P(\mathbf{d}^{\text{obs}})$ should be a relatively large number. If the guessed values are incorrect, then the probability, or *likelihood*, of the observed data will be small. We can imagine sliding the cloud of probability in Fig. 5.2 up along the line and adjusting its diameter until its probability at the point \mathbf{d}^{obs} is maximized.

This procedure defines a method of estimating the unknown parameters in the distribution, the *method of maximum likelihood*. It asserts that the optimum values of the parameters maximize the probability that the observed data are in fact observed. In other words, the probability at the point \mathbf{d}^{obs} is made as large as possible. The maximum is located by differentiating $P(\mathbf{d}^{\text{obs}})$ with respect to mean and variance

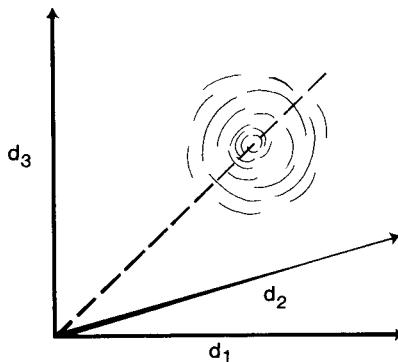


Fig. 5.2. If the data are assumed to be uncorrelated with equal mean and uniform variance, their distribution is a spherical cloud centered on the line $d_1 = d_2 = d_3$ (dashed).

and setting the result to zero as

$$\frac{\partial P}{\partial m_1} = \frac{\partial P}{\partial \sigma} = 0 \quad (5.2)$$

Maximizing $\log(P)$ gives the same result as maximizing P , since $\log(P)$ is a monotonic function of P . We therefore compute derivatives of the likelihood function, $L = \log(P)$, (Fig. 5.3). Ignoring the overall normalization of $(2\pi)^{-N/2}$ we have

$$\begin{aligned} L &= \log(P) = -N \log(\sigma) - \frac{1}{2}\sigma^{-2} \sum_{i=1}^N (d_i^{\text{obs}} - m_1)^2 \\ \frac{\partial L}{\partial m_1} &= 0 = -\frac{1}{2}\sigma^{-2} 2m_1 \sum_{i=1}^N (d_i^{\text{obs}} - m_1) \\ \frac{\partial L}{\partial \sigma} &= 0 = -\frac{N}{\sigma} + \sigma^{-3} \sum_{i=1}^N (d_i^{\text{obs}} - m_1)^2 \end{aligned} \quad (5.3)$$

These equations can be solved for the estimated mean and variance as

$$\begin{aligned} m_1^{\text{est}} &= \frac{1}{N} \sum_{i=1}^N d_i \\ \sigma^{\text{est}} &= \left[\frac{1}{N} \sum_{i=1}^N (d_i^{\text{obs}} - m_1^{\text{est}})^2 \right]^{1/2} \end{aligned} \quad (5.4)$$

These estimates are just the usual formulas for the sample arithmetic mean and sample standard deviation. We note that they arise as a

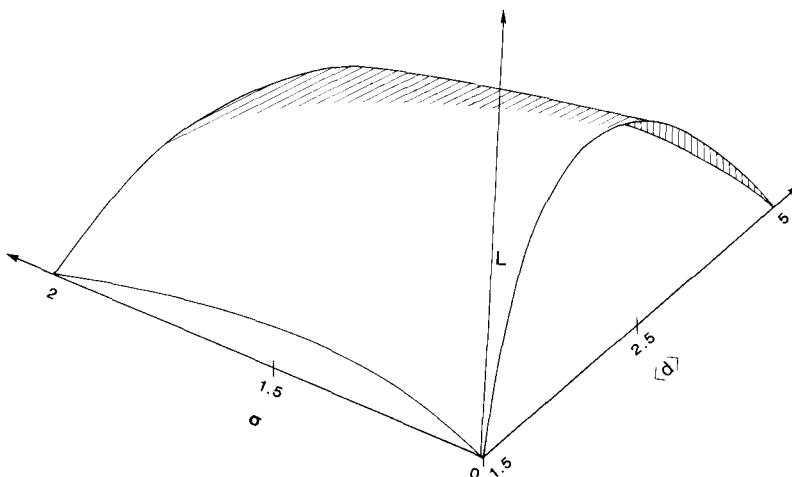


Fig. 5.3. Likelihood surface for 20 realizations of random variables with equal mean $\langle d \rangle = 2.5$ and uniform variance $\sigma_d = 1.5$. The maximum in the direction of the mean is sharper than the maximum in the direction of σ , indicating that the former can be determined with greater certainty.

direct consequence of the assumption that the data possess a Gaussian distribution. If the data distribution were not Gaussian, then the arithmetic mean might not be an appropriate estimate of the mean of the distribution. (As we shall see in Section 8.2, the sample median is the maximum likelihood estimate of the mean of an exponential distribution.)

5.2 Maximum Likelihood Solution of the Linear Inverse Problem

Assume that the data in the linear inverse problem $\mathbf{Gm} = \mathbf{d}$ have a multivariate Gaussian distribution, as given by

$$P(\mathbf{d}) \propto \exp\left[-\frac{1}{2} (\mathbf{d} - \mathbf{Gm})^T [\text{cov } \mathbf{d}]^{-1} (\mathbf{d} - \mathbf{Gm})\right] \quad (5.5)$$

We assume that the model parameters are unknown but (for the sake of simplicity) that the data covariance is known. We can then apply the method of maximum likelihood to estimate the model parameters. The optimum values for the model parameters are the ones that

maximize the probability that the observed data are in fact observed. Clearly, the maximum of $P(\mathbf{d}^{\text{obs}})$ occurs when the argument of the exponential is a maximum, or when the quantity given by

$$(\mathbf{d} - \mathbf{Gm})^T [\text{cov } \mathbf{d}]^{-1} (\mathbf{d} - \mathbf{Gm}) \quad (5.6)$$

is a minimum. But this expression is just a weighted measure of prediction length. The maximum likelihood estimate of the model parameters is nothing but the weighted least squares solution, where the weighting matrix is the inverse of the covariance matrix of the data (in the notation of Chapter 3, $\mathbf{W}_e = [\text{cov } \mathbf{d}]^{-1}$). If the data happen to be uncorrelated and all have equal variance, then $[\text{cov } \mathbf{d}] = \sigma_d^2 \mathbf{I}$, and the maximum likelihood solution is the simple least squares solution. If the data are uncorrelated but their variances are all different (say, $\sigma_{d_i}^2$), then the prediction error is given by

$$E = \sum_{i=1}^N \sigma_{d_i}^{-2} e_i^2 \quad (5.7)$$

where $e_i = (d_i^{\text{obs}} - d_i^{\text{pre}})$ is the prediction error for each datum. Each measurement is weighted by the reciprocal of its variance; the most certain data are weighted most.

We have justified the use of the L_2 norm through the application of probability theory. The least squares procedure for minimizing the L_2 norm of the prediction error makes sense if the data are uncorrelated, have equal variance, and obey Gaussian statistics. If the data are not Gaussian, then other measures of prediction error may be more appropriate.

5.3 A Priori Distributions

If the linear problem is underdetermined, then the least squares inverse does not exist. From the standpoint of probability theory, the distribution of the data $P(\mathbf{d}^{\text{obs}})$ has no well-defined maximum with respect to variations of the model parameters. At best, it has a ridge of maximum probability (Fig. 5.4).

To solve this underdetermined problem we must add a priori information that causes the distribution to have a well-defined peak. One way to accomplish this is to write the a priori information about the model parameters as a probability distribution $P_A(\mathbf{m})$, where the

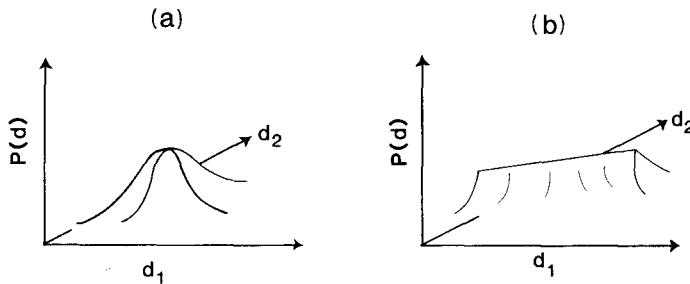


Fig. 5.4. (a) Probability distribution with well-defined peak. (b) Distribution with ridge.

subscript A means “a priori.” The mean of this distribution is then the value we expect the model parameter vector to have, and the shape of the distribution reflects the certainty of this expectation. If we can combine this a priori distribution for the model parameters with $P(\mathbf{d}^{\text{obs}})$, the joint distribution will have a distinct maximum (provided there is enough information in $P_A(\mathbf{m})$ to resolve the underdeterminacy).

A priori distributions for the model parameters can take a variety of forms. For instance, if we expected that the model parameters are close to $\langle \mathbf{m} \rangle$, we might use a Gaussian distribution with mean $\langle \mathbf{m} \rangle$ and variance that reflects the certainty of our knowledge (Fig. 5.5). If the a priori value of one model parameter were more certain than another, we might use different variances for the different model parameters (Fig. 5.6). Equality constraints can be implemented with a distribution

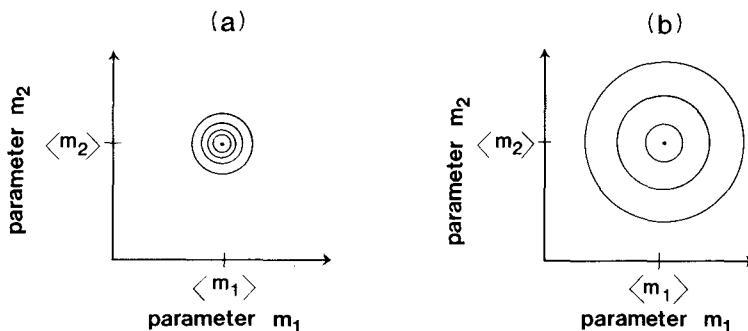


Fig. 5.5. A priori information about model parameters represented by probability distribution (contoured). Most probable values are given by means $\langle m_1 \rangle$ and $\langle m_2 \rangle$. Width of distribution reflects certainty of knowledge: (a) certain, (b) uncertain.

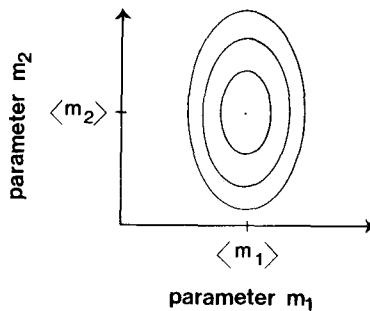


Fig. 5.6. A priori information about model parameters m represented as probability distribution. The model parameters are thought to be near $\langle \mathbf{m} \rangle$, and the certainty in m_1 is greater than the certainty of m_2 .

that contains a ridge (Fig. 5.7). This distribution is non-Gaussian but might be approximated by a Gaussian distribution with nonzero covariance if the expected range of the model parameters were small. Inequality constraints can also be represented by an a priori distribution but are inherently non-Gaussian (Fig. 5.8).

We can summarize the state of knowledge about the inverse problem *before* it is solved by first defining an a priori distribution for the data $P_A(\mathbf{d})$ and then combining this with the a priori distribution for the model $P_A(\mathbf{m})$. An a priori data distribution simply summarizes the observations, so its mean is \mathbf{d}^{obs} and its variance is equal to the expected variance of the data. Since the a priori model distribution is

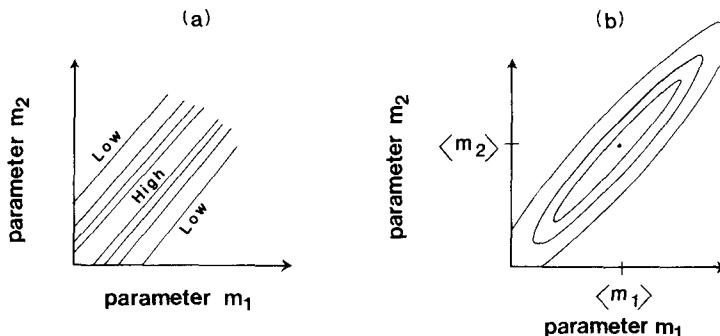


Fig. 5.7. A priori information about model parameters m represented by probability distribution. (a) Distribution when the values of m_1 and m_2 are unknown, but are known to be correlated. (b) Approximation of (a) by Gaussian distribution with finite variance.

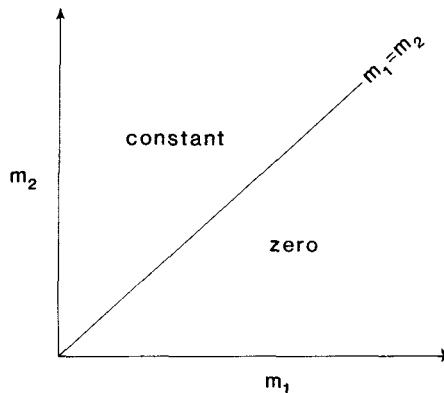


Fig. 5.8. A priori information about model parameters m represented as a probability distribution (contoured) when $m_1 \leq m_2$. Such distributions are inherently non-Gaussian.

completely independent of the actual values of the data, we can form the joint a priori distribution simply by multiplying the two as

$$P_A(m, d) = P_A(m)P_A(d) \quad (5.8)$$

This distribution can be depicted graphically as a “cloud” of probability centered on the observed data and a priori model, with a width that reflects the certainty of these quantities (Fig. 5.9). Note that, if we

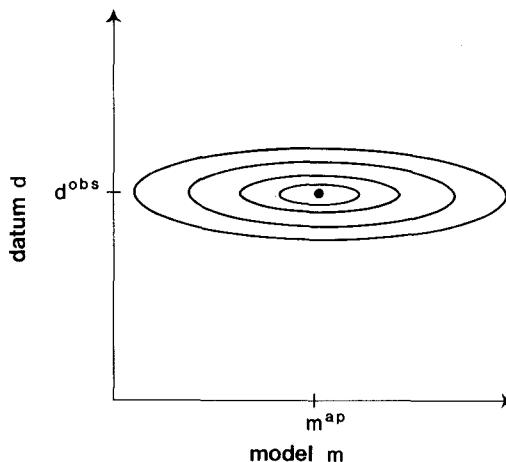


Fig. 5.9. Joint probability distribution for observed data and a priori model parameters.

apply the maximum likelihood method to this distribution, we simply recover the data and a priori model. We have not yet applied our knowledge of the model (the relationship between data and model parameters).

5.4 Maximum Likelihood for an Exact Theory

Suppose that the model is the rather general equation $\mathbf{g}(\mathbf{m}) = \mathbf{d}$ (which may or may not be linear). This equation defines a surface in the space of model parameters and data along which the solution must lie (Fig. 5.10). The maximum likelihood problem then translates into finding the maximum of the joint distribution $P_A(\mathbf{m}, \mathbf{d})$ on the surface $\mathbf{d} = \mathbf{g}(\mathbf{m})$ [Ref. 18]. Note that if the a priori distribution for the model parameters is much more certain than that of the observed data (that is, if $\sigma_m \ll \sigma_d$), then the estimate of the model parameters (the maximum likelihood point) tends to be close to the a priori model parameters (Fig. 5.11). On the other hand, if the data are far more

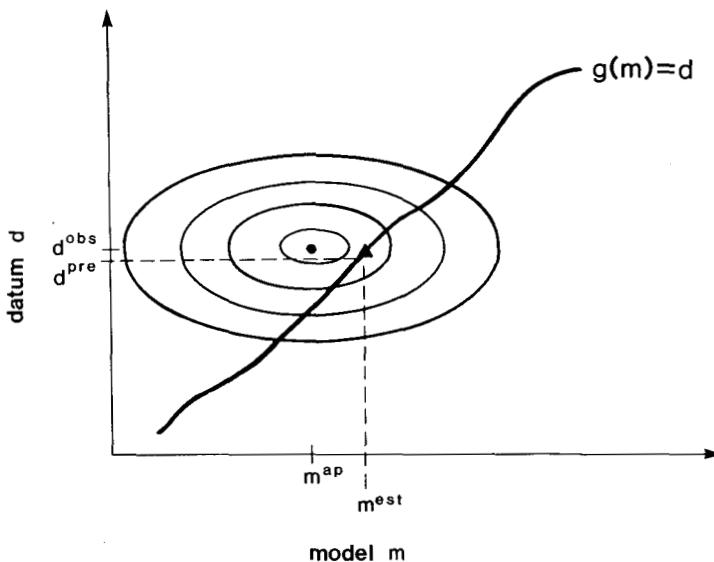


Fig. 5.10. The estimated model parameters m^{est} (triangle) and predicted data d^{pre} fall on the theoretical curve $g(m) = d$ (solid curve).

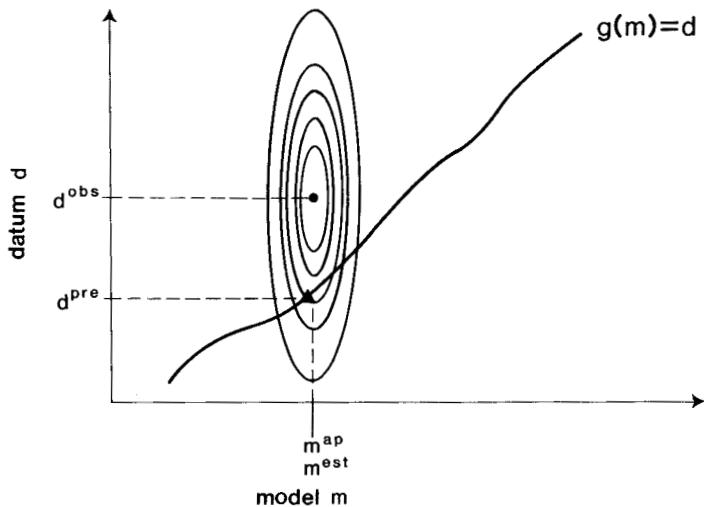


Fig. 5.11. If the a priori model parameters m^{ap} are much more certain than the observed data d^{obs} , the solution (triangle) is close to the a priori model parameters ($m^{ap} = m^{est}$) but may be far from the observed data ($d^{obs} \neq d^{pre}$).

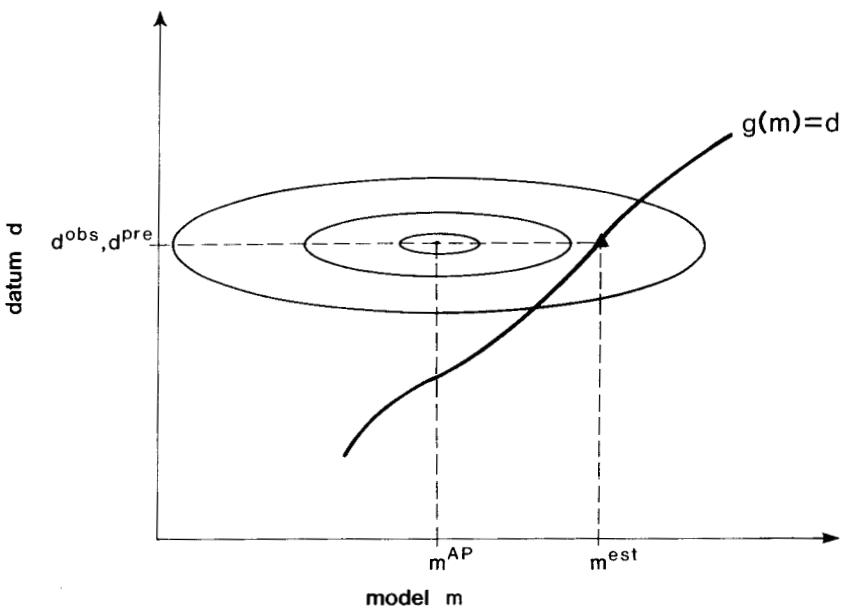


Fig. 5.12. If the a priori model parameters m^{ap} are much less certain than the observed data d^{obs} , then the solution (triangle) is close to the observed data ($d^{obs} = d^{pre}$) but may be far from the a priori model ($m^{ap} \neq m^{est}$).

certain than the model parameters (i.e., $\sigma_d \ll \sigma_m$), then the estimates of the model parameters primarily reflect information contained in the data (Fig. 5.12).

5.5 Inexact Theories

In many realistic problems there are errors associated with the theory. Some of the assumptions that go into the theory may be somewhat unrealistic; or, it may be an approximate form of a clumsier but exact theory. The model equation $g(\mathbf{m}) = \mathbf{d}$ can no longer be represented by a simple surface. It has become “fuzzy” because there are now errors associated with it (Fig. 5.13; Ref. 19). Instead of a surface, one might envision a distribution $P_g(\mathbf{m}|\mathbf{d})$ centered about $g(\mathbf{m}) = \mathbf{d}$, with width proportional to the uncertainty of the theory. (The vertical bar in the expression $P_g(\mathbf{m}|\mathbf{d})$ indicates that this distribution is a *conditional probability distribution*, the probability of the theory predicting a set of data \mathbf{d} given model parameters \mathbf{m} . It is therefore not a joint distribution of \mathbf{m} and \mathbf{d} .) Rather than find the maximum likelihood point of $P_A(\mathbf{m}, \mathbf{d})$ on the surface, we combine $P_A(\mathbf{m}, \mathbf{d})$ and $P_g(\mathbf{m}|\mathbf{d})$ into a single distribution and find its maximum likelihood point (Fig. 5.13). Since the theory is assumed to be indepen-

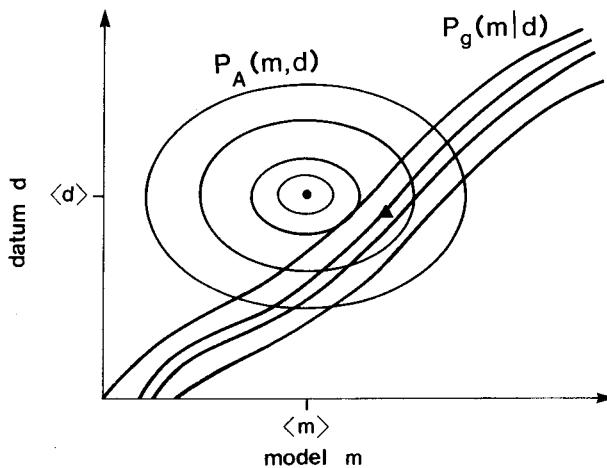


Fig. 5.13. When the theory is inexact and represented by the conditional distribution $P_g(m|d)$, the solution (triangle) is at the maximum likelihood point of the product $P_A(m, d)P_g(m|d)$.

dent of the actual values of the data and model parameters, the combination can be accomplished by simple multiplication of the two component distributions

$$P_T(\mathbf{m}, \mathbf{d}) = P_A(\mathbf{m}, \mathbf{d})P_g(\mathbf{m}|\mathbf{d}) \quad (5.9)$$

Here the subscript T means the combined or total distribution. Note that as the error associated with the theory increases, the maximum likelihood point moves back toward the a priori values of model parameters and observed data (Fig. 5.14). The limiting case in which the theory is infinitely accurate is equivalent to the case in which the distribution is replaced by a distinct surface.

The maximum likelihood point of $P_T(\mathbf{m}, \mathbf{d})$ is specified by both a set of model parameters \mathbf{m}^{est} and a set of data \mathbf{d}^{pre} . The model parameters and data are estimated simultaneously. This approach is somewhat different from that of the least squares problem examined in Section 5.2. In that problem we maximized the distribution with respect to the model parameters only. The least squares procedure found the most probable model parameters; this procedure finds the most probable combination of model parameters and predicted data. These methods do not necessarily yield the same estimates for the model parameters. To find the likelihood point of $P_T(\mathbf{m}, \mathbf{d})$ with respect to model parameters only, we must sum all the probabilities along lines of equal model parameter. This summing can be thought of as projecting the distribution onto the $\mathbf{d} = 0$ plane (Fig. 5.15) and then finding the maximum. The projected distribution $P_P(\mathbf{m})$ is then

$$P_P(\mathbf{m}) = \int P_T(\mathbf{m}, \mathbf{d}) d\mathbf{d} \quad (5.10)$$

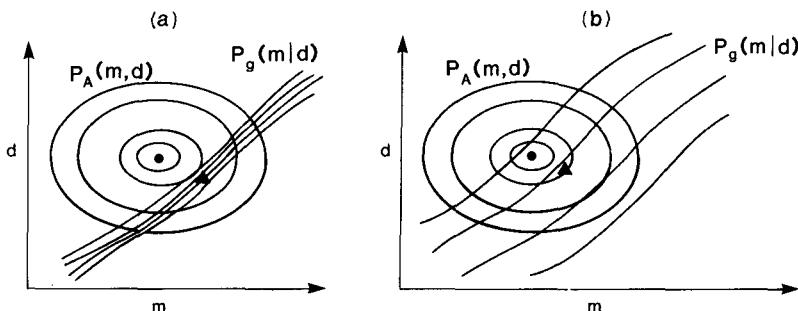


Fig. 5.14. If the theory is made more and more inexact [compare (a) and (b)], the solution (triangle) moves toward the maximum likelihood point of the a priori distribution.

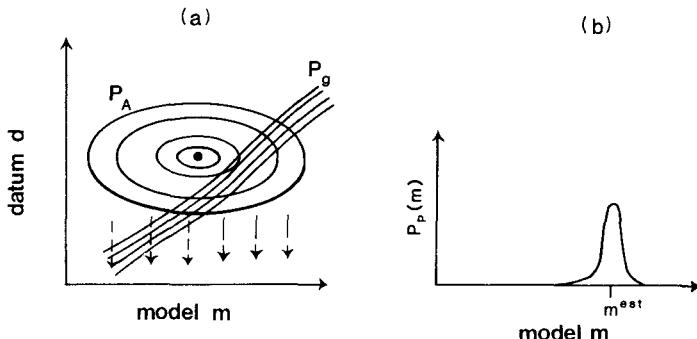


Fig. 5.15. (a) The joint distribution $P_T(m, d) = P_A(m, d)P_g(m|d)$ is projected onto the m axis to form the distribution for model parameters alone, $P_p(m)$. (b) The projected distribution $P_p(m)$.

where the integration is performed over the entire range of d_i . Note that as long as most of the probability is concentrated along the surface $g(\mathbf{m}) = \mathbf{d}$, the process of projecting does not change the maximum likelihood point of the model parameters. Therefore, the distinction between estimating both data and model parameters and estimating only model parameters is important only when the theory is inexact. Furthermore, it turns out that the two types of estimates are always equal if all the distributions involved are Gaussian.

5.6 The Simple Gaussian Case with a Linear Theory

To illustrate this method we again derive the least squares solution. We assume that there is no a priori information, that $P_A(\mathbf{m}) \propto \text{constant}$, and that the data have a Gaussian distribution.

$$P_A(\mathbf{d}) \propto \exp\left[-\frac{1}{2}(\mathbf{d} - \mathbf{d}^{\text{obs}})^T [\text{cov } \mathbf{d}]^{-1} (\mathbf{d} - \mathbf{d}^{\text{obs}})\right] \quad (5.11)$$

If there are no errors in the theory, then its distribution is a Dirac delta function

$$P_g(\mathbf{m}|\mathbf{d}) = \delta[\mathbf{G}\mathbf{m} - \mathbf{d}] \quad (5.12)$$

where we assume that we are dealing with the linear theory $\mathbf{G}\mathbf{m} = \mathbf{d}$. The total distribution is then given by

$$P_T(\mathbf{m}, \mathbf{d}) \propto \exp\left[-\frac{1}{2}(\mathbf{d} - \mathbf{d}^{\text{obs}})^T [\text{cov } \mathbf{d}]^{-1} (\mathbf{d} - \mathbf{d}^{\text{obs}})\right] \delta[\mathbf{G}\mathbf{m} - \mathbf{d}] \quad (5.13)$$

Performing the projection “integrates away” the delta function

$$P_p(\mathbf{m}) \propto \exp\left[-\frac{1}{2}(\mathbf{G}\mathbf{m} - \mathbf{d}^{obs})^T[\text{cov } \mathbf{d}]^{-1}(\mathbf{G}\mathbf{m} - \mathbf{d}^{obs})\right] \quad (5.14)$$

This projected distribution is exactly the one we encountered in the least squares problem, and the position of its maximum likelihood point is given by the least squares solution.

5.7 The General Linear, Gaussian Case

In the general linear, Gaussian case we assume that all the component distributions are Gaussian and that the theory is the linear equation $\mathbf{G}\mathbf{m} = \mathbf{d}$, so that

$$\begin{aligned} P_A(\mathbf{m}) &\propto \exp\left[-\frac{1}{2}(\mathbf{m} - \langle \mathbf{m} \rangle)^T[\text{cov } \mathbf{m}]^{-1}(\mathbf{m} - \langle \mathbf{m} \rangle)\right] \\ P_A(\mathbf{d}) &\propto \exp\left[-\frac{1}{2}(\mathbf{d} - \mathbf{d}^{obs})^T[\text{cov } \mathbf{d}]^{-1}(\mathbf{d} - \mathbf{d}^{obs})\right] \\ P_g(\mathbf{m}|\mathbf{d}) &\propto \exp\left[-\frac{1}{2}(\mathbf{d} - \mathbf{G}\mathbf{m})^T[\text{cov } \mathbf{g}]^{-1}(\mathbf{d} - \mathbf{G}\mathbf{m})\right] \end{aligned} \quad (5.15)$$

The total distribution is then the product of these three distributions. We now show that the combined distribution is itself Gaussian. We first define a vector $\mathbf{x} = [\mathbf{d}, \mathbf{m}]^T$ that contains the data and model parameters and a covariance matrix.

$$[\text{cov } \mathbf{x}] = \begin{bmatrix} [\text{cov } \mathbf{d}] & \mathbf{0} \\ \mathbf{0} & [\text{cov } \mathbf{m}] \end{bmatrix} \quad (5.16)$$

The first two products in the total distribution can then be combined into an exponential, with the argument given by

$$-\frac{1}{2}[\mathbf{x} - \langle \mathbf{x} \rangle]^T[\text{cov } \mathbf{x}]^{-1}[\mathbf{x} - \langle \mathbf{x} \rangle] \quad (5.17)$$

To express the third product in terms of \mathbf{x} , we define a matrix $\mathbf{F} = [\mathbf{I}, -\mathbf{G}]$ such that $\mathbf{F}\mathbf{x} = \mathbf{d} - \mathbf{G}\mathbf{m} = 0$. The argument of the third product’s exponential is then given by

$$-\frac{1}{2}[\mathbf{F}\mathbf{x}]^T[\text{cov } \mathbf{g}]^{-1}[\mathbf{F}\mathbf{x}] \quad (5.18)$$

The total distribution is proportional to an exponential with argument

$$-\frac{1}{2}[\mathbf{x} - \langle \mathbf{x} \rangle]^T[\text{cov } \mathbf{x}]^{-1}[\mathbf{x} - \langle \mathbf{x} \rangle] - \frac{1}{2}[\mathbf{F}\mathbf{x}]^T[\text{cov } \mathbf{g}]^{-1}[\mathbf{F}\mathbf{x}] \quad (5.19)$$

We shall show that this expression can be manipulated into the form

$$-\frac{1}{2}[\mathbf{x} - \mathbf{x}^*]^T[\text{cov } \mathbf{x}^*]^{-1}[\mathbf{x} - \mathbf{x}^*] \quad (5.20)$$

Here \mathbf{x}^* is some vector and $[\text{cov } \mathbf{x}^*]$ is some matrix. The total distribution $P_T(\mathbf{x})$, therefore, has a Gaussian form and has a maximum likelihood point $\mathbf{x}^{\text{est}} = \mathbf{x}^*$. In Section 5.10, we shall derive \mathbf{x}^* and $[\text{cov } \mathbf{x}^*]$. For now, we simply state the result as

$$\begin{aligned}\mathbf{x}^* &= [\mathbf{I} - [\text{cov } \mathbf{x}] \mathbf{F}^T (\mathbf{F} [\text{cov } \mathbf{x}] \mathbf{F}^T \\ &\quad + [\text{cov } \mathbf{g}])^{-1} \mathbf{F}] \langle \mathbf{x} \rangle \\ [\text{cov } \mathbf{x}^*] &= [\mathbf{I} - [\text{cov } \mathbf{x}] \mathbf{F}^T (\mathbf{F} [\text{cov } \mathbf{x}] \mathbf{F}^T \\ &\quad + [\text{cov } \mathbf{g}])^{-1} \mathbf{F}] [\text{cov } \mathbf{x}]\end{aligned}\tag{5.21}$$

Note that the formula for \mathbf{x}^* has the form of a transformation or projection of the a priori vector $\langle \mathbf{x} \rangle$, that is $\mathbf{x}^* = \mathbf{T} \langle \mathbf{x} \rangle$. Note also that nothing in this derivation requires the special forms of \mathbf{F} and $[\text{cov } \mathbf{x}]$ assumed above that made $\mathbf{F}\mathbf{x} = 0$ separable into an *explicit* linear inverse problem. Equation (5.21) is in fact the solution to the completely general, *implicit*, linear inverse problem.

When $\mathbf{F}\mathbf{x} = 0$ is an explicit equation, the formula for \mathbf{x}^{est} in Eqn. (5.21) can be decomposed into its component vectors \mathbf{d}^{pre} and \mathbf{m}^{est} . An explicit formula for the estimated model parameters is given by

$$\begin{aligned}\mathbf{m}^{\text{est}} &= \langle \mathbf{m} \rangle + \mathbf{G}^{-\mathbf{g}} [\mathbf{d}^{\text{obs}} - \mathbf{G} \langle \mathbf{m} \rangle] = \mathbf{G}^{-\mathbf{g}} \mathbf{d} + [\mathbf{I} - \mathbf{R}] \langle \mathbf{m} \rangle \\ \mathbf{G}^{-\mathbf{g}} &= [\text{cov } \mathbf{m}] \mathbf{G}^T \{ [\text{cov } \mathbf{d}] + [\text{cov } \mathbf{g}] + \mathbf{G} [\text{cov } \mathbf{m}] \mathbf{G}^T \}^{-1} \\ &= (\mathbf{G}^T \{ [\text{cov } \mathbf{d}] + [\text{cov } \mathbf{g}] \}^{-1} \mathbf{G} \\ &\quad + [\text{cov } \mathbf{m}]^{-1})^{-1} \mathbf{G}^T \{ [\text{cov } \mathbf{d}] + [\text{cov } \mathbf{g}] \}^{-1}\end{aligned}\tag{5.22}$$

where we have used the generalized inverse notation for convenience. Note that the two forms of the generalized inverse are equivalent (as can be shown by applying the matrix identities derived in Section 5.10).

Since the estimated model parameters are a linear combination of observed data and a priori model parameters, we can therefore calculate its covariance as

$$[\text{cov } \mathbf{m}^{\text{est}}] = \mathbf{G}^{-\mathbf{g}} [\text{cov } \mathbf{d}] \mathbf{G}^{-\mathbf{g}^T} + [\mathbf{I} - \mathbf{R}] [\text{cov } \mathbf{m}] [\mathbf{I} - \mathbf{R}]^T\tag{5.23}$$

This expression differs from those derived in Chapters 3 and 4 in that it contains a term dependent on the a priori model parameter covariance $[\text{cov } \mathbf{m}]$.

We can examine a few interesting limiting cases of problems which have uncorrelated a priori model parameters ($[\text{cov } \mathbf{m}] = \sigma_m^2 \mathbf{I}$), data ($[\text{cov } \mathbf{d}] = \sigma_d^2 \mathbf{I}$) and theory ($[\text{cov } \mathbf{g}] = \sigma_g^2 \mathbf{I}$).

5.7.1 EXACT DATA AND THEORY

Suppose $\sigma_d^2 = \sigma_g^2 = 0$. The solution is then given by

$$\mathbf{m}^{\text{est}} = \mathbf{G}^T[\mathbf{G}\mathbf{G}^T]^{-1}\mathbf{d}^{\text{obs}} = [\mathbf{G}^T\mathbf{G}]^{-1}\mathbf{G}^T\mathbf{d}^{\text{obs}} \quad (5.24)$$

Note that the solution does not depend on the a priori model variance, since the data and theory are infinitely more accurate than the a priori model parameters. These solutions are just the minimum length and least squares solutions, which (as we now see) are simply two different aspects of the same solution. The minimum length of the solution, however, exists only when the problem is purely underdetermined; the least squares form exists only when the problem is purely overdetermined.

If the a priori model parameters are not equal to zero, then another term appears in the estimated solution.

$$\begin{aligned} \mathbf{m}^{\text{est}} &= \mathbf{G}^T[\mathbf{G}\mathbf{G}^T]^{-1}\mathbf{d}^{\text{obs}} + (\mathbf{I} - \mathbf{G}^T[\mathbf{G}\mathbf{G}^T]^{-1}\mathbf{G})\langle \mathbf{m} \rangle \\ &= [\mathbf{G}^T\mathbf{G}]^{-1}\mathbf{G}^T\mathbf{d}^{\text{obs}} + (\mathbf{I} - [\mathbf{G}^T\mathbf{G}]^{-1}\mathbf{G}^T\mathbf{G})\langle \mathbf{m} \rangle \quad (5.25) \\ &= \mathbf{G}^{-g}\mathbf{d}^{\text{obs}} + (\mathbf{I} - \mathbf{R})\langle \mathbf{m} \rangle \end{aligned}$$

The minimum-length-type solution has been changed by adding a weighted amount of the a priori model vector, with the weighting factor being $(\mathbf{I} - \mathbf{G}^T[\mathbf{G}\mathbf{G}^T]^{-1}\mathbf{G})$. This term is not zero, since it can also be written as $(\mathbf{I} - \mathbf{R})$. The resolution matrix of the underdetermined problem never equals the identity matrix. On the other hand, the resolution matrix of the overdetermined least squares problem does equal the identity matrix, so the estimated model parameters of the overdetermined problem are not a function of the a priori model parameters. Adding a priori information with finite error to an inverse problem that features exact data and theory only affects the underdetermined part of the solution.

5.7.2 INFINITELY INEXACT DATA AND THEORY

In the case of infinitely inexact data and theory, we take the limit $\sigma_d^2 \rightarrow \infty$ or $\sigma_g^2 \rightarrow \infty$ (or both). The solution becomes

$$\mathbf{m}^{\text{est}} = \langle \mathbf{m} \rangle \quad (5.26)$$

Since the data and theory contain no information, we simply recover the a priori model parameters.

5.7.3 NO A PRIORI KNOWLEDGE OF THE MODEL PARAMETERS

In this case, the limit is $\sigma_m^2 \rightarrow \infty$. The solutions are the same as in Section 5.6.1.

$$\mathbf{m}^{\text{est}} = \mathbf{G}^T[\mathbf{G}\mathbf{G}^T]^{-1}\mathbf{d}^{\text{obs}} + \{\mathbf{I} - \mathbf{G}^T[\mathbf{G}\mathbf{G}^T]^{-1}\mathbf{G}\}\langle\mathbf{m}\rangle \quad (5.27)$$

Infinitely weak a priori information and finite-error data and theory, produce the same results as finite-error a priori information and error-free data and theory.

5.8 Equivalence of the Three Viewpoints

We can arrive at the same general solution to the linear inverse problem by three distinct routes.

Viewpoint 1. The solution is obtained by minimizing a weighted sum of L_2 prediction error and L_2 solution simplicity.

$$\text{Minimize: } \mathbf{e}^T \mathbf{W}_e \mathbf{e} + \alpha [\mathbf{m} - \langle\mathbf{m}\rangle]^T \mathbf{W}_m [\mathbf{m} - \langle\mathbf{m}\rangle] \quad (5.28)$$

where α is a weighting factor.

Viewpoint 2. The solution is obtained by minimizing a weighted sum of three terms: the Dirichlet spreads of model resolution and data resolution and the size of the model covariance.

$$\text{Minimize: } \alpha_1 \text{ spread}(\mathbf{R}) + \alpha_2 \text{ spread}(\mathbf{N}) + \alpha_3 \text{ size}[\text{cov}_u \mathbf{m}] \quad (5.29)$$

Viewpoint 3. The solution is obtained by maximizing the likelihood of the joint Gaussian distribution of data, a priori model parameters, and theory.

$$\text{Maximize: } P_T(\mathbf{m}, \mathbf{d}) \quad (5.30)$$

These derivations emphasize the close relationship among the L_2 norm, the Dirichlet spread function, and the Gaussian distribution.

5.9 The *F* Test of Error Improvement Significance

We sometimes have *two* candidate models for describing an over-determined inverse problem, one of which is more complicated than the other (in the sense that it possesses a greater number of model parameters). Suppose that Model 2 is more complicated than Model 1 and that the total prediction error for Model 2 is less than the total prediction error for Model 1: $E_2 < E_1$. Does Model 2 really fit the data better than Model 1?

The answer to this question depends on the variance of the data. Almost any complicated model will fit data better than a less complicated one. The relevant question is whether the fit is *significantly* better, that is, whether the improvement is too large to be accounted for by random fluctuations in the data. For statistical reasons that will be cited, we pretend, in this case, that the two inverse problems are solved with two different realizations of the data.

Suppose that we estimate the variance of the data d_i from the prediction error e_i of each model as

$$\sigma_d^2 = \sum e_i^2 / (N - M)$$

This estimate will usually be larger than the true variance of the data, since it also includes a contribution from the (possibly) poor fit of the model. If one model fits the data about as well as the other, then the variance σ_{d1}^2 estimated from Model 1 should be about the same as the variance σ_{d2}^2 estimated from Model 2. On the other hand, if Model 2 gives a better fit than Model 1, the estimated variances will differ in such a way that the ratio $\sigma_{d1}^2 / \sigma_{d2}^2$ will be greater than unity. If the ratio is only slightly greater than unity, the difference in fit may be entirely a result of random fluctuations in the data and therefore may not be significant. Nevertheless, there is clearly some value for the ratio that indicates a significant difference between the two fits.

To compute this critical value, we consider the theoretical distribution for the ratio of two variance estimates derived from two different realizations of the *same* data set. Of course, the ratio of the true variance with itself always has the value unity; but the ratio of two estimates of the true variance will fluctuate randomly about unity. We therefore determine whether or not ratios greater than or equal to the observed ratio occur less than, say, 5% of the time. If they do, then there is a 95% probability that the two estimates are derived from data

sets with different true variances. We are justified in concluding that the second model is a significant improvement over the first.

To handle data with nonuniform variance we form a ratio, not of estimated variances, but of the related quantity

$$\chi_v^2 = 1/v \sum_{i=1}^N e_i^2 / \sigma_{di}^{\text{true}}; \quad v = N - M \quad (5.31)$$

This quantity is chosen because it has a χ_v^2 distribution with v degrees of freedom. The ratio of the χ_v^2 for the two models is given by

$$F = \chi_{v_1}^2 / \chi_{v_2}^2 \quad (5.32)$$

Note that the F ratio is not a function of the overall amplitude of the data's true variance but only of the relative error between the different observations. In practice it is sufficient to use approximate estimates of the relative error between the d_i when computing the F ratio.

The distribution of the F ratio has been derived by statisticians and is called the F distribution. It is a unimodal distribution with mean $v_2/(v_2 - 2)$. Its variance is given by

$$\text{var}(F) = \frac{2v_2^2(v_1 + v_2 - 2)}{v_1(v_2 - 2)^2(v_2 - 4)} \quad (5.33)$$

The functional form of the F distribution is given by

$$P(F) = \frac{\Gamma[(v_1 + v_2)/2](v_1/v_2)^{v_1/2} F^{(v_1/2)-1}}{\Gamma(v_1/2)\Gamma(v_2/2)[1 + (v_1 F/v_2)]^{(v_1+v_2)/2}} \quad (5.34)$$

where Γ is the Gamma function.

Most statistical texts provide tables that give the value for which ratios greater than or equal to F occur only 5% of the time. If the F for the two candidate models is greater than this critical value, then we can reasonably assume that the improvement in error is not a result of random fluctuations in the data but of a significant difference between the models.

5.10 Derivation of the Formulas of Section 5.7

We first need to derive two general matrix identities (adapted from [Ref. 18], with permission). Let \mathbf{C}_1 and \mathbf{C}_2 be two symmetric matrices

whose inverses exist, and let \mathbf{M} be a third matrix. Then note that the expression $\mathbf{M}^T + \mathbf{M}^T\mathbf{C}_1^{-1}\mathbf{MC}_2\mathbf{M}^T$ can be written two ways by grouping terms: as $\mathbf{M}^T\mathbf{C}_1^{-1}[\mathbf{C}_1 + \mathbf{MC}_2\mathbf{M}^T]$ or as $[\mathbf{C}_2^{-1} + \mathbf{M}^T\mathbf{C}_1^{-1}\mathbf{M}]\mathbf{C}_2\mathbf{M}^T$. Multiplying by the matrix inverses gives

$$\mathbf{C}_2\mathbf{M}^T[\mathbf{C}_1 + \mathbf{MC}_2\mathbf{M}^T]^{-1} = [\mathbf{C}_2^{-1} + \mathbf{M}^T\mathbf{C}_1^{-1}\mathbf{M}]^{-1}\mathbf{M}^T\mathbf{C}_1^{-1} \quad (5.35)$$

Now consider the symmetric matrix expression $\mathbf{C}_2 - \mathbf{C}_2\mathbf{M}^T[\mathbf{C}_1 + \mathbf{MC}_2\mathbf{M}^T]^{-1}\mathbf{MC}_2$. By Eq. 5.31 this expression equals $\mathbf{C}_2 - [\mathbf{C}_2^{-1} + \mathbf{M}^T\mathbf{C}_1^{-1}\mathbf{M}]^{-1}\mathbf{M}^T\mathbf{C}_1^{-1}\mathbf{MC}_2$. Factoring out the term in brackets gives $[\mathbf{C}_2^{-1} + \mathbf{M}^T\mathbf{C}_1^{-1}\mathbf{M}]^{-1} \{[\mathbf{C}_2^{-1} + \mathbf{M}^T\mathbf{C}_1^{-1}\mathbf{M}]\mathbf{C}_2 - \mathbf{M}^T\mathbf{C}_1^{-1}\mathbf{MC}_2\}$. Canceling terms gives $[\mathbf{C}_2^{-1} + \mathbf{M}^T\mathbf{C}_1^{-1}\mathbf{M}]^{-1}$ from which we conclude

$$\mathbf{C}_2 - \mathbf{C}_2\mathbf{M}^T[\mathbf{C}_1 + \mathbf{MC}_2\mathbf{M}^T]^{-1}\mathbf{MC}_2 = [\mathbf{C}_2^{-1} + \mathbf{M}^T\mathbf{C}_1^{-1}\mathbf{M}]^{-1} \quad (5.36)$$

We now consider the argument of the exponential in the joint distribution (ignoring the factor of $-\frac{1}{2}$):

$$\begin{aligned} & [\mathbf{x} - \langle \mathbf{x} \rangle]^T[\text{cov } \mathbf{x}]^{-1}[\mathbf{x} - \langle \mathbf{x} \rangle] + [\mathbf{F}\mathbf{x}]^T[\text{cov } \mathbf{g}]^{-1}[\mathbf{F}\mathbf{x}] \\ &= \mathbf{x}^T[\text{cov } \mathbf{x}]^{-1}\mathbf{x} + \langle \mathbf{x} \rangle[\text{cov } \mathbf{x}]^{-1}\langle \mathbf{x} \rangle - 2\mathbf{x}^T[\text{cov } \mathbf{x}]^{-1}\langle \mathbf{x} \rangle \\ &\quad + \mathbf{x}^T\mathbf{F}^T[\text{cov } \mathbf{g}]^{-1}\mathbf{F}\mathbf{x} \\ &= \mathbf{x}^T[\mathbf{F}^T[\text{cov } \mathbf{g}]^{-1}\mathbf{F} + [\text{cov } \mathbf{x}]^{-1}]\mathbf{x} - 2\mathbf{x}^T[\text{cov } \mathbf{x}]^{-1}\langle \mathbf{x} \rangle \\ &\quad + \langle \mathbf{x} \rangle^T[\text{cov } \mathbf{x}]^{-1}\langle \mathbf{x} \rangle \end{aligned} \quad (5.37)$$

We want to show that this expression equals

$$[\mathbf{x} - \mathbf{x}^*]^T[\text{cov } \mathbf{x}^*]^{-1}[\mathbf{x} - \mathbf{x}^*] \quad (5.38)$$

By the second matrix identity [Eq. (5.36)], the value of $[\text{cov } \mathbf{x}^*]$ is given by

$$\begin{aligned} [\text{cov } \mathbf{x}^*] &= [\mathbf{I} - [\text{cov } \mathbf{x}]\mathbf{F}^T(\mathbf{F}[\text{cov } \mathbf{x}]\mathbf{F}^T + [\text{cov } \mathbf{g}])^{-1}][\text{cov } \mathbf{x}] \\ &= [\mathbf{F}^T[\text{cov } \mathbf{g}]^{-1}\mathbf{F} + [\text{cov } \mathbf{x}]]^{-1} \end{aligned} \quad (5.39)$$

and \mathbf{x}^* is

$$\mathbf{x}^* = [\text{cov } \mathbf{x}^*][\text{cov } \mathbf{x}]^{-1}\langle \mathbf{x} \rangle \quad (5.40)$$

Substituting Eqs. 5.39 and 5.40 into Eq. 5.38:

$$\begin{aligned} & [\mathbf{x} - \mathbf{x}^*]^T[\text{cov } \mathbf{x}^*]^{-1}[\mathbf{x} - \mathbf{x}^*] \\ &= \mathbf{x}^T[\text{cov } \mathbf{x}^*]^{-1}\mathbf{x} - 2\mathbf{x}^T[\text{cov } \mathbf{x}^*]^{-1}\mathbf{x}^* + \mathbf{x}^{*\top}[\text{cov } \mathbf{x}^*]^{-1}\mathbf{x}^* \\ &= \mathbf{x}^T[\mathbf{F}^T[\text{cov } \mathbf{g}]^{-1}\mathbf{F} + [\text{cov } \mathbf{x}]^{-1}]\mathbf{x} \end{aligned}$$

$$\begin{aligned} & -2\mathbf{x}^T[\text{cov } \mathbf{x}^*]^{-1}[\text{cov } \mathbf{x}^*][\text{cov } \mathbf{x}]\langle \mathbf{x} \rangle \\ & + \langle \mathbf{x} \rangle [\text{cov } \mathbf{x}]^{-1}[\text{cov } \mathbf{x}^*][\text{cov } \mathbf{x}^*]^{-1}[\text{cov } \mathbf{x}^*][\text{cov } \mathbf{x}]^{-1}\langle \mathbf{x} \rangle \quad (5.41) \end{aligned}$$

This differs from the desired result by only a constant. Since this expression is the argument of an exponential, any constant term can be absorbed into the overall normalization. We have, therefore, finished the derivation of \mathbf{x}^* and $[\text{cov } \mathbf{x}^*]$.

This page intentionally left blank

6

NONUNIQUENESS AND LOCALIZED AVERAGES

6.1 Null Vectors and Nonuniqueness

In Chapters 3–5 we presented the basic method of finding estimates of the model parameters in a linear inverse problem. We showed that we could always obtain such estimates but that sometimes in order to do so we had to add a priori information to the problem. We shall now consider the meaning and consequences of nonuniqueness in linear inverse problems and show that it is possible to devise solutions that do not depend at all on a priori information. As we shall show, however, these solutions are not estimates of the model parameters themselves but estimates of *weighted averages* (linear combinations) of the model parameters.

When the linear inverse problem $\mathbf{Gm} = \mathbf{d}$ has nonunique solutions, there exist nontrivial solutions (that is, solutions with some nonzero m_i) to the homogeneous equation $\mathbf{Gm} = 0$. These solutions are called the *null vectors* of the inverse problem since premultiplying them by the data kernel yields zero. To see why nonuniqueness implies null

vectors, suppose that the inverse problem has two distinct solutions \mathbf{m}^1 and \mathbf{m}^2 as

$$\begin{aligned}\mathbf{Gm}^1 &= \mathbf{d} \\ \mathbf{Gm}^2 &= \mathbf{d}\end{aligned}\tag{6.1}$$

Subtracting these two equations yields

$$\mathbf{G}(\mathbf{m}^1 - \mathbf{m}^2) = \mathbf{0}\tag{6.2}$$

Since the two solutions are by assumption distinct, their difference $\mathbf{n}^{\text{null}} = \mathbf{m}^1 - \mathbf{m}^2$ is nonzero. The converse is also true; any linear inverse problem that has null vectors is nonunique. If \mathbf{m}^{par} (where par stands for “particular”) is any nonnull solution to $\mathbf{Gm} = \mathbf{d}$ (for instance, the minimum length solution), then $\mathbf{m}^{\text{par}} + \alpha\mathbf{n}^{\text{null}}$ is also a solution for any choice of α . Note that since $\alpha\mathbf{n}^{\text{null}}$ is a null vector for any nonzero α , null vectors are only distinct if they are linearly independent. If a given inverse problem has q distinct null solutions, then the most general solution is

$$\mathbf{m}^{\text{gen}} = \mathbf{m}^{\text{par}} + \sum_{i=1}^q \alpha_i \mathbf{m}_i^{\text{null}}\tag{6.3}$$

where gen stands for “general.” We shall show that $0 \leq q \leq M$, that is, that there can be no more linearly independent null vectors than there are unknowns.

6.2 Null Vectors of a Simple Inverse Problem

As an example, consider the following very simple equations:

$$\mathbf{Gm} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \end{bmatrix} = [d_1]\tag{6.4}$$

This equation implies that only the mean value of a set of four model parameters has been measured. One obvious solution to this equation is $\mathbf{m} = [d_1, d_1, d_1, d_1]^T$ (in fact, this is the minimum length solution).

The null solutions can be determined by inspection as

$$\mathbf{m}_1^{\text{null}} = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{m}_2^{\text{null}} = \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \end{bmatrix}, \quad \mathbf{m}_3^{\text{null}} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \end{bmatrix} \quad (6.5)$$

The most general solution is then

$$\mathbf{m}^{\text{gen}} = \begin{bmatrix} d_1 \\ d_1 \\ d_1 \\ d_1 \end{bmatrix} + \alpha_1 \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix} + \alpha_2 \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \end{bmatrix} + \alpha_3 \begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \end{bmatrix} \quad (6.6)$$

where the α 's are arbitrary parameters.

Finding a particular solution to this problem now consists of choosing values for the parameters α_i . If one chooses these parameters so that $\|\mathbf{m}\|_2$ is minimized, one obtains the minimum length solution. Since the first vector is orthogonal to all the others, this minimum occurs when $\alpha_i = 0, i = 1, 2, 3$. We shall show in Chapter 7 that this is a general result: the minimum length solution never contains any null vectors. Note, however, that if other definitions of solution simplicity are used (e.g., flatness or roughness), those solutions will contain null vectors.

6.3 Localized Averages of Model Parameters

We have sought to estimate the elements of the solution vector \mathbf{m} . Another approach is to estimate some average of the model parameter $\langle m \rangle = \mathbf{a}^T \mathbf{m}$, where \mathbf{a} is some averaging vector. The average is said to be localized if this averaging vector consists mostly of zeros (except for some group of nonzero elements that multiplies model parameters centered about one particular model parameter). This definition makes particular sense when the model parameters possess some natural ordering in space and time, such as acoustic velocity as a function of depth in the earth. For instance, if $M = 8$, the averaging

vector $\mathbf{a} = [0, 0, \frac{1}{4}, \frac{1}{2}, \frac{1}{4}, 0, 0, 0]^T$ could be said to be localized about the fourth model parameter. The averaging vectors are usually normalized so that the sum of their elements is unity.

The advantage of estimating averages of the model parameters rather than the model parameters themselves is that quite often it is possible to identify unique averages even when the model parameters themselves are not unique. To examine when uniqueness can occur we compute the average of the general solution as

$$\langle m \rangle = \mathbf{a}^T \mathbf{m}^{\text{gen}} = \mathbf{a}^T \mathbf{m}^{\text{par}} + \sum_{i=1}^q \alpha_i \mathbf{a}^T \mathbf{m}_i^{\text{null}} \quad (6.7)$$

If $\mathbf{a}^T \mathbf{m}_i^{\text{null}}$ is zero for all i , then $\langle m \rangle$ is unique. The process of averaging has completely removed the nonuniqueness of the problem. Since \mathbf{a} has M elements and there are $q \leq M$ constraints placed on \mathbf{a} , one can always find at least one vector that cancels (or “annihilates”) the null vectors. One cannot, however, always guarantee that the averaging vector is localized around some particular model parameter. But, if $q < M$, one has some freedom in choosing \mathbf{a} and there is some possibility of making the averaging vector at least somewhat localized. Whether this can be done depends on the structure of the null vectors, which in turn depends on the structure of the data kernel \mathbf{G} . Since the small-scale features of the model are unresolvable in many problems, unique localized averages can often be found.

6.4 Relationship to the Resolution Matrix

During the discussion of the resolution matrix \mathbf{R} (Section 4.3), we encountered in a somewhat different form the problem of determining averaging vectors. We showed that any estimate \mathbf{m}^{est} computed from a generalized inverse \mathbf{G}^{-g} was related to the true model parameters by

$$\mathbf{m}^{\text{est}} = \mathbf{G}^{-g} \mathbf{G} \mathbf{m}^{\text{true}} = \mathbf{R} \mathbf{m}^{\text{true}}$$

The i th row of \mathbf{R} can be interpreted as a unique averaging vector that is centered about m_i . Whether or not the average is truly localized depends on the structure of \mathbf{R} . The spread function discussed previously is a measure of the degree of localization.

Note that the resolution matrix is composed of the product of the generalized inverse and the data kernel. We can interpret this product

as meaning that a row of the resolution matrix is composed of a weighted sum of the rows of the data kernel \mathbf{G} (where the elements of the generalized inverse are the weighting factors) regardless of the generalized inverse's particular form. An averaging vector \mathbf{a} is unique if and only if it can be represented as a linear combination of the rows of the data kernel \mathbf{G} .

The process of forming the generalized inverse is equivalent to "shuffling" the rows of the equation $\mathbf{Gm} = \mathbf{d}$ by forming linear combinations until the data kernel is as close as possible to an identity matrix. Each row of the data kernel can then be viewed as a localized averaging vector, and each element of the shuffled data vector is the estimated value of the average.

6.5 Averages versus Estimates

We can, therefore, identify a type of dualism in inverse theory. Given a generalized inverse \mathbf{G}^{-g} that in some sense solves $\mathbf{Gm} = \mathbf{d}$, we can speak either of estimates of model parameters $\mathbf{m}^{\text{est}} = \mathbf{G}^{-g}\mathbf{d}$ or of localized averages $\langle \mathbf{m} \rangle = \mathbf{G}^{-g}\mathbf{d}$. The numerical values are the same but the interpretation is quite different. When the solution is interpreted as a localized average, it can be viewed as a unique quantity that exists independently of any a priori information applied to the inverse problem. Examination of the resolution matrix may reveal that the average is not especially localized and the solution may be difficult to interpret. When the solution is viewed as an estimate of a model parameter, the location of what is being solved for is clear. The estimate can be viewed as unique only if one accepts as appropriate whatever a priori information was used to remove the inverse problem's underdeterminacy. In most instances, the choice of a priori information is somewhat ad hoc so the solution may still be difficult to interpret.

In the sample problem stated above, the data kernel has only one row. There is therefore only one averaging vector that will annihilate all the null vectors: one proportional to that row

$$\mathbf{a} = [\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}]^T \quad (6.8)$$

This averaging vector is clearly unlocalized. In this problem the structure of \mathbf{G} is just too poor to form good averages. The generalized

inverse to this problem is by inspection

$$\mathbf{G}^{-\mathbf{s}} = [1, 1, 1, 1]^T \quad (6.9)$$

The resolution matrix is therefore

$$\mathbf{R} = \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad (6.10)$$

which is very unlocalized and equivalent to Eq. 6.8.

6.6 Nonunique Averaging Vectors and A Priori Information

There are instances in which even nonunique averages of model parameters can be of value, especially when they are used in conjunction with other a priori knowledge of the nature of the solution [Ref. 21]. Suppose that one simply picks a localized averaging vector that does not necessarily annihilate all the null vectors and that, therefore, does not lead to a unique average. In the above problem, the vector $\mathbf{a} = [\frac{1}{3} \frac{1}{3} \frac{1}{3} 0]^T$ might be such a vector. It is somewhat localized, being centered about the second model parameter. Note that it does not lead to a unique average, since

$$\langle m \rangle = \mathbf{a}^T \mathbf{m}^{\text{gen}} = d_1 + 0 + 0 + \frac{1}{3}\alpha_3 \quad (6.11)$$

is still a function of one of the arbitrary parameters α_i . Suppose, however, that there is a priori knowledge that every m_i must satisfy $0 \leq m_i \leq 2d_1$. Then from the equation for \mathbf{m}^{gen} , α_3 must be no greater than d_1 and no less than $-d_1$. Since $-d_1 \leq \alpha_3 \leq d_1$, the average has bounds $\frac{2}{3}d_1 \leq \langle m \rangle \leq \frac{4}{3}d_1$. These constraints are considerably tighter than the a priori bounds on m_i , which demonstrates that this technique has indeed produced some useful information. This approach works because even though the averaging vector does not annihilate all the null vectors, $\mathbf{a}^T \mathbf{m}^{\text{null}}$ is small compared with the elements of the null vector. Localized averaging vectors often lead to small products since the null vectors often fluctuate rapidly about zero, indicating that small-scale features of the model are the most poorly resolved. A slightly more complicated example of this type is solved in Fig. 6.1.

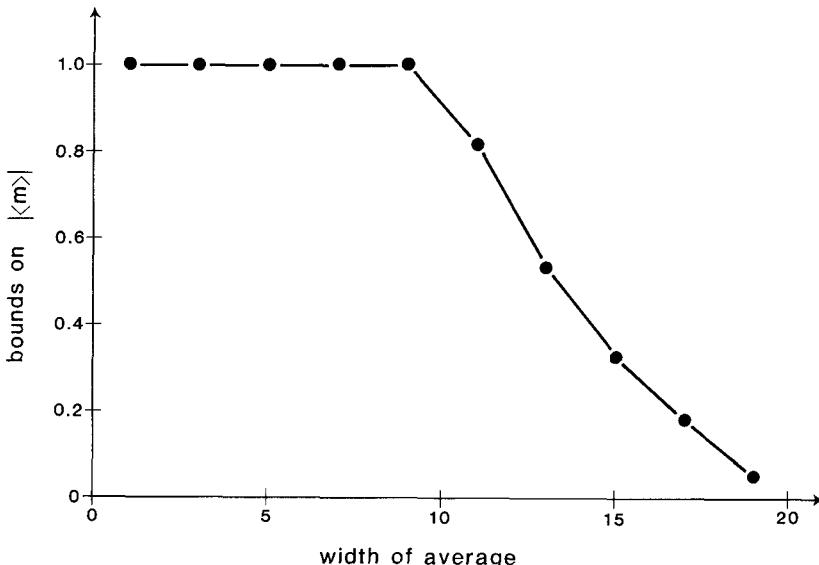


Fig. 6.1. Bounds on weighted averages of model parameters in a problem in which the only datum is that the mean of 20 unknowns is zero. When this observation is combined with the a priori information that each model parameter must satisfy $|m_i| \leq 1$, bounds can be placed on weighted averages of the model parameters. The bounds shown here are for weighted averages centered about the tenth model parameter. Note that the bounds are tighter than the a priori bounds only when the average includes more than 10 model parameters. These bounds were computed according to the method described in Section 7.6.

This approach can be generalized as follows: Given $\mathbf{G}\mathbf{m} = \mathbf{d}$ with solution

$$\mathbf{m}^{\text{gen}} = \mathbf{m}^{\text{par}} + \sum_{i=1}^q \alpha_i \mathbf{m}_i^{\text{null}} \quad (6.12)$$

and given averaging function \mathbf{a} , determine bounds on the average $\langle m \rangle = \mathbf{a}^T \mathbf{m}$ if the solution itself is known to possess bounds $\mathbf{m}^l \leq \mathbf{m} \leq \mathbf{m}^u$ (the superscripts l and u indicate lower and upper limiting values, respectively). The upper bound on $\langle m \rangle$ is found by maximizing $\langle m \rangle$ with respect to α_i , subject to the linear inequality constraints $\mathbf{m}^l \leq \mathbf{m} \leq \mathbf{m}^u$. The lower bound is found by minimizing $\langle m \rangle$ under the same constraints. This procedure is precisely a linear programming problem (see Section 12.8).

This page intentionally left blank

7

APPLICATIONS OF VECTOR SPACES

7.1 Model and Data Spaces

We have used vector notation for the data \mathbf{d} and model parameters \mathbf{m} mainly because it facilitates the algebra. The concept of vectors can also be used to gain insight into the properties of inverse problems. We therefore introduce the idea of vector spaces containing \mathbf{d} and \mathbf{m} , which we shall denote $S(\mathbf{d})$ and $S(\mathbf{m})$. Any particular choice of \mathbf{m} and \mathbf{d} is then represented as a vector in these spaces (Fig. 7.1).

The linear equation $\mathbf{G}\mathbf{m} = \mathbf{d}$ can be interpreted as a mapping of vectors from $S(\mathbf{m})$ to $S(\mathbf{d})$ and its solution $\mathbf{m}^{\text{est}} = \mathbf{G}^{-1}\mathbf{d}$ as a mapping of vectors from $S(\mathbf{d})$ to $S(\mathbf{m})$.

One important property of a vector space is that its coordinate axes are arbitrary. Thus far we have been using axes parallel to the individual model parameters, but we recognize that we are by no means required to do so. Any set of vectors that spans the space will serve as coordinate axes. The M th dimensional space $S(\mathbf{m})$ is spanned by any M vectors, say, \mathbf{m}_i , as long as these vectors are linearly independent.

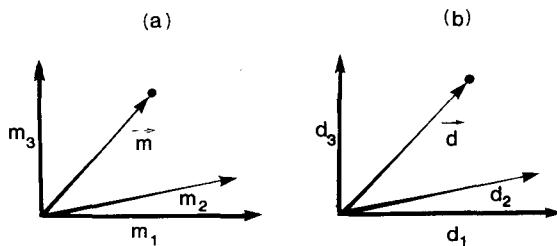


Fig. 7.1. (a) The model parameters represented as a vector \mathbf{m} in the M -dimensional space $S(\mathbf{m})$ of all possible model parameters. (b) The data represented as a vector \mathbf{d} in the N -dimensional space $S(\mathbf{d})$ of all possible data.

An arbitrary vector lying in $S(\mathbf{m})$, say \mathbf{m}^* , can be expressed as a sum of these M basis vectors, written as

$$\mathbf{m}^* = \sum_{i=1}^M \alpha_i \mathbf{m}_i \quad (7.1)$$

where the α 's are the components of the vector \mathbf{m}^* in the new coordinate system. If the \mathbf{m}_i 's are linearly dependent, then the vectors \mathbf{m}_i lie in a subspace, or *hyperplane*, of $S(\mathbf{m})$ and the expansion cannot be made (Fig. 7.2).

We shall consider, therefore, transformations of the coordinate systems of the two spaces $S(\mathbf{m})$ and $S(\mathbf{d})$. If \mathbf{m} is the representation of a vector in one coordinate system and \mathbf{m}' its representation in another,

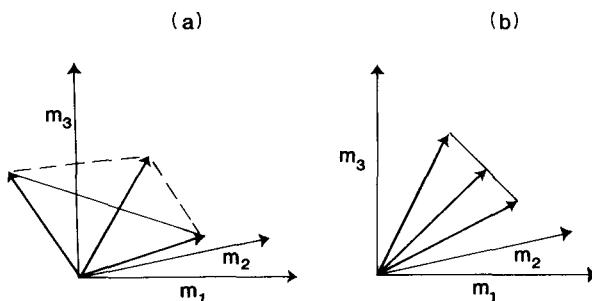


Fig. 7.2. (a) These three vectors span the three-dimensional space $S(\mathbf{m})$. (b) These three vectors do not span the space since they all lie on the same plane.

we can write the transformation as

$$\mathbf{m}' = \mathbf{T}\mathbf{m} \quad \text{and} \quad \mathbf{m} = \mathbf{T}^{-1}\mathbf{m}' \quad (7.2)$$

where \mathbf{T} is the transformation matrix. If the new basis vectors are still unit vectors, then \mathbf{T} represents simple rotations or reflections of the coordinate axes. As we shall show below, however, it is sometimes convenient to choose a new set of basis vectors that are not unit vectors.

7.2 Householder Transformations

In this section we shall show that the minimum length, least squares, and constrained least squares solutions can be found through simple transformations of the equation $\mathbf{G}\mathbf{m} = \mathbf{d}$. While the results are identical to the formulas derived in Chapter 3, the approach and emphasis are quite different and will enable us to visualize these solutions in a new way. We shall begin by considering a purely underdetermined linear problem $\mathbf{G}\mathbf{m} = \mathbf{d}$ with $M > N$. Suppose we want to find the minimum length solution (the one that minimizes $L = \mathbf{m}^T\mathbf{m}$). We shall show that it is easy to find this solution by transforming the model parameters into a new coordinate system $\mathbf{m}' = \mathbf{T}\mathbf{m}$. The inverse problem becomes

$$\mathbf{d} = \mathbf{G}\mathbf{m} = \mathbf{G}\mathbf{I}\mathbf{m} = \mathbf{G}\mathbf{T}^{-1}\mathbf{T}\mathbf{m} = \{\mathbf{G}\mathbf{T}^{-1}\}\{\mathbf{T}\mathbf{m}\} = \mathbf{G}'\mathbf{m}' \quad (7.3)$$

where $\mathbf{G}' = \mathbf{G}\mathbf{T}^{-1}$ is the data kernel in the new coordinate system. The solution length becomes

$$L = \mathbf{m}^T\mathbf{m} = \{\mathbf{T}^{-1}\mathbf{m}'\}^T\{\mathbf{T}^{-1}\mathbf{m}'\} = \mathbf{m}'^T\{\mathbf{T}^{-1T}\mathbf{T}^{-1}\}\mathbf{m}' \quad (7.4)$$

Suppose that we could choose \mathbf{T} so that $\{\mathbf{T}^{-1T}\mathbf{T}^{-1}\} = \mathbf{I}$. The solution length would then have the same form in both coordinate systems, namely, the sum of squares of the vector elements. Minimizing $\mathbf{m}'^T\mathbf{m}'$ would be equivalent to minimizing $\mathbf{m}^T\mathbf{m}$. Transformations of this type that do not change the length of the vector components are called *unitary transformations*. They may be interpreted as rotations and reflections of the coordinate axes. We can see from Eq. (7.4) that unitary transformations satisfy $\mathbf{T}^T = \mathbf{T}^{-1}$.

Now suppose that we could also choose the transformation so that

\mathbf{G}' is the lower triangular matrix

$$\begin{bmatrix} G'_{11} & 0 & 0 & 0 & \cdots & 0 & \cdots & \cdots & 0 \\ G'_{21} & G'_{22} & 0 & 0 & \cdots & 0 & \cdots & \cdots & 0 \\ G'_{31} & G'_{32} & G'_{33} & 0 & \cdots & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ G'_{N1} & G'_{N2} & G'_{N3} & G'_{N4} & \cdots & G'_{NN} & 0 & \cdots & 0 \end{bmatrix} = \begin{bmatrix} m'_1 \\ m'_2 \\ m'_3 \\ \vdots \\ \vdots \\ \vdots \\ m'_M \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ \vdots \\ \vdots \\ \vdots \\ d_N \end{bmatrix} \quad (7.5)$$

Notice that no matter what values we pick for $\{\mathbf{m}'^{\text{est}}, i = N + 1, M\}$, we cannot change the value of $\mathbf{G}'\mathbf{m}'$ since the last $M - N$ columns of \mathbf{G}' are all zero. On the other hand, we can solve for the first N elements of \mathbf{m}'^{est} uniquely as

$$\begin{aligned} m'_1^{\text{est}} &= [d_1]/G'_{11} \\ m'_2^{\text{est}} &= [d_2 - G'_{21}m'_1^{\text{est}}]/G'_{22} \\ m'_3^{\text{est}} &= [d_3 - G'_{31}m'_1^{\text{est}} - G'_{32}m'_2^{\text{est}}]/G'_{33} \\ &\vdots \\ &\vdots \end{aligned} \quad (7.6)$$

This process is known as *back-solving*. Since the first N elements of \mathbf{m}'^{est} are thereby determined, $\mathbf{m}'^T\mathbf{m}'$ can be minimized by setting the remaining m'_i^{est} equal to zero. The solution in the original coordinate system is then $\mathbf{m}^{\text{est}} = \mathbf{T}^{-1}\mathbf{m}'^{\text{est}}$. Since this solution satisfies the data exactly and minimizes the solution length, it is equal to the minimum-length solution (Section 3.7).

One such transformation that can triangularize a matrix is called a *Householder transformation*. We shall discuss in Section 7.3 how such transformations can be found. We employ a transformation process that separates the determined and undetermined linear combinations of model parameters into two distinct groups so that we can deal with them separately. It is interesting to note that we can now easily determine the null vectors for the inverse problem. In the transformed coordinates they are the set of vectors whose first N elements are zero and whose last $M - N$ elements are zero except for one element. There are clearly $M - N$ such vectors, so we have established that there are never more than M null vectors in a purely underdetermined problem.

The null vectors can easily be transformed back into the original coordinate system by premultiplication by T^{-1} . Since all but one element of the transformed null vectors are zero, this operation just selects a column of T^{-1} , (or, equivalently, a row of T).

As we shall see below, Householder transformations have application to a wide variety of methods that employ the L_2 norm as a measure of size. These transformations provide an alternative method of solving such problems and additional insight into their structure. They also have computational advantages that we shall discuss in Section 12.3.

The overdetermined linear inverse problem $\mathbf{G}\mathbf{m} = \mathbf{d}$ with $N > M$ can also be solved through the use of Householder transformations. In this case we seek a solution that minimizes the prediction error $E = \mathbf{e}^T \mathbf{e}$. We seek a transformation with two properties: it must operate on the prediction error \mathbf{e} ($\mathbf{e}' = T\mathbf{e}$) in such a way that minimizing $\mathbf{e}'^T \mathbf{e}'$ is the same as minimizing $\mathbf{e}^T \mathbf{e}$, and it must transform the data kernel into upper-triangularized form. The transformed prediction error is

$$\mathbf{e}' = T\mathbf{e} = T(\mathbf{d} - \mathbf{G}\mathbf{m}) = T\mathbf{d} - TG\mathbf{m} = \mathbf{d}' - \mathbf{G}'\mathbf{m} \quad (7.7)$$

where \mathbf{d}' is the transformed data and \mathbf{G}' is the transformed and triangularized data kernel

$$\begin{bmatrix} e'_1 \\ e'_2 \\ e'_3 \\ \vdots \\ e'_M \\ \vdots \\ e'_N \end{bmatrix} = - \begin{bmatrix} G'_{11} & G'_{12} & G'_{13} & G'_{14} & \cdots & G'_{1M} \\ 0 & G'_{22} & G'_{23} & G'_{24} & \cdots & G'_{2M} \\ 0 & 0 & G'_{33} & G'_{34} & \cdots & G'_{3M} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & G'_{MM} \\ 0 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} m'_1 \\ m'_2 \\ m'_3 \\ \vdots \\ m'_M \\ \vdots \\ m'_N \end{bmatrix} + \begin{bmatrix} d'_1 \\ d'_2 \\ d'_3 \\ \vdots \\ d'_M \\ \vdots \\ d'_N \end{bmatrix} \quad (7.8)$$

We note that no matter what values we choose for \mathbf{m}'^{est} , we cannot alter the last $N - M$ elements of \mathbf{e}' since the last $N - M$ rows of the transformed data kernel are zero. We can, however, set the first M elements of \mathbf{e}' equal to zero by satisfying the first M equations $\mathbf{e}' = \mathbf{d}' - \mathbf{G}'\mathbf{m}' = 0$ exactly. Since the top part of \mathbf{G}' is triangular, we can use the back-solving technique described above. The total error is then the length of the last $N - M$ elements of \mathbf{e}' , written as

$$E = \sum_{i=M+1}^N e_i'^2$$

Again we used Householder transformations to separate the problem into two parts: data that can be satisfied exactly and data that cannot be satisfied at all. The solution is chosen so that it minimizes the length of the prediction error, and the least square solution is thereby obtained.

Finally, we note that the constrained least squares problem can also be solved with Householder transformations. Suppose that we want to solve $\mathbf{G}\mathbf{m} = \mathbf{d}$ in the least squares sense except that we want the solution to obey p linear equality constraints of the form $\mathbf{F}\mathbf{m} = \mathbf{h}$. Because of the constraints, we do not have complete freedom in choosing the model parameters. We therefore employ Householder transformations to separate those linear combinations of \mathbf{m} that are completely determined by the constraints from those that are completely undetermined. This process is precisely the same as the one used in the underdetermined problem and consists of finding a transformation, say, \mathbf{T} , that triangularizes $\mathbf{F}\mathbf{m} = \mathbf{h}$ as

$$\mathbf{h} = \mathbf{F}\mathbf{m} = \mathbf{F}\mathbf{T}^{-1}\mathbf{T}\mathbf{m} = \mathbf{F}'\mathbf{m}'$$

The first p elements of \mathbf{m}'^{est} are now completely determined and can be computed by back-solving the triangular system. The same transformation can be applied to $\mathbf{G}\mathbf{m} = \mathbf{d}$ to yield the transformed inverse problem $\mathbf{G}'\mathbf{m}' = \mathbf{d}$. But \mathbf{G}' will not be triangular since the transformation was designed to triangularize \mathbf{F} , not \mathbf{G} . Since the first p elements of \mathbf{m}'^{est} have been determined by the constraints, we can partition \mathbf{G}' into two submatrices \mathbf{G}'_1 and \mathbf{G}'_2 . The first multiplies the p -determined model parameters and the second multiplies the as yet unknown model parameters.

$$[\mathbf{G}'_1, \mathbf{G}'_2][[m'_1^{\text{est}} \cdots m_p^{\text{est}}], [m'_{p+1} \cdots m'_M]]^T = \mathbf{d} \quad (7.9)$$

The equation can be rearranged into standard form by subtracting the part involving the determined model parameters:

$$\mathbf{G}_2' [m'_{p+1} \dots m'_M]^T = \mathbf{d} - \mathbf{G}_1' [m_1^{\text{est}} \dots m_p^{\text{est}}]^T \quad (7.10)$$

The equation is now a completely overdetermined one in the $M - p$ unknown model parameters and can be solved as described above. Finally, the solution is transformed back into the original coordinate system by $\mathbf{m}^{\text{est}} = \mathbf{T}^{-1} \mathbf{m}'^{\text{est}}$.

7.3 Designing Householder Transformations

For a transformation to preserve length, it must be a unitary transformation (i.e., it must satisfy $\mathbf{T}^T = \mathbf{T}^{-1}$). Any transformation of the form

$$\mathbf{T} = \mathbf{I} - \frac{2\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T \mathbf{v}} \quad (7.11)$$

(where \mathbf{v} is any vector) is a unitary transformation since

$$\mathbf{T}^{-1} \mathbf{T} = \mathbf{T}^T \mathbf{T} = \left[\mathbf{I} - \frac{2\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T \mathbf{v}} \right]^2 = \mathbf{I} - \frac{4\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T \mathbf{v}} + \frac{4\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T \mathbf{v}} = \mathbf{I} \quad (7.12)$$

It turns out that this is the most general form of a unitary transformation. The problem is to find the vector \mathbf{v} such that the transformation triangularizes a given matrix. To do so, we shall begin by finding a sequence of transformations, each of which converts to zeros either the elements beneath the main diagonal of one *column* of the matrix (for premultiplication of the transformation) or the elements to the right of the main diagonal of one *row* of the matrix (for postmultiplication by the transformation). The first i columns are converted to zeros by

$$\mathbf{T} = \mathbf{T}_i \mathbf{T}_{i-1} \mathbf{T}_{i-2} \dots \mathbf{T}_1 \quad (7.13)$$

In the overdetermined problem, applying M of these transformations produces an upper-triangularized matrix. In the $M = 3, N = 4$ case the transformations proceed as

$$\begin{array}{ccccccc}
 \mathbf{G} & \rightarrow & \mathbf{T}_1\mathbf{G} & \rightarrow & \mathbf{T}_2\mathbf{T}_1\mathbf{G} & \rightarrow & \mathbf{T}_3\mathbf{T}_2\mathbf{T}_1\mathbf{G} \\
 \left[\begin{array}{ccc} \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{array} \right] & \rightarrow & \left[\begin{array}{ccc} \times & \times & \times \\ 0 & \times & \times \\ 0 & \times & \times \\ 0 & \times & \times \end{array} \right] & \rightarrow & \left[\begin{array}{ccc} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \\ 0 & 0 & \times \end{array} \right] & \rightarrow & \left[\begin{array}{ccc} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \\ 0 & 0 & 0 \end{array} \right]
 \end{array} \quad (7.14)$$

where the \times 's symbolize nonzero matrix elements. Consider the i th column of \mathbf{G} , denoted by the vector $\mathbf{g} = [G_{1i}, G_{2i}, \dots, G_{Ni}]^T$. We want to construct a transformation such that

$$\mathbf{g}' = \mathbf{T}_i\mathbf{g} = [G'_{1i}, G'_{2i}, \dots, G'_{ii}, 0, 0, \dots, 0]^T \quad (7.15)$$

Substituting in the expression for the transformation yields

$$\mathbf{I}\mathbf{g} - \frac{2\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}}\mathbf{g} = \left[\begin{array}{c} G_{1i} \\ G_{2i} \\ \vdots \\ \vdots \\ G_{Ni} \end{array} \right] - \frac{2\mathbf{v}^T\mathbf{g}}{\mathbf{v}^T\mathbf{v}} \left[\begin{array}{c} v_1 \\ v_2 \\ \vdots \\ \vdots \\ v_N \end{array} \right] = \left[\begin{array}{c} G'_{1i} \\ G'_{2i} \\ \vdots \\ \vdots \\ G_{ii} \\ 0 \\ \vdots \\ \vdots \\ 0 \end{array} \right] \quad (7.16)$$

Since the term $2\mathbf{v}^T\mathbf{g}/\mathbf{v}^T\mathbf{v}$ is a scalar, we can only zero the last $N - i$ elements of \mathbf{g}' if $[2\mathbf{v}^T\mathbf{g}/\mathbf{v}^T\mathbf{v}][v_{i+1}, \dots, v_N] = [G_{i+1,i}, \dots, G_{Ni}]$. We therefore set $[v_{i+1}, \dots, v_N] = [G_{i+1,i}, \dots, G_{Ni}]$ and choose the other i elements of \mathbf{v} so that the normalization is correct. Since this is but a single constraint on i elements of \mathbf{v} , we have considerable flexibility in making the choice. It is convenient to choose the first $i - 1$ elements of \mathbf{v} as zero (this choice is both simple and causes the transformation to leave the first $i - 1$ elements of \mathbf{g} unchanged). This leaves the i th element of \mathbf{v} to be determined by the constraint. It is easy to show that $v_i = G_{ii} - \alpha$, where

$$\alpha^2 = \sum_{j=1}^N G_{ji}^2 \quad (7.17)$$

(Although the sign of α is arbitrary, we shall show in Chapter 12 that one choice is usually better than the other in terms of the numerical stability of computer algorithms.) The Householder transformation is then defined by the vector as

$$\mathbf{v} = [0, 0, \dots, 0, G_{ii} - \alpha, G_{i+1,i}, G_{i+2,i}, \dots, G_{Ni}]^T \quad (7.18)$$

Finally, we note that the $(i + 1)$ st Householder transformation does not destroy any of the zeros created by the i th, as long as we apply them in order of decreasing number of zeros. We can thus apply a succession of these transformations to triangularize an arbitrary matrix. The inverse transformations are also trivial to construct since they are just the transforms of the forward transformations.

7.4 Transformations That Do Not Preserve Length

Suppose we want to solve the linear inverse problem $\mathbf{Gm} = \mathbf{d}$ in the sense of finding a solution \mathbf{m}^{est} that minimizes a weighted combination of prediction error and solution simplicity as

$$\text{Minimize } E + L = \mathbf{e}^T \mathbf{W}_e \mathbf{e} + \mathbf{m}^T \mathbf{W}_m \mathbf{m} \quad (7.19)$$

It is possible to find transformations $\mathbf{m}' = \mathbf{T}_m \mathbf{m}$ and $\mathbf{e}' = \mathbf{T}_e \mathbf{e}$ which, although they do not preserve length, nevertheless change it in precisely such a way that $E + L = \mathbf{e}'^T \mathbf{e}' + \mathbf{m}'^T \mathbf{m}'$ [Ref. 20]. The weighting factors are identity matrices in the new coordinate system.

Consider the weighted measure of length $L = \mathbf{m}^T \mathbf{W}_m \mathbf{m}$. If we could factor the weighting matrix into a product of two equal parts, then we could identify each part with a transformation of coordinates as

$$L = \mathbf{m}^T \mathbf{W}_m \mathbf{m} = \mathbf{m}^T \{\mathbf{T}_m^T \mathbf{T}_m\} \mathbf{m} = \{\mathbf{T}_m \mathbf{m}\}^T \{\mathbf{T}_m \mathbf{m}\} = \mathbf{m}'^T \mathbf{m}' \quad (7.20)$$

This factorization can be accomplished by forming the eigenvalue decomposition of a matrix. Let $\mathbf{\Lambda}_m$ and \mathbf{U}_m be the matrices of eigenvalues and eigenvectors of \mathbf{W}_m , respectively. Then

$$\begin{aligned} \mathbf{W}_m &= \mathbf{U}_m \mathbf{\Lambda}_m \mathbf{U}_m^T = \{\mathbf{U}_m \mathbf{\Lambda}_m^{1/2}\} \{\mathbf{\Lambda}_m^{1/2} \mathbf{U}_m^T\} \\ &= \{\mathbf{\Lambda}_m^{1/2} \mathbf{U}_m^T\}^T \{\mathbf{\Lambda}_m^{1/2} \mathbf{U}_m^T\} = \mathbf{T}_m^T \mathbf{T}_m \end{aligned} \quad (7.21)$$

A similar transformation can be found for the \mathbf{W}_e . The transformed

inverse problem is then $\mathbf{G}'\mathbf{m}' = \mathbf{d}'$, where

$$\begin{aligned}\mathbf{m}' &= (\Lambda_m^{1/2} \mathbf{U}_m^T) \mathbf{m} && \text{and} & \mathbf{m} &= (\mathbf{U}_m \Lambda_m^{-1/2}) \mathbf{m}' \\ \mathbf{d}' &= (\Lambda_e^{1/2} \mathbf{U}_e^T) \mathbf{d} && \text{and} & \mathbf{d} &= (\mathbf{U}_e \Lambda_e^{-1/2}) \mathbf{d}' \\ \mathbf{G}' &= (\Lambda_e^{1/2} \mathbf{U}_e^T) \mathbf{G} (\mathbf{U}_m \Lambda_m^{-1/2}) && \text{and} & \mathbf{G} &= (\mathbf{U}_e \Lambda_e^{-1/2}) \mathbf{G}' (\Lambda_m^{1/2} \mathbf{U}_m)\end{aligned}\tag{7.22}$$

If it is often convenient to transform weighted L_2 problems into this form before proceeding with their solution.

7.5 The Solution of the Mixed-Determined Problem

The concept of vector spaces is particularly helpful in understanding the mixed-determined problem, in which some linear combinations of the model parameters are overdetermined and some are underdetermined.

If the problem is to some degree underdetermined, then the equation $\mathbf{G}\mathbf{m} = \mathbf{d}$ contains information about only some of the model parameters. We can think of these combinations as lying in a subspace $S_p(\mathbf{m})$ of the model parameters space. No information is provided about the part of the solution that lies in the rest of the space, which we shall call the null space $S_0(\mathbf{m})$. The part of the \mathbf{m} that lies in the null space is completely “unilluminated” by the $\mathbf{G}\mathbf{m} = \mathbf{d}$, since the equation contains no information about these linear combinations of the model parameters.

On the other hand, if the problem is to some degree overdetermined, the product $\mathbf{G}\mathbf{m}$ may not be able to span $S(\mathbf{d})$ no matter what one chooses for \mathbf{m} . At best $\mathbf{G}\mathbf{m}$ may span a subspace $S_p(\mathbf{d})$ of the data space. Then no part of the data lying outside of this space, say, in $S_0(\mathbf{d})$, can be satisfied for any choice of the model parameters.

If the model parameters and data are divided into parts with subscript p that lie in the p spaces and parts with subscript 0 that lie in the null spaces, we can write $\mathbf{G}\mathbf{m} = \mathbf{d}$ as

$$\mathbf{G}[\mathbf{m}_p + \mathbf{m}_0] = [\mathbf{d}_p + \mathbf{d}_0]\tag{7.23}$$

The solution length is then

$$L = \mathbf{m}^T \mathbf{m} = [\mathbf{m}_p + \mathbf{m}_0]^T [\mathbf{m}_p + \mathbf{m}_0] = \mathbf{m}_p^T \mathbf{m}_p + \mathbf{m}_0^T \mathbf{m}_0\tag{7.24}$$

(The cross terms $\mathbf{m}_p^T \mathbf{m}_0$ and $\mathbf{m}_0^T \mathbf{m}_p$ are zero since the vectors lie in different spaces.) The prediction error is

$$\begin{aligned} E &= [\mathbf{d}_p + \mathbf{d}_0 - \mathbf{G}\mathbf{m}_p]^T [\mathbf{d}_p + \mathbf{d}_0 - \mathbf{G}\mathbf{m}_p] \\ &= [\mathbf{d}_p - \mathbf{G}\mathbf{m}_p]^T [\mathbf{d}_p - \mathbf{G}\mathbf{m}_p] + \mathbf{d}_0^T \mathbf{d}_0 \end{aligned} \quad (7.25)$$

(since $\mathbf{G}\mathbf{m}_0 = 0$ and \mathbf{d}_p and \mathbf{d}_0 lie in different spaces). We can now define precisely what we mean by a solution to the mixed-determined problem that minimizes prediction error while adding a minimum of a priori information: *a priori information is added to specify only those linear combinations of the model parameters that reside in the null space $S_0(\mathbf{m})$, and the prediction error is reduced to only the portion in the null space $S_0(\mathbf{d})$ by satisfying $\mathbf{e}_p = [\mathbf{d}_p - \mathbf{G}\mathbf{m}_p] = 0$ exactly.* One possible choice of a priori information is $\mathbf{m}_0^{\text{est}} = 0$, which is sometimes called the “natural solution” of the mixed-determined problem. We note that when $\mathbf{G}\mathbf{m} = \mathbf{d}$ is purely underdetermined the natural solution is just the minimum length solution, and when $\mathbf{G}\mathbf{m} = \mathbf{d}$ is purely overdetermined it is just the least squares solution.

7.6 Singular-Value Decomposition and the Natural Generalized Inverse

The p and null subspaces of the linear problem can easily be identified through a type of eigenvalue decomposition of the data kernel that is sometimes called *spectral decomposition* or *singular-value decomposition*. We shall derive this decomposition in Section 7.7, but first we shall state the result and demonstrate its usefulness.

Any $N \times M$ square matrix can be written as the product of three matrices [Refs. 13,16]:

$$\mathbf{G} = \mathbf{U}\Lambda\mathbf{V}^T \quad (7.26)$$

The matrix \mathbf{U} is an $N \times N$ matrix of eigenvectors that span the data space $S(\mathbf{d})$:

$$\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \dots, \mathbf{u}_N] \quad (7.27)$$

where the \mathbf{u}_i 's are the individual vectors. The vectors are orthogonal to one another and can be chosen to be of unit length, so that $\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = \mathbf{I}$ (where the identity matrix is $N \times N$). Similarly, \mathbf{V} is an $M \times M$ matrix of eigenvectors that span the model parameter space

$S(\mathbf{m})$ as

$$\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \dots, \mathbf{v}_M] \quad (7.28)$$

Here the \mathbf{v}_i 's are the individual orthonormal vectors, so that $\mathbf{V}\mathbf{V}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}$ (the identity matrix being $M \times M$). The matrix Λ is an $N \times M$ diagonal eigenvalue matrix whose diagonal elements are non-negative and are called *singular values*. In the $N = 4, M = 3$ case

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \\ 0 & 0 & 0 \end{bmatrix} \quad (7.29)$$

The singular values are usually arranged in order of decreasing size. Some of the singular values may be zero. We therefore partition Λ into a submatrix Λ_p of p nonzero singular values and several zero matrices as

$$\Lambda = \begin{bmatrix} \Lambda_p & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \quad (7.30)$$

where Λ_p is a $p \times p$ diagonal matrix. The decomposition then becomes $\mathbf{G} = \mathbf{U}\Lambda\mathbf{V}^T = \mathbf{U}_p\Lambda_p\mathbf{V}_p^T$, where \mathbf{U}_p and \mathbf{V}_p consist of the first p columns of \mathbf{U} and \mathbf{V} , respectively. The other portions of the eigenvector matrices are canceled by the zeros in Λ . The matrix \mathbf{G} contains no information about the subspaces spanned by these portions of the data and model eigenvectors, which we shall call \mathbf{V}_0 and \mathbf{U}_0 , respectively. As we shall soon prove, these are precisely the same spaces as the p and null spaces defined in the previous section.

The data kernel is not a function of the null eigenvectors \mathbf{V}_0 and \mathbf{U}_0 . The equation $\mathbf{G}\mathbf{m} = \mathbf{d} = \mathbf{U}_p\Lambda_p\mathbf{V}_p^T\mathbf{m}$ contains no information about the part of the model parameters in the space spanned by \mathbf{V}_0 since the model parameters \mathbf{m} are multiplied by \mathbf{V}_p (which is orthogonal to everything in \mathbf{V}_0). The eigenvector \mathbf{V}_p , therefore, lies completely in $S_p(\mathbf{m})$, and \mathbf{V}_0 lies completely in $S_0(\mathbf{m})$. Similarly, no matter what value $\{\Lambda_p\mathbf{V}_p^T\mathbf{m}\}$ attains, it can have no component in the space spanned by \mathbf{U}_0 since it is multiplied by \mathbf{U}_p (and \mathbf{U}_0 and \mathbf{U}_p are orthogonal). Therefore, \mathbf{U}_p lies completely in $S_p(\mathbf{d})$ and \mathbf{U}_0 lies completely in $S_0(\mathbf{d})$.

We have demonstrated that the p and null spaces can be identified through the singular-value decomposition of the data kernel. The full

spaces $S(\mathbf{m})$ and $S(\mathbf{d})$ are spanned by \mathbf{V} and \mathbf{U} , respectively. The p spaces are spanned by the parts of the eigenvector matrices that have nonzero eigenvalues: $S_p(\mathbf{m})$ is spanned by \mathbf{V}_p and $S_p(\mathbf{d})$ is spanned by \mathbf{U}_p . The remaining eigenvectors \mathbf{V}_0 and \mathbf{U}_0 span the null spaces $S_0(\mathbf{m})$ and $S_0(\mathbf{d})$. The p and null matrices are orthogonal and are normalized in the sense that $\mathbf{V}_p^T \mathbf{V}_p = \mathbf{U}_p^T \mathbf{U}_p = \mathbf{I}$, where \mathbf{I} is $p \times p$ in size. However, since these matrices do not in general span the complete data and model spaces, $\mathbf{V}_p \mathbf{V}_p^T$ and $\mathbf{U}_p \mathbf{U}_p^T$ are not in general identity matrices.

The natural solution to the inverse problem can be constructed from the singular-value decomposition. This solution must have an \mathbf{m}^{est} that has no component in $S_0(\mathbf{m})$ and a prediction error \mathbf{e} that has no component in $S_p(\mathbf{d})$. We therefore consider the solution

$$\mathbf{m}^{\text{est}} = \mathbf{V}_p \mathbf{\Lambda}_p^{-1} \mathbf{U}_p^T \mathbf{d} \quad (7.31)$$

which is picked in analogy to the square matrix case. To show that \mathbf{m}^{est} has no component in $S_0(\mathbf{m})$, we take the dot product of the equation with \mathbf{V}_0 , which lies completely in $S_0(\mathbf{m})$, as

$$\mathbf{V}_0^T \mathbf{m}^{\text{est}} = \mathbf{V}_0^T \mathbf{V}_p \mathbf{\Lambda}_p^{-1} \mathbf{U}_p^T \mathbf{d} = 0 \quad (7.32)$$

since \mathbf{V}_0^T and \mathbf{V}_p are orthogonal. To show that \mathbf{e} has no component in $S_p(\mathbf{d})$, we take the dot product with \mathbf{U}_p as

$$\begin{aligned} \mathbf{U}_p^T \mathbf{e} &= \mathbf{U}_p^T [\mathbf{d} - \mathbf{G} \mathbf{m}^{\text{est}}] = \mathbf{U}_p^T [\mathbf{d} - \mathbf{U}_p \mathbf{\Lambda}_p \mathbf{V}_p^T \mathbf{V}_p \mathbf{\Lambda}_p^{-1} \mathbf{U}_p^T \mathbf{d}] \\ &= \mathbf{U}_p^T [\mathbf{d} - \mathbf{U}_p \mathbf{U}_p^T \mathbf{d}] = -\mathbf{U}_p^T \mathbf{d} + \mathbf{U}_p^T \mathbf{d} = 0 \end{aligned} \quad (7.33)$$

The natural solution of the inverse problem is therefore shown to be

$$\mathbf{m}^{\text{est}} = \mathbf{V}_p \mathbf{\Lambda}_p^{-1} \mathbf{U}_p^T \mathbf{d} \quad (7.34)$$

We note that we can define a generalized inverse operator for the mixed-determined problem, the *natural generalized inverse* $\mathbf{G}^{-g} = \mathbf{V}_p \mathbf{\Lambda}_p^{-1} \mathbf{U}_p^T$. (This generalized inverse is so useful that it is sometimes referred to as *the* generalized inverse, although of course there are other generalized inverses that can be designed for the mixed-determined problem that embody other kinds of a priori information.) The natural generalized inverse has model resolution

$$\mathbf{R} = \mathbf{G}^{-g} \mathbf{G} = \{\mathbf{V}_p \mathbf{\Lambda}_p^{-1} \mathbf{U}_p^T\} \{\mathbf{U}_p \mathbf{\Lambda}_p \mathbf{V}_p^T\} = \mathbf{V}_p \mathbf{V}_p^T \quad (7.35)$$

The model parameters will be perfectly resolved only if \mathbf{V}_p spans the complete space of model parameters, that is, if there are no zero

eigenvalues and $p \geq M$. The data resolution matrix is

$$\mathbf{N} = \mathbf{G}\mathbf{G}^{-g} = \{\mathbf{U}_p\Lambda_p\mathbf{V}_p^T\}\{\mathbf{V}_p\Lambda_p^{-1}\mathbf{U}_p^T\} = \mathbf{U}_p\mathbf{U}_p^T \quad (7.36)$$

The data are only perfectly resolved if \mathbf{U}_p spans the complete space of data and $p = N$. Finally, we note that if the data are uncorrelated with uniform variance σ_d^2 , the model covariance is

$$\begin{aligned} [\text{cov } \mathbf{m}^{\text{est}}] &= \mathbf{G}^{-g}[\text{cov } \mathbf{d}]\mathbf{G}^{-gT} = \sigma_d^2(\mathbf{V}_p\Lambda_p^{-1}\mathbf{U}_p^T)(\mathbf{V}_p\Lambda_p^{-1}\mathbf{U}_p^T)^T \\ &= \sigma_d^2\mathbf{V}_p\Lambda_p^{-2}\mathbf{V}_p^T \end{aligned} \quad (7.37)$$

The covariance of the estimated model parameters is very sensitive to the smallest nonzero eigenvalue. (Note that forming the natural inverse corresponds to assuming that linear combinations of the a priori model parameters in the p space have infinite variance and that combinations in the null space have zero variance and zero mean.) The covariance of the estimated model parameters, therefore, does not explicitly contain $[\text{cov } \mathbf{m}]$. If one prefers a solution based on the natural inverse (but with the null vectors chosen to minimize the distance to a set of a priori model parameters with mean $\langle \mathbf{m} \rangle$ and covariance $[\text{cov } \mathbf{m}]$), it is appropriate to use the formula $\mathbf{m}^{\text{est}} = \mathbf{G}^{-g}\mathbf{d} + [\mathbf{I} - \mathbf{R}]\langle \mathbf{m} \rangle$, where \mathbf{G}^{-g} is the natural inverse. The covariance of this estimate is now

$$[\text{cov } \mathbf{m}^{\text{est}}] = \mathbf{G}^{-g}[\text{cov } \mathbf{d}]\mathbf{G}^{-gT} + [\mathbf{I} - \mathbf{R}][\text{cov } \mathbf{m}][\mathbf{I} - \mathbf{R}]^T$$

which is based on the usual rule for computing covariances.

To use the natural inverse one must be able to identify the number p , that is, to count the number of nonzero singular values. Plots of the sizes of the singular values against their index numbers (the *spectrum* of the data kernel) can be useful in this process. The value of p can be easily determined if the singular values fall into two clearly distinguishable groups, one nonzero and one zero (Fig. 7.3a). In realistic inverse problems, however, the situation illustrated by Fig. 7.3b is more typical. The singular values smoothly decline in size, making it hard to distinguish ones that are actually nonzero from ones that are zero but computed somewhat inaccurately owing to round-off error by the computer. Furthermore, if one chooses p so as to include these very small singular values, the solution variance will be very large since it is proportional to Λ_p^{-2} . One solution to this problem is to pick some cutoff size for singular values and then consider any values smaller than this as equal to zero. This process artificially reduces the dimensions of \mathbf{V}_p and \mathbf{U}_p that are included in the generalized inverse. The

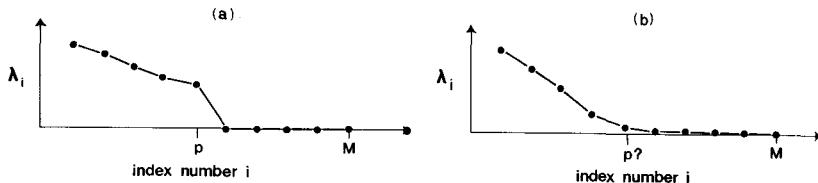


Fig. 7.3. (a) Singular values of a matrix with clearly identifiable p . (b) Singular values of a matrix where p must be chosen in a more arbitrary fashion.

resulting estimates of the model parameters are no longer exactly the natural solution. But, if only small singular values are excluded, the solution is generally close to the natural solution and possesses better variance. On the other hand, its model and data resolution are worse. We recognize that this trade-off is just another manifestation of the trade-off between resolution and variance discussed in Chapter 4.

Instead of choosing a sharp cutoff for the singular values, it is possible to include all the singular values while damping the smaller ones. We let $p = M$ but replace the reciprocals of all the singular values by $1/(\epsilon^2 + \lambda_i)$, where ϵ is some small number. This change has little effect on the larger eigenvalues but prevents the smaller ones from leading to large variances. Of course, the solution is no longer the natural solution. While its variance is improved, its model and data resolution are degraded. In fact, this solution is precisely the damped least squares solution discussed in Chapter 3. The damping of the singular values corresponds to the addition of a priori information that the model parameters are small. The precise value of the number used as the cutoff or damping parameter must be chosen by a trial-and-error process which weighs the relative merits of having a 'solution with small variance against those of having one that fits the data and is well resolved.

In Section 6.6 we discussed the problem of bounding nonunique averages of model parameters by incorporating a priori inequality constraints into the solution of the inverse problem. We see that the singular-value decomposition provides a simple way of identifying the null vectors of $\mathbf{Gm} = \mathbf{d}$. The general solution to the inverse problem [Ref. 21]

$$\mathbf{m}^{\text{gen}} = \mathbf{m}^{\text{par}} + \sum_{i=1}^q \alpha_i \mathbf{m}_i^{\text{null}} \quad (7.38)$$

can be thought of as having the natural solution as its particular

solution and a sum over the null eigenvectors as its null solution:

$$\mathbf{m}^{\text{gen}} = \mathbf{V}_p \Lambda_p^{-1} \mathbf{U}_p^T \mathbf{d} + \mathbf{V}_0 \boldsymbol{\alpha} \quad (7.39)$$

There are $q = M - p$ null vectors in the general solution, each with coefficients given by $\boldsymbol{\alpha}$. In Section 6.6 upper and lower bounds on localized averages of the solution were found by determining the $\boldsymbol{\alpha}$ that maximizes (for the upper bound) or minimizes (for the lower bound) $\langle m \rangle = \mathbf{a}^T \mathbf{m}$ with the constraint that $\mathbf{m}^l \leq \mathbf{m} \leq \mathbf{m}^u$, where \mathbf{m}^l and \mathbf{m}^u are a priori bounds. The use of the natural solution guarantees that the prediction error is minimized in the L_2 sense.

The bounds on the localized average $\langle m \rangle = \mathbf{a}^T \mathbf{m}$ should be treated with some skepticism, since they depend on a particular choice of the solution (in this case one that minimizes the L_2 prediction error). If the total error E increases only slightly as this solution is perturbed and if this additional error can be considered negligible, then the true bounds of the localized average will be larger than those given above. In principle, one can handle this problem by forsaking the eigenvalue decomposition and simply determining the \mathbf{m} that extremizes $\langle m \rangle$ with the constraints that $\mathbf{m}^l \leq \mathbf{m} \leq \mathbf{m}^u$ and that the total prediction error is less than some tolerable amount, say, E_M . For L_2 measures of the prediction error, this is a very difficult nonlinear problem. However, if the error is measured under the L_1 norm, it can be transformed into a linear programming problem.

7.7 Derivation of the Singular-Value Decomposition

We first form an $(N + M) \times (N + M)$ square symmetric matrix \mathbf{S} from \mathbf{G} and \mathbf{G}^T as [Ref. 13]

$$\mathbf{S} = \begin{bmatrix} \mathbf{0} & \mathbf{G} \\ \mathbf{G}^T & \mathbf{0} \end{bmatrix} \quad (7.40)$$

From elementary linear algebra we know that this matrix has $N + M$ real eigenvalues λ_i and a complete set of eigenvectors \mathbf{w}_i which solve $\mathbf{Sw}_i = \lambda_i \mathbf{w}_i$. Partitioning \mathbf{w} into a part \mathbf{u} of length N and a part \mathbf{v} of length M , we obtain

$$\mathbf{Sw} = \begin{bmatrix} \mathbf{0} & \mathbf{G} \\ \mathbf{G}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{u}_i \\ \mathbf{v}_i \end{bmatrix} = \lambda_i \begin{bmatrix} \mathbf{u}_i \\ \mathbf{v}_i \end{bmatrix} \quad (7.41)$$

We shall now show that \mathbf{u}_i and \mathbf{v}_i are the same vectors as those defined in the previous section. We first note that the above equation implies that $\mathbf{G}\mathbf{v}_i = \lambda_i \mathbf{u}_i$ and $\mathbf{G}^T \mathbf{u}_i = \lambda_i \mathbf{v}_i$. Suppose that there is a positive eigenvalue λ_i with eigenvector $[\mathbf{u}_i, \mathbf{v}_i]^T$. Then we note that $-\lambda_i$ is also an eigenvalue with eigenvector $[-\mathbf{u}_i, \mathbf{v}_i]^T$. If there are p positive eigenvalues, then there are $N + M - 2p$ zero eigenvalues. Now by manipulating the above equations we obtain

$$\mathbf{G}^T \mathbf{G} \mathbf{v}_i = \lambda_i^2 \mathbf{v}_i \quad \text{and} \quad \mathbf{G} \mathbf{G}^T \mathbf{u}_i = \lambda_i^2 \mathbf{u}_i \quad (7.42)$$

Since a symmetric matrix can have no more distinct eigenvectors than its dimension, we note that $p \leq \min(N, M)$. Since both matrices are square and symmetric, there are M vectors \mathbf{v}_i that form a complete orthogonal set \mathbf{V} spanning $S(\mathbf{m})$ and N vectors \mathbf{u}_i that form a complete orthogonal set \mathbf{U} spanning $S(\mathbf{d})$. These include p of the \mathbf{w} eigenvectors with distinct nonzero eigenvalues and remaining ones chosen from the eigenvectors with zero eigenvalues. The equation $\mathbf{G}\mathbf{v}_i = \lambda_i \mathbf{u}_i$ can be written in matrix form as $\mathbf{GV} = \mathbf{U}\Lambda$, where Λ is a diagonal matrix of the eigenvalues. Post-multiplying by \mathbf{V}^T gives the singular-value decomposition $\mathbf{G} = \mathbf{U}\Lambda\mathbf{V}^T$.

7.8 Simplifying Linear Equality and Inequality Constraints

The singular-value decomposition can be helpful in simplifying linear constraints.

7.8.1 LINEAR EQUALITY CONSTRAINTS

Consider the problem of solving $\mathbf{G}\mathbf{m} = \mathbf{d}$ in the sense of finding a solution that minimizes the L_2 prediction error subject to the $s < M$ constraints that $\mathbf{F}\mathbf{m} = \mathbf{h}$. This problem can be reduced to the unconstrained problem $\mathbf{G}'\mathbf{m}' = \mathbf{d}'$ in $M' \leq M$ new model parameters. We first find the singular-value decomposition of the constraint matrix $\mathbf{F} = \mathbf{U}_p \Lambda_p \mathbf{V}_p^T$. If $p = s$, the constraints are consistent and determine p linear combinations of the unknowns. The general solution is then $\mathbf{m} = \mathbf{V}_p \Lambda_p^{-1} \mathbf{U}_p^T \mathbf{h} + \mathbf{V}_0 \boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is an arbitrary vector of length $M - p$ and is to be determined by minimizing the prediction error. Substitut-

ing this equation for \mathbf{m} into $\mathbf{Gm} = \mathbf{d}$ and rearranging terms yields

$$\mathbf{V}_p \boldsymbol{\alpha} = \mathbf{d} - \mathbf{GV}_p \boldsymbol{\Lambda}_p^{-1} \mathbf{U}_p^T \mathbf{h} \quad (7.43)$$

which is of the desired form and can be solved as an unconstrained least squares problem. We note that we have encountered this problem in a somewhat different form during the discussion of Householder transformations (Section 7.2). The main advantage of using the singular-value decomposition is that it provides a test of the constraints's consistency.

7.8.2 LINEAR INEQUALITY CONSTRAINTS

Consider the L_2 problem

$$\text{Minimize } \|\mathbf{d} - \mathbf{Gm}\|_2 \quad \text{subject to } \mathbf{Fm} \geq \mathbf{h} \quad (7.44)$$

We shall show that as long as $\mathbf{Gm} = \mathbf{d}$ is in fact overdetermined, this problem can be reduced to the simpler problem [Ref. 14]:

$$\text{Minimize } \|\mathbf{m}'\|_2 \quad \text{subject to } \mathbf{F'm'} \geq \mathbf{h'} \quad (7.45)$$

To demonstrate this transformation, we form the singular-value decomposition of the data kernel $\mathbf{G} = \mathbf{U}_p \boldsymbol{\Lambda}_p \mathbf{V}_p^T$. The prediction error is then

$$\begin{aligned} E &= \left\| \begin{bmatrix} \mathbf{U}_p^T \mathbf{d} \\ \mathbf{U}_0^T \mathbf{d} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\Lambda}_p \mathbf{V}_p^T \mathbf{m} \\ 0 \end{bmatrix} \right\|_2 \\ &= \|\mathbf{d}_p - \boldsymbol{\Lambda}_p \mathbf{V}_p^T \mathbf{m}\|_2 + \|\mathbf{d}_0\|_2 = \|\mathbf{m}'\|_2 + \|\mathbf{d}_0\|_2 \end{aligned} \quad (7.46)$$

where $\mathbf{m}' = \mathbf{d}_p - \boldsymbol{\Lambda}_p \mathbf{V}_p^T \mathbf{m}$. We note that minimizing $\|\mathbf{m}'\|_2$ is the same as minimizing E since the other term is a constant. Inverting this expression for the unprimed model parameters gives $\mathbf{m} = \mathbf{V}_p \boldsymbol{\Lambda}_p^{-1} [\mathbf{d}_p - \mathbf{m}']$. Substituting this expression into the constraint equation and rearranging terms yields

$$\{-\mathbf{FV}_p \boldsymbol{\Lambda}_p^{-1}\} \mathbf{m}' \geq \{\mathbf{h} + \mathbf{V}_p \boldsymbol{\Lambda}_p^{-1} \mathbf{d}_p\} \quad (7.47)$$

which is in the desired form.

7.9 Inequality Constraints

We shall now consider the solution of L_2 minimization problems with inequality constraints of the form

$$\text{Minimize } \|\mathbf{d} - \mathbf{Gm}\|_2 \quad \text{subject to } \mathbf{Fm} \geq \mathbf{h} \quad (7.48)$$

We first note that problems involving $=$ and \leq constraints can be reduced to this form. Equality constraints can be removed by the transformation described in Section 7.9.1, and \leq constraints can be removed by multiplication by -1 to change them into \geq constraints. For this minimization problem to have any solution at all, the constraints $\mathbf{F}\mathbf{m} \geq \mathbf{h}$ must be consistent; there must be at least one \mathbf{m} that satisfies all the constraints. We can view these constraints as defining a volume in $S(\mathbf{m})$. Each constraint defines a hyperplane that divides $S(\mathbf{m})$ into two halfspaces, one in which that constraint is satisfied (the *feasible halfspace*) and one in which it is violated (the *infeasible halfspace*). The set of inequality constraints, therefore, defines a volume in $S(\mathbf{m})$ which might be zero, finite, or infinite in extent. If the region has zero volume, then it is clear that no feasible solution exists (Fig. 7.4b). If it has nonzero volume, then there is at least one solution that minimizes the prediction error (Fig. 7.4a). This volume has the shape of a polyhedron since its boundary surfaces are planes. It can be shown that the polyhedron must be convex: it can have no reentrant angles or grooves.

The starting point for solving the L_2 minimization problem with inequality constraints is the Kuhn–Tucker theorem, which describes the properties that any solution to this problem must possess. For any \mathbf{m} that minimizes $\|\mathbf{d} - \mathbf{G}\mathbf{m}\|_2$ subject to p constraints $\mathbf{F}\mathbf{m} \geq \mathbf{h}$, it is possible to find a vector \mathbf{y} of length p such that

$$\begin{aligned}-\nabla[\mathbf{F}\mathbf{m}]\mathbf{y} &= -\nabla E \\ -\mathbf{F}^T\mathbf{y} &= \mathbf{G}^T[\mathbf{d} - \mathbf{G}\mathbf{m}]\end{aligned}\quad (7.49)$$

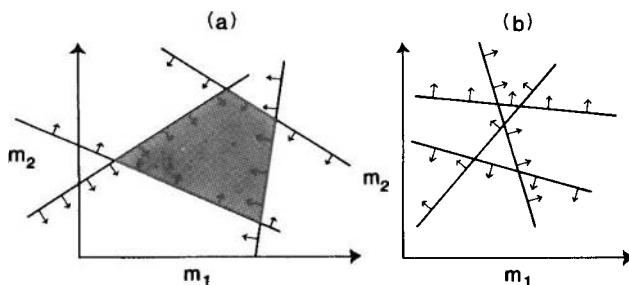


Fig. 7.4. Each linear inequality constraint divides $S(\mathbf{m})$ into two halfspaces, one feasible, the other infeasible (arrows point into feasible halfspace). (a) Consistent constraints form convex polyhedral volume (shaded) of feasible \mathbf{m} . (b) Inconsistent constraints have no feasible volume.

Here \mathbf{y} can be partitioned into two parts \mathbf{y}_E and \mathbf{y}_S (possibly requiring reordering of the constraints) that satisfy

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_E > 0 \\ \mathbf{y}_S = 0 \end{bmatrix} \quad \text{and} \quad \begin{aligned} \mathbf{F}_E \mathbf{m} - \mathbf{h}_E &= 0 \\ \mathbf{F}_S \mathbf{m} - \mathbf{h}_S &> 0 \end{aligned} \quad (7.50)$$

The first group of equality–inequality constraints are satisfied in the equality sense (thus the subscript E for “equality”). The rest are satisfied more loosely in the inequality sense (thus the subscript S for “slack”).

The theorem states that any feasible solution \mathbf{m} is the minimum solution only if the direction in which one would have to perturb \mathbf{m} to decrease the total error E causes the solution to cross some constraint hyperplane and become infeasible. The direction of decreasing error is $-\nabla E = \mathbf{G}^T[\mathbf{d} - \mathbf{G}\mathbf{m}]$. The constraint hyperplanes have normals $+\nabla[\mathbf{F}\mathbf{m}] = \mathbf{F}^T$ which point into the feasible side. Since $\mathbf{F}_E \mathbf{m} - \mathbf{h}_E = 0$, the solution lies exactly on the bounding hyperplanes of the \mathbf{F}_E constraints but within the feasible volume of the \mathbf{F}_S constraints. An infinitesimal perturbation $\delta\mathbf{m}$ of the solution can, therefore, only violate the \mathbf{F}_E constraints. If it is not to violate these constraints, the perturbation must be made in the direction of feasibility, so that it must be expressible as a nonnegative combination of hyperplane normals $\delta\mathbf{m} \cdot \nabla[\mathbf{F}\mathbf{m}] \geq 0$. On the other hand, if it is to decrease the total prediction error it must satisfy $\delta\mathbf{m} \cdot \nabla E \leq 0$. For solutions that satisfy the Kuhn–Tucker theorem these two conditions are incompatible and $\delta\mathbf{m} \cdot \nabla E = \delta\mathbf{m} \cdot \nabla[\mathbf{F}\mathbf{m}] \cdot \mathbf{y} \geq 0$ since both $\delta\mathbf{m} \cdot \nabla[\mathbf{F}\mathbf{m}]$ and \mathbf{y} are positive. These solutions are indeed minimum solutions to the constrained problem (Fig. 7.5).

To demonstrate how the Kuhn–Tucker theorem can be used, we consider the simplified problem

$$\text{Minimize } \|\mathbf{d} - \mathbf{G}\mathbf{m}\|_2 \quad \text{subject to } \mathbf{m} \geq 0 \quad (7.51)$$

We find the solution using an iterative scheme of several steps [Ref. 14].

Step 1. Start with an initial guess for \mathbf{m} . Since $\mathbf{F} = \mathbf{I}$ each model parameter is associated with exactly one constraint. These model parameters can be separated into a set \mathbf{m}_E that satisfies the constraints in the equality sense and a set \mathbf{m}_S that satisfies the constraints in the inequality sense. The particular initial guess $\mathbf{m} = 0$ is clearly feasible and has all its elements in \mathbf{m}_E .

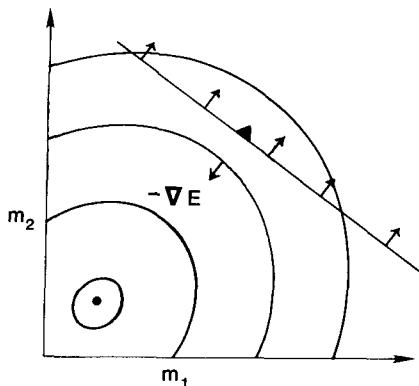


Fig. 7.5. Error $E(\mathbf{m})$ (contoured) has a single minimum (dot). The linear inequality constraint (straight line) divides $S(\mathbf{m})$ into a feasible and infeasible halfspace (arrows point into feasible halfspace). Solution (triangle) lies on the boundary of the halfspaces and therefore satisfies the constraint in the equality sense. At this point the normal of the constraint hyperplane is antiparallel to $-\nabla E$.

Step 2. Any model parameter m_i in \mathbf{m}_E that has associated with it a negative gradient $[\nabla E]_i$ can be changed both to decrease the error and to remain feasible. Therefore, if there is no such model parameter in \mathbf{m}_E , the Kuhn–Tucker theorem indicates that this \mathbf{m} is the solution to the problem.

Step 3. If some model parameter m_i in \mathbf{m}_E has a corresponding negative gradient, then the solution can be changed to decrease the prediction error. To change the solution, we select the model parameter corresponding to the most negative gradient and move it to the set \mathbf{m}_S . All the model parameters in \mathbf{m}_S are now recomputed by solving the system $\mathbf{G}_S \mathbf{m}'_S = \mathbf{d}_S$ in the least squares sense. The subscript S on the matrix indicates that only the columns multiplying the model parameters in \mathbf{m}_S have been included in the calculation. All the \mathbf{m}_E 's are still zero. If the new model parameters are all feasible, then we set $\mathbf{m} = \mathbf{m}'$ and return to Step 2.

Step 4. If some of the \mathbf{m}' 's are infeasible, however, we cannot use this vector as a new guess for the solution. Instead, we compute the change in the solution $\delta\mathbf{m} = \mathbf{m}'_S - \mathbf{m}_S$ and add as much of this vector as

possible to the solution \mathbf{m}_S without causing the solution to become infeasible. We therefore replace \mathbf{m}_S with the new guess $\mathbf{m}_S + \alpha\delta\mathbf{m}$, where $\alpha = \min_i(m_{Si}/[m_{Si} - m'_{Si}])$ is the largest choice that can be made without some \mathbf{m}_S becoming infeasible. At least one of the m_{Si} 's has its constraint satisfied in the equality sense and must be moved back to \mathbf{m}_E . The process then returns to Step 3.

This algorithm contains two loops, one nested within the other. The outer loop successively moves model parameters from the group that is constrained to the group that minimizes the prediction error. The inner loop ensures that the addition of a variable to this latter group has not caused any of the constraints to be violated. Discussion of the convergence properties of this algorithm can be found in Ref. 14.

This algorithm can also be used to solve the problem [Ref. 14]

$$\text{Minimize } \|\mathbf{m}\|_2 \quad \text{subject to } \mathbf{Fm} \geq \mathbf{h} \quad (7.52)$$

and, by virtue of the transformation described in Section 7.9.1, the completely general problem. The method consists of forming the equation

$$\mathbf{G}'\mathbf{m}' = \mathbf{d}' = \begin{bmatrix} \mathbf{F}^T \\ \mathbf{h}^T \end{bmatrix} \mathbf{m}' = \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix} \quad (7.53)$$

and finding the \mathbf{m}' that minimizes $\|\mathbf{d}' - \mathbf{G}'\mathbf{m}'\|_2$ subject to $\mathbf{m}' \geq 0$ by the above algorithm. If the prediction error $\mathbf{e}' = \mathbf{d}' - \mathbf{G}'\mathbf{m}'$ is identically zero, then the constraints $\mathbf{Fm} \geq \mathbf{h}$ are inconsistent. Otherwise, the solution is $m_i = e'_i/e'_{M+1}$.

We shall show that this method does indeed solve the indicated problem [adapted from Ref. 14]. We first note that the gradient of the error is $\nabla E' = \mathbf{G}'^T[\mathbf{d}' - \mathbf{G}'\mathbf{m}'] = -\mathbf{G}'^T\mathbf{e}'$, and that because of the Kuhn–Tucker theorem, \mathbf{m}' and $\nabla E'$ satisfy

$$\begin{aligned} \mathbf{m}'_E &= 0 & [\nabla E']_E &< 0 \\ \mathbf{m}'_S &> 0 & [\nabla E']_S &= 0 \end{aligned} \quad (7.54)$$

The length of the error is therefore

$$\begin{aligned} \mathbf{e}'^T\mathbf{e}' &= [\mathbf{d}' - \mathbf{G}'\mathbf{m}']^T\mathbf{e}' = -\mathbf{m}'^T\mathbf{G}'^T\mathbf{e}' + \mathbf{d}'^T\mathbf{e}' \\ &= \mathbf{m}'^T\nabla E' + e'_{M+1} = e'_{M+1} \end{aligned} \quad (7.55)$$

If the error is not identically zero, e'_{M+1} is greater than zero. We can use this result to show that the constraints $\mathbf{Fm} \geq \mathbf{h}$ are consistent and \mathbf{m} is

feasible since

$$0 \leq -\nabla E' = \mathbf{G}'^T \mathbf{e}' = [\mathbf{F}, \mathbf{h}][\mathbf{m}', -1]^T e'_{M+1} = [\mathbf{F}\mathbf{m} - \mathbf{h}]e'_{M+1} \quad (7.56)$$

Since $e'_{M+1} > 0$, we have $[\mathbf{F}\mathbf{m} - \mathbf{h}] \geq 0$. The solution minimizes $\|\mathbf{m}\|_2$ because the Kuhn–Tucker condition that the gradient $\nabla \|\mathbf{m}\|_2 = \mathbf{m}$ be represented as a nonnegative combination of the rows of \mathbf{F} is satisfied:

$$\nabla \|\mathbf{m}\|_2 = \mathbf{m} = [e'_1, \dots, e'_M]^T / e'_{M+1} = \mathbf{F}^T \mathbf{m}' / e'_{M+1} \quad (7.57)$$

Here \mathbf{m}' and e'_{M+1} are nonnegative. Finally, assume that the error is identically zero but that a feasible solution exists. Then

$$0 = \mathbf{e}'^T \mathbf{e}' / e'_{M+1} = [\mathbf{m}^T, -1][\mathbf{d}' - \mathbf{G}' \mathbf{m}'] = -1 - [\mathbf{F}\mathbf{m} - \mathbf{h}]^T \mathbf{m}' \quad (7.58)$$

Since $\mathbf{m}' \geq 0$, the relationship $\mathbf{F}\mathbf{m} < \mathbf{h}$ is implied. This contradicts the constraint equations $\mathbf{F}\mathbf{m} \geq \mathbf{h}$, so that an identically zero error implies that no feasible solution exists and that the constraints are inconsistent.

This page intentionally left blank

8

LINEAR INVERSE PROBLEMS AND NON- GAUSSIAN DISTRIBUTIONS

8.1 L_1 Norms and Exponential Distributions

In Chapter 5 we showed that the method of least squares and the more general use of L_2 norms could be rationalized through the assumption that the data and a priori model parameters followed Gaussian statistics. This assumption is not always appropriate, however; some data sets follow other distributions. The *exponential distribution* is one simple alternative. When Gaussian and exponential distributions of the same mean $\langle d \rangle$ and variance σ^2 are compared, the exponential distribution is found to be much longer-tailed (Fig. 8.1 and Table 8.1):

$$\begin{aligned} P(d) &= \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[\frac{-(d - \langle d \rangle)^2}{2\sigma^2}\right] && \text{Gaussian} \\ P(d) &= \frac{1}{2^{1/2}\sigma} \exp\left[\frac{-|d - \langle d \rangle|}{\sigma}\right] && \text{Exponential} \end{aligned} \tag{8.1}$$

TABLE 8.1

The area beneath a portion of the Gaussian and exponential distributions, centered on the mean and having the given half-width^a

Gaussian		Exponential	
Half-width	Area (%)	Half-width	Area (%)
1σ	68.2	1σ	76
2σ	95.4	2σ	94
3σ	99.7	3σ	98.6
4σ	99.999+	4σ	99.7

^a Note that the exponential distribution is longer-tailed.

Note that the probability of realizing data far from $\langle d \rangle$ is much higher for the exponential distribution than for the Gaussian distribution. A few data several standard deviations from the mean are reasonably probable in a data set of, say, 1000 samples drawn from an exponential distribution but very improbable for data drawn from a Gaussian distribution. We therefore expect that methods based on the exponential distribution will be able to handle a data set with a few “bad” data (outliers) better than Gaussian methods. Methods that can tolerate a few outliers are said to be *robust* [Ref. 4].

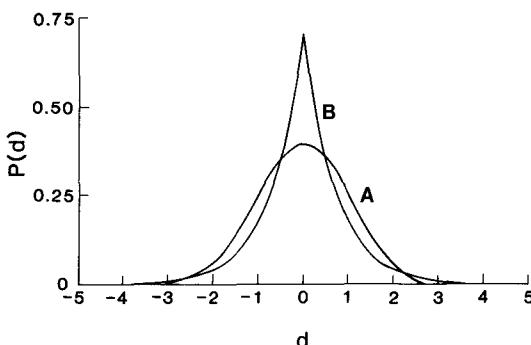


Fig. 8.1. Gaussian (curve A) and exponential (curve B) distribution with zero mean and unit variance. The exponential distribution has the longer tail.

8.2 Maximum Likelihood Estimate of the Mean of an Exponential Distribution

Exponential distributions bear the same relationship to L_1 norms as Gaussian distributions bear to L_2 norms. To illustrate this relationship, we consider the joint distribution for N independent data, each with the same mean $\langle d \rangle$ and variance σ^2 . Since the data are independent, the joint distribution is just the product of N univariate distributions:

$$P(d) = 2^{-N/2} \sigma^{-N} \exp\left[-\frac{2^{1/2}}{\sigma} \sum_{i=1}^N |d_i - \langle d \rangle|\right] \quad (8.2)$$

To maximize the likelihood of $P(d)$ we must maximize the argument of the exponential, which involves minimizing the sum of absolute residuals as

$$\text{Minimize } E = \sum_{i=1}^N |d_i - \langle d \rangle| \quad (8.3)$$

This is the L_1 norm of the prediction error in a linear inverse problem of the form $\mathbf{G}\mathbf{m} = \mathbf{d}$, where $M = 1$, $\mathbf{G} = [1, 1, \dots, 1]^T$, and $\mathbf{m} = [\langle d \rangle]$. Applying the principle of maximum likelihood, we obtain

$$\text{Maximize } L = \log P = -\frac{N}{2} \log 2 - N \log \sigma - \frac{2^{1/2}}{\sigma} \sum_{i=1}^N |d_i - \langle d \rangle| \quad (8.4)$$

Setting the derivatives to zero yields

$$\begin{aligned} \frac{\partial L}{\partial \langle d \rangle} &= 0 = -\frac{2^{1/2}}{\sigma} \sum_{i=1}^N \text{sign}(d_i - \langle d \rangle) \\ \frac{\partial L}{\partial \sigma} &= 0 = \frac{N}{\sigma} - \frac{2^{1/2}}{\sigma^2} \sum_{i=1}^N |d_i - \langle d \rangle| \end{aligned} \quad (8.5)$$

where the sign function $\text{sign}(x)$ equals $+1$ if $x > 0$, -1 if $x < 0$ and 0 if $x = 0$. The first equation yields the implicit expression for $\langle d \rangle^{\text{est}}$ for which $\sum_i \text{sign}(d_i - \langle d \rangle) = 0$. The second equation can then be solved for an estimate of the variance as

$$\sigma^{\text{est}} = \frac{2^{1/2}}{N} \sum_{i=1}^N |d_i - \langle d \rangle^{\text{est}}| \quad (8.6)$$

The equation for $\langle d \rangle^{\text{est}}$ is exactly the sample median; one finds a $\langle d \rangle$ such that half the d_i 's are less than $\langle d \rangle$ and half are greater than $\langle d \rangle$. There is then an equal number of negative and positive signs and the sum of the signs is zero. The median is a robust property of a set of data. Adding one outlier can at worst move the median from one central datum to another nearby central datum. While the maximum likelihood estimate of a Gaussian distribution's true mean is the sample arithmetic mean, the maximum likelihood estimate of an exponential distribution's true mean is the sample median.

The estimate of the variance also differs between the two distributions: in the Gaussian case it is the square of the sample standard deviation but in the exponential case it is not. If there is an odd number of samples, then $\langle d \rangle^{\text{est}}$ equals the middle d_i ; if there is an even number of samples, any $\langle d \rangle^{\text{est}}$ between the two middle data maximizes the likelihood. In the odd case the error E attains a minimum only at the middle sample, but in the even case it is flat between the two middle samples (Fig. 8.2). We see, therefore, that L_1 problems of minimizing the prediction error of $\mathbf{Gm} = \mathbf{d}$ can possess nonunique solutions that are distinct from the type of nonuniqueness encountered in the L_2 problems. The L_1 problems can still possess nonuniqueness owing to the existence of null solutions since the null solutions cannot change the prediction error under any norm. That kind of nonuniqueness leads to a completely unbounded range of estimates. The new type of nonuniqueness, on the other hand, permits the solution to take on any values between *finite* bounds.

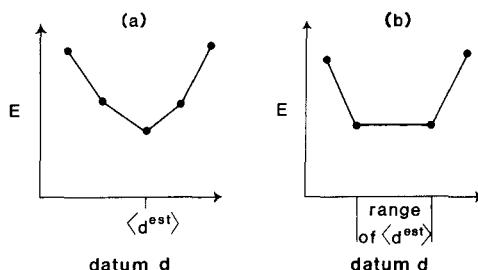


Fig. 8.2. The L_1 error in determining the mean of (a) an odd number of data and (b) an even number of data.

We also note that regardless of whether N is even or odd, we can choose $\langle d \rangle^{\text{est}}$ so that one of the equations $\mathbf{Gm} = \mathbf{d}$ is satisfied exactly (in this case, $\langle d \rangle^{\text{est}} = d_{\text{mid}}$, where mid denotes “middle”). This can be shown to be a general property of L_1 norm problems. Given N equations and M unknowns related by $\mathbf{Gm} = \mathbf{d}$, it is possible to choose \mathbf{m} so that the L_1 prediction error is minimized and so that M of the equations are satisfied exactly.

8.3 The General Linear Problem

Consider the linear inverse problem $\mathbf{Gm} = \mathbf{d}$ in which the data and a priori model parameters are uncorrelated with known means \mathbf{d}^{obs} and $\langle \mathbf{m} \rangle$ and known variances σ_d^2 and σ_m^2 , respectively. The joint distribution is then

$$P(\mathbf{d}, \mathbf{m}) = 2^{-(N+M)/2} \prod_{i=1}^N \sigma_{di}^{-1} \prod_{i=1}^M \sigma_{mi}^{-1} \times \exp \left[-2^{-1/2} \sum_{i=1}^N |e_i| \sigma_{di}^{-1} - 2^{-1/2} \sum_{i=1}^M |\ell_i| \sigma_{mi}^{-1} \right] \quad (8.7)$$

where the prediction error is given by $\mathbf{e} = \mathbf{d} - \mathbf{Gm}$ and the solution length by $\ell = \mathbf{m} - \langle \mathbf{m} \rangle$. The maximum likelihood estimate of the model parameters occurs when the exponential is a minimum, that is, when the sum of the weighted L_1 prediction error and the weighted L_1 solution length is minimized:

$$\text{Minimize } E + L = \sum_{i=1}^N |e_i| \sigma_{di}^{-1} + \sum_{i=1}^M |\ell_i| \sigma_{mi}^{-1} \quad (8.8)$$

In this case the weighting factors are the reciprocals of the standard deviations—this in contrast to the Gaussian case, in which they are the reciprocals of the variances. Note that linear combinations of exponentially distributed random variables are not themselves exponential (unlike Gaussian variables, which give rise to Gaussian combinations). The covariance matrix of the estimated model parameters is, therefore, difficult both to calculate and to interpret since the manner in which it is related to confidence intervals varies from distribution to distribution. We shall focus our discussion on estimating only the model parameters themselves.

8.4 Solving L_1 Norm Problems

We shall show that this problem can be transformed into a “linear programming” problem, a type that has been studied intensively and for which means of solution are known (e.g., the so-called Simplex algorithm; see Section 12.7). The linear programming problem can be stated as follows:

Find the \mathbf{x} that maximizes $z = \mathbf{c}^T \mathbf{x}$ subject to the constraints

$$\mathbf{A}\mathbf{x} \begin{bmatrix} \geq \\ = \\ \leq \end{bmatrix} \mathbf{b} \quad \text{and} \quad \mathbf{x} \geq 0 \quad (8.9)$$

This problem was first studied by economists and business operations analysts. For example, z might represent the profit realized by a factory producing a product line, where the number of each product is given by \mathbf{x} and the profit on each item given by \mathbf{c} . The problem is to maximize the total profit $z = \mathbf{c}^T \mathbf{x}$ without violating the constraint that one can produce only a positive amount of each product, or any other linear inequality constraints that might represent labor laws, union regulations, physical limitation of machines, etc. Computer programs that can solve the linear programming problem are readily available.

First, we shall consider the completely underdetermined linear problem with a priori model parameters, mean $\langle \mathbf{m} \rangle$, and variance σ_m^2 . The problem is to minimize the weighted length

$$L = \sum |m_i - \langle m_i \rangle| \sigma_{m_i}^{-1}$$

subject to the constraint $\mathbf{Gm} = \mathbf{d}$. We first introduce $5M$ new variables m'_i , m''_i , α_i , x_i , and x'_i , where $i = 1, \dots, M$. The linear programming problem may be stated as follows [Ref. 6]:

Minimize $\sum \alpha_i \sigma_{m_i}^{-1}$ subject to the constraints

$$\mathbf{G}[\mathbf{m}' - \mathbf{m}''] = \mathbf{d}$$

$$[m'_i - m''_i] + x_i - \alpha_i = \langle m_i \rangle \quad (8.10)$$

$$[m'_i - m''_i] - x'_i + \alpha_i = \langle m_i \rangle$$

$$m'_i \geq 0 \quad m''_i \geq 0 \quad \alpha_i \geq 0 \quad x_i \geq 0 \quad x'_i \geq 0$$

This linear programming problem has $5M$ unknowns and

$N + 2M$ constraints. If one makes the identification $\mathbf{m} = \mathbf{m}' - \mathbf{m}''$, signs of the elements of \mathbf{m} are not constrained even though those of \mathbf{m}' and \mathbf{m}'' are. Note that the remaining constraints can be rewritten as

$$\begin{aligned}\alpha_i - x_i &= [m_i - \langle m_i \rangle] \\ \alpha_i - x'_i &= -[m_i - \langle m_i \rangle]\end{aligned}\tag{8.11}$$

where the α_i , x_i , and x'_i are nonnegative. Now if $[m_i - \langle m_i \rangle]$ is positive, the first equation requires $\alpha_i \geq [m_i - \langle m_i \rangle]$ since x_i cannot be negative. The second constraint can always be satisfied by choosing some appropriate x'_i . On the other hand, if $[m_i - \langle m_i \rangle]$ is negative, then the first constraint can always be satisfied by choosing some appropriate x_i but the second constraint requires that $\alpha_i \geq -[m_i - \langle m_i \rangle]$. Taken together, these two constraints imply that $\alpha_i \geq |[m_i - \langle m_i \rangle]|$. Minimizing $\sum \alpha_i \sigma_{m_i}^{-1}$ is therefore equivalent to minimizing the weighted solution length L . The L_1 minimization problem has been converted to a linear programming problem.

The completely overdetermined problem can be converted into a linear programming problem in a similar manner. We introduce $2M + 3N$ new variables, m'_i , m''_i , $i = 1, M$ and α_i , x_i , x'_i , $i = 1, N$ and $2N$ constraints. The equivalent linear programming problem is [Ref. 6]:

$$\begin{aligned}&\text{Minimize } \sum \alpha_i \sigma_{d_i}^{-1} \quad \text{subject to the constraints} \\ &\sum_j G_{ij} [m'_j - m''_j] + x_i - \alpha_i = d_i \\ &\sum_j G_{ij} [m'_j - m''_j] - x'_i + \alpha_i = d_i \\ &m'_i \geq 0 \quad m''_i \geq 0 \quad \alpha_i \geq 0 \quad x_i \geq 0 \quad x'_i \geq 0\end{aligned}\tag{8.12}$$

The mixed-determined problem can be solved by any of several methods. By analogy to the L_2 methods described in Chapters 3 and 7, we could either pick some a priori model parameters and minimize $E + L$, or try to separate the overdetermined model parameters from the underdetermined ones and apply a priori information to the underdetermined ones only. The first method leads to a linear programming problem similar to the two cases stated above but with even

more variables ($5M + 3N$) and constraints ($2M + 2N$):

$$\begin{aligned}
 \text{Minimize} \quad & \sum \alpha_i \sigma_{m_i}^{-1} + \sum \alpha'_i \sigma_{d_i}^{-1} \quad \text{subject to the constraints} \\
 & [m'_i - m''_i] + x_i - \alpha_i = \langle m_i \rangle \\
 & [m'_i - m''_i] - x'_i + \alpha'_i = \langle m_i \rangle \\
 & \sum_j G_{ij} [m'_j - m''_j] + x''_i - \alpha'_i = d_i \\
 & \sum_j G_{ij} [m'_j - m''_j] - x'''_i + \alpha'_i = d_i \\
 & m'_i \geq 0 \quad m''_i \geq 0 \quad \alpha_i \geq 0 \quad x_i \geq 0 \quad x'_i \geq 0 \\
 & x''_i \geq 0 \quad x'''_i \geq 0
 \end{aligned} \tag{8.13}$$

The second method is more interesting. We can use the singular-value decomposition to identify the null space of \mathbf{G} . The solution then has the form

$$\mathbf{m}^{\text{est}} = \sum_{i=1}^p a_i \mathbf{v}_{pi} + \sum_{i=p+1}^M b_i \mathbf{v}_{oi} = \mathbf{V}_p \mathbf{a} + \mathbf{V}_0 \mathbf{b} \tag{8.14}$$

where the \mathbf{v} 's are the model eigenvectors from the singular-value decomposition and \mathbf{a} and \mathbf{b} are vectors with unknown coefficients. Only the vector \mathbf{a} can affect the prediction error, so one uses the overdetermined algorithm to determine it:

find \mathbf{a}^{est} that minimizes

$$E = \|\mathbf{d} - \mathbf{G}\mathbf{m}\|_1 = \sum_i |d_i - [\mathbf{U}_p \mathbf{\Lambda}_p \mathbf{a}]_i| \sigma_{d_i}^{-1} \tag{8.15}$$

Next, one uses the underdetermined algorithm to determine \mathbf{b} :

find \mathbf{b}^{est} that minimizes

$$L = \|\mathbf{m} - \langle \mathbf{m} \rangle\| = \sum_{i=1}^M \sigma_{m_i}^{-1} |[\mathbf{V}_0 \mathbf{b}]_i - (m_i - [\mathbf{V}_p \mathbf{a}^{\text{est}}]_i)| \tag{8.16}$$

It is possible to implement the basic underdetermined and overdetermined L_1 algorithms in such a manner that the many extra variables are never explicitly calculated [Ref. 6]. This procedure vastly decreases the storage and computation time required, making these algorithms practical for solving moderately large ($M = 100$) inverse problems.

8.5 The L_∞ Norm

While the L_1 norm weights “bad” data less than the L_2 norm, the L_∞ norm weights it more:

$$\begin{aligned} \text{Minimize } L + E &= \|\mathbf{e}\|_\infty + \|\boldsymbol{\ell}\|_\infty \\ &= \max_i(|e_i|/\sigma_{d_i}) + \max_i(|\ell_i|/\sigma_{m_i}) \end{aligned} \quad (8.17)$$

The prediction error and solution length are weighted by the reciprocal of their a priori standard deviations. Normally one does not want to emphasize outliers, so the L_∞ form is useful mainly in that it can provide a “worst-case” estimate of the model parameters for comparison with estimates derived on the basis of other norms. If the estimates are in fact close to one another, one can be confident that the data are highly consistent. Since the L_∞ estimate is controlled only by the worst error or length, it is usually nonunique (Fig. 8.3).

The general linear equation $\mathbf{Gm} = \mathbf{d}$ can be solved in the L_∞ sense by transformation into a linear programming problem using a variant of the method used to solve the L_1 problem. In the underdetermined problem, we again introduce new variables, m'_i , m''_i , x_i , and x'_i , $i = 1, M$ and a single parameter α ($4M + 1$ variables). The linear program-

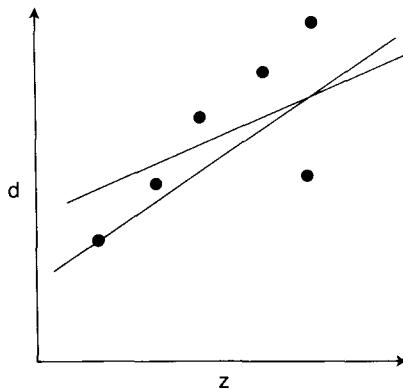


Fig. 8.3. The L_∞ solution to a linear inverse problem can be nonunique. Both of these lines have the same error $E = \max_i |e_i|$.

ming problem is

Minimize α subject to the constraints

$$\mathbf{G}[\mathbf{m}' - \mathbf{m}''] = \mathbf{d}$$

$$[m'_i - m''_i] + x_i - \alpha\sigma_{m_i} = \langle m_i \rangle \quad (8.18)$$

$$[m'_i - m''_i] - x'_i + \alpha\sigma_{m_i} = \langle m_i \rangle$$

$$m'_i \geq 0 \quad m''_i \geq 0 \quad \alpha \geq 0 \quad x_i \geq 0 \quad x'_i \geq 0$$

where $\mathbf{m} = \mathbf{m}' - \mathbf{m}''$. We note that the new constraints can be written as

$$\begin{aligned} \alpha - x_i/\sigma_{m_i} &= [m_i - \langle m_i \rangle]/\sigma_{m_i} \\ \alpha - x'_i/\sigma_{m_i} &= -[m_i - \langle m_i \rangle]/\sigma_{m_i} \end{aligned} \quad (8.19)$$

where α , x_i , and x'_i are nonnegative. Using the same argument as was applied in the L_1 case, we conclude that these constraints require that $\alpha \geq |[m_i - \langle m_i \rangle]/\sigma_{m_i}|$ for all i . Since this problem has but a single parameter α , it must therefore satisfy

$$\alpha \geq \max_i \{|[m_i - \langle m_i \rangle]/\sigma_{m_i}|\} \quad (8.20)$$

Minimizing α yields the L_∞ solution. The linear programming problem for the overdetermined case is

Minimize α subject to the constraints

$$\sum_j G_{ij}[m'_j - m''_j] + x_i - \alpha\sigma_{m_i} = d_i \quad (8.21)$$

$$\sum_j G_{ij}[m'_j - m''_j] - x'_i + \alpha\sigma_{m_i} = d_i$$

$$m'_i \geq 0 \quad m''_i \geq 0 \quad \alpha \geq 0 \quad x_i \geq 0 \quad x'_i \geq 0$$

The mixed-determined problem can be solved by applying these algorithms and either of the two methods described for the L_1 problem.

9

NONLINEAR INVERSE PROBLEMS

9.1 Parameterizations

In setting up any inverse problem, it is necessary to choose variables to represent data and model parameters (to select a parameterization). In many instances this selection is rather ad hoc; there might be no strong reasons for selecting one parameterization over another. This can become a substantial problem, however, since the answer obtained by solving an inverse problem is dependent on the parameterization. In other words, the solution is not invariant under transformations of the variables. The exception is the linear inverse problem with Gaussian statistics, in which solutions are invariant for any linear reparameterization of the data and model parameters.

As an illustration of this difficulty, consider the problem of fitting a straight line to the data pairs $(1, 1), (2, 2), (3, 3), (4, 5)$. Suppose that we regard these data as (z, d) pairs where z is an auxiliary variable. The least squares fit is $d = -0.500 + 1.300z$. On the other hand, we might regard them as (d', z') pairs where z' is the auxiliary variable. Least

squares then gives $d' = 0.309 + 0.743z'$, which can be rearranged as $z' = -0.416 + 1.345d'$. These two straight lines have slopes and intercepts that differ by about 20%.

This discrepancy arises from two sources. The first is an inconsistent application of probability theory. In the example above we alternately assumed that z was exactly known and d followed Gaussian statistics and that $z = d'$ followed Gaussian statistics and $d = z'$ was exactly known. These are two radically different assumptions about the distribution of errors, so it is no wonder that the solutions are different.

This first source of discrepancy can in theory be avoided completely by recognizing and taking into account the fact that a reparameterization introduces a kind of distribution that differs from that of the original parameterization. For instance, consider an inverse problem in which there is a datum d that is known to possess a white distribution $P(d)$ on the interval $[0, 1]$ (Fig. 9.1a). If the inverse problem is reparameterized in terms of a new model parameter $d' = d^2$, then the distribution $P(d')$ can be calculated as

$$\begin{aligned} P(d) \partial d &= P[d(d')] |\partial d / \partial d'| \partial d' = P(d') \partial d' \\ P(d') &= 1/(2\sqrt{d'}) \end{aligned} \quad (9.1)$$

The distribution of d' is not white (Fig. 9.1b) and any inverse method developed to solve the problem under this parameterization must account for this fact. [Note that Eq. (9.1) is a special case of the general rule for transforming probability distributions, $P(\mathbf{d}) dV_d = P[\mathbf{d}(\mathbf{d}')]/|\det(\mathbf{J})| dV_{d'}$, where $J_{ij} = \partial d_i / \partial d'_j$ is the Jacobian matrix and dV is the volume element.]

The second source of discrepancy is more serious. Suppose that we could use some inverse theory to calculate the distribution of the model

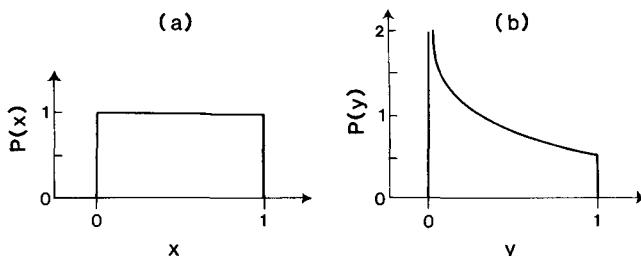


Fig. 9.1. (a) The probability distribution for x is white on the interval $[0, 1]$. (b) The distribution for $y = x^2$.

parameters under a particular parameterization. We could then use Eq.(9.1) to find their distribution under any arbitrary parameterization. Insofar as the distribution of the model parameters is the answer to the inverse problem, we would have the correct answer in the new parameterization. Probability distributions are invariant under changes of parameterization. However, a distribution is not always the answer for which we are looking. More typically, we need an estimate (a single number) based on a probability distribution (for instance, its maximum likelihood point or mean value).

Estimates are not invariant under changes in the parameterization. For example, suppose $P(m)$ has a white distribution as above. Then if $m' = m^2$, $P(m') = 1/(2\sqrt{m'})$. The distribution in m has no maximum likelihood point, whereas the distribution in m' has one at $m' = m = 0$. The distributions also have different means (expectations) E :

$$\begin{aligned} E[m] &= \int_0^1 m P(m) dm = \int_0^1 m dm = \frac{1}{2} \\ E[m'] &= \int_0^1 m' P(m') dm' = \int_0^1 \frac{1}{2}\sqrt{m'} dm' = \frac{1}{3} \end{aligned} \quad (9.2)$$

Even though m' equals the square of the model parameter m , the expectation of m' is not equal to the square of the expectation of m : $\frac{1}{3} \neq (\frac{1}{2})^2$.

There is some advantage, therefore, in working explicitly with probability distributions as long as possible, forming estimates only at the last step. If \mathbf{m} and \mathbf{m}' are two different parameterizations of model parameters, we want to avoid as much as possible sequences like

distribution for \mathbf{m} → estimate of \mathbf{m} → estimate of \mathbf{m}'

in favor of sequences like

distribution for \mathbf{m} → distribution for \mathbf{m}' → estimate of \mathbf{m}'

Note, however, that the mathematics for this second sequence is typically much more difficult than that for the first.

There are objective criteria for the “goodness” of a particular estimate of a model parameter. Suppose that we are interested in the value of a model parameter \mathbf{m} . Suppose further that this parameter either is deterministic with a true value or (if it is a random variable) has a well-defined distribution from which the true expectation could be calculated if the distribution were known. Of course, we cannot know the true value; we can only perform experiments and then apply

inverse theory to derive an estimate of the model parameter. Since any one experiment contains noise, the estimate we derive will not coincide with the true value of the model parameter. But we can at least expect that if we perform the experiment enough times, the estimated values will scatter about the true value. If they do, then the method of estimating is said to be *unbiased*. Estimating model parameters by taking nonlinear combinations of estimates of other model parameters almost always leads to bias.

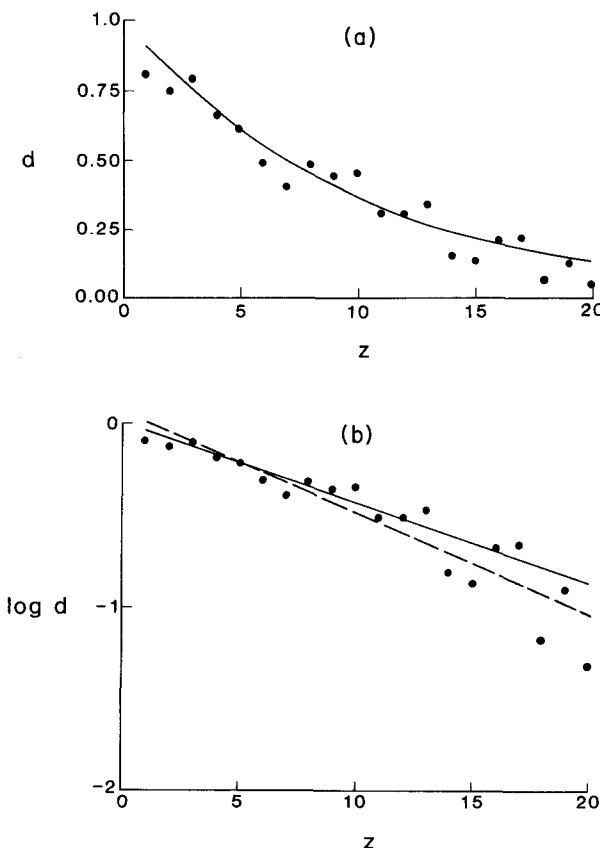


Fig. 9.2. (a) The best-fit exponential curve (solid curve) to (z, d) data (dots). (b) The best-fit curve from (a) (solid) is not the same as the best-fit line to the data in the $(z, \log(d))$ domain (dashed line). Note that, while the scatter of the data is independent of z in (a), it increases with z in (b).

9.2 Linearizing Parameterizations

One of the reasons for changing parameterizations is that it is sometimes possible to transform an inverse problem into a form that can be solved by a known method. The problems that most commonly benefit from such transformations involve fitting exponential and power functions to data. Consider a set of (z, d) data pairs that are thought to obey the model $d_i = m_1 \exp(m_2 z_i)$. By making the transformation $m'_1 = \log(m_1)$, $m'_2 = m_2$, and $d'_i = \log(d_i)$, we can write the model as the linear equation $d'_i = m'_1 + m'_2 z_i$, which can be solved by simple least squares techniques. To justify rigorously the application of least squares techniques to this problem, we must assume that the d'_i are independent random variables with a Gaussian distribution of uniform variance. The distribution of the data in their original parameterization must therefore be non-Gaussian.

For example, if the exponential decays with increasing z for all $m_2 < 0$, then the process of taking a logarithm amplifies the scattering of the near-zero points that occurs at large z . The assumption that the d'_i have uniform variance, therefore, implies that the data \mathbf{d} were measured with an accuracy that increases with z (Fig. 9.2). This assumption may well be inconsistent with the facts of the experiment. Linearizing transformations must be used with some caution.

9.3 The Nonlinear Inverse Problem with Gaussian Data

Linearizing transformations cannot be found for most inverse problems. We must consider other methods for directly solving the nonlinear problem. We shall begin by considering the very general implicit equation $\mathbf{f}(\mathbf{d}, \mathbf{m}) = 0$ (where \mathbf{f} is of length $p \leq M + N$). We simplify by assuming that the data \mathbf{d} and a priori model parameters $\langle \mathbf{m} \rangle$ have Gaussian distributions with covariance $[\text{cov } \mathbf{d}]$ and $[\text{cov } \mathbf{m}]$, respectively. If we let $\mathbf{x} = [\mathbf{d}, \mathbf{m}]^T$, we can think of the a priori distribution of the data and model as a cloud in the space $S(\mathbf{x})$ centered about the observed data and mean a priori model parameters, with a shape determined by the covariance matrix $[\text{cov } \mathbf{x}]$ (Fig. 9.3). This matrix $[\text{cov } \mathbf{x}]$ contains $[\text{cov } \mathbf{d}]$ and $[\text{cov } \mathbf{m}]$ on diagonal blocks. In principle the off-diagonal blocks could be made nonzero, indicating some cor-

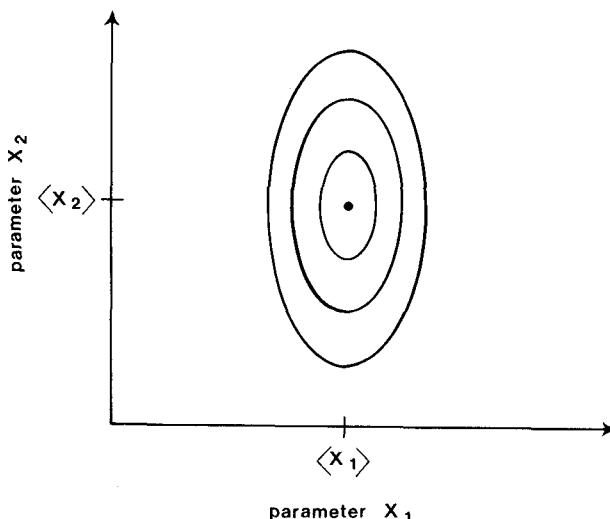


Fig. 9.3. The data and model parameters are grouped together in a vector \mathbf{x} . The a priori information for \mathbf{x} is then represented as a distribution in the $(M + N)$ -dimensional space $S(\mathbf{x})$.

relation between observed data and a priori model parameters. However, specifying a priori constraints is typically such an ad hoc procedure that one can seldom find motivation for introducing such a correlation. The a priori distribution is therefore

$$P_A(\mathbf{x}) \propto \exp\left(-\frac{1}{2}[\mathbf{x} - \langle \mathbf{x} \rangle]^T [\text{cov } \mathbf{x}]^{-1} [\mathbf{x} - \langle \mathbf{x} \rangle]\right) \quad (9.3)$$

where $\langle \mathbf{x} \rangle = [\mathbf{d}^{\text{obs}}, \langle \mathbf{m} \rangle]^T$ are the mean observed data and a priori model parameters.

The theory $\mathbf{f}(\mathbf{x}) = 0$ defines a surface in $S(\mathbf{x})$ on which the predicted data and estimated model parameters $\mathbf{x}^{\text{est}} = [\mathbf{d}^{\text{pre}}, \mathbf{m}^{\text{est}}]^T$ must lie. The probability distribution for \mathbf{x}^{est} is, therefore, $P_A(\mathbf{x})$, evaluated on this surface (Fig. 9.4). If the surface is plane, this is just the linear case described in Chapter 5 and the final distribution is Gaussian. On the other hand, if the surface is very “bumpy,” the distribution on the surface will be very non-Gaussian and may even possess several maxima.

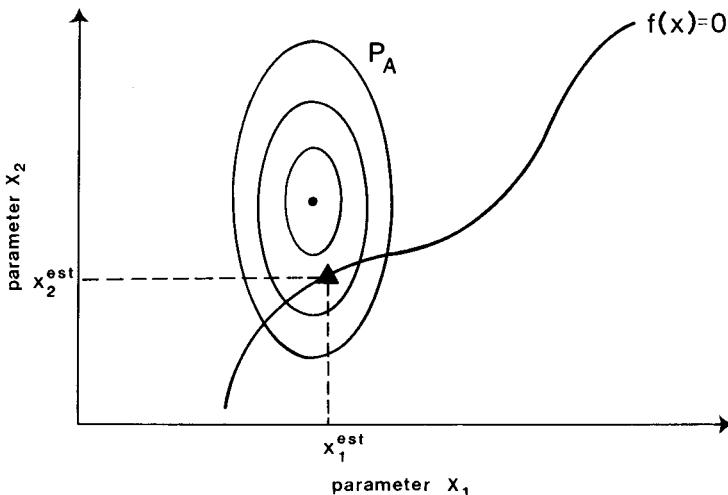


Fig. 9.4. The estimated solution \mathbf{x}^{est} is at the point on the surface $f(\mathbf{x}) = 0$ where the a priori distribution $P_A(\mathbf{x})$ attains its largest value.

One approach to estimating the solution is to find the maximum likelihood point of $P_A(\mathbf{x})$ on the surface $f(\mathbf{x}) = 0$ (Fig. 9.5). This point can be found without explicitly determining the distribution on the surface. One just maximizes $P_A(\mathbf{x})$ with the constraint that $f(\mathbf{x}) = 0$. One should keep in mind, however, that the maximum likelihood point of a non-Gaussian distribution may not be the most sensible estimate that can be made from that distribution. Gaussian distributions are symmetric, so their maximum likelihood point always coincides with their mean value. In contrast, the maximum likelihood point can be arbitrarily far from the mean of a non-Gaussian distribution (Fig. 9.6). Computing the mean, however, requires one to compute explicitly the distribution on the surface and then take its expectation (a much more difficult procedure).

These caveats aside, we proceed with the calculation of the maximum likelihood point by minimizing the argument of the exponential in $P_A(\mathbf{x})$ with the constraint that $f(\mathbf{x}) = 0$ [adapted from Ref. 18]:

$$\begin{aligned} \text{minimize } & \Phi = [\mathbf{x} - \langle \mathbf{x} \rangle]^T [\text{cov } \mathbf{x}]^{-1} [\mathbf{x} - \langle \mathbf{x} \rangle] \\ \text{subject to the constraint } & f(\mathbf{x}) = 0 \end{aligned} \quad (9.4)$$

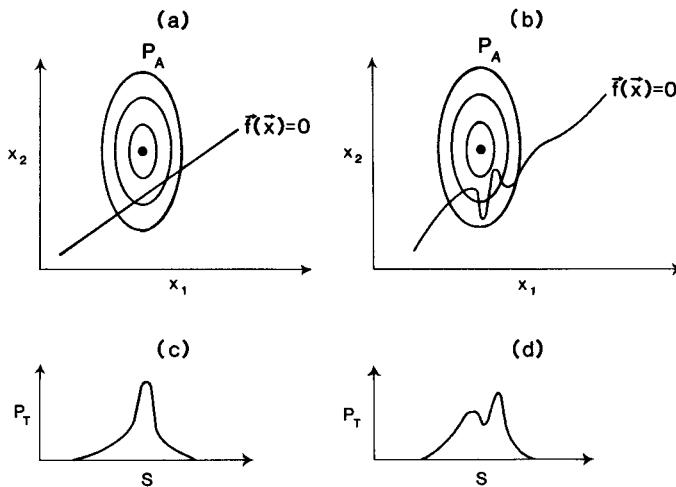


Fig. 9.5. (a) The linear theory $f(\mathbf{x}) = 0$ is a hyperplane (solid line) in the space $S(\mathbf{x})$. (c) The total distribution $P_T(\mathbf{x})$ is the a priori distribution $P_A(\mathbf{x})$ evaluated along this hyperplane (with arclength s). It has a single maximum. The nonlinear theory (b) can lead to distributions with several maxima (d).

The Lagrange multiplier equations are

$$\partial\Phi/\partial x_i - \sum_{j=1}^p 2\lambda_j \partial f_j / \partial x_i = 0 \quad \text{or} \quad [\mathbf{x} - \langle \mathbf{x} \rangle]^T [\text{cov } \mathbf{x}]^{-1} = \boldsymbol{\lambda}^T \mathbf{F} \quad (9.5)$$

where $\boldsymbol{\lambda}$ is a vector of Lagrange multipliers and \mathbf{F} the matrix of derivatives ∇f . The Lagrange multipliers can be eliminated by premul-

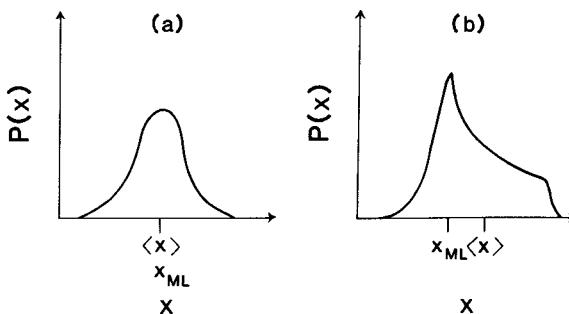


Fig. 9.6. (a) The maximum likelihood point x_{ML} and mean $\langle x \rangle$ of the Gaussian distribution are the same point. (b) They can be unequal for non-Gaussian distributions.

tiplying the transformed Lagrange equation by \mathbf{F} as

$$\mathbf{F}[\mathbf{x} - \langle \mathbf{x} \rangle] = \mathbf{F}[\text{cov } \mathbf{x}] \mathbf{F}^T \lambda \quad (9.6)$$

and then premultiplying by $\{\mathbf{F}[\text{cov } \mathbf{x}] \mathbf{F}^T\}^{-1}$ as

$$\lambda = \{\mathbf{F}[\text{cov } \mathbf{x}] \mathbf{F}^T\}^{-1} \mathbf{F}[\mathbf{x} - \langle \mathbf{x} \rangle] \quad (9.7)$$

Substitution into the original equation yields

$$[\mathbf{x} - \langle \mathbf{x} \rangle] = [\text{cov } \mathbf{x}] \mathbf{F}^T \{\mathbf{F}[\text{cov } \mathbf{x}] \mathbf{F}^T\}^{-1} \mathbf{F}[\mathbf{x} - \langle \mathbf{x} \rangle] \quad (9.8)$$

which must be solved simultaneously with the constraint equation $\mathbf{f}(\mathbf{x}) = 0$. These two equations are equivalent to the single equation

$$[\mathbf{x} - \langle \mathbf{x} \rangle] = [\text{cov } \mathbf{x}] \mathbf{F}^T \{\mathbf{F}[\text{cov } \mathbf{x}] \mathbf{F}^T\}^{-1} (\mathbf{F}[\mathbf{x} - \langle \mathbf{x} \rangle] - \mathbf{f}(\mathbf{x})) \quad (9.9)$$

since the original two equations can be recovered by premultiplying this equation by \mathbf{F} . Since the unknown \mathbf{x} appears on both sides of the equation and since \mathbf{f} and \mathbf{F} are functions of \mathbf{x} , this equation may be difficult to solve explicitly. Only in the linear case when $\mathbf{f}(\mathbf{x}) = \mathbf{F}\mathbf{x}$ (\mathbf{F} being constant) is its solution comparatively simple, as was shown in Chapter 5. We now examine an iterative method of solving it. This method consists of starting with some initial guess, say, $\mathbf{x}_n^{\text{est}}$ where $n = 0$, and then generating successive approximations as

$$\mathbf{x}_{n+1}^{\text{est}} = \langle \mathbf{x} \rangle + [\text{cov } \mathbf{x}] \mathbf{F}_n^T \{\mathbf{F}_n [\text{cov } \mathbf{x}] \mathbf{F}_n^T\}^{-1} (\mathbf{F}_n [\mathbf{x}_n^{\text{est}} - \langle \mathbf{x} \rangle] - \mathbf{f}(\mathbf{x}_n^{\text{est}})) \quad (9.10)$$

The subscript on \mathbf{F}_n implies that it is evaluated at $\mathbf{x}_n^{\text{est}}$. If the initial guess is close enough to the maximum likelihood point, the successive approximations will converge to the desired value. We shall discuss the matter of convergence in more detail.

If the theory is explicit (i.e., if $\mathbf{f}(\mathbf{x}) = \mathbf{d} - \mathbf{g}(\mathbf{m}) = 0$) and if the data and a priori model parameters are uncorrelated, the iterative formula can be rewritten as

$$\begin{aligned} \mathbf{m}_{n+1}^{\text{est}} &= \langle \mathbf{m} \rangle + \mathbf{G}_n^{-\mathbf{g}} (\mathbf{d} - \mathbf{g}(\mathbf{m}_n^{\text{est}}) + \mathbf{G}_n [\mathbf{m}_n^{\text{est}} - \langle \mathbf{m} \rangle]) \\ \mathbf{G}_n^{-\mathbf{g}} &= [\text{cov } \mathbf{m}] \mathbf{G}_n^T \{[\text{cov } \mathbf{d}] + \mathbf{G}_n [\text{cov } \mathbf{m}] \mathbf{G}_n^T\}^{-1} \quad (9.11) \\ &= \{[\text{cov } \mathbf{m}]^{-1} + \mathbf{G}_n^T [\text{cov } \mathbf{d}]^{-1} \mathbf{G}_n^T\}^{-1} \mathbf{G}_n^T [\text{cov } \mathbf{d}]^{-1} \end{aligned}$$

where $[\mathbf{G}_n]_{ij} = \partial g_i / \partial m_j$ is evaluated at $\mathbf{m}_n^{\text{est}}$ and where the generalized inverse notation has been used for convenience. These formulas are the nonlinear, iterative analogues to the linear, noniterative formulas

stated for the linear inverse problem in Section 5.7, except that in this case the theory has been assumed to be exact. If the theory is inexact with a Gaussian distribution of error $P_g(\mathbf{m}|\mathbf{d})$ described by covariance $[\text{cov } \mathbf{g}]$ (see Section 5.5), then one must find the maximum likelihood point of the total distribution $P_T(\mathbf{x}) = P_A(\mathbf{x})P_g(\mathbf{m}|\mathbf{d})$, as was done in Section 5.6 for the linear case. The formulas one obtains are just the iterative formulas stated above with each occurrence of $[\text{cov } \mathbf{d}]$ replaced by $[\text{cov } \mathbf{d}] + [\text{cov } \mathbf{g}]$.

Since the distribution for \mathbf{m}^{est} is non-Gaussian, its covariance is both difficult to calculate and difficult to interpret in terms of confidence intervals. If the problem is not too nonlinear, the covariance might be estimated using the linear formula

$$[\text{cov } \mathbf{m}^{\text{est}}] \simeq \mathbf{G}_n^{-\mathbf{g}}[\text{cov } \mathbf{d}]\mathbf{G}_n^{-\mathbf{g}^T} + [\mathbf{I} - \mathbf{R}_n][\text{cov } \mathbf{m}][\mathbf{I} - \mathbf{R}_n]^T \quad (9.12)$$

where $\mathbf{R}_n = \mathbf{G}_n^{-\mathbf{g}}\mathbf{G}_n$ and the last n in the iteration is used. The same restriction applies to interpretations of the resolution matrices \mathbf{N} and \mathbf{R} . Since the problem is nonlinear, they do not describe the true resolution of the problem. On the other hand, they give the resolution of a linear problem that is in some sense close to the nonlinear one.

The iterative formulas for \mathbf{m}^{est} can also be derived from a somewhat different point of view. The nonlinear equation $\mathbf{g}(\mathbf{m}) = \mathbf{d}$ can be linearized by expanding it about some point, say, $\mathbf{m}_n^{\text{est}}$ using Taylor's theorem as

$$\mathbf{g}(\mathbf{m}) \simeq \mathbf{g}(\mathbf{m}_n^{\text{est}}) + \nabla \mathbf{g}[\mathbf{m} - \mathbf{m}_n^{\text{est}}] = \mathbf{g}(\mathbf{m}_n^{\text{est}}) + \mathbf{G}_n[\mathbf{m} - \mathbf{m}_n^{\text{est}}] \quad (9.13)$$

where we have ignored terms of second order and higher. Defining $\Delta \mathbf{m}_{n+1} = [\mathbf{m} - \mathbf{m}_n^{\text{est}}]$, we can write the approximate equations as

$$\begin{aligned} \mathbf{G}_n \Delta \mathbf{m}_{n+1} &= \mathbf{d} - \mathbf{g}(\mathbf{m}_n^{\text{est}}) \\ \mathbf{m}_{n+1}^{\text{est}} &= \mathbf{m}_n^{\text{est}} + \Delta \mathbf{m}_{n+1} \end{aligned} \quad (9.14)$$

If one makes an initial guess of $\mathbf{m}_0^{\text{est}}$, the first equation can be solved for $\Delta \mathbf{m}_1$ by any of the methods of Chapters 3–5, and a new guess $\mathbf{m}_1^{\text{est}}$ can be computed from the second equation. The process can be iterated until a solution to the inverse problem is obtained. Note, however, that measures of solution length should be based on the length of $\mathbf{m}_{n+1}^{\text{est}}$, not just on the length of the perturbation $\Delta \mathbf{m}_{n+1}$. The application to this problem of the linear formulas derived in Section 5.7 leads to exactly the nonlinear iterative equation derived in Eq. (9.11).

9.4 Special Cases

9.4.1 LEAST SQUARES SOLUTION

We assume that the data are independent, that they have uniform variance σ_d^2 , and that no a priori information has been added to the problem:

$$\mathbf{m}_{n+1}^{\text{est}} = [\mathbf{G}_n^T \mathbf{G}_n]^{-1} \mathbf{G}_n^T [\mathbf{d} - \mathbf{g}(\mathbf{m}_n^{\text{est}})] + \mathbf{m}_n^{\text{est}} \quad (9.15)$$

9.4.2 DAMPED LEAST SQUARES

We assume that the a priori model parameters are independent with zero mean and uniform variance σ_m^2 :

$$\mathbf{m}_{n+1}^{\text{est}} = \left[\mathbf{G}_n^T \mathbf{G}_n + \frac{\sigma_d^2}{\sigma_m^2} \mathbf{I} \right]^{-1} \mathbf{G}_n^T [\mathbf{d} - \mathbf{g}(\mathbf{m}_n^{\text{est}})] + \mathbf{m}_n^{\text{est}} \quad (9.16)$$

9.4.3 MINIMUM-LENGTH SOLUTION

We assume that the a priori model parameters are independent with infinite variance and zero mean:

$$\mathbf{m}_{n+1}^{\text{est}} = \mathbf{G}_n^T [\mathbf{G}_n \mathbf{G}_n^T]^{-1} [\mathbf{d} - \mathbf{g}(\mathbf{m}_n^{\text{est}})] + \mathbf{m}_n^{\text{est}} \quad (9.17)$$

Note that, in the minimum-length case, the perturbation $\Delta \mathbf{m}$ is minimized (as is \mathbf{m}). This is a mathematical coincidence and is not true for other choices of $[\text{cov } \mathbf{m}]$.

9.5 Convergence and Nonuniqueness of Nonlinear L_2 Problems

In the linear case we found simple means for deciding whether a problem has a unique solution. In contrast, there are no simple means for deciding whether a nonlinear inverse problem has a unique solution that minimizes prediction error in the absence of a priori information. Consider the very simple nonlinear model $d_i = m_1^2 + m_1 m_2 z_i$. This problem can be linearized by the transformation of variables $m'_1 = m_1^2$, $m'_2 = m_1 m_2$ and can therefore be solved by the least squares method if $N > 2$. Nevertheless, even if the primed parameters are

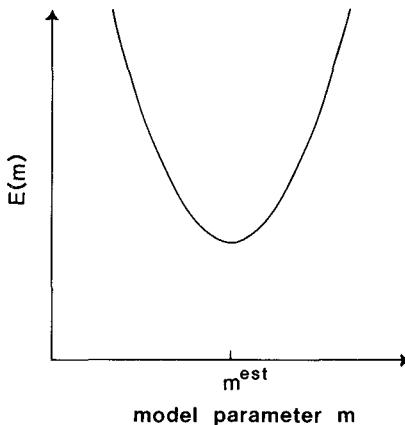


Fig. 9.7. Error E for the linear Gaussian case is always a paraboloid.

unique, the unprimed ones are not: if \mathbf{m}^{est} is a solution that minimizes the prediction error, then $-\mathbf{m}^{\text{est}}$ is also a solution with the same error. In this instance the error $E(\mathbf{m})$ has two minima of equal depth.

To investigate the question of nonuniqueness, we must examine the global properties of the prediction error. If the surface has but a single minimum point, then the solution is unique. If it has more than one minimum point, the solution is nonunique, and a priori information must be added to resolve the indeterminacy. The error surface of a linear problem is always a paraboloid (Fig. 9.7), which can have only a simple range of shapes. An arbitrarily complex nonlinear inverse problem can have an arbitrarily complicated error. If $M = 2$ or 3 , it may be possible to investigate the shape of the surface by graphical techniques. For most realistic problems this is infeasible [Fig. 9.8].

Even if an inverse problem is known to have a unique solution, there is no guarantee that the iterative technique described above will converge to the solution. The difficulty arises from the fact that the linearized method does not "see" the entire error surface $E(\mathbf{m})$. Instead, it sees only the part of $E(\mathbf{m})$ in the vicinity of $\mathbf{m}_n^{\text{est}}$ and approximates the rest of the surface as a paraboloid tangent to the actual surface at that point. The new estimate $\mathbf{m}_{n+1}^{\text{est}}$ is the minimum of the paraboloid (Fig. 9.9). Since any minimum with continuous derivatives is locally paraboloid in shape, the method will converge to the minimum of the error function if the initial guess is close enough. If it is not close enough, the process may not converge at all or may converge to a

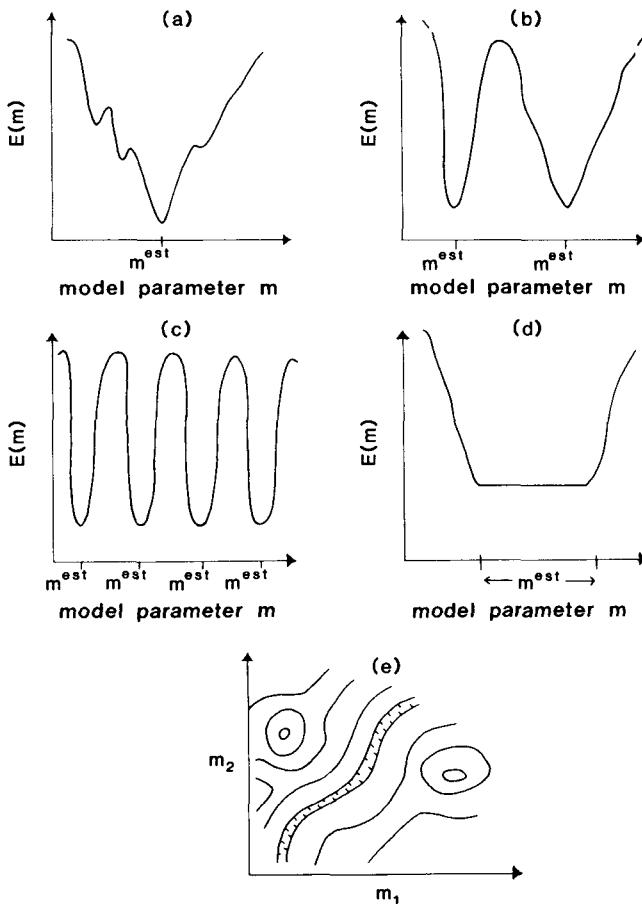


Fig. 9.8. Prediction error E as a function of model parameters \mathbf{m} . (a) Single minimum corresponds to an inverse problem with a unique solution. (b) Two solutions. (c) Many well-separated solutions. (d) Finite range of solutions. (e) Contour plot of E . There is an infinite range of solutions that can lie anywhere in the valley.

local minimum or even to a maximum (Fig. 9.10). The convergence properties are therefore seen to depend on the geometry of the error surfaces. Iterative methods can only find solutions that are *linearly close* to the initial guess. Except for experimentation with a variety of initial guesses (which is usually inconclusive since one can never examine enough), there is no general method for determining whether

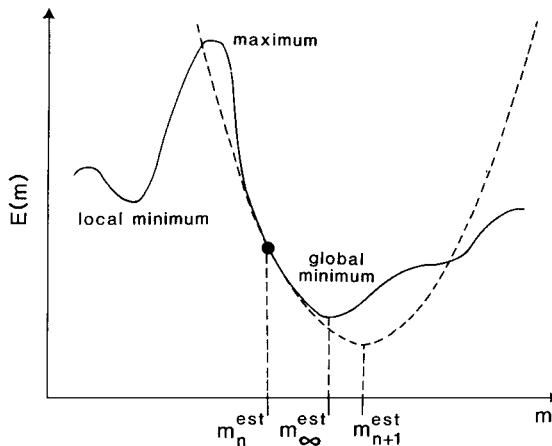


Fig. 9.9. The iterative method locates the global minimum of the error E (solid curve) by locating the minimum of the paraboloid (dashed curve) that is tangent to E at the trial solution m_n^{est} .

a solution obtained by the iterative method really does minimize error in the global sense.

9.6 Non-Gaussian Distributions

In Section 5.7 the general distribution of model parameters and data for a linear, Gaussian theory was derived. In Section 9.2 this analysis was extended to nonlinear, Gaussian theories. In both cases the answer to the inverse problem was based on a distribution—either the projected distribution for the model parameters or some estimate computed from that distribution.

To extend this analysis to the completely general case requires a more careful analysis of the way in which the component probability distributions are defined and combined [Ref. 19]. We shall begin, as we did in the other cases, by assuming that we can define an a priori distribution $P_A(\mathbf{d}, \mathbf{m}) = P_A(\mathbf{x})$ for observed data and a priori model parameters. We use this distribution to characterize all that is known about model parameters and data before applying the actual theory.

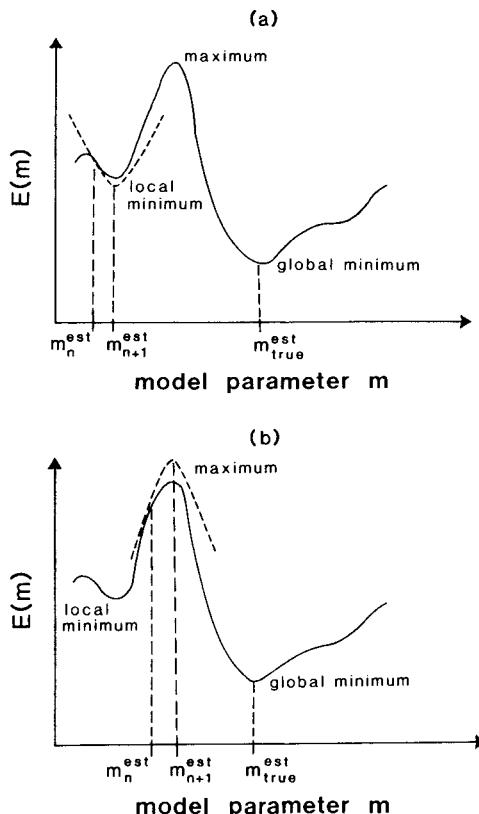


Fig. 9.10. If the trial solution is too far from the global minimum, the method may converge to (a) a local minimum or to (b) a maximum.

The information content of $P_A(\mathbf{x})$ can be evaluated by comparing it to a distribution that characterizes the state of complete ignorance about \mathbf{x} . This distribution is called the *null distribution* $P_N(\mathbf{x})$ (not to be confused with *null vectors* of Chapters 6 and 7, which are completely different). If we know nothing about \mathbf{x} , we might be inclined to say that it can take on any value with equal probability: $P_N(\mathbf{x}) \propto \text{constant}$. This is the choice we implicitly made in all the previous discussions, and it is usually adequate. However, this definition is sometimes inadequate. For instance, if \mathbf{x} is the Cartesian coordinates of an object in three-dimensional space, then $P_N \propto \text{constant}$ means that the object could be

anywhere with equal probability. This is an adequate definition of the null distribution. On the other hand, if the position of the object is specified by the spherical coordinates $\mathbf{x} = [r, \theta, \phi]^T$, then the statement $P_N \propto \text{constant}$ actually implies that the object is near the origin. The statement that the object could be anywhere is $P_N(\mathbf{x}) \propto r^2 \sin \theta$. The null distribution must be chosen with the physical significance of the vector \mathbf{x} in mind.

Unfortunately, it is sometimes difficult to find a guiding principle with which to choose the null distribution. Consider the case of an acoustics problem in which a model parameter is the acoustic velocity v . At first sight it may seem that a reasonable choice for the null distribution is $P_N(v) \propto \text{constant}$. Acousticians, however, often work with the acoustic slowness $s = 1/v$, and the distribution $P_N(v) \propto \text{constant}$ implies $P_N(s) \propto s^2$. This is somewhat unsatisfactory, since one could, with equal plausibility, argue that $P_N(s) \propto \text{constant}$, in which case $P_N(v) \propto v^2$. One possible solution to this dilemma is to choose a null solution whose *form* is invariant under the reparameterization. The distribution that works in this case is $P_N(v) \propto 1/v$, since this leads to $P_N(s) \propto 1/s$.

Once one has defined the null solution one can compare any other distribution to it to measure the information content of that distribution. One commonly used quantitative measure of the information in a distribution P is the scalar information I , defined by

$$I(P, P_N) = \int P(\mathbf{x}) \log[P(\mathbf{x})/P_N(\mathbf{x})] d\mathbf{x} \quad (9.18)$$

The information I has the following properties [Ref. 19]: (1) the information of the null distribution is zero; (2) all distributions except the null distribution have positive information; (3) the more sharply peaked $P(\mathbf{x})$ becomes, the more its information increases; and (4) the information is invariant under reparameterizations. We can therefore measure the amount of information added to the inverse problem by imposing a priori constraints.

To solve the inverse problem we must combine the a priori distribution with a distribution that represents the (possibly erroneous) theory. We shall call this composite distribution $P_f(\mathbf{x})$. In the Gaussian case we performed this combination simply by multiplying the distributions. In the general case we shall find that the process of combining two distributions is more complicated. Suppose we let $P_3 = C(P_1, P_2)$ mean that distributions 1 and 2 are combined into distribution 3.

Then, clearly, the process of combining must have the following properties [adapted from Ref. 19]:

- (a) $C(P_1, P_2)$ should be invariant under reparameterizations.
- (b) $C(P_1, P_2)$ should be commutative: $C(P_1, P_2) = C(P_2, P_1)$.
- (c) $C(P_1, P_2)$ should be associative: $C(P_1, C(P_2, P_3)) = C(C(P_1, P_2), P_3)$.
- (d) Combining a distribution with the null distribution should return the same distribution: $C(P_1, P_N) = P_1$.
- (e) $C(P_1, P_2)$ should be everywhere zero if and only if P_1 or P_2 is everywhere zero.

These conditions can be shown to be satisfied by the choice [Ref. 19]

$$C(P_1, P_2) = P_1 P_2 / P_N \quad (9.19)$$

Note that if the null distribution is constant, one combines distributions simply by multiplying them.

The distribution for the theory $P_f(\mathbf{x})$ gives the probability that the theory will simultaneously predict data \mathbf{d} and model parameters \mathbf{m} . In the case of an implicit theory, this distribution may be exceedingly difficult to state. On the other hand, if the theory is explicit (has form $\mathbf{f}(\mathbf{x}) = \mathbf{d} - \mathbf{g}(\mathbf{m}) = 0$), then the distribution for the theory can be constructed from two component distributions: the distribution $P_g(\mathbf{d}|\mathbf{m})$ for data \mathbf{d} given model parameters \mathbf{m} , and the distribution for the model parameters, which, since they are completely unknown, is just the null distribution:

$$P_f(\mathbf{x}) = P_g(\mathbf{d}|\mathbf{m})P_N(\mathbf{m}) \quad (9.20)$$

The distribution $P_g(\mathbf{d}|\mathbf{m})$ is sometimes called a conditional probability distribution; it gives the probability of \mathbf{d} , given a value for \mathbf{m} .

The total distribution is then

$$P_T(\mathbf{x}) = P_A(\mathbf{x})P_f(\mathbf{x})/P_N(\mathbf{x}) = P_A(\mathbf{d}, \mathbf{m})P_g(\mathbf{d}|\mathbf{m})/P_N(\mathbf{d}) \quad (9.21)$$

where the second form can be used only in the case of explicit theories. The joint probability of the data and model parameters is $P_T(\mathbf{x})$. We are therefore free to consider this expression the answer to the inverse problem. Alternatively, we could derive estimates from it, such as the maximum likelihood point or mean, and call these the answer. On the other hand, the probability of the model parameters alone might be considered the answer, in which case we must project the distribution $P_P(\mathbf{m}) = \int P_T(\mathbf{x}) d\mathbf{d}$. We could also derive answers in the form of

estimates drawn from this projected distribution [Ref. 19]. In general, each of these answers will be different from the others. Which one is the best depends on one's point of view.

9.7 Maximum Entropy Methods

Note that in the previous section we first had to postulate the form of the various probability distributions to solve the inverse problem. In some cases there is good theoretical reason for using a particular distribution (Gaussian, Poisson, exponential, white, etc). In other cases the choice seems more arbitrary. We might ask whether we could somehow determine the form of the distribution directly, rather than having to assume it.

A guiding principle is needed to accomplish this. One such principle is the assertion that the best distributions are the ones with the most information, as measured by the scalar information I [Ref. 17]. As an example, consider the underdetermined, linear problem $\mathbf{Gm} = \mathbf{d}$ where $P(\mathbf{m})$ is taken as unknown. Let us further assume that the equation is interpreted to mean that the data are exactly equal to the mean (expected value) of the model parameters $\mathbf{GE}(\mathbf{m}) = \mathbf{d}$. We can then formulate the following problem for the distribution of the model parameters:

Find the $P(\mathbf{m})$ that maximizes $I(P(\mathbf{m}), P_N(\mathbf{m}))$

subject to the constraints $\mathbf{GE}(\mathbf{m}) = \mathbf{d}$ and $\int P(\mathbf{m}) \partial\mathbf{m} = 1$. (9.22)

This approach is called the *maximum entropy method* because it was first used in statistical mechanics. The problem can be solved by the use of a combination of Lagrange multipliers and the calculus of variations. At first sight the method may seem less arbitrary than previously discussed methods that require the form of the distributions to be known *a priori*. In fact it is no less arbitrary since the definition of information I is by no means unique. The form of I stated in Eq. (9.18) forces $P(\mathbf{m})$ to be an exponential distribution.

10

FACTOR ANALYSIS

10.1 The Factor Analysis Problem

Consider an ocean whose sediment is derived from the simple mixing of continental source rocks A and B (Fig. 10.1). Suppose that the concentrations of three elements are determined for many samples of sediment and then plotted on a graph whose axes are percentages of those elements. Since all the sediments are derived from only two source rocks, the sample compositions lie on the triangular portion of a plane bounded by the compositions of A and B (Fig. 10.2).

The factor analysis problem is to deduce the number of the source rocks (called *factors*) and their composition from observations of the composition of the sediments (called *samples*). It is therefore a problem in inverse theory. We shall discuss it separately, since it provides an interesting example of the use of some of the vector space analysis techniques developed in Chapter 7.

In the basic model the samples are simple mixtures (linear combinations) of the factors. If there are N samples containing M elements and if there are p factors, we can state this model algebraically with the

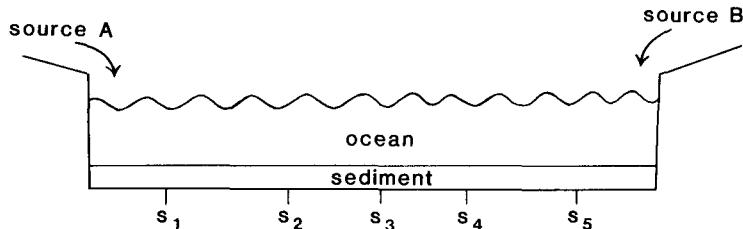


Fig. 10.1. Material from sources *A* and *B* are eroded into the ocean and mix to form sediment. Samples s_i of the sediment are collected and their chemical composition is determined. The data are used to infer the composition of the sources.

equation

$$\mathbf{S} = \mathbf{C}\mathbf{F} \quad (10.1)$$

where S_{ij} is the fraction of element j in sample i :

$$\mathbf{S} = \begin{bmatrix} \text{element 1} & \text{element 2} & \dots & \text{element } M \\ \text{in sample 1} & \text{in sample 1} & & \text{in sample 1} \\ \vdots & \vdots & & \vdots \\ \text{element 1} & \text{element 2} & \dots & \text{element } M \\ \text{in sample } N & \text{in sample } N & & \text{in sample } N \end{bmatrix} \quad (10.2)$$

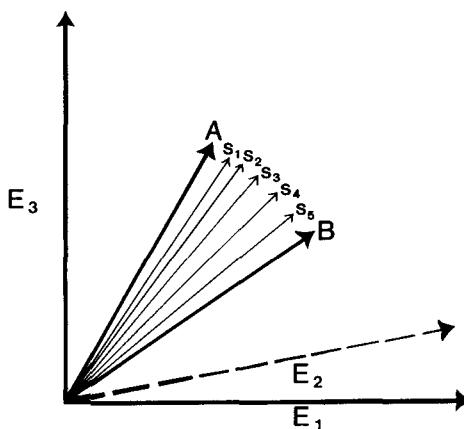


Fig. 10.2. The composition of the sample s_i lies on a triangular sector of a plane bounded by the composition of the sources *A* and *B*.

Similarly, F_{ij} is the fraction of element j in factor i :

$$\mathbf{F} = \begin{bmatrix} \text{element 1} & \text{element 2} & \dots & \text{element } M \\ \text{in factor 1} & \text{in factor 1} & & \text{in factor 1} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \text{element 1} & \text{element 2} & \dots & \text{element } M \\ \text{in factor } p & \text{in factor } p & & \text{in factor } p \end{bmatrix} \quad (10.3)$$

c_{ij} is the fraction of factor i in sample j :

$$\mathbf{C} = \begin{bmatrix} \text{factor 1} & \text{factor 2} & \dots & \text{factor } p \\ \text{in sample 1} & \text{in sample 1} & & \text{in sample 1} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \text{factor 1} & \text{factor 2} & \dots & \text{factor } p \\ \text{in sample } N & \text{in sample } N & & \text{in sample } N \end{bmatrix} \quad (10.4)$$

The elements of the matrix \mathbf{C} are sometimes referred to as the *factor loadings*.

The inverse problem is to factor the matrix \mathbf{S} into \mathbf{C} and \mathbf{F} . Each sample (each row of \mathbf{S}) is represented as a linear combination of the factors (rows of \mathbf{F}), with the elements of \mathbf{C} giving the coefficients of the combination. It is clear that, as long as we pick an \mathbf{F} whose rows span the space spanned by the rows of \mathbf{S} , we can perform the factorization. For $p \geq M$ any linearly independent set of factors will do, so in this sense the factor analysis problem is completely nonunique. It is much more interesting to ask what the minimum number of factors is that can be used to represent the samples. Then the factor analysis problem is equivalent to examining the space spanned by \mathbf{S} and determining its dimension. This problem can easily be solved by expanding the samples with the singular-value decomposition as

$$\mathbf{S} = \mathbf{U}_p \Lambda_p \mathbf{V}_p^T = (\mathbf{U}_p \Lambda_p) (\mathbf{V}_p^T) = \mathbf{CF} \quad (10.5)$$

Only the eigenvectors with nonzero singular values appear in the decomposition. The number of factors is given by the number of nonzero eigenvalues. One possible set of factors is the p eigenvectors. This set of factors is not unique. Any set of factors that spans the p space will do.

If we write out the above equations, we find that the composition of the i th sample, say, s_i , is related to the eigenvectors v_i and singular values λ_i by

$$\begin{aligned} s_1 &= [\mathbf{U}_p]_{11}\lambda_1 v_1 + [\mathbf{U}_p]_{12}\lambda_2 v_2 + \cdots + [\mathbf{U}_p]_{1p}\lambda_p v_p \\ &\quad \cdot \\ &\quad \cdot \\ &\quad \cdot \\ s_N &= [\mathbf{U}_p]_{N1}\lambda_1 v_1 + [\mathbf{U}_p]_{N2}\lambda_2 v_2 + \cdots + [\mathbf{U}_p]_{Np}\lambda_p v_p \end{aligned} \tag{10.6}$$

If the singular values are arranged in descending order, then most of each sample is composed of factor 1, with a smaller contribution from factor 2, etc. Because \mathbf{U}_p and \mathbf{V}_p are vectors of unit length, on average their elements are of equal size. We have identified the most “important” factors (Fig. 10.3). Even if $p = M$, it might be possible to neglect some of the smaller singular values and still achieve a reasonably good prediction of the sample compositions.

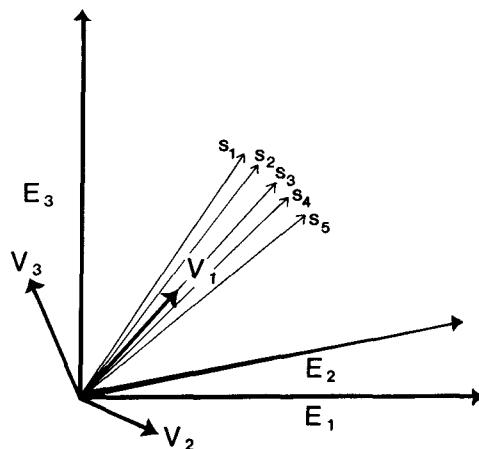


Fig. 10.3. Eigenvectors v_1 and v_2 lie in the plane of the samples (v_1 is close to the mean sample). Eigenvector v_3 is normal to this plane.

The eigenvector with the largest singular value is near the mean of the sample vectors. It is easy to show that the sample mean $\langle \mathbf{s} \rangle$ maximizes the sum of dot products with the data $\sum_i [\mathbf{s}_i \cdot \langle \mathbf{s} \rangle]$, while the eigenvector with largest singular value maximizes the sum of squared dot products $\sum_i [\mathbf{s}_i \cdot \mathbf{v}]^2$. (To show this, maximize the given functions using Lagrange multipliers, with the constraint that $\langle \mathbf{s} \rangle$ and \mathbf{v} are unit vectors.) As long as most of the samples are in the same quadrant, these two functions have roughly the same maximum.

10.2 Normalization and Physicality Constraints

In many instances an element can be important even though it occurs only in trace quantities. In such cases one cannot neglect factors simply because they have small singular values. They may contain an important amount of the trace elements. It is therefore appropriate to normalize the matrix \mathbf{S} so that there is a direct correspondence between singular value size and importance. This is usually done by defining a diagonal matrix of weights \mathbf{W} (usually proportional to the reciprocal of the standard deviations of measurement of each of the elements) and then forming a new weighted sample matrix $\mathbf{S}' = \mathbf{SW}$.

The singular-value decomposition enables one to determine a set of factors that span, or approximately span, the space of samples. These factors, however, are not unique in the sense that one can form linear combinations of factors that also span the space. This transformation is typically a useful thing to do since, ordinarily, the singular-value decomposition eigenvectors violate *a priori* constraints on what “good” factors should be like. One such constraint is that the factors should have a unit L_1 norm, that is, their elements should sum to one. If the components of a factor represent fractions of chemical elements, for example, it is reasonable that the elements should sum to 100%. Another constraint is that the elements of both the factors and the factor loadings should be nonnegative. Ordinarily a material is composed of a positive combination of components. Given an initial representation of the samples $\mathbf{S} = \mathbf{CFW}^{-1}$, we could imagine finding a new representation consisting of linear combinations of the old factors, defined by $\mathbf{F}' = \mathbf{TF}$, where \mathbf{T} is an arbitrary $p \times p$ transformation matrix. The problem can then be stated.

Find \mathbf{T} subject to the following constraints:

$$\begin{aligned} \sum_j [\mathbf{F}'\mathbf{W}^{-1}]_{ij} &= 1 && \text{for all } i \\ [\mathbf{C}\mathbf{T}^{-1}]_{ij} &\geq 0 && \text{for all } i \text{ and } j \\ [\mathbf{F}'\mathbf{W}^{-1}]_{ij} &\geq 0 && \text{for all } i \text{ and } j \end{aligned} \quad (10.7)$$

These conditions do not uniquely determine \mathbf{T} , as can be seen from Fig. 10.4. Note that the second constraint is nonlinear in the elements of \mathbf{T} . This is a very difficult constraint to implement and in practice is often ignored.

To find a unique solution one must add some a priori information. One possibility is to find a set of factors that maximize some measure of simplicity. Commonly used measures are based on the L_4 norm of the elements of the factor loadings. Maximizing this norm tends to select factors that have one especially large factor loading element. For instance, one might choose the transformation to maximize the measure of simplicity.

$$L = \sum_{i=1}^N \sum_{j=1}^p C_{ij}^4 / \left[\sum_{i=1}^N \sum_{j=1}^p C_{ij}^2 \right]^2 \quad (10.8)$$

(in which case the factors are called *quartimax factors*) or the measure of simplicity

$$L' = p^{-2} \sum_{i=1}^N \left[p \sum_{j=1}^p C_{ij}^4 - \left(\sum_{j=1}^p C_{ij}^2 \right)^2 \right] \quad (10.9)$$

(in which case the factors are called *varimax factors*).

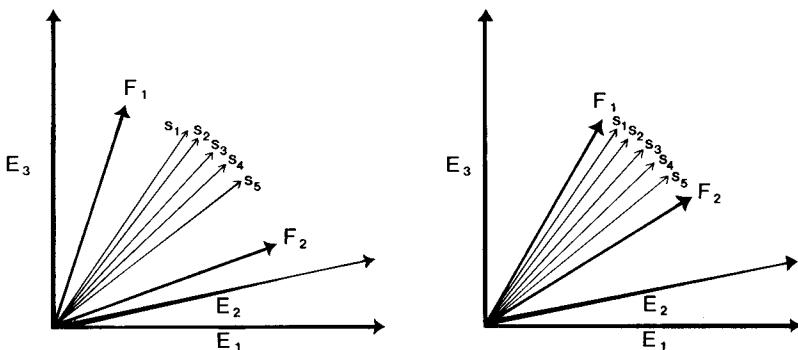


Fig. 10.4. Any two factors \mathbf{F}_i that lie in the plane of the samples and that bound the range of sample compositions are acceptable.

Another possible way of adding a priori information is to find factors that are in some sense close to a set of a priori factors. If closeness is measured by the L_1 or L_2 norm and if the constraint on the positivity of the factor loadings is omitted, then this problem can be solved using the techniques of Chapters 7 and 12. One advantage of this latter approach is that it permits one to test whether a particular set of a priori factors can be factors of the problem (that is, whether or not the distance between a priori factors and actual factors can be reduced to an insignificant amount).

10.3 *Q*-Mode and *R*-Mode Factor Analysis

In addition to the normalization of the sample matrix \mathbf{S} described above (which is based on precision of measurement), several other kinds of normalizations are commonly employed. One, given the name *Q mode*, is used to counter the common problem that measured data do not always have a unit L_1 norm because some of their components are not included in the analysis. The individual samples are therefore normalized before any calculations are made, which has the effect of making the factors dependent on only the ratios of the observed elements. They are usually normalized to unit L_2 norm, however; the eigenvectors of the singular-value decomposition then have the physical interpretation of extremizing the sum of squared cosines of *angles* between the samples, instead of just extremizing the dot products (see Section 10.1).

Another type of normalization is used when the data possess only small variability. It is then appropriate to remove the mean sample from the data before performing the factor analysis so that perturbations about the mean sample are directly analyzed. When this kind of normalization is applied, the name *R-mode analysis* is used.

10.4 Empirical Orthogonal Function Analysis

Factor analysis need not be limited to data that contain actual mixtures of components. Given any set of vectors \mathbf{s}_i , one can perform the singular-value decomposition and represent \mathbf{s}_i as a linear combina-

tion of a set of orthogonal factors. Even when the factors have no obvious physical interpretation, the decomposition can be useful as a tool for quantifying the similarities between the s_i vectors. This kind of factor analysis is often called *empirical orthogonal function analysis*.

As an example of this application of factor analysis, consider the set of $N = 14$ shapes shown in Fig. 10.5. These shapes might represent profiles of mountains or other subjects of interest. The problem we shall consider is how these profiles might be ordered to bring out the similarities and differences between the shapes. A geomorphologist might desire such an ordering because, when combined with other kinds of geological information, it might reveal the kinds of erosional processes that cause the shape of mountains to evolve with time.

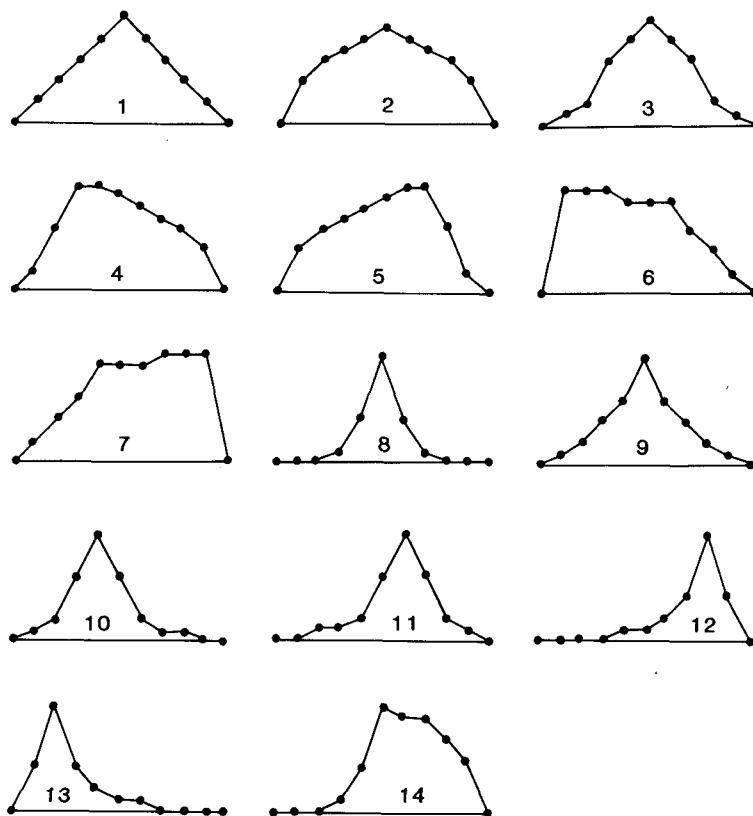


Fig. 10.5. A set of hypothetical mountain profiles. The variability of shape will be determined using factor analysis.

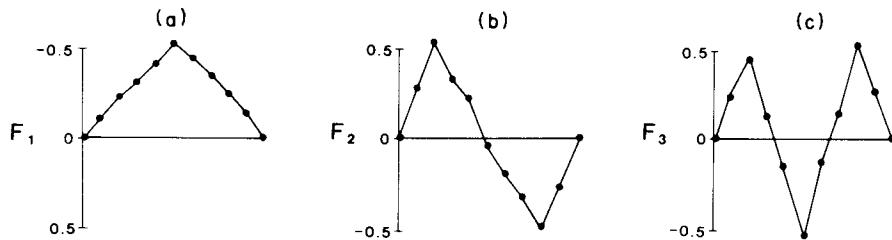


Fig. 10.6. The three largest factors F_i in the representation of the profiles in Fig. 10.5, with their corresponding singular values λ_i . (a) The first factor represents an average mountain profile, with $\lambda_1 = 3.29$; (b) the second, the asymmetry of the mountain, with $\lambda_2 = 1.19$; (c) and the third, the sharpness of the mountain's summit, with $\lambda_3 = 1.01$.

We begin by discretizing each profile and representing it as a unit vector (in this case of length $M = 11$). These unit vectors make up the matrix S , on which we perform factor analysis. Since the factors do not represent any particular physical object, there is no need to impose any positivity constraints on them, and we use the untransformed singular-value decomposition factors. The three most important factors (that is, the ones with the three largest singular values) are shown in Fig. 10.6. The first factor, as expected, appears to be simply an “average” mountain; the second seems to control the skewness, or

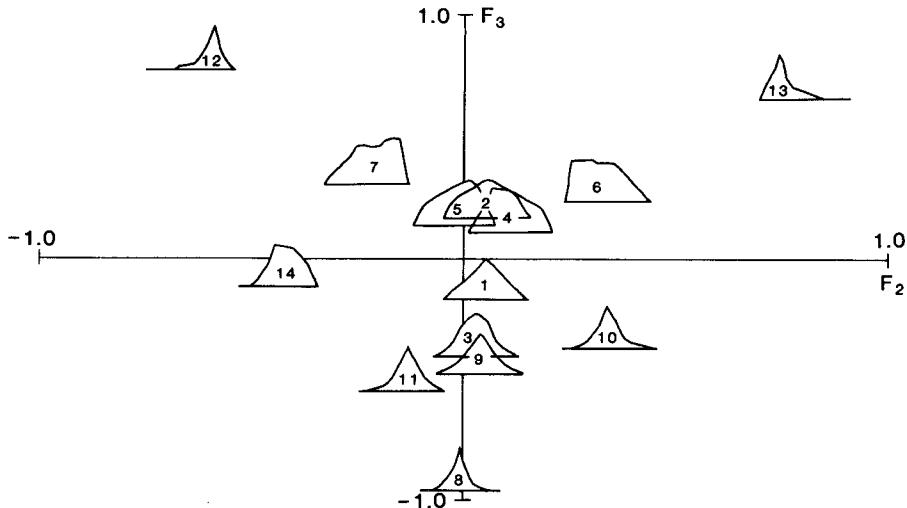


Fig. 10.7. The mountain profiles of Fig. 10.5 arranged according to the relative amounts of factors 2 and 3 contained in each profile's orthogonal decomposition.

degree of asymmetry, of the mountain; and the third, the sharpness of the mountain's summit. We emphasize, however, that this interpretation was made after the factor analysis and was not based on any a priori notions of how mountains might differ. We can then use the factor loadings as a measure of the similarities between the mountains. Since the amount of the first factor does not vary much between mountains, we use a two-dimensional ordering based on the relative amounts of the second and third factors in each of the mountain profiles (Fig. 10.7).

11

CONTINUOUS INVERSE THEORY AND TOMOGRAPHY

11.1 The Backus–Gilbert Inverse Problem

While continuous inverse problems are not the main subject of this book, we will cover them briefly to illustrate their relationship to discrete problems. Discrete and continuous inverse problems differ in their assumptions about the model parameters. Whereas the model parameters are treated as a finite-length vector in discrete inverse theory, they are treated as a continuous function in continuous inverse theory. The standard form of the continuous inverse problem is

$$d_i = \int_a^b G_i(z)m(z) dz \quad (11.1)$$

when the model function $m(z)$ varies only with one parameter, such as depth z . When the model function depends on several variables, then Eq. (11.1) must be generalized to

$$d_i = \int_V G_i(\mathbf{x})m(\mathbf{x}) dV_x \quad (11.2)$$

where dV_x is the volume element in the space of \mathbf{x} .

The “solution” of a discrete problem can be viewed as either an estimate of the model parameter vector \mathbf{m}^{est} or a series of weighted averages of the model parameters, $\mathbf{m}^{\text{avg}} = \mathbf{R}\mathbf{m}^{\text{true}}$, where \mathbf{R} is the resolution matrix (see Section 4.3). If the discrete inverse problem is very underdetermined, then the interpretation of the solution in terms of weighted averages is most sensible, since a single model parameter is

very poorly resolved. Continuous inverse problems can be viewed as the limit of discrete inverse problems as the number of model parameters becomes infinite, and they are inherently underdetermined. Attempts to estimate the model function $m(\mathbf{x})$ at a specific point $\mathbf{x} = \mathbf{x}_0$ are futile. All determinations of the model function must be made in terms of local averages, which are simple generalizations of the discrete case, $m_i^{\text{avg}} = \sum_j G_{ij}^{-g} d_j = \sum_j R_{ij} m_j^{\text{true}}$ where $R_{ij} = \sum_k G_{ik}^{-g} G_{kj}$:

$$m^{\text{avg}}(\mathbf{x}_0) = \sum_{i=1}^N G_i^{-g}(\mathbf{x}_0) d_i = \int_V R(\mathbf{x}_0, \mathbf{x}) m^{\text{true}}(\mathbf{x}) dV_x \quad (11.3)$$

where

$$R(\mathbf{x}_0, \mathbf{x}) = \sum_{i=1}^N G_i^{-g}(\mathbf{x}_0) G_i(\mathbf{x})$$

Here $G_j^{-g}(\mathbf{x}_0)$ is the continuous analogy to the generalized inverse G_{ij}^{-g} and the averaging function $R(\mathbf{x}_0, \mathbf{x})$ (often called the resolving kernel) is the analogy to R_{ij} . The average is localized near the target point \mathbf{x}_0 if the resolving kernel is peaked near \mathbf{x}_0 . The solution of the continuous inverse problem involves constructing the most peaked resolving kernel possible with a given set of measurements (that is, with a given set of data kernels, $G_i(\mathbf{x})$). The spread of the resolution function is quantified by [compare with Eq. (4.23)]:

$$J(\mathbf{x}_0) = \int_V w(\mathbf{x}_0, \mathbf{x}) R^2(\mathbf{x}_0, \mathbf{x}) dV_x \quad (11.4)$$

Here $w(\mathbf{x}_0, \mathbf{x})$ is a nonnegative function that is zero at the point \mathbf{x}_0 and that grows monotonically away from that point. One commonly used choice is the quadratic function $w(\mathbf{x}_0, \mathbf{x}) = |\mathbf{x}_0 - \mathbf{x}|^2$. Other, more complicated functions can be meaningful if the elements of \mathbf{x} have an interpretation other than spatial position. After inserting the definition of the resolving kernel [Eq. (11.3), second line] into the definition of the spread [Eq. (11.4)], we find

$$\begin{aligned} J(\mathbf{x}_0) &= \int_V w(\mathbf{x}_0, \mathbf{x}) R(\mathbf{x}_0, \mathbf{x}) R(\mathbf{x}_0, \mathbf{x}) dV_x \\ &= \int_V w(\mathbf{x}_0, \mathbf{x}) \sum_{i=1}^N G_i^{-g}(\mathbf{x}_0) G_i(\mathbf{x}) \sum_{j=1}^N G_j^{-g}(\mathbf{x}_0) G_j(\mathbf{x}) dV_x \\ &= \sum_{i=1}^N \sum_{j=1}^N G_i^{-g}(\mathbf{x}_0) G_j^{-g}(\mathbf{x}_0) [S_{ij}(\mathbf{x}_0)] \end{aligned} \quad (11.5)$$

where

$$[S_{ij}(\mathbf{x}_0)] = \int_V w(\mathbf{x}_0, \mathbf{x}) G_i(\mathbf{x}) G_j(\mathbf{x}) dV_x \quad (11.6)$$

The continuous spread function has now been manipulated into a form completely analogous to the discrete spread function in Eq. (4.24). The generalized inverse that minimizes the spread of the resolution is the precise analogy of Eq. (4.32).

$$G_k^{-g}(\mathbf{x}_0) = \frac{\sum_{i=1}^N [S_{ik}(\mathbf{x}_0)]^{-1} u_i}{\sum_{i=1}^N \sum_{j=1}^N [S_{ij}(\mathbf{x}_0)]^{-1} u_i u_j} \quad \text{where } u_i = \int_V G_i(\mathbf{x}) dV_x \quad (11.7)$$

11.2 Resolution and Variance Trade-Off

Since the data \mathbf{d} are determined only up to some error quantified by the covariance matrix $[\text{cov } \mathbf{d}]$, the localized average $m^{\text{avg}}(\mathbf{x}_0)$ is determined up to some corresponding error

$$\text{var}[m^{\text{avg}}(\mathbf{x}_0)] = \sum_{i=1}^N \sum_{j=1}^N G_i^{-g}(\mathbf{x}_0) [\text{cov } \mathbf{d}]_{ij} G_j^{-g}(\mathbf{x}_0) \quad (11.8)$$

As in the discrete case, the generalized inverse that minimizes the spread of resolution may lead to a localized average with large error bounds. A slightly less localized average may be desirable because it may have much less error. This generalized inverse may be found by minimizing a weighted average of the spread of resolution and size of variance

$$\text{minimize } \mathbf{J}'(\mathbf{x}_0) = \alpha \int_V w(\mathbf{x}_0, \mathbf{x}) R^2(\mathbf{x}_0, \mathbf{x}) dV_x + (1 - \alpha) \text{var}[m^{\text{avg}}(\mathbf{x}_0)] \quad (11.9)$$

The parameter α , which varies between 0 and 1, quantifies the relative weight given to spread of resolution and size of variance. As in the discrete case (see Section 4.10), the corresponding generalized inverse can be found by using Eq. (11.7), where all instances of $[S_{ij}(\mathbf{x}_0)]$ are replaced by

$$[S'_{ij}(\mathbf{x}_0)] = \alpha \int_V w(\mathbf{x}_0, \mathbf{x}) G_i(\mathbf{x}) G_j(\mathbf{x}) dV_x + (1 - \alpha) [\text{cov } \mathbf{d}]_{ij} \quad (11.10)$$

Backus and Gilbert [Ref. 2] prove a number of important properties of the trade-off curve (Fig. 11.1), which is a plot of size of variance against spread of resolution, including the fact that the variance decreases monotonically with spread.

11.3 Approximating Continuous Inverse Problems as Discrete Problems

A continuous inverse problem can be converted into a discrete one with the assumption that the model function can be represented by a finite number M of coefficients, that is,

$$m(\mathbf{x}) \approx \sum_{j=1}^M m_j f_j(\mathbf{x}) \quad (11.11)$$

The type of discretization of the model determines the choice of the functions $f_i(\mathbf{x})$. One commonly used assumption is that the model is constant within certain subregions V_i of V (e.g., box models). In this case $f_i(\mathbf{x})$ is unity inside V_i and zero outside it (step functions in the one-dimensional case), and m_i is the value of the model in each subregion. Many other choices of $f_i(\mathbf{x})$ are encountered, based on polynomial approximations, splines, and truncated Fourier series representations of $m(\mathbf{x})$.

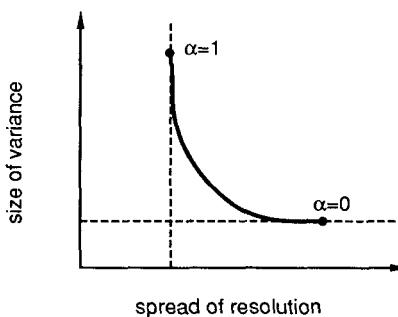


Fig. 11.1. Typical trade-off of resolution and variance for a linear continuous inverse problem. Note that the size (spread) function decreases monotonically with spread and that it is tangent to two asymptotes at the endpoints $\alpha = 0$ and $\alpha = 1$.

Inserting Eq. (11.11) into the canonical continuous problem [Eq. (11.2)] leads to a discrete problem

$$\begin{aligned} d_i &= \int_V G_i(\mathbf{x}) \sum_{j=1}^M m_j f_j(\mathbf{x}) \, dV_x \\ &= \sum_{j=1}^M \left[\int_V G_i(\mathbf{x}) f_j(\mathbf{x}) \, dV_x \right] m_j = \sum_{j=1}^M G_{ij} m_j \end{aligned} \quad (11.12)$$

where the discrete data kernel is

$$G_{ij} = \int_V G_i(\mathbf{x}) f_j(\mathbf{x}) \, dV_x \quad (11.13)$$

In the special case of the model function presumed to be constant in subregions V_i of V (centered, say, on the points \mathbf{x}_i), Eq. (11.13) becomes

$$G_{ij} = \int_{V_j} G_i(\mathbf{x}) \, dV_x \quad (11.14)$$

If the data kernel varies slowly with position so that it is approximately constant in the subregion, then we may approximate the integral as just its integrand evaluated at \mathbf{x}_j times the volume V_j :

$$G_{ij} = \int_{V_j} G_i(\mathbf{x}) \, dV_x \approx G_i(\mathbf{x}_j) V_j \quad (11.15)$$

Two different factors control the choice of the size of the subregions. The first is dictated by an *a priori* assumption of the smoothness of the model. The second becomes important only when one uses Eq. (11.15) in preference to Eq. (11.14). Then the subregion must be small enough that the data kernels $G_i(\mathbf{x})$ are approximately constant in the subregion. This second requirement often forces the subregions to be much smaller than dictated by the first requirement, so Eq. (11.14) should be used whenever the data kernels can be analytically integrated. Equation (11.15) fails completely whenever a data kernel has an integrable singularity within the subregion. This case commonly arises in problems involving the use of seismic rays to determine acoustic velocity structure.

11.4 Tomography and Continuous Inverse Theory

The term “tomography” has come to be used in geophysics almost synonymously with the term “inverse theory.” Tomography is derived from the Greek word *tomos*, that is, slice, and denotes forming an image of an object from measurements made from slices (or rays) through it. We consider tomography a subset of inverse theory, distinguished by a special form of the data kernel that involves measurements made along rays. The model function in tomography is a function of two or more variables and is related to the data by

$$d_i = \int_{C_i} m[x(s), y(s)] ds \quad (11.16)$$

Here the model function is integrated along a ray C_i having arc length s . This integral is equivalent to the one in a standard continuous problem [Eq. (11.2)] when the data kernel is $G_i(x, y) = \delta\{x(s) - x_i[y(s)]\} ds/dy$, where $\delta(x)$ is the Dirac delta function:

$$\begin{aligned} d_i &= \int \int m(x, y) \delta\{x(s) - x_i[y(s)]\} \frac{ds}{dy} dx dy \\ &= \int_{C_i} m[x(s), y(s)] ds \end{aligned} \quad (11.17)$$

Here x is supposed to vary with y along the curve C_i , and y is supposed to vary with arc length s .

While the tomography problem is a special case of a continuous inverse problem, several factors limit the applicability of the formulas of the previous sections. First, the Dirac delta functions in the data kernel are not square integrable, so that the S_{ij} (“overlap” integrals; see Eq. (11.6)] have nonintegrable singularities at points where rays intersect. Furthermore, in three-dimensional cases the rays may not intersect at all, so that all the S_{ij} may be identically zero. Neither of these problems is insurmountable, and they can be overcome by replacing the rays with tubes of finite cross-sectional width. (Rays are often an idealization of a finite-width process anyway, as in acoustic wave propagation, where they are an infinitesimal wavelength approximation.) Since this approximation is equivalent to some statement about the smoothness of the model function $m(x, y)$, it often suffices to discretize the continuous problem by dividing it into constant m subregions, where the subregions are large enough to guarantee a reasonable num-

ber containing more than one ray. The discrete inverse problem is then of the form $d_i = \sum_j G_{ij} m_j$, where the data kernel G_{ij} gives the arc length of the i th ray in the j th subregion. The concepts of resolution and variance, now interpreted in the discrete fashion of Chapter 4, are still applicable and of considerable importance.

11.5 Tomography and the Radon Transform

The simplest tomography problem involves straight-line rays and a two-dimensional model function $m(x,y)$ and is called Radon's problem. By historical convention, the straight-line rays C_i in Eq. (11.6) are parametrized by their perpendicular distance u from the origin and the angle θ (Fig. 11.2) that the perpendicular makes with the x axis. Position (x,y) and ray coordinates (u,s) , where s is arc length, are related by

$$\begin{aligned}(x) &= \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} u \\ s \end{pmatrix} \\(u) &= \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}\end{aligned}\quad (11.18)$$

The tomography problem is then

$$d(u,\theta) = \int_{-\infty}^{+\infty} m(x = u \cos \theta - s \sin \theta, y = u \sin \theta + s \cos \theta) ds \quad (11.19)$$

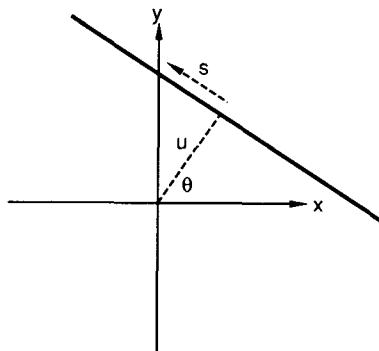


Fig. 11.2. The Radon transform is performed by integrating a function of (x,y) along straight lines (bold) parameterized by their arc length s , perpendicular distance u , and angle θ .

In realistic experiments, $d(u,\theta)$ is sampled only at discrete points $d_i = d(u_i, \theta_i)$. Nevertheless, much insight can be gained into the behavior of Eq. (11.19) by regarding (u,θ) as continuous variables. Equation (11.19) is then an integral transform that transforms variables (x,y) to two new variables (u,θ) and is called a Radon transform.

11.6 The Fourier Slice Theorem

The Radon transform is similar to another integral transform, the Fourier transform, which transforms spatial position x to spatial wave number k_x :

$$\begin{aligned}\hat{f}(k_x) &= \int_{-\infty}^{+\infty} f(x) \exp(ik_x x) dx \quad \text{and} \\ f(x) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{f}(k_x) \exp(-ik_x x) dx\end{aligned}\quad (11.20)$$

In fact, the two are quite closely related, as can be seen by Fourier transforming Eq. (11.19) with respect to $u \rightarrow k_u$:

$$\begin{aligned}\hat{d}(k_u, \theta) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} m(x = u \cos \theta - s \sin \theta, y \\ &\quad = u \sin \theta + s \cos \theta) ds \exp(ik_u u) du\end{aligned}\quad (11.21)$$

We now transform the double integral from $ds du$ to $dx dy$, using the fact that the Jacobian determinant $|\partial(x,y)/\partial(u,s)|$ is unity [see Eq. (11.18)]

$$\begin{aligned}\hat{d}(k_u, \theta) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} m(x, y) \exp(ik_u \cos \theta x + ik_u \sin \theta y) du ds \\ &= \hat{m}(k_x = k_u \cos \theta, k_y = k_u \sin \theta)\end{aligned}\quad (11.22)$$

This result, called the Fourier slice theorem, provides a method of inverting the Radon transform. The Fourier-transformed quantity $\hat{d}(k_u, \theta)$ is simply the Fourier-transformed image $\hat{m}(k_x, k_y)$ evaluated along radial lines in the (k_x, k_y) plane (Fig. 11.3). If the Radon transform is known for all values of (u, θ) , then the Fourier-transformed image is known for all (k_x, k_y) . The space domain image $m(x, y)$ is found by taking the inverse Fourier transform.

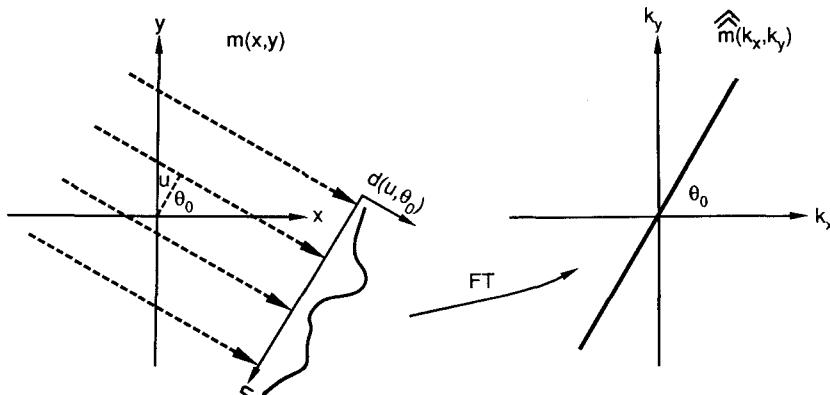


Fig. 11.3. (Left) The function $m(x,y)$ is integrated along a set of parallel lines (dashed) in a Radon transform to form the function $d(u,\theta_0)$. This function is called the projection of m at angle θ_0 . (Right) The Fourier slice theorem states that the Fourier transform of the projection (here denoted FT) is equal to the Fourier-transformed image $\hat{m}(k_x,k_y)$ evaluated along a line (bold) in the (k_x,k_y) plane.

Since the Fourier transform and its inverse are unique, the Radon transform can be uniquely inverted if it is known for all possible (u,θ) . Furthermore, the Fourier slice theorem can be used to invert the Radon transform in practice by using discrete Fourier transforms in place of integral Fourier transforms. However, u must be sampled sufficiently evenly that the $u \rightarrow k_u$ transform can be performed and θ must be sampled sufficiently finely that $\hat{m}(k_x,k_y)$ can be sensibly interpolated onto a rectangular grid of (k_x,k_y) to allow the $k_x \rightarrow x$ and $k_y \rightarrow y$ transforms to be performed (Fig. 11.4).

11.7 Backprojection

While the Fourier slice theorem provides a method of solving tomography problems, it relies on the rays being straight lines and the measurements being available for a complete range of ray positions and orientations. These two requirements are seldom satisfied in geophysical tomography, where the rays (for example, the ray paths of seismic waves) are curved and the data coverage is poor (for example, determined by the natural occurrence of earthquake sources). In these cases, the Fourier slice theorem in particular—and the theory of integral transforms in general—has limited applicability, and the tomography

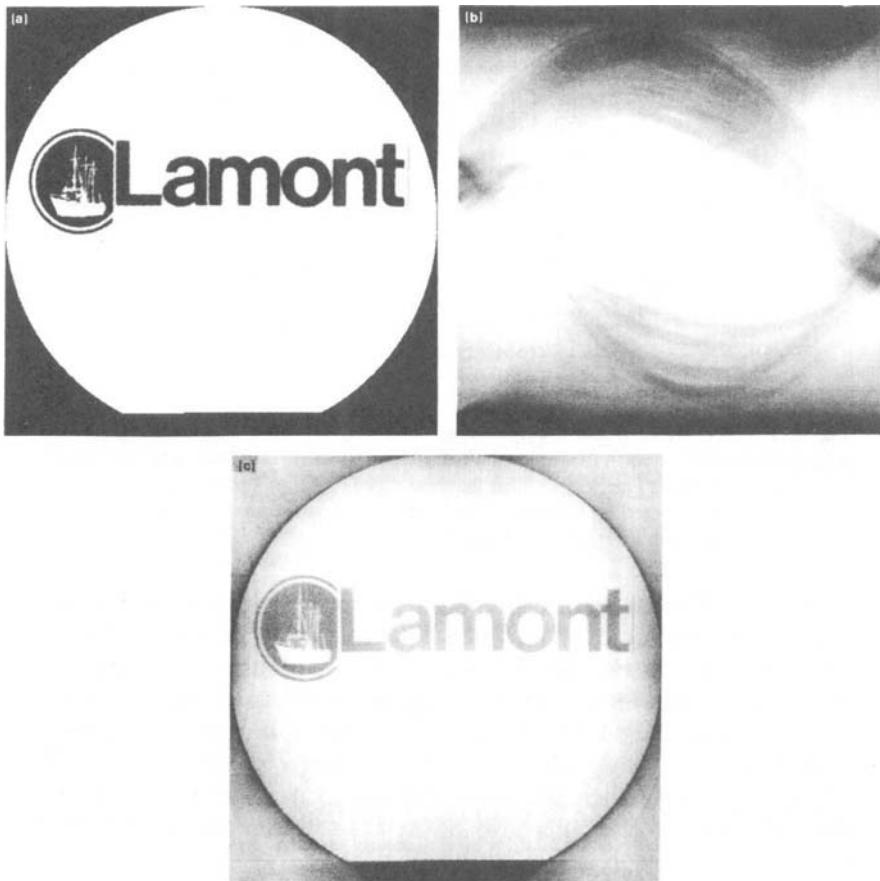


Fig. 11.4. (a) A test image consisting of 262,144 discrete values, or pixels, of (x,y) . (b) The Radon transform of the image in (a). The parameter θ increases from left to right and the parameter u increases from bottom to top. This function has also been evaluated at 262,144 combinations of (u,θ) . (c) The image reconstructed from its Radon transform by direct application of the Fourier slice theorem. Small errors in the reconstruction arise from the interpolation of the Fourier-transformed image to a rectangular grid.

problem must be solved with direct application of the general principles of inverse theory.

With the assumption that the model function $m(x,y)$ can be approximated as having the constant value m_j in a set of M subregions V_j , the discrete inverse theory formulation of the tomography problem be-

comes $d_i = \sum_j G_{ij} m_j$, where the data kernel G_{ij} gives the arc length of the i th ray in the j th subregion. In a typical tomography problem, there may be tens of thousands of subregions ($M = 10^4$) and millions of rays ($N = 10^6$), so the data kernel is a very large matrix. (The data kernel is, however, sparse, since a given ray will not traverse very many of the M subregions.) The solution of the problem directly by, say, the least squares formula $\mathbf{m} = [\mathbf{G}^T \mathbf{G}]^{-1} \mathbf{G}^T \mathbf{d}$ may be impractical, owing to the difficulty of directly inverting the large $M \times M$ matrix $[\mathbf{G}^T \mathbf{G}]$. One alternative to direct inversion is “backprojection,” an iterative method for solving a system of linear equations that is similar to the well-known Gauss–Seidel and Jacobi methods of linear algebra.

We begin by writing the data kernel as $G_{ij} = h_i F_{ij}$, where h_i is the length of the i th ray and F_{ij} is the fractional length of the i th ray in the j th subregion. The basic equation of the tomography problem is then $\sum_j F_{ij} m_j = d'_i$, where $d'_i = d_i/h_i$. The least squares solution is

$$\begin{aligned} [\mathbf{F}^T \mathbf{F}] \mathbf{m} &= \mathbf{F}^T \mathbf{d}' \\ [\mathbf{I} - \mathbf{I} + \mathbf{F}^T \mathbf{F}] \mathbf{m} &= \mathbf{F}^T \mathbf{d}' \\ \mathbf{m}^{\text{est}} &= \mathbf{F}^T \mathbf{d}' - [\mathbf{I} - \mathbf{F}^T \mathbf{F}] \mathbf{m}^{\text{est}} \end{aligned} \quad (11.23)$$

Note that we have both added and subtracted the identity matrix in the second step. The trick is now to use this formula iteratively, with the previous estimate of the model parameters on the right-hand side of the equation:

$$\mathbf{m}^{\text{est}(i)} = \mathbf{F}^T \mathbf{d}' - [\mathbf{I} - \mathbf{F}^T \mathbf{F}] \mathbf{m}^{\text{est}(i-1)} \quad (11.24)$$

Here $\mathbf{m}^{\text{est}(i)}$ is the \mathbf{m}^{est} after the i th iteration. When the iteration is started with $\mathbf{m}^{\text{est}(0)} = 0$, then

$$m_j^{\text{est}(1)} = \sum_{i=1}^N \frac{F_{ij} d_i}{h_i} \quad (11.25)$$

This formula has a simple interpretation which earns it the name backprojection. In order to bring out this interpretation, we will suppose that this is an acoustic tomography problem, so that the model parameters correspond to acoustic slowness (reciprocal velocity) and the data to acoustic travel time. Suppose that there are only one ray ($N = 1$) and one subregion ($M = 1$). Then the slowness of this subregion would be estimated as $m_1^{\text{est}} = d_1/h_1$, that is, as travel time divided by ray length. If there are several subregions, then the travel time is distributed equally (“backprojected”) among them, so that those with

the shortest ray lengths h_i are assigned the largest slowness. Finally, if there are several rays, a given model parameter's total slowness is just the sum of the estimates for the individual rays. This last step is quite unphysical and causes the estimates of the model parameters to grow with the number of rays. Remarkably, this problem introduces only long-wavelength errors into the image, so that a high-pass-filtered version of the image can often be quite useful. In many instances the first approximation in Eq. (11.24) is satisfactory, and further iterations need not be performed. In other instances, iterations are necessary.

A further complication arises in practical tomography problems when the ray path is itself a function of the model parameters (such as in acoustic tomography, where the acoustic velocity structure determines the ray paths). The procedure here is to write the model as a small perturbation $\delta m(x,y)$ about a trial or “background” $m(x,y) = m_0(x,y) + \delta m(x,y)$. The tomography problem for δm and rays based on m_0 is solved using standard methods, new rays are determined for the new structure, and the whole process is iterated. Little is known about the convergence properties of such an iteration, except perhaps when there is an underlying reason to believe that the data are relatively insensitive to small errors in the model (such as is given by Fermat's principle in acoustics).

12

SAMPLE INVERSE PROBLEMS

12.1 An Image Enhancement Problem

Suppose that a camera moves slightly during the exposure of a photograph (so that the picture is blurred) and that the amount and direction of motion are known. Can the photograph be “unblurred”?

For simplicity we shall consider a one-dimensional camera (that is, one that records brightness along a line) and assume that the camera moves parallel to the line. The camera is supposed to be of the digital type; it consists of a row of light-sensitive elements that measure the total amount of light received during the exposure. The data d_i are a set of numbers that represent the amount of light recorded at each of N camera “pixels.” Because the scene’s brightness varies continuously along the line of the camera, this is properly a problem in continuous inverse theory. We shall discretize it, however, by assuming that the scene can be adequately approximated by a row of small square elements, each with a constant brightness. These elements form M model parameters m_i . Since the camera’s motion is known, it is

possible to calculate each scene element's relative contribution to the light recorded at a given camera pixel. For instance, if the camera moves through three scene elements during the exposure, then each camera element records the average of three neighboring scene brightnesses (Fig. 12.1):

$$d_i = m_{i-1}/3 + m_i/3 + m_{i+1}/3 \quad (12.1)$$

Note that this is a linear equation and can be written in the form $\mathbf{Gm} = \mathbf{d}$, where

$$\mathbf{G} = \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 1 & 1 & 0 & \cdots & 0 \\ & & & \ddots & & & \\ & & & & \ddots & & \\ 0 & \cdots & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \quad (12.2)$$

In general, there will be several more model parameters than data, so the problem will be underdetermined. In this case $M = N + 2$, so there will be at least two null vectors. In fact, the problem is purely underdetermined, and there are only two null vectors, which can be identified by inspection as

$$\begin{aligned} \mathbf{m}_1^{\text{null}} &= [\frac{1}{3}, 0, -\frac{1}{3}, \frac{1}{3}, 0, -\frac{1}{3}, \dots, \frac{1}{3}, 0, -\frac{1}{3}]^T \\ \mathbf{m}_2^{\text{null}} &= [0, \frac{1}{3}, -\frac{1}{3}, 0, \frac{1}{3}, -\frac{1}{3}, \dots, 0, \frac{1}{3}, -\frac{1}{3}]^T \end{aligned} \quad (12.3)$$

These null vectors have rapidly fluctuating elements, which indicates that at best only the longer wavelength features can be recovered. To find a solution to the inverse problem, we must add a priori information. We shall use a simplicity constraint and find the scene of shortest

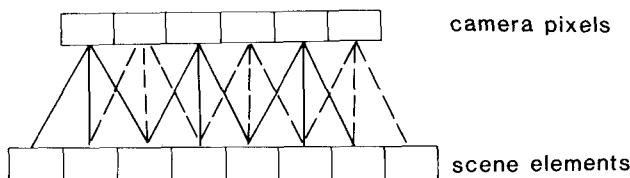


Fig. 12.1. When the blur is three pixels wide, each camera pixel records the average brightness of three neighboring scene elements. There are two more unknowns than data.

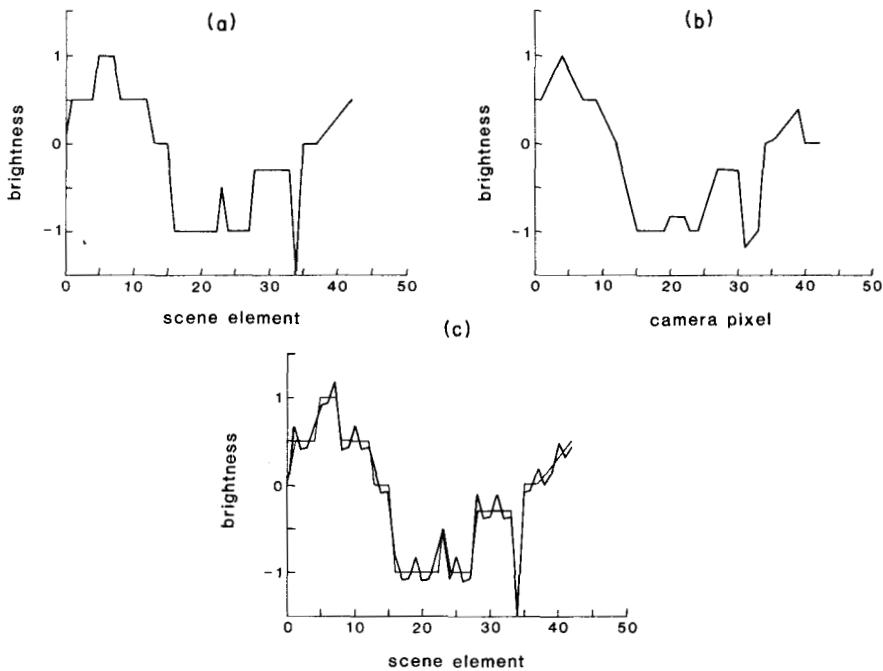


Fig. 12.2. (a) Actual scene (true model); (b) blurred photograph of (a) (synthetic data); (c) recovered scene (jagged curve, estimated model) compared with true scene. The estimated model differs from the true model because the null vectors of the system (which are jagged) have been incorrectly estimated.

length. If the data are normalized so that zero represents gray (with white negative and black positive), then this procedure in effect finds the scene of least contrast that fits the data. We therefore estimate the solution with the minimum length generalized inverse as

$$\mathbf{m}^{\text{est}} = \mathbf{G}^T [\mathbf{G}\mathbf{G}^T]^{-1} \mathbf{d} \quad (12.4)$$

As an example, we shall solve the problem for the data kernel given above (for a blur width of 3) with $M = 40$. We first select a true scene (Fig. 12.2a) and blur it by multiplication with the data kernel to produce synthetic data (Fig. 12.2b). Note that the blurred image is much smoother than the true scene. We now try to invert back for the true scene by premultiplying by the generalized inverse. The result (Fig. 12.2c) has correctly captured some of the sharp features in the original scene but also contains a short wavelength oscillation not

present in the true scene. This error results from creating a reconstructed scene containing an incorrect combination of null vectors.

In this example, the generalized inverse is computed by simply performing the given matrix multiplications on a computer and then inverting $[GG^T]$ by Gauss–Jordan reduction with partial pivoting (see Section 13.2). It is possible, however, to deduce the form of $[GG^T]$ analytically since \mathbf{G} has a simple structure. A typical row contains the sequence $\{ \dots, 1, 2, 3, 2, 1, \dots \}$. In this problem, which deals with small (40×42) matrices, the effort saved is negligible. In more complicated inverse problems, however, considerable savings can result from a careful analytical treatment of the structures of the various matrices.

Each row of the generalized inverse states how a particular model parameter is constructed from the data. We might expect that this blurred image problem would have a localized generalized inverse, meaning that each model parameter would be constructed mainly from a few neighboring data. By examining \mathbf{G}^{-g} , we see that this is not the case (Fig. 12.3). We also note that the process of blurring is a integrating process; it sums neighboring data. The process of unblurring should be a sort of differentiating process, subtracting neighboring data. The generalized inverse is, in fact, just that.

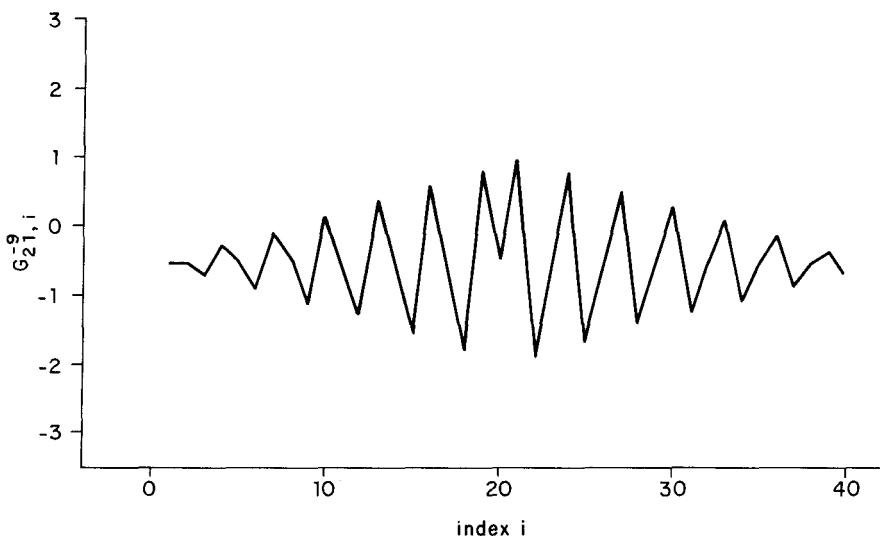


Fig. 12.3. Twenty-first row of the generalized inverse for the blurred image problem. Note that neighboring data are subtracted from one another to form the model estimate.

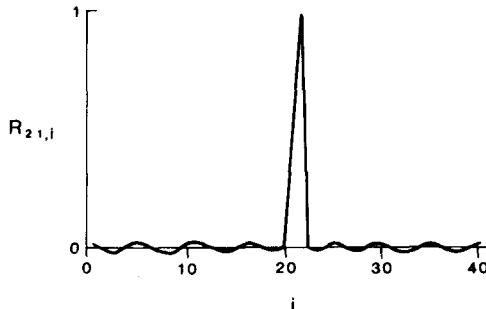


Fig. 12.4. Model resolution for twenty-first scene element. The maximum resolution is 0.98. The off-diagonal elements oscillate about ± 0.02 . This problem has large sidelobes, in the sense that the resolution does not decline with distance from the main diagonal.

The resolution matrix for this problem is seen to be quite “spiky” (Fig. 12.4); the diagonal elements are several orders of magnitude larger than the off-diagonal elements. On the other hand, the off-diagonal elements are all of uniform size, indicating that if the estimated model parameters are interpreted as localized averages, they are in fact not completely localized. It is interesting to note that the Backus–Gilbert inverse, which generally gives very localized resolution kernels, returns only the blurred image and a resolution with a peak width of 3. The solution to this problem contains an unavoidable trade-off between the width of resolution and the presence of sidelobes.

12.2 Digital Filter Design

Suppose that two signals $a(t)$ and $b(t)$ are known to be related by convolution with a filter $f(t)$:

$$a(t) = f(t) * b(t) = \int f(\tau)b(t - \tau) d\tau \quad (12.5)$$

where τ is a dummy integration variable. Can $f(t)$ be found if $a(t)$ and $b(t)$ are known?

Since the signals and filters are continuous functions, this is a problem in continuous inverse theory. We shall analyze it, however, by approximating the functions as *time series*. Each function will be represented by its value at a set of points spaced equally in time (with interval Δt). We shall assume that the signals are transient (have a definite beginning and end) so that $a(t)$ and $b(t)$ can be represented by

time series of length N . Typically, the advantage of relating two signals by a filter is realized only when the filter length M is shorter than either signal, so $M < N$ is assumed. The convolution integral can then be approximated by the sum

$$a_i = \Delta t \sum_{j=1}^M f_j b_{i-j+1} \quad (12.6)$$

where $b_i = 0$ if $i < 1$ or $i > N$. This equation is linear in the unknown filter coefficients and can be written in the form $\mathbf{Gm} = \mathbf{d}$, where

$$\mathbf{G} = \Delta t \begin{bmatrix} b_1 & 0 & 0 & \cdots & 0 \\ b_2 & b_1 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \ddots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ b_N & b_{N-1} & b_{N-2} & \cdots & b_{N-M+1} \end{bmatrix} \quad (12.7)$$

and the time series a_i is identified with the data and the filter f_i with the model parameters. The equation is therefore an overdetermined linear system for $M < N$ filter coefficients and may be solved using simple least squares:

$$\mathbf{m}^{\text{est}} = [\mathbf{G}^T \mathbf{G}]^{-1} \mathbf{G}^T \mathbf{d} \quad (12.8)$$

For this solution to be consistent with the tenets of probability theory, however, \mathbf{b} must be known exactly, while the a_i 's contain uncorrelated Gaussian noise of uniform variance.

While we can compute the solution by “brute force” multiplication of the matrices in the generalized inverse, analytic treatment greatly simplifies the computation. We restate the equation for the estimated model parameters as $[\mathbf{G}^T \mathbf{G}] \mathbf{m}^{\text{est}} = \mathbf{G}^T \mathbf{d}$ and note that

$$\mathbf{G}^T \mathbf{G} = (\Delta t)^2 \begin{bmatrix} \sum_{i=1}^N b_i^2 & \sum_{i=1}^{N-1} b_i b_{i-1} & \cdots \\ \sum_{i=1}^{N-1} b_i b_{i-1} & \sum_{i=1}^{N-2} b_i^2 & \cdots \\ \cdot & \cdot & \ddots \\ \cdot & \cdot & \cdot \end{bmatrix} \quad \mathbf{G}^T \mathbf{d} = \Delta t \begin{bmatrix} \sum_{i=1}^N a_i b_i \\ \sum_{i=1}^{N-1} a_i b_{i-1} \\ \cdot \\ \cdot \end{bmatrix} \quad (12.9)$$

where $\mathbf{G}^T \mathbf{G}$ contains the autocorrelation of \mathbf{b} and $\mathbf{G}^T \mathbf{d}$ the crosscorrelation of \mathbf{a} with \mathbf{b} . Furthermore, if the time series are assumed to be zero outside the interval $[1, N]$, then the upper limit on all the sums can be changed to N . Each diagonal of $\mathbf{G}^T \mathbf{G}$ then contains elements all of which are equal, so that only M autocorrelations and M crosscorrelations need be computed. The equation $[\mathbf{G}^T \mathbf{G}] \mathbf{m}^{\text{est}} = \mathbf{G}^T \mathbf{d}$ can then be solved by Gauss–Jordan reduction (see Section 13.1). It turns out, however, that the special form of $\mathbf{G}^T \mathbf{G}$ (*Toeplitz*, or symmetric with all equal diagonal elements) permits solution of these equations by an even faster method called *Levinson recursion*, the details of which we shall not discuss.

As an example of filter construction, we shall consider a time series $b(t)$, which represents a recording of the sound emitted by a seismic exploration airgun (Fig. 12.5a). Signals of this sort are used to detect layering at depth in the earth through echo sounding. In principle one wants a very spiky sound from the airgun so that echoes from layers at depth can be easily detected. Engineering constraints, however, limit the airgun signal to a series of pulses. We shall attempt, therefore, to find a filter that, when applied to the airgun pulse, produces a signal spike or delta function $\mathbf{a} = [0, 0, 0, \dots, 0, 1, 0, \dots, 0]^T$ centered on the largest pulse in the original signal (Fig. 12.5b). This filter can then be applied to the recorded echo soundings to remove the rever-

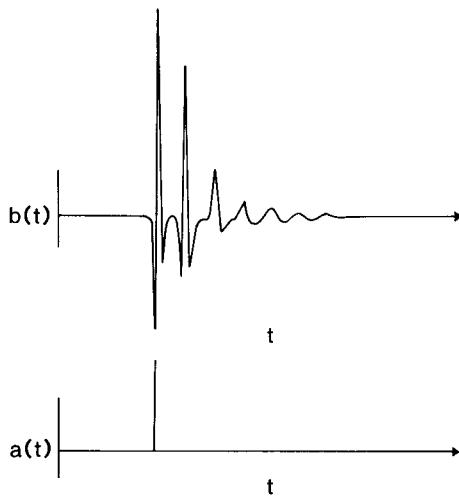


Fig. 12.5. The airgun signal $b(t)$ and delta function $a(t)$ are related by the convolution $a(t) = f(t) * b(t)$.

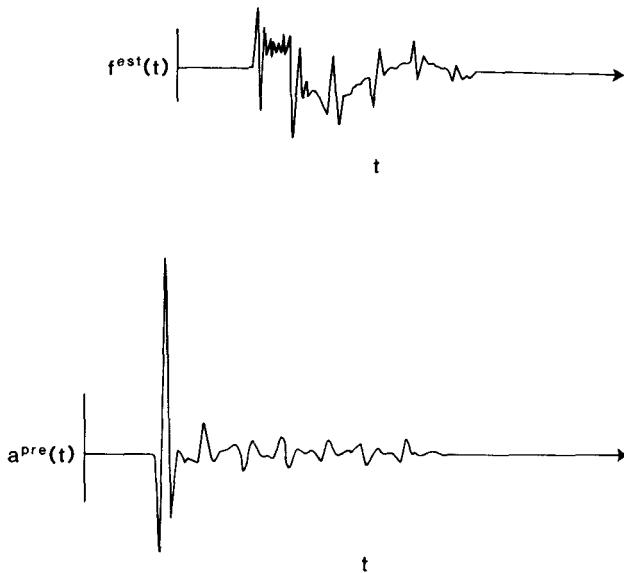


Fig. 12.6. (a) The least squares filter $f(t)$. (b) The predicted signal $a^{\text{pre}} = f^{\text{est}} * b$ is not exactly a delta function.

beration of the airgun and reveal the layering of the earth. The airgun signal has $N = 240$ samples. We choose a filter length of $M = 100$, which is shorter than the airgun pulse but longer than the time between its component reverberations. The least squares filter (computed for this example by Gauss–Jordan reduction) is shown in Fig. 12.6a and the resulting signal $a(t) = f(t) * b(t)$ in Fig. 12.6b. Note that, although the reverberations are reduced in amplitude, they are by no means completely removed.

12.3 Adjustment of Crossover Errors

Consider a set of radar altimetry data from a remote-sensing satellite such as SEASAT. These data consist of measurements of the distance from the satellite to the surface of the earth directly below the satellite. If the altitude of the satellite with respect to the earth's center were known, then these data could be used to measure the elevation of the surface of the earth. Unfortunately, while the height of the satellite

during each orbit is approximately constant, its exact value is unknown. Since the orbits criss-cross the earth, one can try to solve for the satellite height in each orbit by minimizing the overall crossover error [Ref. 12].

Suppose that there are M orbits and that the unknown altitude of the satellite during the i th orbit is m_i . We shall divide these orbits into two groups (Fig. 12.7), the ascending orbits (when the satellite is traveling north) and the descending orbits (when the satellite is traveling south). The ascending and descending orbits intersect at N points. At one such point ascending orbit number A_i intersects with descending orbit D_i (where the numbering refers to the ordering in \mathbf{m}). At this point the two orbits have measured a satellite-to-earth distance of, say, s_{A_i} and s_{D_i} , respectively. The elevation of the ground is $m_{A_i} - s_{A_i}$ according to the data collected on the ascending orbit and $m_{D_i} - s_{D_i}$ according to the data from the descending orbit. The crossover error at the i th intersection is $e_i = (m_{A_i} - s_{A_i}) - (m_{D_i} - s_{D_i}) = (m_{A_i} - m_{D_i}) - (s_{A_i} - s_{D_i})$. The assertion that the crossover error should be zero leads to a linear equation of the form $\mathbf{Gm} = \mathbf{d}$, where

$$\begin{aligned} G_{ij} &= \delta_{jA_i} - \delta_{jD_i} \\ d_i &= s_{A_i} - s_{D_i} \end{aligned} \quad (12.10)$$

Each row of the data kernel contains one 1, one -1 and $M - 2$ zeros.

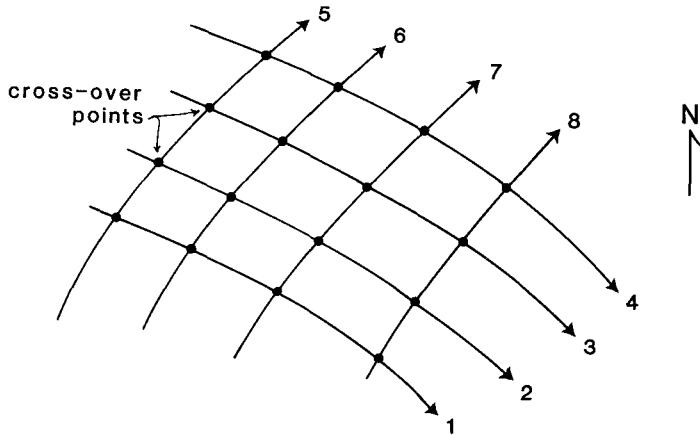


Fig. 12.7. Descending tracks 1–4 intersect ascending tracks 5–8 at 16 points. The height of the satellite along each track is determined by minimizing the cross-over error at the intersections.

At first sight we might assume that we can use simple least squares to solve this problem. We note, however, that the solution is always to a degree underdetermined. Any constant can be added to all the m_i 's without changing the crossover error since the error depends only on the difference between the elevations of the satellite during the different orbits. This problem is therefore mixed-determined. We shall handle this problem by imposing the a priori constraint that $\sum m_i = 0$. While this constraint is not physically realistic (implying as it does that the satellite has on average zero altitude), it serves to remove the underdeterminacy. Any desired constant can subsequently be added to the solution.

We shall implement this constrained least squares problem using the method of Lagrange multipliers in Section 3.10. We shall need to compute the matrices $\mathbf{G}^T \mathbf{G}$ and $\mathbf{G}^T \mathbf{d}$. In a realistic problem there may be several thousand orbits. The data kernel will therefore be very large, with dimensions on the order of $1,000,000 \times 1000$. Solving the problem by “brute force” calculation is impractical, and we must perform a careful analysis if we are to be able to solve the problem at all:

$$\begin{aligned} [\mathbf{G}^T \mathbf{G}]_{rs} &= \sum_{i=1}^N G_{ir} G_{is} \\ &= \sum_{i=1}^N [\delta_{rA_i} - \delta_{rD_i}][\delta_{sA_i} - \delta_{sD_i}] \\ &= \sum_{i=1}^N [\delta_{rA_i} \delta_{sA_i} - \delta_{rA_i} \delta_{sD_i} - \delta_{rD_i} \delta_{sA_i} + \delta_{rD_i} \delta_{sD_i}] \quad (12.11) \end{aligned}$$

The diagonal elements of $\mathbf{G}^T \mathbf{G}$ are

$$[\mathbf{G}^T \mathbf{G}]_{rr} = \sum_{i=1}^N [\delta_{rA_i} \delta_{rA_i} - 2\delta_{rA_i} \delta_{rD_i} + \delta_{rD_i} \delta_{rD_i}] \quad (12.12)$$

The first term contributes to the sum whenever the ascending orbit is r , and the third term contributes whenever the descending orbit is r . The second term is zero since an orbit never intersects itself. The r th element of the diagonal is the number of times the r th orbit is intersected by other orbits.

Only the two middle terms of the sum in the expression for $[\mathbf{G}^T \mathbf{G}]_{rs}$ contribute to the off-diagonal elements. The second term contributes whenever $A_i = r$ and $D_i = s$, and the third when $A_i = s$ and $D_i = r$. The

(r, s) off-diagonal element is the number of times the r th and s th orbits intersect, multiplied by -1 .

The other matrix product is

$$[\mathbf{G}^T \mathbf{d}]_r = \sum_{i=1}^N G_{ir} d_i = \sum_{i=1}^N [\delta_{rA_i} - \delta_{rD_i}] d_i \quad (12.13)$$

We note that the delta functions can never both equal 1 since an orbit can never intersect itself. Therefore $[\mathbf{G}^T \mathbf{d}]_r$ is the sum of all the d_i 's that have ascending orbit number $A_i = r$ minus the sum of all the d_i 's that have descending orbit number $D_i = r$.

We can then compute the matrix products. We first prepare a table that gives the ascending orbit number A_i , descending orbit number D_i , and measurement difference d_i for each of the N orbital intersections. We then start with $[\mathbf{G}^T \mathbf{G}]$ and $[\mathbf{G}^T \mathbf{d}]$ initialized to zero and, for each i th row of the table, execute the following steps:

- (1) Add 1 to the $r = A_i, s = A_i$ element of $[\mathbf{G}^T \mathbf{G}]_{rs}$.
- (2) Add 1 to the $r = D_i, s = D_i$ element of $[\mathbf{G}^T \mathbf{G}]_{rs}$.
- (3) Subtract 1 from the $r = A_i, s = D_i$ element of $[\mathbf{G}^T \mathbf{G}]_{rs}$.
- (4) Subtract 1 from the $r = D_i, s = A_i$ element of $[\mathbf{G}^T \mathbf{G}]_{rs}$.
- (5) Add d_i to the $r = A_i$ element of $[\mathbf{G}^T \mathbf{d}]_r$.
- (6) Subtract d_i from the $r = D_i$ element of $[\mathbf{G}^T \mathbf{d}]_r$.

The final form for $\mathbf{G}^T \mathbf{G}$ is relatively simple. If two orbits intersect at most once, then it will contain only zeros and ones on its off-diagonal elements. The solution is given by the solution to the $M + 1 \times M + 1$

TABLE 12.1

Solutions of a Sample Crossover-error Adjustment Showing True Solution, Least Squares Solution with Mean Constrained to Zero (CLS), and Damped Least Squares Solution (DLS) with Various Choices of the Damping Parameter σ^2

Type	σ^2	Solution									Reduction in error (%)
True	—	[1.00	-1.00	1.00	-1.00	1.00	-1.00	1.00	-1.00] ^T	—	
CLS	—	[1.05	-0.95	1.01	-1.00	0.80	-1.11	1.15	-0.96] ^T	98.4	
DLS	0.01	[1.05	-0.95	1.01	-1.00	0.80	-1.11	1.15	-0.96] ^T	98.4	
DLS	0.10	[1.04	-0.93	0.99	-0.98	0.78	-1.09	1.13	-0.94] ^T	98.3	
DLS	1.00	[0.85	-0.76	0.81	-0.80	0.64	-0.90	0.92	-0.78] ^T	94.4	

^a Damped least squares should be viewed as an approximate solution that can be formed without detailed knowledge of the underdeterminacy of the problem. Note that for some choices of the damping parameter it gives a solution close to the exact, constrained solution.

Lagrange multiplier equation

$$\begin{bmatrix} [\mathbf{G}^T \mathbf{G}] & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{m} \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{G}^T \mathbf{d} \\ 0 \end{bmatrix} \quad (12.14)$$

This system of equations is typically on the order of 1000×1000 in dimension. While it can be solved by Gauss–Jordan reduction, it is usually faster to solve it by one of the approximate, iterative techniques that do not require triangularizing the matrix (e.g., Gauss–Seidel iteration). The simple structure of the matrix makes such a method of solution particularly easy to implement.

For example, let us consider $M = 8$ orbits with $N = 16$ intersections (Fig. 12.7). The matrix product for this case is

$$\mathbf{G}^T \mathbf{G} = \begin{bmatrix} 4 & 0 & 0 & 0 & -1 & -1 & -1 & -1 \\ 0 & 4 & 0 & 0 & -1 & -1 & -1 & -1 \\ 0 & 0 & 4 & 0 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 4 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & 4 & 0 & 0 & 0 \\ -1 & -1 & -1 & -1 & 0 & 4 & 0 & 0 \\ -1 & -1 & -1 & -1 & 0 & 0 & 4 & 0 \\ -1 & -1 & -1 & -1 & 0 & 0 & 0 & 4 \end{bmatrix} \quad (12.15)$$

Some synthetic data were constructed for the case $\mathbf{m} = [1, -1, 1, -1, 1, -1, 1, -1]^T$ containing 5% Gaussian random noise. The inverse problem for \mathbf{m} was then solved by both the constrained least squares and damped least square methods (Table 12.1). The damped least squares method only approximately minimizes the error, but is slightly faster to compute.

12.4 An Acoustic Tomography Problem

An acoustic tomography problem was discussed previously in Section 1.1.3. Travel times d_i of sound are measured through the rows and columns of a square grid of bricks, each having a constant acoustic slowness m_i . The equation is therefore of the linear form $\mathbf{Gm} = \mathbf{d}$. For

instance, for $M = 16$ bricks of width h ,

$$\begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_8 \end{bmatrix} = h \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_{16} \end{bmatrix} \quad (12.16)$$

We note that if $M > 4$, then $M > N$ and the problem is underdetermined. But it is not purely underdetermined, as can be seen by adding all the rows of \mathbf{G} that correspond to measurements through rows of bricks and comparing this sum to the sum of all the rows of \mathbf{G} that correspond to measurements through columns of bricks. The two results are equal: $h[1, 1, 1, 1, \dots, 1]$. Physically, this equality corresponds to the fact that either sum has measured the travel time

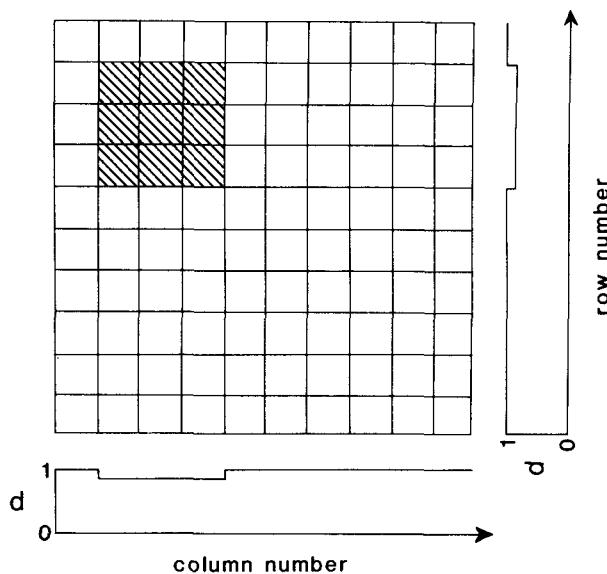


Fig. 12.8. Square array of 100 bricks, 91 of which have acoustic slowness $s = \frac{1}{10}$ and 9 of which (shaded) have $s = \frac{1}{12}$. The travel time d is measured along the rows and columns of the bricks (adjoining graphs).

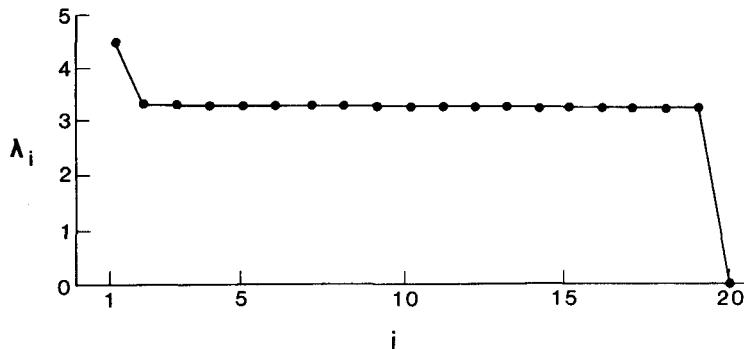


Fig. 12.9. Singular values λ_i for the acoustic tomography problem. Note that one singular value is zero, indicating that the problem is mixed-determined.

through the same complete set of bricks. The data kernel, therefore, contains redundant information.

Consider the $M = 100$ problem, where we have chosen a true model consisting of a 3×3 square of uniformly high (12) velocity bricks embedded near the upper left hand corner of a uniformly slow (10) velocity material (Fig. 12.8). The data kernel of this problem has $p = 19$ nonzero singular values, as computed by the method of Section 13.9 (Fig. 12.9), indicating that we have identified the only source of redundancy. To solve the problem we must add some a priori information. In this instance it is reasonable to find a model that is close to some a priori slowness (say, $\langle m_i \rangle = \frac{1}{10}$). The solution can then be

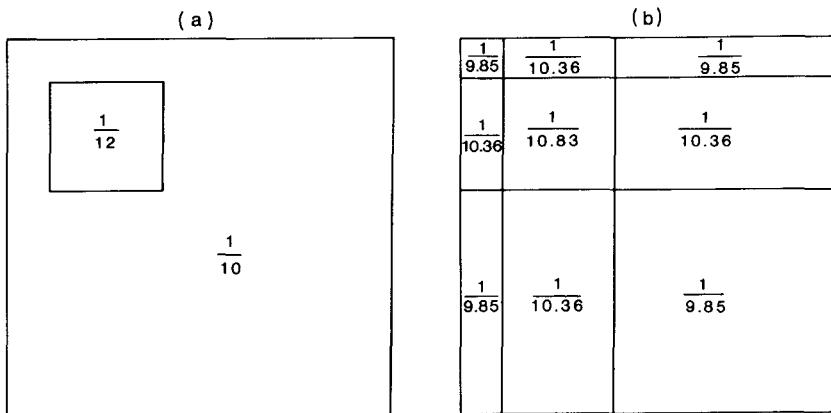


Fig. 12.10. (a) True model; (b) estimated model.

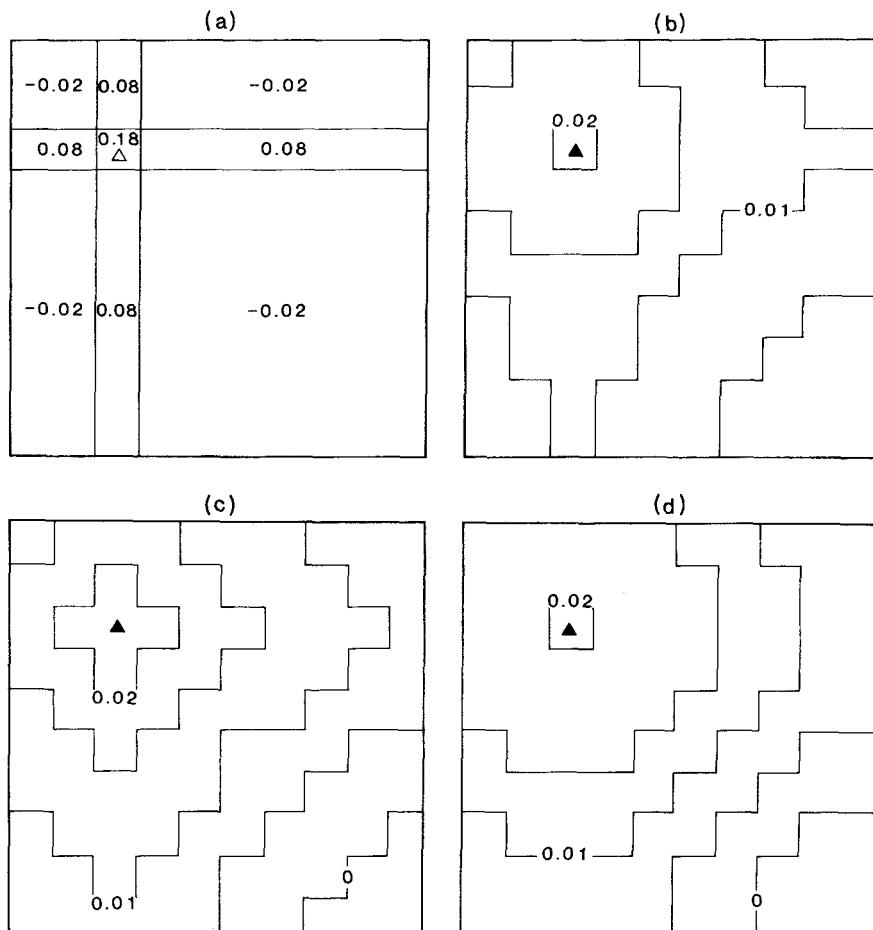


Fig. 12.11. Model resolution for brick in row 3, column 3 (triangle) using (a) singular-value decomposition generalized inverse, and (b-d) Backus–Gilbert generalized inverses with weighting functions based on the first, second, and fourth power of the Euclidian distance between model parameters, respectively. Note that, as the exponent of the weighting increases, the sidelobes decrease in magnitude, but the central peak becomes wider and wider.

constructed from the natural generalized inverse:

$$\mathbf{m}^{\text{est}} = \langle \mathbf{m} \rangle + \mathbf{G}^{-s}[\mathbf{d} - \mathbf{G}\langle \mathbf{m} \rangle]$$

The estimated solution (Fig. 12.10) recovers the main features of the true solution but stretches the high velocity box out along its rows and

columns. This behavior can be understood by examining the model resolution matrix \mathbf{R} . One row of this matrix is shown in Fig. 12.11a (arranged in a 10×10 grid for ease of interpretation). We note that the diagonal element has a value of only 0.18 and that elements that correspond to boxes on the same row and column also have values nearly as large (0.08). The geometry of measurements in this acoustic tomography problem is not especially suitable for resolving the exact location of a feature, so the resolution matrix has large sidelobes.

One might be tempted to try to reduce the sidelobes by using the Backus–Gilbert generalized inverse (Section 4.8), which weights off-diagonal elements of the resolution matrix preferentially. For instance, one might employ a weighting function that grows with some power of the Euclidian distance between the bricks. Since this problem is mixed–determined, some care must be taken to avoid singular matrices when computing the Backus–Gilbert generalized inverse. We have handled this problem by adding a small amount of the covariance size to the measure of goodness [Eq. (4.36)], which effectively damps the singularity. Unfortunately, while this method does reduce the amplitude of the most distant sidelobes (Fig. 12.11b–d), it severely widens the central part of the peak and the diagonal elements are of the order of only 0.02.

The fact that this inverse problem has inherently poor resolution cannot be circumvented by any application of inverse theory. To improve the resolution, one would have to make additional measurements of the travel time through the blocks, especially in the diagonal directions.

12.5 Temperature Distribution in an Igneous Intrusion

Suppose that a slab of igneous material (a dike) is injected into an infinite expanse of country rock of originally uniform temperature (Fig. 12.12). Can the temperature distribution within the slab be reconstructed by making measurements of the temperature in the surrounding rock?

We shall treat this problem as one in experimental design and attempt to determine the optimum placement of thermal sensors. For simplicity we shall assume that the country rock and dike both have a

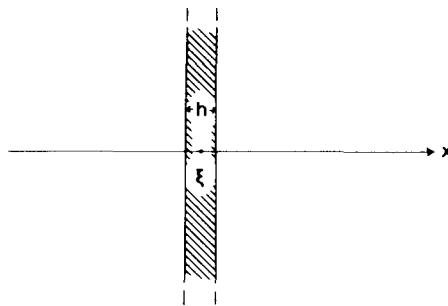


Fig. 12.12. Single dike of thickness h located at position $x = \xi$.

thermal diffusivity of unity and that the country rock initially has zero temperature. Since the temperature distribution in the dike is a continuous function, this is properly a problem in continuous inverse theory. We shall discretize it by assuming that the dike can be subdivided into M thinner dikes, each of thickness h and uniform initial temperature m_i . The temperature distribution in the country rock can then be shown to be

$$T(x, t) = \sum_{i=1}^M \frac{m_i}{2} \left\{ \operatorname{erf} \left[\frac{x - (\xi_i - h/2)}{t^{1/2}} \right] - \operatorname{erf} \left[\frac{x - (\xi_i + h/2)}{t^{1/2}} \right] \right\} \quad (12.17)$$

where the temperature T is measured at position x and time t after emplacement and the dikes are at position ξ_i , and where erf is the error function.

We shall first consider the problem of determining the temperature distribution with a single thermometer located a distance $h/2$ to the right of the rightmost dike in the case $M = 20$ (Fig. 12.13). If the

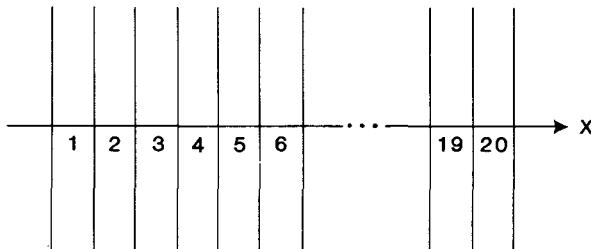


Fig. 12.13. A large dike is subdivided in 20 smaller ones to discretize this continuous inverse problem.

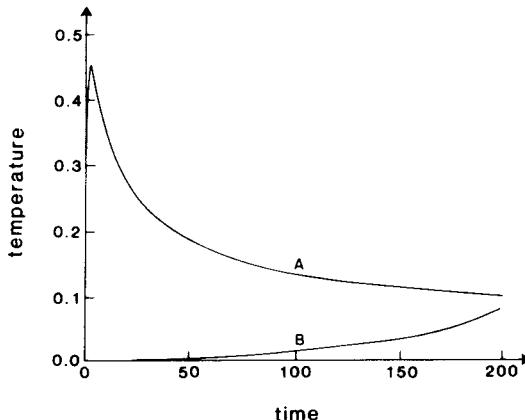


Fig. 12.14. Curve A is temperature observed near single dike ($x - \xi = 1$); curve B is temperature observed far from dike ($x - \xi = 20$).

temperature is measured at regular intervals, what should the sampling rate and total number of samples be to resolve the temperature of the dike? We shall approach this question by first computing the temperature time series for a single dike at various distances from the observation point (Fig. 12.14). It is clear that to record the shape of the temperature variation for nearby dikes the sampling interval must be on the order of $\Delta t = 1\text{s}$, and to record the shape of the most distant dikes the total duration must be about $t_{\max} = 1000\text{s}$. To refine these values, we first note that the model parameters and data $d_i = T(x, t_i)$ are related by a linear equation of the form $\mathbf{Gm} = \mathbf{d}$. We therefore

TABLE 12.2

Number of Significant Singular Values for Various Choices of Experimental Parameters

Δt	t_{\max}	No. of singular values about cutoff
1	10	3
1	50	6
1	100	7
1	500	8
1	1000	8
1	2000	8
0.1	50	6

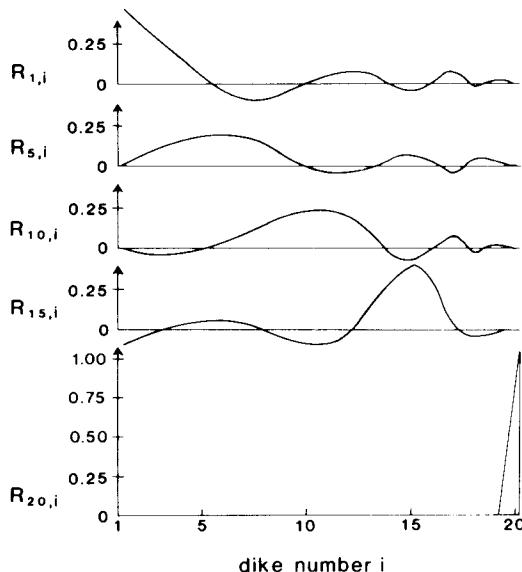


Fig. 12.15. Resolution for selected dikes when there is one observer, located a unit to the right of dike 20.

compute the singular values of the data kernel for a variety of Δt 's and t_{\max} 's. The strategy is to make as many singular values as large as possible, so that one can choose a relatively large p without amplifying too greatly the errors mapped from data to model parameters. Since variance is proportional to the squared reciprocal of the singular value, a cutoff of 10^{-3} of the largest singular value is a reasonable choice. Table 12.2 shows the results for a variety of choices.

Decreasing Δt below 1 does not add any new eigenvalues, nor does increasing t_{\max} past 500. We therefore choose $\Delta t = 1$, $t_{\max} = 500$ as a reasonable set of experimental parameters and we then examine the **model resolution matrix** (Fig. 12.15) to determine how well one can resolve the temperature of the individual dikes. The nearby dikes are almost completely resolved, but the more distant ones have wide peaks and large sidelobes. The unit variances of the closer dikes imply a decrease in error of about 50% but those of the distant dikes are very poor, implying amplification of the error in the data by a factor of 10.

Next, we consider whether a second observation point improves the resolution of the temperature distribution. We first put the point near

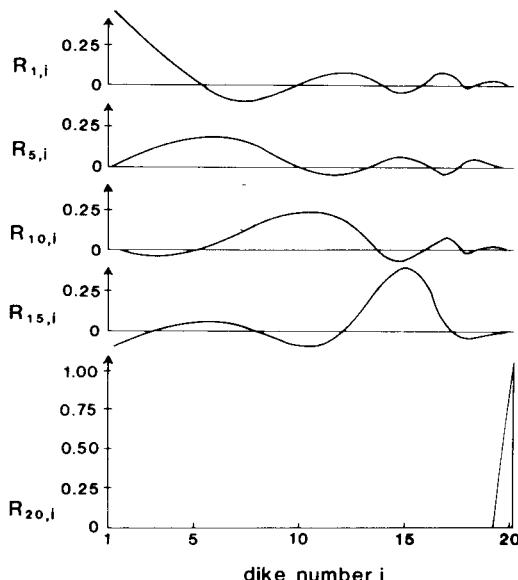


Fig. 12.16. (a) Resolution for selected dikes when there are two slightly separated observers ($\xi_1 - \xi_2 = 1$), both to the right of dike 20. (b) The resolution shows negligible improvement when compared with the one-observer case (Fig. 12.15).

the first, say, at a distance h to its right. In effect, we are now measuring both temperature and temperature gradient at the right-hand edge of the intrusion. The resolution matrix and unit variance show only slight improvement (Fig. 12.16). On the other hand, putting the observation points on opposite sides of the intrusion leads, as one would expect, to a substantial improvement in resolution (Fig. 12.17). Nevertheless, the middle dikes are still very poorly resolved. The variance of the dikes has been substantially improved; the variance of even the middle dikes is only about half as large as that of the data.

12.6 L_1 , L_2 , and L_∞ Fitting of a Straight Line

The L_1 , L_2 , and L_∞ problem is to fit the straight line $d_i = m_1 + m_2 z_i$ to a set of (z, d) pairs by minimizing the prediction error under a

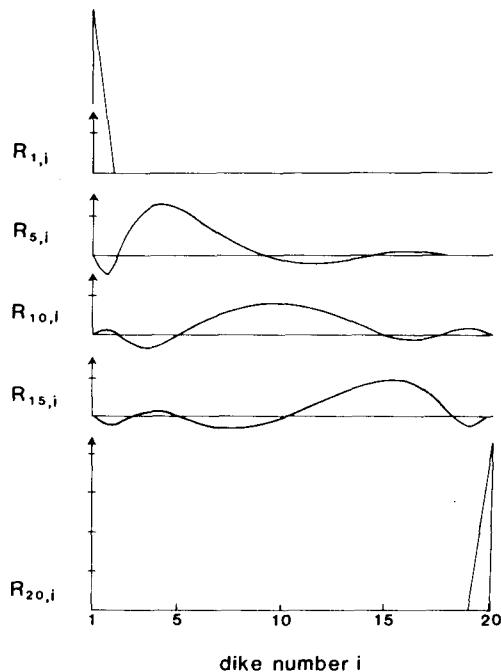


Fig. 12.17. Resolution of selected dikes when there are 2 observers, one near dike 1 and one near dike 20. The resolution is significantly improved from the one-observer case (Fig. 12.15).

variety of norms. This is a linear problem with an $N \times 2$ data kernel:

$$\mathbf{G} = \begin{bmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ \vdots & \vdots \\ 1 & z_N \end{bmatrix} \quad (12.18)$$

The L_2 norm is the simplest to implement. It implies that the data possess a Gaussian distribution with $[\text{cov } \mathbf{d}] = \sigma_d^2 \mathbf{I}$. The simple least squares solution $\mathbf{m}^{\text{est}} = [\mathbf{G}^T \mathbf{G}]^{-1} \mathbf{G}^T \mathbf{d}$ is adequate since the problem typically is very overdetermined. Since the L_2 problem has been discussed, we shall not treat it in detail here. However, it is interesting to compute the data resolution matrix $\mathbf{N} = \mathbf{G} \mathbf{G}^{-\text{est}}$:

$$\mathbf{N} = \begin{bmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \ddots \\ \vdots & \ddots \\ 1 & z_N \end{bmatrix} \frac{1}{N \sum z_i^2 - \left(\sum z_i \right)^2} \begin{bmatrix} \sum z_i^2 & \sum z_i \\ \sum z_i & N \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_N \end{bmatrix}$$

$$N_{ij} = \frac{\sum z_i^2 - (z_i + z_j) \sum z_i + z_i z_j N}{N \sum z_i^2 - \left(\sum z_i \right)^2}$$

If the data have equally spaced auxillary variable z and if they are ordered according to increasing z , then each row of the resolution matrix N_{ij} is a linear function of the column index j . The resolution is not at all localized (Fig. 12.18a). The points with most extreme z_i control the orientation of the straight line and therefore make the largest contribution to the data resolution for all the data. The resolution is not even centered about the main diagonal.

The L_1 and L_∞ estimates can be determined by using the transformation to a linear programming problem described in Chapter 8. Although more efficient algorithms exist, we shall set up the problems so that they can be solved with a standard Simplex algorithm computer program. This algorithm determines a vector \mathbf{y} that minimizes $\mathbf{c}^\top \mathbf{y}$ subject to $\mathbf{A}\mathbf{y} = \mathbf{b}$ and $c_i \geq 0$. The first step is to define two new variables \mathbf{m}' and \mathbf{m}'' such that $\mathbf{m} = \mathbf{m}' - \mathbf{m}''$. This definition relaxes the positivity constraints on the model parameters. For the L_1 problem we define three additional vectors $\boldsymbol{\alpha}$, \mathbf{x} , and \mathbf{x}' and then arrange them in the form of a linear programming problem in $2M + 3N$ variables and $2N$ constraints as

$$\begin{aligned} \mathbf{y}^\top &= [[m'_1, \dots, m'_M], [m''_1, \dots, m''_M], [\alpha_1, \dots, \alpha_N], \\ &\quad [x_1, \dots, x_N], [x'_1, \dots, x'_N]] \\ \mathbf{c}^\top &= [[0, \dots, 0], [0, \dots, 0], [1, \dots, 1], \\ &\quad [0, \dots, 0], [0, \dots, 0]] \quad (12.19) \\ \mathbf{A} &= \begin{bmatrix} \mathbf{G}_{N \times M} & -\mathbf{G}_{N \times M} & -\mathbf{I}_{N \times N} & \mathbf{I}_{N \times N} & \mathbf{O}_{N \times N} \\ \mathbf{G}_{N \times M} & -\mathbf{G}_{N \times M} & \mathbf{I}_{N \times N} & \mathbf{O}_{N \times N} & -\mathbf{I}_{N \times N} \end{bmatrix} \\ \mathbf{b}^\top &= [[d_1, \dots, d_N], [d_1, \dots, d_N]] \end{aligned}$$

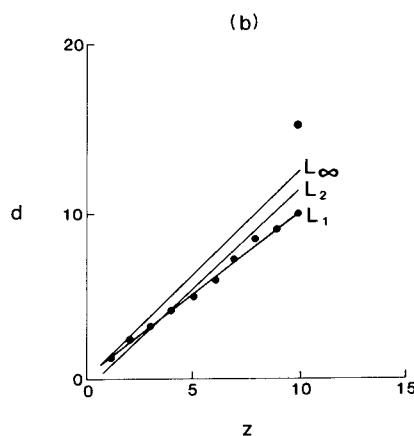
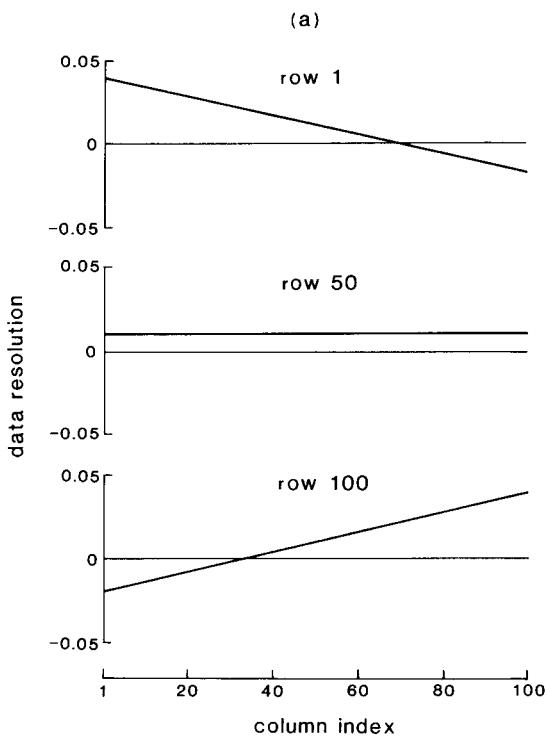


Fig. 12.18. (a) Selected rows of the data resolution matrix N for the L_2 fitting of a straight line to data. There are 100 data with equally spaced z_i 's. Note that the resolution is not localized and not centered about the main diagonal. (b) L_1 , L_2 , and L_∞ fits of a straight line to data.

The L_∞ problem is transformed into a linear programming problem with additional variables \mathbf{x} and \mathbf{x}' and a scalar parameter α . The transformed problem in $2M + 2N + 1$ unknowns and $2N$ constraints is

$$\begin{aligned}\mathbf{y}^T &= [[m'_1, \dots, m'_M], [m''_1, \dots, m''_M], [\alpha], \\ &\quad [x_1, \dots, x_N], [x'_1, \dots, x'_N]] \\ \mathbf{c}^T &= [[0, \dots, 0], [0, \dots, 0], [1], \\ &\quad [0, \dots, 0], [0, \dots, 0]] \\ \mathbf{A} &= \begin{bmatrix} \mathbf{G}_{N \times M} - \mathbf{G}_{N \times M} - \mathbf{I}_N & \mathbf{I}_{N \times N} & \mathbf{O}_{N \times N} \\ \mathbf{G}_{N \times M} - \mathbf{G}_{N \times M} & \mathbf{I}_N & \mathbf{O}_{N \times N} - \mathbf{I}_{N \times N} \end{bmatrix} \\ \mathbf{b}^T &= [[d_1, \dots, d_N], [d_1, \dots, d_N]]\end{aligned}\tag{12.20}$$

To illustrate the results of using the different norms; we fit lines to the following data:

$$\begin{aligned}\mathbf{z}^T &= [1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.00, 10.0, 10.0] \\ \mathbf{d}^T &= [1.1, 2.2, 3.2, 4.1, 4.9, 5.8, 7.1, 8.3, 9.05, 9.9, 15.0]\end{aligned}\tag{12.21}$$

Note that the last point falls far from the trend defined by the other points. The results are shown in Fig. 12.18b and Table 12.3. Note that the L_1 line is by far the least sensitive to the presence of the outlier. This good fit results from the implicit assumption that the data follow the long-tailed exponential distribution. This distribution predicts that outliers are reasonably probable and therefore gives them little weight. Both the L_1 and L_∞ fits may be nonunique. Most versions of the Simplex algorithm will find only one solution, so the process of identifying the complete range of minimum solutions may be difficult.

TABLE 12.3

Results of Fitting a Straight Line to Data by
Minimizing the Prediction Error Under a Variety
of Norms

Norm	Intercept	Slope
L_1	0.14	0.99
L_2	-0.60	1.19
L_∞	0.00	1.25

12.7 Finding the Mean of a Set of Unit Vectors

Suppose that a set of measurements of direction (defined by unit vectors in a three-dimensional Cartesian space) are thought to scatter randomly about a mean direction (Fig. 12.19). How can the mean vector be determined?

This problem is similar to that of determining the mean of a group of scalar quantities (Sections 5.1 and 8.2) and is solved by direct application of the principle of maximum likelihood. In scalar mean problems we assumed that the data possessed a Gaussian or exponential distribution and then applied the principle of maximum likelihood to estimate a single model parameter, the mean. Neither of these distributions is applicable to the distribution of directions since they are defined on the wrong interval ($[-\infty, +\infty]$, instead of $[0, \pi]$).

One distribution suitable for directions is the Fisher distribution (Ref. 7). Its vectors are clumped near the mean direction with no preferred azimuthal direction. It proposes that the probability of finding a vector in an increment of solid angle $d\Omega = \sin(\theta) d\theta d\phi$, located at an angle with inclination θ and azimuth ϕ from the mean direction (Fig. 12.20) is

$$P(\theta, \phi) = [\kappa/4\pi \sinh(\kappa)] \exp[\kappa \cos(\theta)] \quad (12.22)$$

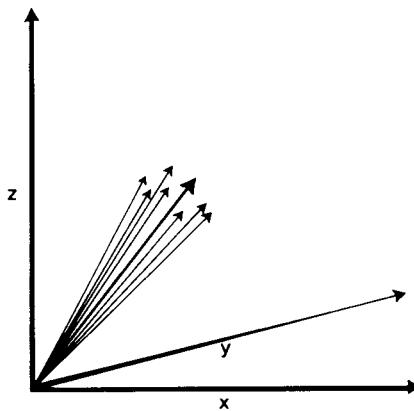


Fig. 12.19. Several unit vectors (light) scattering about a central direction (bold).

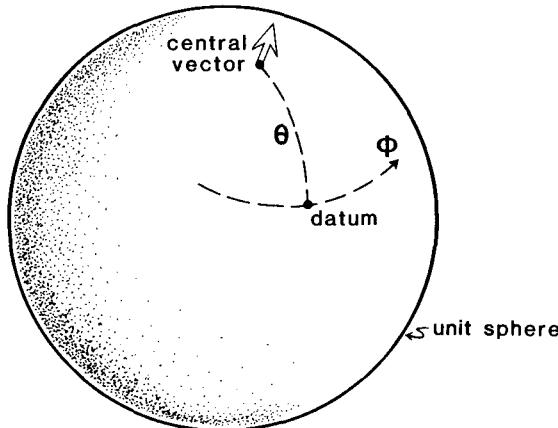


Fig. 12.20. Coordinate system used in Fisher distribution.

This distribution is peaked near the mean direction ($\theta = 0$), and its width depends on the value of the *precision parameter* κ (Fig. 12.21). The reciprocal of this parameter serves a role similar to the variance in the Gaussian distribution. When $\kappa = 0$ the distribution is completely white or random on the sphere, but when $\kappa \gg 1$ it becomes very peaked near the mean direction.

If the data are specified by N Cartesian unit vectors (x_i, y_i, z_i) and the mean by its unit vector (m_1, m_2, m_3) , then the cosine of the inclination for any data is just the dot product $\cos(\theta_i) = [x_i m_1 + y_i m_2 + z_i m_3]$. The joint distribution for the data $P(\theta, \phi)$ is then the product of N distributions of the form $P(\theta_i, \phi_i) \sin(\theta_i)$. The $\sin(\theta_i)$ term must be included since we shall be maximizing the joint probability with respect to m_i and not with respect to solid angle. The joint distribution is then

$$P(\theta) = [\kappa/4\pi \sinh(\kappa)]^N \exp \left[\kappa \sum_{i=1}^N \cos(\theta_i) \right] \prod_{i=1}^N \sin(\theta_i) \quad (12.23)$$

where $\cos(\theta_i) = x_i m_1 + y_i m_2 + z_i m_3$ and the likelihood function is

$$\begin{aligned} L = \log(P) &= N \log(\kappa) - N \log(4\pi) - N \log[\sinh(\kappa)] \\ &\quad + \kappa \sum_{i=1}^N [x_i m_1 + y_i m_2 + z_i m_3] \\ &\quad + \sum_{i=1}^N \log[\sin(\theta_i)] \end{aligned} \quad (12.24)$$

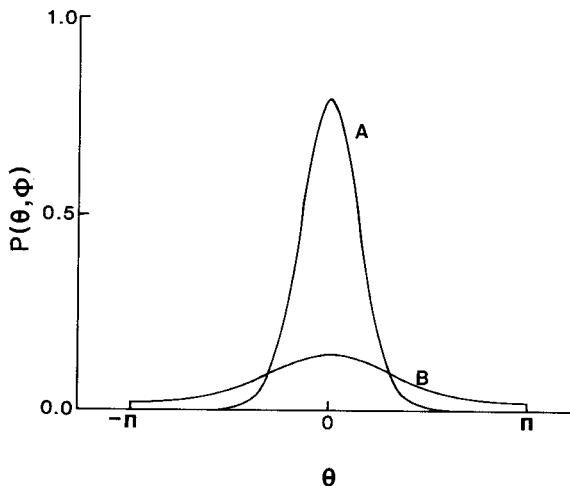


Fig. 12.21. Fisher distribution $p(\theta, \phi)$ for (a) large precision parameter ($\kappa = 5$), (b) small precision parameter ($\kappa = 1$).

The best estimate of model parameters occurs when the likelihood is maximized. However, since the solution is assumed to be a unit vector, L must be maximized under the constraint that $\sum_i m_i^2 = 1$. We first simplify this maximization by making an approximation. As long as κ is reasonably large (say, $\kappa > 5$), any variations in the magnitude of the joint probability that are caused by varying the m_i will come mainly from the exponential, since it varies much faster than the sine. We therefore ignore the last term in the likelihood function when computing the derivatives $\partial L / \partial m_g$. The Lagrange multiplier equations for the problem are then approximately

$$\begin{aligned} \kappa \sum x_i - 2\lambda m_1 &= 0 \\ \kappa \sum y_i - 2\lambda m_2 &= 0 \\ \kappa \sum z_i - 2\lambda m_3 &= 0 \\ \frac{N}{\kappa} - N \frac{\cosh(\kappa)}{\sinh(\kappa)} + \sum_{i=1}^N [x_i m_1 + y_i m_2 + z_i m_3] &= 0 \end{aligned} \quad (12.25)$$

where λ is the Lagrange multiplier. The first three equations can be solved simultaneously along with the constraint equation for the

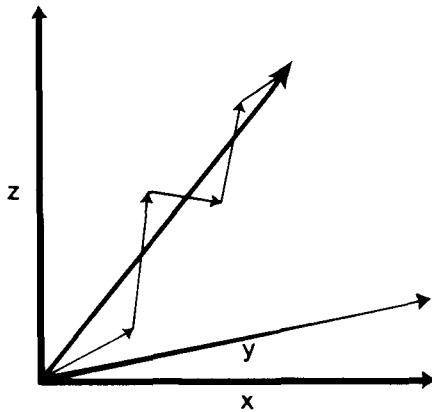


Fig. 12.22. Maximum likelihood estimate of the Fisher distribution's central vector (bold) is parallel to the vector sum of the data.

model parameters as

$$[m_1, m_2, m_3]^T = \left[\sum x_i, \sum y_i, \sum z_i \right]^T / \sqrt{\left(\sum x_i \right)^2 + \left(\sum y_i \right)^2 + \left(\sum z_i \right)^2} \quad (12.26)$$

and the fourth equation is then an implicit transcendental equation for κ . Since we have assumed that $\kappa > 5$, we can use the approximation $\cosh(\kappa)/\sinh(\kappa) \approx 1$, and the fourth equation yields

$$\kappa = N / \left[N - \sum \cos(\theta_i) \right].$$

The mean vector is the normalized vector sum of the individual observed unit vectors (Fig. 12.22).

12.8 Gaussian Curve Fitting

Many types of spectral data consist of several overlapping peaks, each of which has a Gaussian shape (Fig. 12.23). The problem is to determine the location, area, and width of each peak through least squares curve-fitting.

Suppose that the data consist of $N(z, d)$ pairs, where the auxiliary variable z represents spectral frequency. Each of, say, q peaks is

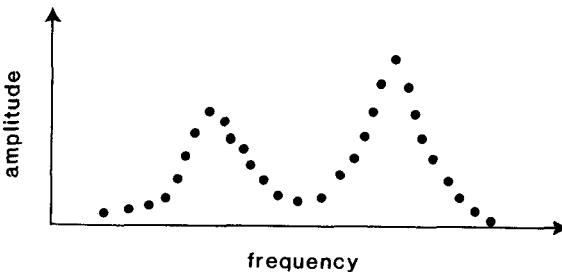


Fig. 12.23. Hypothetical x-ray fluorescence data containing two spectral peaks.

parameterized by its center frequency f_i , area A_i , and width σ_i . There are then $M = 3q$ model parameters $\mathbf{m}^T = [f_1, A_1, \sigma_1, \dots, f_q, A_q, \sigma_q]$. The model is nonlinear and of the form $\mathbf{d} = \mathbf{g}(\mathbf{m})$:

$$d_i = \sum_{j=1}^q \frac{A_j}{\sqrt{2\pi}\sigma_j} \exp\left[-\frac{(z_i - f_j)^2}{2\sigma_j^2}\right] \quad (12.27)$$

If the data have Gaussian error, it is appropriate to use an iterative L_2 method to solve this problem. Furthermore, this problem will typically be overdetermined, at least if $N > M$ and if the peaks do not overlap completely. The equation is linearized around an initial guess using Taylor's theorem, as in Section 9.3. This linearization involves computing the derivatives $\nabla \mathbf{g}$. In this problem the model is simple enough for the derivatives to be computed analytically as

$$\begin{aligned} \partial g_i / \partial A_j &= [1/\sqrt{2\pi}\sigma_j] \exp[-(z_i - f_j)^2/2\sigma_j^2] \\ \partial g_i / \partial \sigma_j &= [A_j/\sqrt{2\pi}\sigma_j][(z_i - f_j)/\sigma_j^2] \exp[-(z_i - f_j)^2/2\sigma_j^2] \\ \partial g_i / \partial f_j &= [A_j/\sqrt{2\pi}\sigma_j^2][((z_i - f_j)^2/\sigma_j^2) - 1] \exp[-(z_i - f_j)^2/2\sigma_j^2] \end{aligned} \quad (12.28)$$

These derivatives are evaluated at an initial guess $\mathbf{m}_0^{\text{est}}$ based on visual inspection of the data. Improved estimates of the model parameters are found using the recursions $\nabla \mathbf{g}_n \Delta \mathbf{m}_{n+1}^{\text{est}} = \mathbf{d} - \mathbf{g}(\mathbf{m}_n^{\text{est}})$ and $\mathbf{m}_{n+1}^{\text{est}} = \mathbf{m}_n^{\text{est}} + \Delta \mathbf{m}_{n+1}^{\text{est}}$, where the matrix equation can be solved with the simple least squares method. The iterations are terminated when the correction factor $\Delta \mathbf{m}_{n+1}^{\text{est}}$ becomes negligibly small (for instance, when the absolute value of each component becomes less than some given tolerance).

To illustrate the procedure, we fit Gaussian curves to 30 synthetic data containing two triangular peaks (Fig. 12.24). The iterative method converges in 7 iterations.

Occasionally there is a priori information that separation between

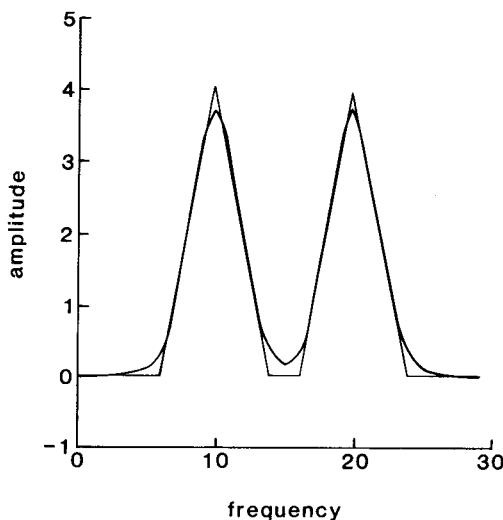


Fig. 12.24. Nonlinear least squares fit of two Gaussians to data containing two triangles.

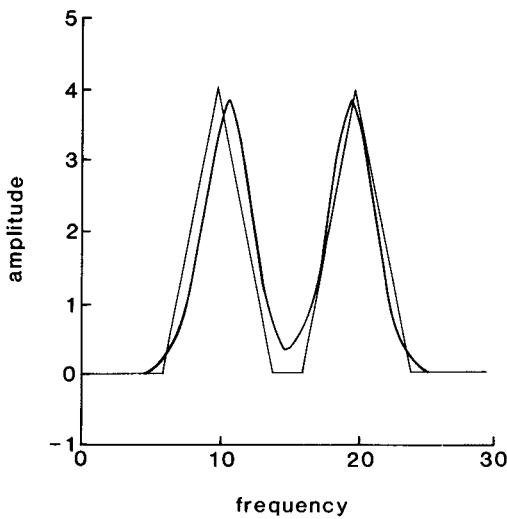


Fig. 12.25. Nonlinear least squares fit of Gaussians to data, where the peak separation has been constrained to be less than the peak separation in the data (8 and 10, respectively).

two peaks must equal a given constant $f_i - f_j = \langle s_{ij} \rangle$. This constraint can be implemented by imposing it at each iteration; $\mathbf{F} \Delta \mathbf{m}_{n+1}^{\text{est}} = \mathbf{h}$, where

$$\begin{aligned}\mathbf{F} &= [\dots, 1, 0, 0, \dots, -1, 0, 0] \\ \mathbf{h} &= [\langle s_{ij} \rangle - f_{ni} + f_{nj}]\end{aligned}\quad (12.29)$$

These constraints can be implemented either by Lagrange multipliers or by Householder transformations. An example computed by the latter method is shown in Fig. 12.25. The peak separation is constrained to be somewhat larger than the peak separation in the data, so the fit is not particularly good. This calculation also converges in seven iterations.

One of the drawbacks to these iterative methods is that if the initial guess is too far off, the solution may oscillate wildly or diverge from one iteration to the next. There are several possible remedies for this difficulty. One is to force the perturbation $\Delta \mathbf{m}_{n+1}$ to be less than a given length. This result can be achieved by examining the perturbation on each iteration and, if it is longer than some empirically derived limit, decreasing its length (but not changing its direction). This procedure will prevent the method from wildly “overshooting” the true minimum but will also slow the convergence. Another possibility is to constrain some of the model parameters to equal a priori values for the first few iterations, thus allowing only some of the model parameters to vary. The constraints are relaxed after convergence, and the iteration is continued until the unconstrained solution converges.

12.9 Earthquake Location

When a fault ruptures within the earth, seismic compressional P and shear S waves are emitted. These waves propagate through the earth and can be recorded by instruments on the earth's surface. The earthquake location problem is to determine the location (x_0, y_0, z_0) and time t_0 of the rupture on the basis of measurements of the time of arrival of the waves at the instruments.

The forward problem—that of predicting the travel time of seismic waves given a particular source (fault), receiver geometry, and seismic velocity structure in the earth—is quite formidable. We shall not discuss it here, but note that it usually requires finding the path, or *ray*,

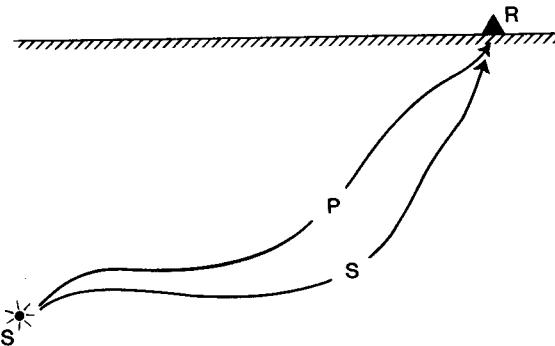


Fig. 12.26. Compressional P and shear S waves travel along rays from earthquake source S to receiver R . The ray path is determined by the position of source and receiver and the velocity structure of the medium.

that the waves followed as they traveled from source to receiver (Fig. 12.26). The process of *ray tracing* is a lengthy numerical task typically performed with a computer.

Assuming that the travel time $T_i(\mathbf{m}) = T(\mathbf{m}, x_i, y_i, z_i)$ of a P and S wave from an earthquake at $\mathbf{m}^T = (x_0, y_0, z_0, t_0)$ to a receiver at (x_i, y_i, z_i) can be calculated, the model can then be written as

$$t_{P_i} = T_P(\mathbf{m}, x_i, y_i, z_i) + t_0 \quad t_{S_i} = T_S(\mathbf{m}, x_i, y_i, z_i) + t_0 \quad (12.30)$$

where t_i is the arrival time of the given wave at the i th receiver. These equations are nonlinear and of the form $\mathbf{d} = \mathbf{g}(\mathbf{m})$. If many observations are made so that the equations are overdetermined, an iterative least squares approach may be tried. This method requires that the derivatives ∇T_i be computed for various locations of the earthquake. Unfortunately, there is no simple, differentiable analytic formula for travel time. One possible solution is to calculate this derivative numerically using the finite difference formula

$$\begin{aligned} \frac{\partial T_i}{\partial m_1} &= \frac{[T_i(\mathbf{m} + [\epsilon, 0, 0, 0]^T) - T_i(\mathbf{m})]}{\epsilon} \\ \frac{\partial T_i}{\partial m_2} &= \frac{[T_i(\mathbf{m} + [0, \epsilon, 0, 0]^T) - T_i(\mathbf{m})]}{\epsilon} \\ \frac{\partial T_i}{\partial m_3} &= \frac{[T_i(\mathbf{m} + [0, 0, \epsilon, 0]^T) - T_i(\mathbf{m})]}{\epsilon} \end{aligned} \quad (12.31)$$

Note that $\partial T_i / \partial m_4 = 0$ since the travel time depends only on the location of the earthquake and not on its time of occurrence. These equations represent moving the location of the earthquake a small distance ϵ along the directions of the coordinate axes and then computing the change in travel time. This approach has two disadvantages. First, if ϵ is made very small so that the finite difference approximates a derivative very closely, the terms in the numerator become nearly equal and computer round-off error can become very significant. Second, this method requires that the travel time be computed for three additional earthquake locations and, therefore, is four times as expensive as computing travel time alone.

In some inverse problems there is no alternative but to use finite element derivatives. Fortunately, it is possible in this problem to deduce the gradient of travel time by examining the geometry of a ray as it leaves the source (Fig. 12.27). If the earthquake is moved a small distance s parallel to the ray in the direction of the receiver, then the travel time is simply decreased by an amount $s/v(\mathbf{m})$, where $v(\mathbf{m})$ is the appropriate wave velocity in the vicinity of the source. If it is moved a small distance perpendicular to the ray, then the change in travel time is negligible since the new ray path will have nearly the same length as the old. The gradient is therefore $\nabla T = \mathbf{s}/v(\mathbf{m})$, where \mathbf{s} is a unit vector tangent to the ray at the source that points away from the receiver. Since we must calculate the ray path to find the travel time, no extra computational effort is required to find the gradient. If $\mathbf{s}^P(\mathbf{m})$ is the

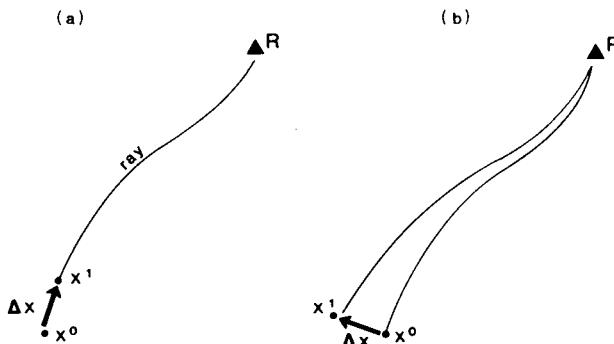


Fig. 12.27. (a) Moving the source from x^0 to x^1 parallel to the ray path leads to a large change in travel time. (b) Moving the source perpendicular to the ray path leads to no (first-order) change in travel time. The partial derivative of travel time with respect to source position can therefore be determined with minimal extra effort.

direction of the P wave ray to the i th receiver (of a total of N receivers), the linearized problem is

$$\begin{bmatrix} s_1^{P_i}(\mathbf{m}_n)/v_P(\mathbf{m}_n) & s_2^{P_i}(\mathbf{m}_n)/v_P(\mathbf{m}_n) & s_3^{P_i}(\mathbf{m}_n)/v_P(\mathbf{m}_n) & 1 \\ \vdots & \vdots & \vdots & \vdots \\ s_1^{P_N}(\mathbf{m}_n)/v_P(\mathbf{m}_n) & s_2^{P_N}(\mathbf{m}_n)/v_P(\mathbf{m}_n) & s_3^{P_N}(\mathbf{m}_n)/v_P(\mathbf{m}_n) & 1 \\ s_1^{S_i}(\mathbf{m}_n)/v_S(\mathbf{m}_n) & s_2^{S_i}(\mathbf{m}_n)/v_S(\mathbf{m}_n) & s_3^{S_i}(\mathbf{m}_n)/v_S(\mathbf{m}_n) & 1 \\ \vdots & \vdots & \vdots & \vdots \\ s_1^{S_N}(\mathbf{m}_n)/v_S(\mathbf{m}_n) & s_2^{S_N}(\mathbf{m}_n)/v_S(\mathbf{m}_n) & s_3^{S_N}(\mathbf{m}_n)/v_S(\mathbf{m}_n) & 1 \end{bmatrix} \begin{bmatrix} \Delta x_0 \\ \Delta y_0 \\ \Delta z_0 \\ \Delta t_0 \end{bmatrix}_{n+1}$$

$$= \begin{bmatrix} t_{P_1} - T_{P_1}(\mathbf{m}_n) \\ \vdots \\ t_{P_N} - T_{P_N}(\mathbf{m}_n) \\ t_{S_1} - T_{S_1}(\mathbf{m}_n) \\ \vdots \\ t_{S_N} - T_{S_N}(\mathbf{m}_n) \end{bmatrix} \quad (12.32)$$

This equation can then be solved iteratively using the least squares method. There are some instances in which the matrix can become underdetermined and the least squares method will fail. This possibility is especially likely if the problem contains only P wave arrival times. The matrix equation consists of only the top half of (12.32). If all the rays leave the source in such a manner that one or more components of their unit vectors are all equal, then the corresponding column of the matrix will be proportional to the vector $[1, 1, 1, \dots, 1]^T$ and will therefore be linearly dependent on the fourth column of the matrix. The earthquake location is then nonunique and can be traded off with origin time (Fig. 12.28). This problem occurs when the

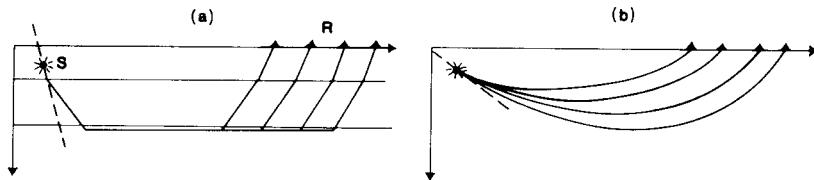


Fig. 12.28. (a) In layered media, rays follow “refracted” paths, such that rays to all receivers (triangles) leave the source (star) at the same angle. The position and travel time of the source on the dashed line therefore trade off. (b) This phenomenon also occurs in nonlayered media if the earthquake is far from the receivers.

earthquake is far from all of the stations. The addition of S wave arrival times resolves the underdeterminacy (the columns are then proportional to $[1, 1, 1, \dots, 1, v_S/v_P, \dots]^T$ and are not linearly dependent on the fourth column). It is therefore wise to use singular-value decomposition and the natural inverse to solve earthquake location problems, permitting easy identification of underdetermined cases.

The earthquake location problem can also be extended to locate the earthquake and determine the velocity structure of the medium simultaneously. It then becomes similar to the tomography problem (Section 12.4) except that the orientations of the ray paths are unknown. To ensure that the medium is crossed by a sufficient number of rays to resolve the velocity structure, one must locate simultaneously a large set of earthquakes—on the order of 100 [Ref. 5].

12.10 Vibrational Problems

There are many inverse problems that involve determining the structure of an object from measurements of its characteristic frequencies of vibration (eigenfrequencies). For instance, the solar and terrestrial vibrational frequencies can be inverted for the density and elastic structure of those two bodies. These inverse problems are typically nonlinear. Special effort must therefore be made to find simple means of calculating the partial derivatives of eigenfrequency with respect to the structure. We shall describe a general procedure for calculating these derivatives that is based on perturbation theory and that is applicable whenever the vibrations satisfy a linear, second-order

differential equation with homogeneous boundary conditions. For one-dimensional cases, this *Sturm–Liouville* problem has the form

$$\frac{\partial}{\partial x} \left[p(x) \frac{\partial y}{\partial x} \right] + q(x)y + \omega^2 r(x)y = 0 \quad (12.33)$$

where x is position, y a measure of deformation in the body (pressure, particle displacement, etc.), ω frequency and p , q , and r are functions that define the structure of the body. For any given structure, there is an infinite set of frequencies ω_i and associated wavefunctions $y_i(x)$ that satisfy the equation with homogeneous boundary conditions (at, say, $x = a$ and $x = b$). The inverse problem is to determine the structure ($p(x)$, $q(x)$, and $r(x)$) through observations of a finite number of frequencies. Since the structure is described by continuous functions, this is properly a problem in continuous inverse theory. However, we shall assume that it has been discretized in some appropriate fashion.

To establish a relationship between the characteristic frequency and the wave function, we multiply the differential equation by y and integrate over the interval $[a, b]$ as

$$\int_a^b y [py']' dx + \int_a^b qy^2 dx + \omega^2 \int_a^b ry^2 dx = 0 \quad (12.34)$$

where primes indicate differentiation with respect to x . Integrating the first term by parts yields

$$[pyy']_a^b - \int_a^b py'^2 dx + \int_a^b qy^2 dx + \omega^2 \int_a^b ry^2 dx = 0 \quad (12.35)$$

Since the boundary conditions are homogeneous, either y or y' is zero at the boundary, and the first term vanishes. The remaining terms imply

$$\omega^2 = I_1/I_2 = \frac{\int_a^b [py'^2 - qy^2] dx}{\int_a^b ry^2 dx} \quad (12.36)$$

where I_1 and I_2 refer to the two integrals. When ω is a characteristic frequency of the body, this expression is stationary with respect to small variations in the wave function y :

$$\begin{aligned}
 \delta(\omega^2) &= \frac{1}{I_2} [\delta I_1 - \omega^2 \delta I_2] \\
 &= \frac{2}{I_2} \left[[py' \delta y]_a^b - \int_a^b [(py')' + qy] \delta y \, dx - \omega^2 \int_a^b ry \delta y \, dx \right] \\
 &= \frac{-2}{I_2} \left[\int_a^b [(py')' + qy + \omega^2 ry] \delta y \, dx \right]
 \end{aligned} \tag{12.37}$$

If the variation is to satisfy the boundary conditions, it must vanish at the endpoints. The expression $[py' \delta y]_a^b$ is therefore zero, and the equations contain the original differential equation under the integral sign. Since the differential equation is assumed to be identically satisfied, $\delta(\omega^2) = 0$. These two relationships can be rewritten as

$$\begin{aligned}
 \delta I_1 - \omega^2 \delta I_2 &= 0 \\
 I_1 - \omega^2 I_2 &= 0
 \end{aligned} \tag{12.38}$$

Suppose that the structure of the body is perturbed by δp , δq , δr , causing a change in frequency $\delta\omega$ and a change in wave function δy . By the formula $I_1 - \omega^2 I_2 = 0$ the perturbed problem satisfies

$$\begin{aligned}
 &(\omega + \delta\omega)^2 \int_a^b (r + \delta r)(y + \delta y)^2 \, dx \\
 &= \int_a^b [(p + \delta p)(y' + \delta y')^2 - (q + \delta q)(y + \delta y)^2] \, dx
 \end{aligned} \tag{12.39}$$

Keeping only first-order terms in the perturbations and subtracting out $I_1 - \omega^2 I_2 = 0$ for the unperturbed problem lead to the expression

$$\begin{aligned}
 &\omega^2 \int_a^b (\delta r y^2 + 2ry \delta y) \, dx + 2\omega \delta\omega \int_a^b ry^2 \, dx \\
 &= \int_a^b (2py' \delta y + \delta p y'^2 - \delta q y^2 - 2qy \delta y) \, dx
 \end{aligned} \tag{12.40}$$

This expression still involves the perturbations of the wave function δy . Fortunately, these terms can be subtracted out using the relationship $\delta I_1 - \omega^2 \delta I_2 = 0$, yielding

$$\delta\omega = \frac{\int_a^b [y'^2\delta p - y^2\delta q - \omega^2 y^2\delta r] dx}{2\omega \int_a^b ry^2 dx} \quad (12.41)$$

This formula permits the calculation of the perturbation in frequency resulting from arbitrary perturbations in structure. Only the unperturbed frequency and wave function need be known. If the problem is discretized by dividing the body into many small boxes of constant properties, then the calculation of the derivative requires integrating (possibly by numerical means) y^2 and y'^2 , over the length of the box.

13

NUMERICAL ALGORITHMS

Most applications of inverse theory require substantial computations that can be performed only by a digital computer. The practitioner of inverse theory must therefore have some familiarity with numerical methods for implementing the necessary matrix operations. This chapter describes some commonly used algorithms for performing these operations and some of their limitations. FORTRAN 77 programs are included to demonstrate how some of these algorithms can be implemented.

There are three considerations in designing useful numerical algorithms: accuracy, speed, and storage requirements. Accuracy is the most important. Since computers store numbers with only a finite precision, great care must be taken to avoid the accumulation of round-off error and other problems. It is clearly desirable to have an algorithm that is fast and makes efficient use of memory; however, speed and memory requirements usually trade off. The speed of an algorithm can usually be improved by reusing (and storing) intermediate results. The algorithms we shall discuss below reach some compromise between these two goals.

13.1 Solving Even-Determined Problems

If the solution of an over- or underdetermined L_2 inverse problem is of interest (and not its covariance or resolution), then it can be found by solving a square, even-determined system:

$$\begin{aligned} [\mathbf{G}^T \mathbf{G}] \mathbf{m} &= \mathbf{G}^T \mathbf{d} && \text{for overdetermined problems} \\ [\mathbf{G} \mathbf{G}^T] \boldsymbol{\lambda} &= \mathbf{d} \quad \text{and} \quad \mathbf{m} = \mathbf{G}^T \boldsymbol{\lambda} && \text{for underdetermined systems} \end{aligned} \quad (13.1)$$

These systems have the form $\mathbf{A}\mathbf{x} = \mathbf{b}$, where \mathbf{A} is a square matrix, \mathbf{b} is a vector, and \mathbf{x} is unknown. As will be shown, less work is required to solve this system directly than to form \mathbf{A}^{-1} and then multiply to obtain $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.

A good strategy to solve the $N \times N$ system $\mathbf{A}\mathbf{x} = \mathbf{b}$ is to transform these equations into a system $\mathbf{A}'\mathbf{x}' = \mathbf{b}'$ that is upper triangular in the sense that \mathbf{A}' has zeros beneath its main diagonal. In the $n = 5$ case

$$\mathbf{A}' = \begin{bmatrix} \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & 0 & \times \end{bmatrix} \quad (13.2)$$

where \times signifies a nonzero matrix element. These equations can be easily solved starting with the last row: $x_N = b'_N/A'_{NN}$. The next to the last row is then solved for x_{N-1} , and so forth. In general, this *back-solving* procedure gives

$$x_i = \frac{1}{A'_{ii}} \left[b'_i - \sum_{j=i+1}^N A'_{ij} x_j \right] \quad (13.3)$$

There are many ways to triangularize a square matrix. One of the simplest is Gauss–Jordan reduction. It consists of subtracting rows of \mathbf{A} from each other in such a way as to convert to zeros (annihilate) the part of a column below the main diagonal and then repeating the process until enough columns are partially annihilated to form the triangle. The left-most column is partially annihilated by subtracting A_{21}/A_{11} times the first row from the second row, A_{32}/A_{11} times the first row from the third row, and so forth. The second column is partially

annihilated by subtracting A_{32}/A_{22} times the second row from the third row, A_{42}/A_{22} times the second row from the fourth row, and so forth. The process is repeated $N - 1$ times, at which point the matrix is triangularized. The vector \mathbf{b} must also be modified during the triangularization process. Between 1 and N multiplications are needed to annihilate each of approximately $N^2/2$ elements, so the effort needed to triangularize the matrix grows with order N^3 (in fact, it is proportional to $N^3/3$).

One problem with this algorithm is that it requires division by the diagonal elements of the matrix. If any of these elements are zero, the method will fail. Furthermore, if any are very small compared to the typical matrix element, significant round-off error can occur because of the computer's limited mathematical precision and the results can be seriously in error. One solution to this problem is to examine all the elements of the untriangularized portion of the matrix before partially annihilating a column, select the one with largest absolute value, and move it to the appropriate diagonal position. The element is moved by exchanging two rows of \mathbf{A} (and two corresponding elements of \mathbf{b}) to move it to the correct row and then exchanging two columns of \mathbf{A} (and reordering the unknowns) to move it to the correct column. This procedure is called *pivoting* (Fig. 13.1), and the new diagonal element is called the *pivot*.

Pivoting guarantees that the matrix will be properly triangularized. In fact, if at any stage no nonzero pivot can be found, then the matrix is singular. We note that pivoting does not really require the reordering of unknowns, since the columns of the triangle need not be in order. There is no special reason for the column with $N - 1$ zeros to be first. It is sufficient to leave the columns in place and keep track of their order in the triangle. Pivoting does require reordering the elements of \mathbf{b} . This process can be a bit messy, so one often simplifies the pivoting to avoid this step. *Partial pivoting* selects the largest element in the same row as the diagonal element under consideration and transfers it to the diagonal by column-exchanging operations alone.

Triangularization by partial pivoting has the further advantage of providing an easy means of separating the process of triangularizing \mathbf{A} from that of transforming \mathbf{b} . Separating these steps is useful when solving several matrix equations that all have the same \mathbf{A} but have different \mathbf{b} 's. The effort of performing the triangularization is expended only once. To separate the operations, two types of information must be retained: (1) which columns were exchanged during each

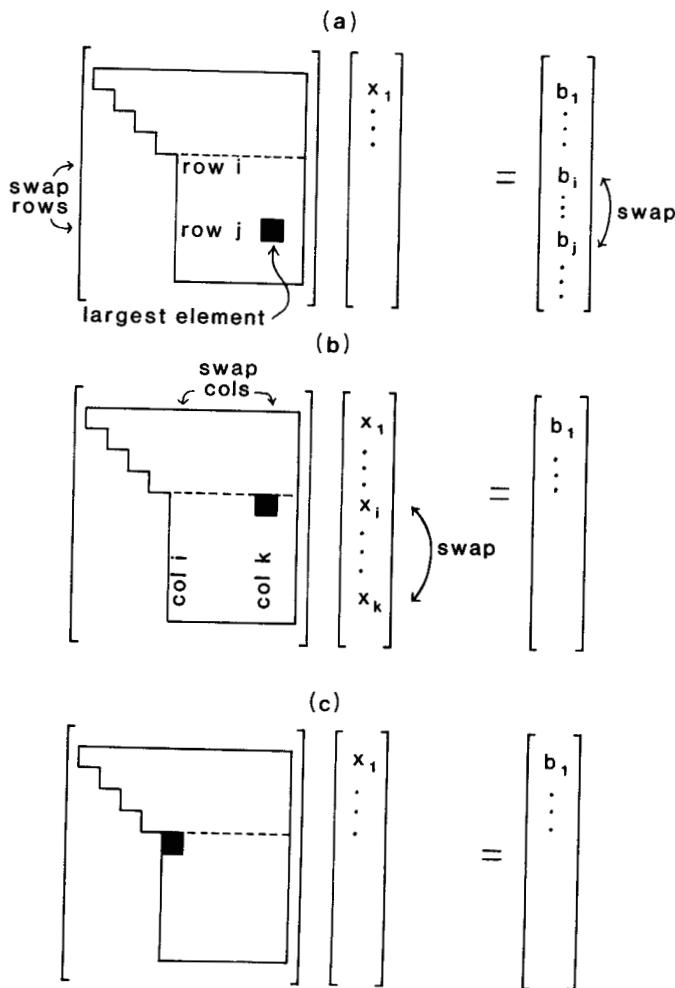


Fig. 13.1. Pivoting steps: (a) Locate largest element below and to the right of the current diagonal element; exchange rows and two data so that largest element is brought to desired row. (b) Exchange columns and two unknowns. (c) The largest element is now on the main diagonal.

pivot and (2) the coefficients that transform the vector \mathbf{b} . The first requirement can be satisfied by keeping a list of length N of row numbers. The second requires saving about $N^2/2$ coefficients: $N - 1$ for the operations that partially annihilated the first column, $N - 2$ for those that partially annihilated the second column, etc. Fortunately, the first column contains $N - 1$ zeros, the second ($N - 2$), etc., so that the coefficients can be stored in the annihilated elements of \mathbf{A} and require no extra memory.

FORTRAN Program (GAUSS) for Solving Even-Determined Linear Problems

```

subroutine gauss(a,vec,n,nstore,test,ierror,itriag)
real a(nstore,nstore), vec(nstore), test
integer n, nstore, ierror, itriag

c subroutine gauss, by William Menke, July 1978 (modified Feb 1983)

c a subroutine to solve a system of n linear equations in n
c unknowns, where n doesn't exceed 100. all parameters real.

c gaussian reduction with partial pivoting is used. the
c original matrix a(n,n) is first triangularized. its zero
c elements are replaced by coefficients which allow the trans-
c forming of the vector vec(n).

c if itriag is set to 1 before the subroutine call, the tri-
c angularization procedure is omitted and a is assumed to
c have been triangularized on a previous subroutine call.
c the variables a, isub, and l1 must be preserved from
c the call that triangularized a.

c the matrix a (nxn) and the vector vec (n) are altered
c during the subroutine call. the solution is returned in vec.

c test is a real positive number specified by the user which is
c used in testing for division by near zero numbers. if the absolute
c value of a number is .le. test an error condition
c will result.

c the error conditions are returned in ierror. they are :
c           0 : no error
c           1 : division condition violated
c                 during triangularization of a
c           2 : division condition violated
c                 during back-solving
c           3 : division condition violated
c                 at both places

c nstore is the size to which a and vec were dimensioned in the main
c program, whereas n is the size of the used portion of a and vec

c currently set up for maximum n of 100

dimension line(100),isub(100)
save isub, l1
c
c     iet=0      /* initial error flags, one for triangularization,
c     ieb=0      /*                      one for back-solving
c
c ***** triangularize the matrix a, replacing the zero elements
c          of the triangularized matrix with the coefficients needed
c          to transform the vector vec
c
c     if( itriag .eq. 0) then      /* triangularize matrix
c
c         do 1 j=1,n            /* line is an array of flags.
c                           line(j)=0 /* elements of a are not moved during
c                           continue /* pivoting. line=0 flags unused lines
1

```

Program GAUSS, continued.

```

c
      do 30 j=1,n-1      /* triangularize matrix by partial pivoting
      big = 0.0      /* find biggest element in j-th column
      do 40 l1=1,n   /* of unused portion of matrix
          if( line(l1) .eq. 0 ) then
              testa=abs(a(l1,j))
              if(testa.gt.big) then
                  i=l1
                  big=testa
              end if
          end if
      continue
40
      if( big .le. test) then    /* test for div by 0
          iet=1
      end if

      line(i)=1 /* selected unused line becomes used line
      isub(j)=i /* isub points to j-th row of triang. matrix

      sum=1.0/a(i,j) /* reduce matrix toward triangle
      do 10 k=1,n
          if( line(k) .eq. 0 ) then
              b=a(k,j)*sum
              do 20 l=j+1,n
                  a(k,l)=a(k,l)-b*a(i,l)
              continue
20
              a(k,j)=b
          end if
      continue
10
      continue

30
      continue

      do 32 j=1,n    /* find last unused row and set its pointer
                      /* this row contains the apex of the triangle
          if( line(j) .eq. 0 ) then
              l1 = j /* apex of triangle
              isub(n)= j
              goto 35 /* break loop
          end if
      continue
32
      continue
35
      end if /* done triangularizing

c ***** start back-solving

      do 100 i=1,n    /* invert pointers. line(i) now gives
                      /* row # in triang. matrix
                      /* of i-th row of actual matrix
          line(isub(i)) = i
      continue
100

      do 320 j=1, n-1 /* transform the vector to match triang. matrix
          b=vec(isub(j))
          do 310 k=1, n

```

Program GAUSS, continued.

```

        if(line(k).le.j) then
            continue /* skip elements outside of triangle
        else
            vec(k)=vec(k)-a(k,j)*b
            end if
310      continue
320      continue

b=a(11,n) /* apex of triangle
if( abs(b) .le. test) then /* check for div by zero in back-solving
    ieb=2
    end if
vec(isub(n))=vec(isub(n))/b

do 50 j = n-1, 1, -1 /* back-solve rest of triangle
    sum=vec(isub(j))
    do 60 j2 = j+1, n
        sum = sum - vec(isub(j2)) * a(isub(j),j2)
        continue
60      b = a(isub(j)),j)
      if( abs(b) .le. test ) then /* test for div by 0 in bksolv.
          ieb=2
          end if
          vec(isub(j))=sum/b /* solution returned in vec
50      continue
c
c ***** put the solution vector into the proper order
c
do 230 i=1,n /* reorder solution
    do 210 k=i,n /* search for i-th solution element
        if( line(k) .eq. i ) then
            j=k
            go to 220 /* break loop
            end if
210      continue
220      b = vec(j) /* swap solution and pointer elements
        vec(j) = vec(i)
        vec(i) = b
        line(j) = line(i)
        continue

ierror = iet + ieb /* set final error flag

return
end

```

13.2 Inverting a Square Matrix

The inverse of the square $N \times N$ matrix \mathbf{A} must satisfy $\mathbf{AA}^{-1} = \mathbf{I}$. This equation can be interpreted as N equations of the form $\mathbf{Ax} = \mathbf{b}$, where \mathbf{x} is a column of \mathbf{A}^{-1} and \mathbf{b} is the corresponding column of the identity matrix. The inverse can therefore be found by solving N square systems of equations using Gauss–Jordan reduction or any other convenient method. Since all the equations have the same matrix, it needs to be triangularized only once. The triangularization requires $N^3/3$ multiplications, and the back-solving requires $N^2/2$ for each of the N columns, so about $5N^3/6$ operations are required to invert the matrix.

FORTRAN Code for Inverting a Square Matrix

```

subroutine gjinv( a, b, n, nstore, test, ierror )
real*4 a(nstore, nstore), b(nstore, nstore), test
integer n, nstore, ierror

c subroutine gjinv, by William Menke, January, 1982
c to compute the inverse of a square matrix using
c Gauss-Jordan reduction with partial pivoting

c      a      input matrix
c      b      inverse of a
c      n      size of a and b
c      nstore   dimensioned size of a and b
c      test    division by numbers smaller than this generate
c              a divide by zero error
c      ierror   error flag, zero on no error

c currently limited to max 100 by 100 matrix
dimension c(100)

do 10 icol=1,n /* build inverse columnwise by solving AB=I.

      do 1 i=1,n /* c is icol column of identity matrix
          c(i)=0.0
1      continue
      c(icol)=1.0

      if( icol .eq. 1 ) then /* triangularize a on first call
          itriag = 0
      else
          itriag = 1
      end if
      call gauss( a, c, n, nstore, test, ierror, itriag )

      if( ierror .ne. 0 ) then /* return on error
          return
      end if

      do 2 i=1,n /* move solution into matrix inverse
          b(i,icol)=c(i)
2      continue

10     continue

      return
end

```

13.3 Solving Underdetermined and Overdetermined Problems

The under- or overdetermined linear system $\mathbf{G}\mathbf{m} = \mathbf{d}$ can be easily solved through the use of Householder transformations as was described in Section 7.2. The technique, sometimes called Golub's method, consists of rotating these systems to triangular form by multiplication by a sequence of transformations T_j , each of which annihilates the appropriate elements in one column of the matrix. The transformation that annihilates the j th column (or row) of \mathbf{G} is

$$T_j = \left\{ \mathbf{I} - \frac{1}{\alpha(\alpha - G_{jj})} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ G_{j,j} - \alpha \\ G_{j+1,j} \\ \vdots \\ G_{N,j} \end{bmatrix} \begin{bmatrix} 0, \dots, 0, G_{jj} - \alpha, G_{j+1,j}, \dots, G_{N,j} \end{bmatrix}^T \right\} \quad (13.4)$$

where $\alpha = \pm \sqrt{\sum_{i=j}^N G_{ij}^2}$. The transformation is correct regardless of the sign of α . To avoid numerical problems, however, the sign should be chosen to be opposite that of G_{jj} . This choice prevents the denominator from ever being near zero, except in the unlikely case that all the elements from j to N are zero. Golub's method therefore works well without a pivoting strategy.

Consider the overdetermined case, where the transformation is being applied to columns. When the transformation is applied to the j th column of \mathbf{G} that column becomes $[G_{1,j}, G_{2,j}, \dots, G_{j-1,j}, \alpha, 0, 0, \dots, 0]^T$. Notice that of all the nonzero elements of the column, only G_{jj} is actually changed. In practice, the transformation is not applied to this column (it is not multiplied out). The result is merely substituted in. (This is a general principle of programming; one never computes any result for which the answer is known beforehand.) Note also that the vector in the transformation has $N - j + 1$ nonzero elements, and

that the transformed column of \mathbf{G} has $N - J$ zeros. All but one of the transformation coefficients can be stored in the zeros. The algorithm can be formulated so that \mathbf{G} is triangularized first and the vector \mathbf{d} or \mathbf{m} is transformed later. When the transformation is applied to the k th column of \mathbf{G} , where $k < j$, no elements are altered. In practice, the transformation is simply not applied to these columns. For $k > j$ the first $j - 1$ elements are unchanged, and the remaining ones become

$$G'_{pk} = G_{pk} - [G_{pj} - \alpha\delta_{pj}] \left[\frac{(G_{jj} - \alpha)G_{jk} + \sum_{i=j+1}^N G_{ij}G_{ik}}{\alpha(\alpha - G_{jj})} \right];$$

$$p \geq j, \quad k > j \quad (13.5)$$

FORTRAN Code (UNDERC) for Finding the Minimum-Length Solution to a Purely Underdetermined System

```

subroutine underc( g,y, n,nstore, m,mstore,u, test,ier,itr )
c
c subroutine underc, by william menke, january, 1980.
c
c a subroutine to solve a system of m equations in n unknowns,
c written as the matrix equation g(n,m) * x(m) = y(n),
c where m is greater than or equal to n (that is, under/even determined)
c the equations are satisfied exactly with shortest L-2 length solution.
c golub's method, employing householder transformations is used.
c parameters are :
c
c g ..... the n by m matrix of coefficients to the equations.
c           a is changed by the subroutine, being replaced by the
c           triangularized version.
c
c y ..... the vector of constants. this is destroyed during
c           the computation. the solution is returned in the
c           first m locations of y.
c
c n ..... the number of equations.
c
c m ..... the number of unknowns.
c
c nstore the number of rows g was dimensioned to in the
c           main program.
c
c mstore the number of columns g was dimensioned to in the main
c           program.
c
c u ..... a temporary vector at least n in length.
c
c test .. if the reciprocal of a number smaller than test is
c           computed, an error flag is set.
c
c ier ... an error flag which is zero on no error. errors
c           can be caused by inconsistant input or by division
c           by near zero numbers.
c
c itr ... if several systems of equations sharing the same
c           coefficient matrix are to be solved, the coefficient
c           matrix need be triangularized only once. set itr=0
c           to include triangularization in the call, itr=1
c           to assume that it was triangularized on a previous call
c           and that a and u have been preserved.
c
c           dimension g(nstore,mstore), y(nstore), u(1)

c check for bad parms
  if( n .gt. m .or. n .lt. 1 .or. m .lt. 2 ) then
    ier = 1
    return

```

Program UNDERC, continued.

```

    else
        ier = 0
    end if
c one less rotation if even-determined
    if( n .eq. m ) then
        nlast = n-1
    else
        nlast = n
    end if

    if( itr .eq. 0 ) then
c **** post-multiply a by householder transformations to
c      zero elements to the right of the main diagonal of
c      the first mc rows.

        do 50 irow=1,nlast

            alfa = 0.0
            do 10 i=irow,m
                t = g(irow,i)
                u(i) = t
                alfa = alfa + t*t
                continue
            10   alfa = sqrt( alfa )
            if( u(irow) .lt. 0.0 ) then
                alfa=-alfa
            end if

                t = u(irow) + alfa
                u(irow) = t

                beta = alfa * t
c
c ***** apply transformation to remaining rows.
c
            do 20 j=irow+1, n
                gama = 0.0
                do 30 i=irow, m
                    gama = gama + g(j,i)*u(i)
                    continue
                30   if( abs(beta) .le. test) then
                    ier=2
                    end if
                    gama = gama / beta
                    do 40 i=irow, m
                        g(j,i) = g(j,i) - gama*u(i)
                        continue
                    40   continue
                20

```

Program UNDERC, continued.

```

g(irow,irow) = t
u(irow) = -alfa

50          continue

end if

c ***** back solve for solution.
c first parameter
if( abs(u(1)) .le. test ) then
    ier=4
    end if
y(1) = y(1) / u(1)
c rest of determined parameters
if( n .eq. m ) then
    u(m) = g(m,m)
    end if
do 60 irow=2, n
    sum = 0.0
    do 70 i=1,irow-1
        sum = sum + y(i) * g(irow,i)
70          continue
    if( abs(u(irow)) .le. test) then
        ier=5
        end if
        y(irow) = ( y(irow) - sum ) / u(irow)
60          continue
c undetermined parameters
do 80 irow=n+1, m
    y(irow) = 0.0
80          continue

c ***** apply post-multiplication transformations to solution.

do 90 irow=nlast,1,-1
    beta = -u(irow) * g(irow,irow)
    gama = 0.0
    do 100 i=irow,m
        gama = gama + y(i)*g(irow,i)

100         continue
    gama = gama / beta
    do 120 i=irow,m
        y(i) = y(i) - gama*g(irow,i)
120         continue
90          continue

return
end

```

FORTRAN Code (CLS) for Solving a Purely Overdetermined Linear System with Optional Linear Equality Constraints

```

subroutine cls(g,xm,d, n,m,nd,md, f,h, nc,ncd,mcd, test,
&           itr, sq,ierr)
real g(nd,md), xm(md), d(nd), f(ncd,mcd), h(ncd), test, sq
integer n, m, nd, md, nc, ncd, mcd, itr, ierr

/* subroutine cls (constrained least squares) by William Menke

/* solves g * xm = d for xm with linear constraints f * xm = h
/* g is n by m   (dimensioned nd by md)
/* f is nc by m  (dimensioned ncd by mcd)
/* division by zero if denominator smaller than test
/* system triangularized only if itr=0
/* sq returned as rms error
/* ierr returned as zero on no error

/* problem must be overdetermined with no redundant constraints
/* solution in several steps:
/*      1) lower-left triangularization of constraint matrix by
/*          Householder reduction and simultaneous rotation of
/*          least squares matrix.
/*      2) upper-right triangularization of least squares matrix
/*          by Householder reduction, ignoring first nc columns
/*      3) rotation of least squares vector to match matrix
/*      4) back-solution of constraint matrix for first nc unknowns
/*      5) back-solution of least squares matrix for rest of unknowns
/*      6) computation of residual error
/*      7) back-rotation of unknowns

dimension u(100), v(100)    /* temp vectors
save u, v, nlast, mlast    /* must be saved if itr=1 option used

if( m .lt. 1 ) then      /* not enough unknowns
  ierr = 1
  return
else if( m .gt. (n+nc) ) then /* not enough data
  ierr = 1
  return
else if( nc .gt. m ) then /* too many constraints
  ierr = 1
  return
end if

if( itr .eq. 0 ) then    /* triangularize systems

  if( nc .eq. m ) then    /* number of Householder rotations one
    nlast = nc-1          /* less if problem is even-determined
  else
    nlast = nc
  end if
  if( n+nc .eq. m ) then
    mlast = m-1
  else
    mlast = m
  end if

```

Program CLS, continued.

```

do 240 irow=1, nlast      /* triangularize constraint matrix
    alfa = 0.0
    do 10 i=irow, m      /* build transformation to put zeros
        t = f(irow,i)          /* in row of h
        u(i) = t
        alfa = alfa + t*t
        continue
10     alfa = sqrt( alfa )
        if( u(irow) .lt. 0.0 ) then
            alfa = -alfa
            end if
        t = u(irow) + alfa
        u(irow) = t
        beta = alfa * t
        if( abs(beta) .le. test ) then /* div by 0 test
            ierr = 2
            end if
        do 200 j = irow+1, nc      /* rotate remaining rows of f
            gama = 0.0             /* (if any)
            do 210 i = irow, m
                gama = gama + f(j,i)*u(i)
                continue
210     gama = gama / beta
            do 220 i = irow, m
                f(j,i) = f(j,i) - gama*u(i)
                continue
220     continue
200     do 1200 j = 1, n      /* rotate rows of g (if any)
            gama = 0.0
            do 1210 i = irow, m
                gama = gama + g(j,i)*u(i)
                continue
1210     gama = gama / beta
            do 1220 i = irow, m
                g(j,i) = g(j,i) - gama*u(i)
                continue
1220     continue
1200     continue
f(irow,irow) = t
u(irow) = -alfa
continue

240     if( nlast .ne. nc ) then      /* set last u if loop short
        u(nc) = f(nc,nc)
        end if

        if( n .gt. 0 ) then
do 480 icol = nc+1, mlast /* zero columns of g
        alfa = 0.0             /* starting with nc+1 column
        do 260 i=icol-nc, n      /* (if any)
            t = g(i,ic平)
            v(i) = t
            alfa = alfa + t*t
            continue
260     alfa = sqrt( alfa )
        if( v(icol-nc) .lt. 0.0 ) then
            alfa = - alfa
            end if

```

Program CLS, continued.

```

t = v(icol-nc) + alfa
v(icol-nc) = t
beta = alfa * t
if( abs(beta) .le. test ) then /* div by zero check
   ierr = 3
   end if
do 270 j=1, nc /* rotate first nc columns (if any)
   gama = 0.0
   do 280 i=icol-nc, n
      gama = gama + g(i,j)*v(i)
      continue
   gama = gama / beta
   do 290 i=icol-nc, n
      g(i,j) = g(i,j) - gama*v(i)
      continue
   continue
do 300 j = icol+1, m /* rotate last columns (if any)
   gama = 0.0
   do 320 i=icol-nc, n
      gama = gama + g(i,j)*v(i)
      continue
   gama = gama / beta
   do 330 i=icol-nc, n
      g(i,j) = g(i,j) - gama*v(i)
      continue
   continue
g(icol-nc,icol) = t
v(icol-nc) = -alfa
continue
end if

if( mlast .ne. m ) then /* set last u if loop short
   v(m-nc) = g(m-nc,m)
   end if
end if      /* done with triangularizations

do 580 icol= nc+1, mlast /* rotate d vector (if necessary)
   beta = -v(icol-nc) * g(icol-nc,icol)
   gama = 0.0
   do 510 i=icol-nc, n
      gama = gama + d(i)*g(i,icol)
      continue
   gama = gama / beta
   do 520 i=icol-nc, n
      d(i) = d(i) - gama*g(i,icol)
      continue
   continue

if( nc .gt. 0 ) then /* back-solve for first nc unknowns (if any)
   do 620 irow=1, nc
      sum = 0.0
      do 610 i=1, irow-1
         sum = sum + xm(i) * f(irow,i)
         continue
      if( abs(u(irow)) .le. test ) then
         ierr = 5
         end if

```

Program CLS, continued.

```

620           xm(irow) = ( h(irow) - sum ) / u(irow)
               continue
           end if

           if( m .gt. nc ) then /* back-solve for last m-nc unknowns (if any)
               do 680 irow=m, nc+1, -1
                   sum = 0.0
                   do 640 i=1, nc
                       sum = sum + xm(i) * g(irow-nc,i)
                   continue
                   do 660 i=m, irow+1, -1
                       sum = sum + xm(i) * g(irow-nc,i)
                   continue
                   if( abs(v(irow-nc)) .le. test ) then
                       ierr = 6
                   end if
                   xm(irow) = ( d(irow-nc) - sum ) / v(irow-nc)
               continue
           end if

           sq = 0.0      /* compute residual error
           do 930 irow = m-nc+1, n
               sum = 0.0
               do 1640 i=1, nc
                   sum = sum + xm(i) * g(irow,i)
               continue
               sq = sq + (d(irow)-sum)**2
           continue
           if( (n+nc-m) .gt. 0 ) then
               sq = sqrt( sq / float(n+nc-m) )
           else
               sq = 0.0
           end if

           do 890 irow=nlast, 1, -1      /* rotate solution (if necessary)
               beta = -u(irow) * f(irow,irow)
               gama = 0.0
               do 810 i=irow, m
                   gama = gama + xm(i)*f(irow,i)
               continue
               gama = gama / beta
               do 820 i=irow, m
                   xm(i) = xm(i) - gama*f(irow,i)
               continue
           continue

           return
       end

```

13.4 L_2 Problems with Inequality Constraints

The most general problem involving both inequality and equality constraints can be reduced to the simpler problem of solving $\mathbf{Gm} = \mathbf{d}$ in the least squares sense, with the constraint that \mathbf{m} is nonnegative, by a succession of three transformations (Section 7.8). We shall discuss only this transformed problem, the basic algorithm for which was described in Section 7.9.

The algorithm is straightforward but requires the solution of a succession of least squares problems, each involving one more or one less unknown than the preceding one. Solving so many problems would be time consuming if each were solved separately. Fortunately, the solution of one problem can use intermediate calculations performed in the preceding one [Ref. 14].

Suppose that the least squares problem is being solved by Householder reduction to upper-triangular form, where the Householder transformations are stored in the lower part of the triangularized matrix. Then adding one new unknown requires that the triangle be increased in width by one column. This column can always be added to the right of all the other columns (at the expense of reordering the unknowns). Thus, the enlarged matrix can be triangularized by applying the previous Householder transformations to the new column, finding a new transformation that annihilates the subdiagonal elements of the new column, and applying this new transformation to the

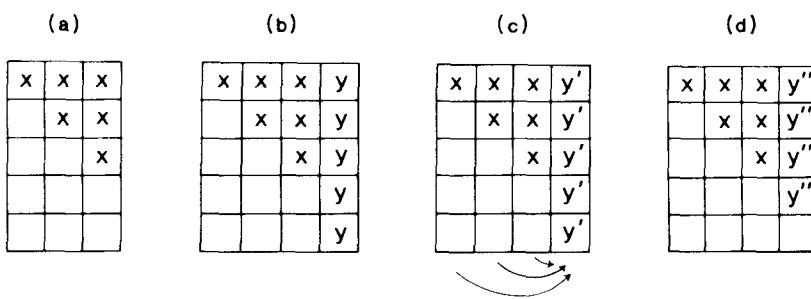


Fig. 13.2. Steps in adding a column to a triangularize matrix. (a) Original matrix. (b) Reorder unknowns so that column is adjoined on right. (c) Transform adjoined column using transformations previously applied to other columns. (d) Annihilate elements below main diagonal of last column (and apply transformation to data vector).

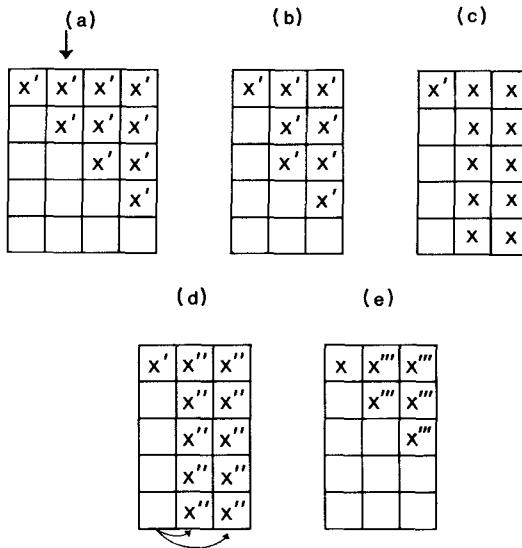


Fig. 13.3. Steps in deleting a column (in this case the second) from a triangularized matrix. (a) Original matrix. (b) Delete column and reorder unknowns. (c) Replace columns to right of deleted columns (and data vector) by original untriangularized ones. (d) Apply transformations from left-hand columns to new columns (and data vector). (e) Apply transformations that annihilate elements below the main diagonal of the right-hand columns.

column and to the right-hand side of the equation (Fig. 13.2). Only one new Householder transformation is required to solve the enlarged problem. Removing a column is more difficult than adding one, since in general it may be anywhere in the triangle. One must remove the effect that partially annihilating this column had on all the columns to its right. The columns can be “untransformed” by reverse application of the appropriate Householder transformations, but this is time consuming. On the other hand, one might simply replace all the right-hand columns by their original, unrotated values and apply all the transformations from the left-hand columns. This second method is faster and less sensitive to the accumulation of round-off error, but it requires that a copy of the unrotated matrix be retained (Fig. 13.3). Once the column is deleted, the right-hand columns must be rediagonalized by the application of new Householder transformations. On average the $(M/2)$ th variable is deleted, so about M applications of $M/2$ old and $M/2$ new transformations are needed to rediagonalize the matrix.

FORTRAN Code (PLS) for the Least Squares Problem with Positivity Constraints

```

subroutine pls(g,xm,d,n,m,nst,mst,gp,dp,e,npst,mpst,test,ierr)
real g(nst,mst), xm(mst), d(nst), test
real dp(npst), e(npst), gp(npst,mpst)
integer n, m, nst, mst, npst, mpst, ierr

/* subroutine PLS solves the overdetermined (or even-determined)
/* linear system g*xm=d in the least squares sense with the
/* constraint that all the unknowns are nonnegative
/* protocol:
/*
/* g: (sent) the n by m data kernel matrix
/*
/* xm: (returned) a vector of length m of unknowns
/*
/* d: (sent) a vector of length n of data
/*
/* n: (sent) the number of data (rows of g, currently limited to be
/* less than or equal to 100)
/*
/* m: (sent) the number of unknowns (columns of g)
/*
/* nst: (sent) the row dimension of g in the calling routine
/*
/* mst: (sent) the column dimension of g in the calling routine
/*
/* gp: (scratch) a scratch matrix at least n by m
/*
/* dp: (scratch) a scratch vector at least n in length
/*
/* e: (scratch) a scratch vector at least n in length
/*
/* npst: (sent) the row dimension of gp in the calling routine
/*
/* mpst: (sent) the column dimension of gp in the calling routine
/*
/* test: (sent) division by numbers smaller than test generates
/* an error condition, solutions within test of zero are considered
/* feasible
/*
/* ierr: (returned) an error flag, zero on no error
/* subroutines needed: ADD, SUB

integer list(100)
real grade(100), z(100)
common /plscom/ mcur, itr(100), igr(100), u(100)
/* these arrays must be redimensioned if n is greater than 100
/* grade is the error gradient, z is a temp vector of unknowns
/* list is for bookkeeping
/* mcu is the current number of unpinned unknowns
/* itr(i) is the triangle column index of the i-th column of gp
/* igr(i) is the gp column index of the i-th column of the triangle
/* u hold the extra coefficient of the Householder transformations

do 100 i=1, n /* initialize temporary vector

```

Program PLS, continued.

```

dp(i) = d(i)
100    continue

mcur = 0
do 1 i=1, m /* at start all variables are pinned to zero
     itr(i)=0
     igr(i)=0
     xm(i)=0.0
     z(i) = 0.0
1      continue

do 16 i=1, 3*m /* main loop over adding variables. There is no
     /* guarantee that 3*m iterations are enough, but
     /* they usually are

     do 3 j=1, n /* error vector
         e(j) = d(j)
         do 2 k=1, m
             e(j) = e(j) - g(j,k)*xm(k)
2         continue
3     continue

     do 5 j=1, m /* gradient of error
         grade(j) = 0.0
         do 4 k=1, n
             grade(j) = grade(j) + g(k,j)*e(k)
4         continue
5     continue

gmax=0.0 /* find maximum gradient
iadd = 0 /* add (unpin) this variable
do 6 j=1, m
    if( itr(j).eq.0 .and. grade(j).gt.gmax ) then
        iadd = j
        gmax = grade(j)
    end if
6    continue

if( gmax .le. 0.0 ) then /* done. there is no variable left
    ierr=0           /* that can be perturbed to reduce
    return           /* the error and remain feasible
end if

/* unpin this variable and solve least
/* squares problem for all unpinned variables
call add(g,z,d,n,m,mst,nst,gp,dp,mpst,npst,test,iadd,ierr)
if( ierr .ne. 0 ) then
    return
end if

zmin = 0.0 /* any negative unknowns?
izmin = 0
do 7 j=1, m
    if( itr(j).ne.0 .and. z(j).lt.zmin) then
        izmin=j
        zmin = z(j)
7

```

Program PLS, continued.

```

    end if
7   continue

if( zmin. ge. -test ) then /* no unknowns became infeasible
   do 8 k=1, m /* swap temp unknowns to unknowns
      if( itr(k).ne.0 ) then
         xm(k) = z(k)
      end if
   continue

8 else /* some became infeasible
   do 14 j=1, m /* get rid of infeasible variables

      alfa = 1.e30 /* a big number for comparison
      do 9 k=1, m /* find largest feasible perturbation
         if( itr(k).ne.0 ) then
            t = xm(k) / (xm(k)-z(k))
            if( t.lt.alfa ) then
               alfa=t
            end if
         end if
      continue

9      do 10 k=1, m /* add perturbation to solution
         if( itr(k).ne.0 ) then
            xm(k)=xm(k)+alfa*(z(k)-xm(k))
         end if
      continue

10     nlist=0 /* find all newly pinned unknowns
      do 11 k=1, m
         if(itr(k).ne.0.and.xm(k).le.test) then
            nlist=nlist+1
            list(nlist)=k
         end if
      continue

11     do 12 k=1, nlist /* pin them
        call sub(g,z,d,n,m,mst,nst,gp,
        &          dp,mpst,npst,test,list(k),ierr)
        if( ierr .ne. 0 ) then
           return
        end if
      continue

12     zmin = 0.0 /* any negative unknowns?
      izmin = 0
      do 13 k=1, m
         if( itr(k).ne.0 .and.
             z(k) .le. zmin ) then
            izmin=k
            zmin=z(k)
         end if
      continue

13     if( izmin.eq.0 .or.
      &       zmin.gt.-test ) then /*break loop
         goto 15 /* all remaining unknowns feasible

```

Program PLS, continued.

```

        end if

14           continue
15       continue
      end if

      do 17 k=1, m
          if( itr(k).eq.0 ) then
              xm(k) = 0.0
          else
              xm(k) = z(k)
          end if
17       continue
16       continue

ierr=2 /* process didn't converge in 3m iterations
return
end

subroutine add(g,xm,d,n,m,nst,mst,gp,dp,npst,mpst,test,iv,ierr)
real g(nst,mst), xm(mst), d(nst), test, gp(npst,mpst), dp(npst)
integer n, m, nst, mst, npst, mpst, ierr

/* subroutine ADD adds unpinned variable iv and finds the least
/* squares solution

common /plscom/ mcur, itr(100), igr(100), u(100)

if( iv.lt.1 .or. iv.gt.m ) then /* no such variable
    ierr=1
    return
end if

if( itr(iv) .ne. 0 ) then /* check that column not already in
    ierr=1
    return
end if

mcur = mcur + 1 /* add one column to triangle
igr(mcur)=iv
itr(iv)=mcur

do 1 i=1, n /* copy column from original data kernel to scratch
    gp(i,iv) = g(i,iv)
1     continue

do 4 j=1, mcur-1 /* transform new column to match rest of
    /* triangle
    k = igr(j) /* location of j-th col of triangle
    beta = -u(k) * gp(j,k)
    gama = 0.0
    do 2 i=j, n
        gama = gama + gp(i,iv)*gp(i,k)
        continue
2     gama = gama / beta
    do 3 i=j, n
        gp(i,iv) = gp(i,iv) - gama*gp(i,k)
3

```

Program PLS, continued.

```

3           continue
4           continue

if( n.gt.mcur ) then /* if overdetermined, zero bottom elements of
/* the new column, and transform d vector

    alfa = 0.0 /* transform column
    do 5 i=mcur, n
        alfa = alfa + gp(i,iv)**2
5       continue
    alfa = sqrt( alfa )
    if( gp(mcur,iv) .lt. 0.0 ) then /* choose sign
        alfa = -alfa
        end if
    u(iv) = -alfa
    t = gp(mcur,iv) + alfa
    if( abs(t) .lt. test ) then /* div by 0?
        ierr=1
        return
        end if
    gp(mcur,iv) = t

    beta = -u(iv) * gp(mcur,iv) /* transform vector
    gama = 0.0
    do 6 i = mcur, n
        gama = gama + dp(i)*gp(i,iv)
6       continue
    gama = gama / beta
    do 7 i = mcur, n
        dp(i) = dp(i) - gama*gp(i,iv)
7       continue
else
    u(iv) = gp(mcur,iv)
end if

do 9 i=mcur,1,-1 /* back-solve for unknowns
    t = dp(i)
    do 8 j=i+1,mcur
        k = igp(j) /* location of j-th column of triangle
        t = t - gp(i,k) * xm(k)
8       continue
    xm(igp(i)) = t / u(igp(i))
    continue

return
end

subroutine sub(g,xm,d,n,m,nst,mst,gp,dp,npst,mpst,test,iv,ierr)
real g(nst,mst), xm(mst), d(nst), test, gp(npst,mpst), dp(npst)
integer n, m, nst, mst, npst, mpst, ierr

/* subroutine SUB subtracts out newly pinned column iv and solves
/* least squares problem

common /plscm/ mcur, itr(100), igp(100), u(100)

```

Program PLS, continued.

```

if( iv.lt.1 .or. iv.gt.m ) then /* no such variable
   ierr=1
   return
end if

ntr = itr( iv ) /* the location of the column in the triangle
if( ntr .eq. 0 ) then /* check that column was in triangle
   ierr = 1
   return
end if

do 2 i=ntr+1, mcur /* throw away columns to right of iv and copy
   /* in untransformed ones
      k = igp(i) /* location of i-th column of triangle
      do 1 j=1, n
         gp(j,k) = g(j,k)
         continue
      1 continue
      2 continue

do 3 i=1, n /* throw away data vector and copy in untransformed one
      dp(i) = d(i)
      continue
  3 continue

do 9 j=1, ntr-1 /* apply transformation from left-hand columns
      k = igp(j) /* location of j-th col (left)
      beta = -u(k) * gp(j,k)

      do 6 jj=ntr+1, mcur /* transform right columns
         kk = igp(jj) /* location of jj-th col (right)
         gama = 0.0
         do 4 i=j, n
            gama = gama + gp(i,kk)*gp(i,k)
            continue
        4 continue
         gama = gama / beta
         do 5 i=j, n
            gp(i,kk) = gp(i,kk) - gama*gp(i,k)
            continue
        5 continue
       6 continue

      gama = 0.0 /* transform data vector
      do 7 i=j, n
         gama = gama + dp(i)*gp(i,k)
         continue
      7 continue
      gama = gama / beta
      do 8 i=j, n
         dp(i) = dp(i) - gama*gp(i,k)
         continue
      8 continue
     9 continue

do 10 i=ntr+1, mcur /* resequence pointers to exclude iv column
      igp(i-1) = igp(i)
      itr(igp(i-1)) = i-1
      continue
  10 continue

itr(iv) = 0
igp(mcur) = 0
mcur = mcur - 1

```

Program PLS, continued.

```

do 17 j = ntr, mcur /* triangularize new columns

    alfa = 0.0 /* build transformation
    k = igp(j) /* location of j-th column of triangle
    do 11 i=j, n
        alfa = alfa + gp(i,k)**2
11     continue
    alfa = sqrt( alfa )
    if( gp(j,k) .lt. 0.0 ) then /* choose sign
        alfa = -alfa
    end if
    u(k) = -alfa
    t = gp(j,k) + alfa
    if( abs(t) .lt. test ) then /* div by 0?
        ierr=1
        return
    end if
    gp(j,k) = t
    beta = -u(k) * gp(j,k)

    do 14 jj = j+1, mcur /* apply transformation to columns
        kk = igp(jj)
        gama = 0.0
        do 12 i=j, n
            gama = gama + gp(i,kk)*gp(i,k)
12     continue
        gama = gama / beta
        do 13 i=j, n
            gp(i,kk) = gp(i,kk) - gama*gp(i,k)
13     continue
14     continue

    gama = 0.0 /* apply transformation to data vector
    do 15 i = j, n
        gama = gama + dp(i)*gp(i,k)
15     continue
    gama = gama / beta
    do 16 i = j, n
        dp(i) = dp(i) - gama*gp(i,k)
16     continue
17     continue

    xm(iv) = 0.0
    do 19 i=mcur,1,-1 /* back-solve for unknowns
        t = dp(i)
        do 18 j=i+1,mcur
            k = igp(j) /* location of j-th column of triangle
            t = t - gp(i,k) * xm(k)
18         continue
        xm(igp(i)) = t / u(igp(i))
19         continue

    return
end

```

FORTRAN Code (MLIC) for the Minimum-Length Solution with Inequality Constraints

```

subroutine mlic(f,xm,h,n,m,nst,mst,fp,test,ierr)
real f(nst,mst), xm(mst), h(nst), test, fp((m+1)*n)
integer n, m, nst, mst, npst, mpst, ierr

/* subroutine MLIC finds the minimum L2 solution length
/* subject to linear inequality constraints  $f^*xm \geq h$ 

/* protocol:
/*
/* h: (sent, altered) the n by m constraint matrix. This matrix
/* must be dimensioned at least  $(m+1)*n$  in size
/*
/* xm: (returned) a vector of length m of unknowns. Must be
/* dimensioned at least n in length
/*
/* h: (sent, altered) a vector of length n of constraints. must be
/* at least m+1 in length
/*
/* n: (sent) the number of data (rows of g, currently limited to be
/* less than or equal to 99)
/*
/* m: (sent) the number of unknowns (columns of g)
/*
/* fp: (scratch) a scratch vector at least  $n*(m+1)$  in length
/*
/* test: (sent) division by numbers smaller than test generates
/* an error condition, solutions within test of zero are considered
/* feasible
/*
/* ierr: (returned) an error flag, zero on no error, -1 on
/* no feasible solution, positive on other errors.

/* subroutines needed: ADD, SUB, PLS

real hp(100), e(100) /* must be redimensioned if n gt 99

do 2 i=1, n /* build (m+1) by n matrix with f-transpose at top
               /* and h-transpose at bottom
    j = (m+1)*(i-1)
    do 1 k=1, m
        fp(j+k) = f(i,k)
        continue
    fp(j+m+1) = h(i)
    continue

1   do 3 i=1, m /* build m+1 vector hp
    hp(i) = 0.0
    continue
3   hp(m+1) = 1.0

/* solve positive least squares problem  $fp^*xm=hp$ 
call pls(fp,xm,hp,m+1,n,m+1,n,f,h,e,m+1,n,test,ierr)
if ( ierr .ne. 0 ) then
    return
end if

xnorm = 0.0

```

Program MLIC, continued.

```
do 5 i=1, m+1 /* compute residuals and their L-2 norm
    e(i) = -hp(i)
    do 4 j=1, n
        e(i) = e(i) + fp((j-1)*(m+1)+i)*xm(j)
4        continue
        xnorm = xnorm + e(i)**2
5        continue
        xnorm = sqrt( xnorm / float(m+1) )

if( xnorm .lt. test ) then
    ierr=-1 /* constraints incompatible
else
    do 6 i=1, m
        xm(i) = -e(i)/e(m+1)
6        continue
    ierr=0
end if

return
end
```

13.5 Finding the Eigenvalues and Eigenvectors of a Real Symmetric Matrix

If the real symmetric matrix A is diagonal, then the process of finding its eigenvalues and eigenvectors is trivial: the eigenvalues are just the diagonal elements, and the eigenvectors are just unit vectors in the coordinate directions. If it were possible to find a transformation that diagonalized a matrix while preserving lengths, then the eigenvalue problem would essentially be solved. The eigenvalues in both coordinate systems would be the same, and the eigenvectors could easily be rotated back into the original coordinate system. (In fact, since they are unit vectors in the transformed coordinate system, they simply *are* the elements of the inverse transform.)

Unfortunately, there is no known method of finding a finite number of unitary transformations that accomplish this goal. It is possible, however, to reduce an arbitrary matrix to either bi-diagonal or tri-diagonal form by pre- and postmultiplying by a succession of Householder transformations. The zeros are placed alternately in the rows and columns, starting two elements from the main diagonal (Fig. 13.4). The tri-diagonal form is symmetric, and therefore simpler to work with. We shall assume that this triangularization has been performed on the matrix A .

At this point there are two possible ways to proceed. One method is to try to further reduce the tri-diagonal matrix into a diagonal one. Although this cannot be done with a finite number of transformations, it can be accomplished with an infinite succession of *Givens rotations*. In practice, only a finite number of rotations need be applied, since at some point the off-diagonal elements become sufficiently small to be ignored. We shall discuss this method in the section on singular-value decomposition (Section 13.6).

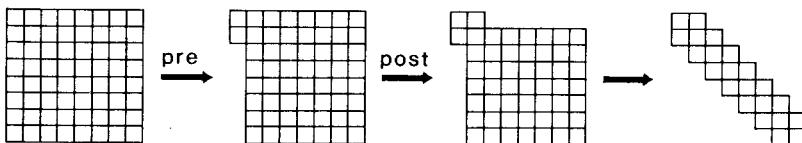


Fig. 13.4. Tri-diagonalization of a square matrix by alternate pre- and post-multiplication by Householder transformations.

The second method is to find the eigenvalues of the tri-diagonal matrix, which are the roots of its characteristic polynomial

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0.$$

When \mathbf{A} is tri-diagonal, these roots can be found by a recursive algorithm.

To derive this algorithm, we consider the 4×4 tri-diagonal matrix

$$\mathbf{A} - \lambda \mathbf{I} = \begin{bmatrix} a_1 - \lambda & b_2 & 0 & 0 \\ c_2 & a_2 - \lambda & b_3 & 0 \\ 0 & c_3 & a_3 - \lambda & b_4 \\ 0 & 0 & c_4 & a_4 - \lambda \end{bmatrix} \quad (13.6)$$

where a_i , b_i , and c_i are the three nonzero diagonals. When the determinant is computed by expanding in minors about the fourth row, there are only two nonzero terms:

$$\det(\mathbf{A} - \lambda \mathbf{I}) = (a_4 - \lambda) \det \begin{bmatrix} a_1 - \lambda & b_2 & 0 \\ c_2 & a_2 - \lambda & b_3 \\ 0 & c_3 & a_3 - \lambda \end{bmatrix} - c_4 \det \begin{bmatrix} a_1 - \lambda & b_2 & 0 \\ c_2 & a_2 - \lambda & 0 \\ 0 & c_3 & b_4 \end{bmatrix} \quad (13.7)$$

Because all but one element in the last column of the second matrix are zero, this expression can be reduced to

$$\det(\mathbf{A} - \lambda \mathbf{I}) = (a_4 - \lambda) \det \begin{bmatrix} a_1 - \lambda & b_2 & 0 \\ c_2 & a_2 - \lambda & b_3 \\ 0 & c_3 & a_3 - \lambda \end{bmatrix} - b_4 c_4 \det \begin{bmatrix} a_1 - \lambda & b_2 \\ c_2 & a_2 - \lambda \end{bmatrix} \quad (13.8)$$

The determinant of the 4×4 matrix \mathbf{A} is related to the determinants of the 3×3 and 2×2 submatrices in the upper left-hand corner of \mathbf{A} . This relationship can easily be generalized to a matrix of any size as

$$p_N = (a_N - \lambda)p_{N-1} - b_N c_N p_{N-2} \quad (13.9)$$

where p_N is the determinant of the $N \times N$ submatrix. By defining $p_{-1} = 0$ and $p_0 = 1$, this equation provides a recursive method for computing the value of the characteristic polynomial for any specific value of λ . This sequence of polynomials is a *Sturm sequence*, which

has two properties that facilitate the locating of the roots: (1) the roots of p_N are separated by the roots of p_{N-1} ; (2) the number of roots of $p_N(\lambda)$ less than λ_0 is equal to the number of sign changes in the sequence $[p_0(\lambda_0), p_1(\lambda_0), \dots, p_N(\lambda_0)]$. The first property is proved in most elementary numerical analysis texts. The second follows from the fact that for large positive λ , p_N and p_{N-1} have opposite signs, since $p_N = (a_N - \lambda)p_{N-1}$ (Fig. 13.5).

Bounds on the locations of the roots of this polynomial can be found by computing the norm of the eigenvalue equation $\lambda\mathbf{x} = \mathbf{Ax}$:

$$|\lambda| \|\mathbf{x}\|_2 = \|\mathbf{Ax}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{x}\|_2 \quad (13.10)$$

The roots must lie on the interval $[-\|\mathbf{A}\|_2, \|\mathbf{A}\|_2]$. If the Sturm sequence is evaluated for some λ_0 on this interval, then the number of roots in the subintervals $[-\|\mathbf{A}\|_2, \lambda_0]$ and $[\lambda_0, \|\mathbf{A}\|_2]$ can be determined by counting the sign changes in the Sturm sequence. By successive subdivision of intervals the position of the roots (the eigenvalues) can be determined to any desired accuracy. This method of solving the eigenvalue problem is known as the Givens–Householder algorithm.

Once the eigenvalues have been determined, the eigenvectors \mathbf{x} can be computed by solving the equation $(\mathbf{A} - \lambda_i \mathbf{I})\mathbf{x}_i = 0$ for each eigenvalue. These equations are underdetermined, since only the directions and not the overall magnitudes of the eigenvectors are determined. If Gauss–Jordan reduction is used to solve the equations, then some pivoting algorithm must be employed to avoid division by near-zero numbers. Another possibility is to solve the equation $[\mathbf{A} - (\lambda_i + \delta)\mathbf{I}]\mathbf{x}_i^j = \mathbf{x}_i^{j-1}$ iteratively, where the initial guess of the eigenvector \mathbf{x}_i^0 is any random vector and $\delta \ll \lambda_i$ some small number,

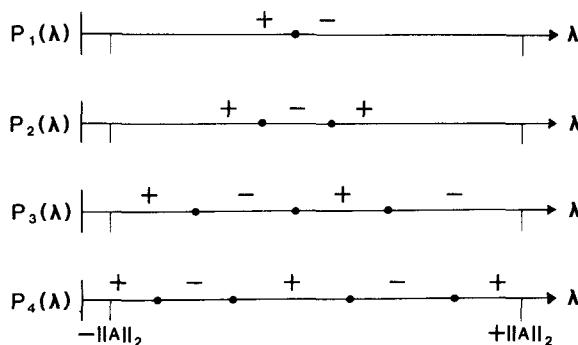


Fig. 13.5. Roots (dots) of the Sturm polynomials $P_n(\lambda)$. Note that the roots of P_{n-1} separate (interleave with) the roots of P_n .

to prevent the equations from becoming completely singular. This iteration typically converges to the eigenvector in only three iterations. Once the eigenvectors have been determined, they are transformed back into the original coordinate system by inverting the sequence of Householder transformations.

13.6 The Singular-Value Decomposition of a Matrix

The first step in this decomposition process is to apply Householder transformations to reduce the matrix \mathbf{G} to bi-diagonal form (in the 4×4 case) (see Section 13.5):

$$\mathbf{G} = \begin{bmatrix} q_1 & e_2 & 0 & 0 \\ 0 & q_2 & e_3 & 0 \\ 0 & 0 & q_3 & e_4 \\ 0 & 0 & 0 & q_4 \end{bmatrix} \quad (13.11)$$

where q_i and e_i are the nonzero diagonals. The problem is essentially solved, since the squares of the singular values of \mathbf{G} are the eigenvalues of $\mathbf{G}^T\mathbf{G}$, which is tri-diagonal with diagonal elements $a_i = q_i^2$ and off-diagonal elements $b_i = c_i = e_i q_{i-1}$. We could therefore proceed by the symmetric matrix algorithm of Section 13.5. However, since this algorithm would be computing the squares of the singular values, it would not be very precise. Instead, it is appropriate to deal with the matrix \mathbf{G} directly, using the *Golub–Reinch method*.

We shall examine the computation of the eigenvalues of a tri-diagonal matrix by transformation into a diagonal matrix first. Suppose that the symmetric, tri-diagonal $N \times N$ matrix \mathbf{A} is transformed by a unitary transformation into another symmetric tri-diagonal matrix $\mathbf{A}' = \mathbf{T}\mathbf{A}\mathbf{T}^T$. It is possible to choose this transformation so that the $b'_N = c'_N$ elements become smaller in absolute magnitude. Repeated application of this procedure eventually reduces these elements to zero, leaving a matrix consisting of an $N - 1 \times N - 1$ tri-diagonal block and a 1×1 block. The latter is its own eigenvalue. The whole process can then be repeated on the $N - 1 \times N - 1$ matrix to yield an $N - 2 \times N - 2$ block and another eigenvalue. The transformation $\mathbf{A}' = \mathbf{T}\mathbf{A}\mathbf{T}^T$ is accomplished by a coordinate shift. For any *shift parameter* σ , $\mathbf{A}' = \mathbf{T}\mathbf{A}\mathbf{T}^T = \mathbf{T}[\mathbf{A} - \sigma\mathbf{I}]\mathbf{T}^T + \sigma\mathbf{I}$. It can be shown that if σ is

chosen to be the eigenvalue of the lower-right 2×2 submatrix of \mathbf{A} that is closest in value to A_{NN} and if \mathbf{T} is chosen to make $\mathbf{T}[\mathbf{A} - \sigma\mathbf{I}]$ upper triangular, then \mathbf{A}' is triangular and the $c'_N = b'_N$ elements of \mathbf{A}' are smaller than the corresponding elements of \mathbf{A} .

The singular values of \mathbf{G} are found by using the Householder–Reinch algorithm to diagonalize $\mathbf{A} = \mathbf{G}^T \mathbf{G}$. All of the operations are performed on the matrix \mathbf{G} , and \mathbf{A} is never explicitly computed. The shift parameter is first computed by finding the eigenvalues of the 2×2 matrix

$$\mathbf{M} = \begin{bmatrix} q_{N-1}^2 + e_{N-1}^2 & e_N q_{N-1} \\ e_N q_{N-1} & q_N^2 + e_N^2 \end{bmatrix} \quad (13.12)$$

and selecting the eigenvalue closest to $(q_N^2 + e_N^2)$. A transformation must then be found that makes $\mathbf{T}(\mathbf{A} - \sigma\mathbf{T})$ upper-triangular. It can be shown [Ref. 14] that if \mathbf{A} and $\mathbf{T}\mathbf{A}\mathbf{T}^T$ are symmetric, tri-diagonal matrices and if all but the first element of the first column of $\mathbf{T}(\mathbf{A} - \sigma\mathbf{I})$ are zero, then $\mathbf{T}(\mathbf{A} - \sigma\mathbf{I})$ is upper triangular. The effect \mathbf{T} has on \mathbf{G} can be controlled by introducing another unitary transformation \mathbf{R} and noting that if $\mathbf{G}' = \mathbf{R}^T \mathbf{G} \mathbf{T}^T$, then $\mathbf{A}' = \mathbf{G}'^T \mathbf{G}' = \mathbf{T} \mathbf{G}^T \mathbf{R} \mathbf{R}^T \mathbf{G} \mathbf{T}^T = \mathbf{T} \mathbf{G}^T \mathbf{G} \mathbf{T}^T$ is not a function of \mathbf{R} .

The transformations \mathbf{T} and \mathbf{R} are constructed out of a series of Givens transformations. These transformations are a special case of Householder transformations designed to change only two elements of a vector, one of which is annihilated. Suppose that the elements v_i and v_j are to be altered, v_j being annihilated. Then the transformation $T_{ii} = T_{jj} = c$, $T_{ij} = -T_{ji} = s$ (all other elements being zero) will transform $v'_i = (v_i^2 + v_j^2)^{1/2}$, $v'_j = 0$ if $c = v_i/(v_i^2 + v_j^2)^{1/2}$ and $s = v_j/(v_i^2 + v_j^2)^{1/2}$. The first Givens transformation in \mathbf{T} is chosen to annihilate the subdiagonal element of the first column of $\mathbf{T}(\mathbf{A} - \sigma\mathbf{I})$, $[q_1^2 - \sigma, q_1 e_1, 0, 0, \dots, 0]^T$. Now $\mathbf{G}\mathbf{T}_1^T$ is no longer bi-diagonal, so a series of pre- and postmultiplied Givens transformations are applied which bi-diagonalize it. These transformations define the rest of \mathbf{T} and \mathbf{R} . Their application is sometimes called *chasing* since, after each application of a transformation, a nonzero element appears at a new (off-bi-diagonal) position in \mathbf{G} . This step assures that $\mathbf{T}\mathbf{A}\mathbf{T}^T$ is tri-diagonal. The whole process is iterated until the last subdiagonal element of $\mathbf{T}\mathbf{A}\mathbf{T}^T$ becomes sufficiently small to be ignored. At this stage e'_N must also be zero, so one singular value q_N has been found. The whole process is then repeated on a \mathbf{G} the size of which has been reduced by one row and one column.

13.7 The Simplex Method and the Linear Programming Problem

In the linear programming problem, we must determine a vector \mathbf{x} that maximizes the *objective function* $z = \mathbf{c}^T \mathbf{x}$ with the equality and inequality constraints $\mathbf{Ax}[\leq, =, \geq] \mathbf{b}$ and with the nonnegativity constraint $\mathbf{x} \geq 0$. As discussed in Section 7.9, the constraints (presuming they are consistent) define a convex polyhedron in solution space, within which the solution must lie. Since the objective function is linear, it can have no maxima or minima. The solution must therefore lie on the boundary of the polyhedron, at one of its vertices (or “extreme points”). In many instances there is a unique solution, but there may be no solution if the constraints are inconsistent, an infinite solution if the polyhedron is not closed in the direction of increasing z , or an infinite number of finite solutions if the solution lies on an edge or face of the polyhedron that is normal to ∇z (Fig. 13.6).

The Simplex method is typically used to solve the linear programming problem. It has two main steps. The first step (which, for historical reasons, is usually called the “second phase”) consists of locating one extreme point on the polyhedron defined by $\mathbf{Ax} [\leq, =, \geq] \mathbf{b}$ that is also feasible in the sense that it satisfies $\mathbf{x} \geq 0$. The second step (the “first phase”) consists of jumping from this extreme point to another neighboring one that is also feasible and that has a larger value of the objective function. When no such extreme point can be found, the solution of the linear programming problem has been found. There is no guarantee that this search procedure will eventually reach

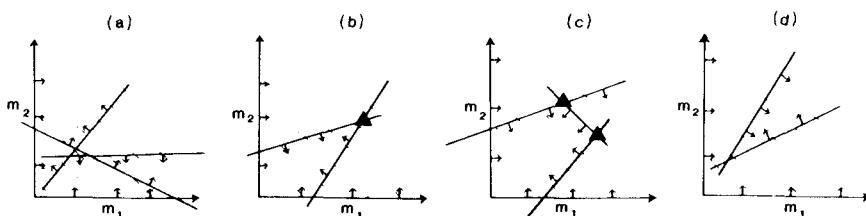


Fig. 13.6. Depending on the objective function (dashed contours) and the shape of the feasible volume defined by the inequality constraints (solid lines with arrows pointing into feasible halfspace), linear programming problems can have (a) no solution, (b) one solution (triangle), (c) an infinite number of finite solutions (on line segment bounded by triangles), or (d) an infinite solution.

the “optimum” extreme point (closed loops are possible), but it usually does.

The inequality constraints $\mathbf{Ax} \leq \mathbf{b}$ can be converted to equality constraints by changing all the \geq constraints to \leq constraints by multiplication by -1 , and then changing all the \leq constraints to equality constraints by increasing the number of unknowns. For example, if the problem originally has two unknowns and a constraint $x_1 + 2x_2 \leq 4$, it can be converted into the equality constraint $x_1 + 2x_2 + x_3 = 4$ by adding the unknown x_3 and requiring that $x_3 \geq 0$. The transformed linear programming problem therefore has, say, M unknowns (original ones plus the new “slack” variables) \mathbf{x} , objective function $z = \mathbf{c}^T \mathbf{x}$ (where \mathbf{c} has been extended with zeros), and, say, N linear equality constraints $\mathbf{Ax} = \mathbf{b}$. Since one slack variable was added for each inequality constraint, these equations will usually be under-determined ($M > N$).

If the i th column of \mathbf{A} is denoted by the vector \mathbf{v}_i , the equations $\mathbf{Ax} = \mathbf{b}$ can be rewritten as

$$\mathbf{v}_1 x_1 + \mathbf{v}_2 x_2 + \cdots + \mathbf{v}_M x_M = \mathbf{b} \quad (13.13)$$

The vector \mathbf{b} is of length N . It can be represented by a linear combination of only $N \leq M$ of the \mathbf{v}_i . We can find one solution to the constraint equation by choosing N linearly independent columns of \mathbf{A} , expanding \mathbf{b} on this basis, and noting that N components of the solution vector are the expansion coefficients (the rest being zero):

$$x_1 \mathbf{v}_1 + x_2 \mathbf{v}_2 + \cdots + x_N \mathbf{v}_N + 0 \mathbf{v}_{N+1} + \cdots + 0 \mathbf{v}_M = \mathbf{b} \quad (13.14)$$

Any solution that can be represented by a sum of precisely M columns of \mathbf{A} is said to be a *basic solution*, and the N vectors \mathbf{v}_i are said to form the *basis* of the problem. If all the x_i 's are also nonnegative, the problem is then said to be *basic-feasible*.

The key observations used by the Simplex algorithm are that all basic-feasible solutions are extreme points of the polyhedron and that all extreme points of the polyhedron are basic-feasible solutions. The first assertion is proved by contradiction. Assume that \mathbf{x} is both basic-feasible and *not* an extreme point. If \mathbf{x} is not an extreme point, then it lies along a line segment bounded by two distinct vectors \mathbf{x}_1 and \mathbf{x}_2 that are feasible points (that is, are within the polyhedron):

$$\mathbf{x} = \lambda \mathbf{x}_1 + (1 - \lambda) \mathbf{x}_2 \quad \text{where } 0 < \lambda < 1 \quad (13.15)$$

Since \mathbf{x} is basic, all but N of its elements are zero. Let the unknowns be

ordered so that the nonzero elements are first. Then

$$\mathbf{x}_i = 0 = \lambda \mathbf{x}_{1i} + (1 - \lambda) \mathbf{x}_{2i}; \quad i = (N+1), \dots, M \quad (13.16)$$

but since $\mathbf{x}_1 \geq 0$, $\mathbf{x}_2 \geq 0$, $\lambda > 0$, and $(1 - \lambda) > 0$, the last $M - N$ elements of both \mathbf{x}_1 and \mathbf{x}_2 must be identically zero. Since \mathbf{x} , \mathbf{x}_1 , and \mathbf{x}_2 are all assumed to be feasible (that is, to solve $\mathbf{Ax} = \mathbf{b}$) we have

$$\begin{aligned} \mathbf{b} &= \mathbf{v}_1 x_1 + \dots + \mathbf{v}_N x_N = \mathbf{v}_1 x_{11} + \dots + \mathbf{v}_N x_{1N} \\ &= \mathbf{v}_1 x_{21} + \dots + \mathbf{v}_N x_{2N} \end{aligned}$$

This equation represents the same vector \mathbf{b} being expanded in three seemingly different combinations of the vectors \mathbf{v}_i . But since the \mathbf{v}_i 's are linearly independent, their expansion coefficients are unique and $\mathbf{x} = \mathbf{x}_1 = \mathbf{x}_2$. The solution is therefore not within the polyhedron and must lie on its surface: it must be an extreme point. The converse assertion is proved in a similar manner.

Once one basic-feasible solution of the linear programming problem has been found, other basic solutions are scanned to see if any are feasible *and* have larger objective functions. This is done by removing one vector from the basis and replacing it with a vector formerly outside the basis. The decision on which two vectors to exchange must be based on three considerations: (1) the basis must still span an N th dimensional space, (2) \mathbf{x} must still be feasible, and (3) the objective function must be increased in value. Suppose that \mathbf{v}_q is to be exchanged and \mathbf{v}_p is to replace it. Condition (1) requires that \mathbf{v}_p and \mathbf{v}_q are not orthogonal. If \mathbf{v}_p is expanded in the old basis, the result is

$$\mathbf{v}_p = \alpha_{pq} \mathbf{v}_q + \sum_{\substack{i=1 \\ i \neq q}}^N \alpha_{ip} \mathbf{v}_i \quad (13.17)$$

Noting that $\mathbf{b} = \sum_i x_i \mathbf{v}_i$ and solving for \mathbf{v}_q yields

$$\mathbf{b} = \frac{x_q}{\alpha_{qp}} \mathbf{v}_q + \sum_{\substack{i=1 \\ i \neq q}}^N \left[x_i - \frac{x_q \alpha_{ip}}{\alpha_{qp}} \right] \mathbf{v}_i \quad (13.18)$$

Condition (2) implies that all the coefficients of the \mathbf{v}_i must be positive. Since all the x_i 's of the old solution were positive, condition (2) becomes

$$\alpha_{pq} \geq 0 \quad \text{and} \quad x_i - x_q \alpha_{ip} / \alpha_{qp} \geq 0 \quad (13.19)$$

The first inequality means that we must choose a positive α_{pq} . The

second is only important if α_{ip} is negative, since otherwise the equality must be true. Therefore, given a v_p to be exchanged into the basis, one can remove any v_q for which $\alpha_{pq} \geq 0$. It is convenient to pick the vector with smallest x_q/α_{qp} , since the coefficients of the other vectors experience the smallest change. Condition (2) has identified which vectors can be exchanged into the basis and which one would then be removed. Typically, there will be several possible choices for v_q (if there are none, then the problem is solved).

Condition (3) selects which one will be used. If z_0 and z_1 are the values of the objective function before and after the exchange, respectively, then the change in the objective function can be computed by inserting the above expansions into the formula $z = \mathbf{c}^T \mathbf{x}$ as

$$z_1 - z_0 = x_q/\alpha_{qp} \left[-\sum_{i=1}^N c_i \alpha_{ip} + c_q \right] = \theta_p(c_q - z_p) \quad (13.20)$$

where $\theta_p = x_q/\alpha_{qp}$ and $z_p = \sum_i c_i \alpha_{ip}$ and $\theta_p \geq 0$ since only feasible solutions are considered. Any vector for which $(c_q - z_p) > 0$ improves the objective function. Typically, the vector with largest improvement is exchanged in. If there are vectors that can be exchanged in but none improves the objective function, then the linear programming problem has been solved. If there are no vectors that can be exchanged in but $(c_q - z_p) > 0$ for some other vector, then it can be shown that the solution to the linear programming problem is infinite.

The second phase of the Simplex algorithm consists of finding one basic-feasible solution with which to start the search algorithm. This initial solution is usually found by multiplying the equations $\mathbf{Ax} = \mathbf{b}$ by 1 or -1 to make \mathbf{b} nonnegative and then adding N new artificial variables x_A to the problem (\mathbf{A} being augmented by an identity matrix). One basic-feasible solution to this modified problem is then just $\mathbf{x} = 0$, $\mathbf{x}_A = \mathbf{b}$. The modified problem is then solved with an objective function that tends to drive the artificial variables out of the basis (for instance, $\mathbf{c} = [0, \dots, 0, -1, \dots, -1]^T$). If at any point all the artificial variables have left the basis, then \mathbf{x} is a basic-feasible solution to the original problem. Special cases arise when the optimal solution of the modified problem is either infinite or contains artificial variables. Careful examination shows that these cases correspond to the original problem having no feasible solution or having an infinite solution.

This page intentionally left blank

14

APPLICATIONS OF INVERSE THEORY TO GEOPHYSICS

14.1 Earthquake Location and Determination of the Velocity Structure of the Earth from Travel Time Data

The problem of determining the source parameters of an earthquake (that is, its location and origin time) and the problem of determining the velocity structure of the earth are very closely coupled in geophysics, because they both use measurements of the arrival times of earthquake-generated elastic waves at seismic stations as their primary data.

In areas where the velocity structure is already known, earthquakes can be located using Geiger's method (see Section 12.9). Most commonly, the velocity structure is assumed to be a stack of homogeneous layers (in which case a simple formula for the partial derivatives of travel time with source parameters can be used) and each earthquake is located separately. The data kernel of the linearized problem is an $N \times 4$ matrix and is typically solved by singular-value decomposition. A very commonly used computer program is HYPOINVERSE [Ref. 43].

Arrival time data often determine the relative locations of a group of neighboring earthquakes better than their mean location, because inaccuracies in the velocity model cause similar errors in the predicted arrival time of elastic waves from all the earthquakes at a given station.

Since the pattern of location is critical to the detection of geological features such as faults, clusters of earthquakes are often simultaneously located. The earthquake locations are parameterized in terms of the mean location (or a constant, reference location) and their deviation from the mean [Ref. 41].

At the other extreme is the case where the source parameters of the earthquakes are known and the travel time of elastic waves through the earth is used to determine the elastic wave velocity $c(x)$ in the earth. The simplest case occurs when the velocity can be assumed to vary only with depth z (we are ignoring the sphericity of the earth). Then the travel time T and distance X of an elastic ray are

$$\begin{aligned} T(p) &= 2 \int_0^{z_p} c^{-1}(1 - c^2 p^2)^{-1/2} dz \quad \text{and} \\ X(p) &= 2 \int_0^{z_p} c p(1 - c^2 p^2)^{-1/2} dz \end{aligned} \tag{14.1}$$

Here z_p solves $c(z_p) = 1/p$ and is called the turning point of the ray (since the seismic ray has descended from the surface of the earth down to its deepest depth and is now beginning to ascend). The travel time and distance are parameterized in terms of the ray parameter p , which is constant along the ray and related to the angle θ from the vertical by $p = \sin(\theta)/c$. One problem that immediately arises is that the T and X are not ordinarily measured at a suite of known p . Instead, travel time $T(X)$ is measured as a function of distance X and $p(X)$ is estimated from the relation $p = dT/dX$. This estimation is a difficult inverse problem in its own right, since both T and X can have considerable measurement error. Current methods of solution, based on requiring *a priori* smoothness in the estimated $p(X)$, have major failings.

The integrals in Eq. (14.1) form a nonlinear continuous inverse problem for model function $c(z)$. Note that the nonlinearity extends to the upper limit of integration, which is itself a function of the unknown velocity structure. This limit is often replaced by an approximate, constant limit based on an assumed $v(z)$. Fermat's principle, which states that first-order errors in the ray path cause only second-order errors in the travel time, justifies this approximation. A second problem arises from the singularities in Eq. (14.1) that occur at the turning point. The singularities can be eliminated by working with the tau function $\tau(p) = T(p) - pX(p)$, since

$$\tau(p) = 2 \int_0^{z_p} (1 - c^2 p^2)^{1/2} dz \tag{14.2}$$

Bessonova *et al.* [Ref. 26] describe a method for determining $c(z)$ based on extremal estimates of tau. A different possibility, described by Johnson and Gilbert [Ref. 40], is to linearize Eq. (14.2) by assuming that the velocity can be written $c(z) = c_0(z) + \delta c(z)$, where c_0 is known *a priori* and δc is small. Then Eq. (14.2) becomes a linear continuous inverse problem in δc :

$$\delta\tau = \tau(p) - 2 \int_0^{z_p} (1 - c_0^2 p^2)^{1/2} dz = 2 \int_0^{z_p} c_0^{-3} (1 - c_0^2 p^2)^{-1/2} \delta c dz \quad (14.3)$$

The integrand is singular, but this problem can be eliminated either by integrating Eq. (14.3) by parts, so that the model function becomes a depth derivative of δc , or by discretizing the problem by representing the velocity perturbation as $\delta c = \sum_i^M m_i f_i(z)$, where $f_i(z)$ are known functions (see Section 11.3) and performing the integral analytically. Kennett and Orcutt [Ref. 42] examine the model estimates and resolution kernels for this type of inversion.

Equation (14.1) can also be linearized by a transformation of variables, presuming that $c(z)$ is a monotonically increasing function of z [Refs. 25 and 32].

$$\begin{aligned} T(p) &= 2 \int_{c(z=0)}^{c=1/p} c^{-1} (1 - c^2 p^2)^{-1/2} \frac{dz}{dc} dc \quad \text{and} \\ X(p) &= 2 \int_{c(z=0)}^{c=1/p} cp (1 - c^2 p^2)^{-1/2} \frac{dz}{dc} dc \end{aligned} \quad (14.4)$$

The model function is the inverse velocity gradient as a function of velocity. The singularities in the integrals are removed through discretization $dz/dc = \sum_i^M m_i f_i(c)$ and analytic integration, as in the previous problem. Dorman and Jacobson [Ref. 32] also discuss the improvement in solution that can be obtained by inverting both the tau and the zeta function $\zeta(p) = T + pX$.

The assumption of vertical stratification of the velocity function is often violated, the velocity structure varying in all three dimensions. The inverse problem is then properly one of tomographic inversion, since travel time is given by

$$T = \int_{\text{ray}} c^{-1} ds \quad \text{or} \quad T - \int_{\substack{\text{unperturbed} \\ \text{ray}}} c_0^{-1} ds \approx - \int_{\substack{\text{unperturbed} \\ \text{ray}}} c_0^{-2} \delta c ds \quad (14.5)$$

where s is arc length along the ray. The second expression in Eq. (14.5) has been linearized by assuming that the velocity is some small perturbation δc about a reference velocity c_0 . (Note that Fermat's principle has also been used to write the line integral over the ray path of the ray in the unperturbed velocity structure.) This type of inversion has been used to image the earth at many scales, including small areas of the crust [Ref. 66], the mantle [Refs. 33 and 38], and the core–mantle boundary [Ref. 29]. An alternative to model estimation is presented by Vasco [Ref. 65], who uses extremal inversion to determine quantities such as the maximum difference between the velocities in two different parts of the model, assuming that the velocity perturbation is everywhere within a given set of bounds (see Chapter 6).

Tomographic inversions are hindered by the fact that the earthquake source parameters are often imprecisely known, leading to errors in the estimation of travel time along the ray. Aki *et al.* [Ref. 24] eliminate this problem by limiting their data to rays from very distant earthquakes (teleseisms). All the rays from a single earthquake travel along very similar paths, except in the area being imaged, and all have the same unknown error in travel time. The authors show that if the primary data are the time differences between these rays, as contrasted to their absolute travel times, then horizontal variations in velocity can be determined uniquely even when the origin times of the earthquakes are completely unknown. Vertical variations in travel time are not determined. This method has been applied in many areas of the world [e.g., Refs. 35 and 47].

When both the velocity structure and earthquake source parameters are poorly known, they must be solved for simultaneously. A very useful algorithm due to Pavlis and Booker [Ref. 57] simplifies the inversion of the very large matrices that result from simultaneously inverting for the source parameter of many hundreds of earthquakes and the velocity within many thousands of blocks in the earth model. Suppose that the problem has been converted into a standard discrete linear problem $\mathbf{d} = \mathbf{Gm}$, where the model parameters can be arranged into two groups $\mathbf{m} = [\mathbf{m}_1, \mathbf{m}_2]^T$, where \mathbf{m}_1 is a vector of the earthquake source parameters and \mathbf{m}_2 is a vector of the velocity parameters. Then the inverse problem can be written $\mathbf{d} = \mathbf{Gm} = [\mathbf{G}_1, \mathbf{G}_2][\mathbf{m}_1, \mathbf{m}_2]^T = \mathbf{G}_1\mathbf{m}_1 + \mathbf{G}_2\mathbf{m}_2$. Now suppose that \mathbf{G}_1 has singular-value decomposition $\mathbf{G}_1 = \mathbf{U}_1 \Lambda_1 \mathbf{V}_1^T$, with p nonzero singular values, so that $\mathbf{U}_1 = [\mathbf{U}_{1p}, \mathbf{U}_{10}]$ can be partitioned into two corresponding submatrices, where $\mathbf{U}_{10}^T \mathbf{U}_{1p} = 0$. Then the data can be written as $\mathbf{d} = \mathbf{d}_1 + \mathbf{d}_2 =$

$\mathbf{N}_1\mathbf{d} + (\mathbf{I} - \mathbf{N}_1)\mathbf{d}$, where $\mathbf{N}_1 = \mathbf{U}_{1p}\mathbf{U}_{1p}^T$ is the data resolution matrix associated with \mathbf{G}_1 . The inverse problem is then of the form $\mathbf{d}_1 + \mathbf{d}_2 = \mathbf{G}_1\mathbf{m}_1 + \mathbf{G}_2\mathbf{m}_2$. Premultiplying the inverse problem by \mathbf{U}_{10}^T yields $\mathbf{U}_{10}^T\mathbf{d}_2 = \mathbf{U}_{10}^T\mathbf{G}_2\mathbf{m}_2$ since $\mathbf{U}_{10}^T\mathbf{N}_1 = \mathbf{U}_{10}^T\mathbf{U}_{1p}\mathbf{U}_{1p}^T = 0$ and $\mathbf{U}_{10}^T\mathbf{G}_1 = \mathbf{U}_{10}^T\mathbf{U}_{1p}\Lambda_{1p}\mathbf{V}_{1p}^T = 0$. Premultiplying the inverse problem by \mathbf{U}_{1p}^T yields $\mathbf{U}_{1p}^T\mathbf{d}_2 = \mathbf{U}_{1p}^T\mathbf{G}_1\mathbf{m}_1 + \mathbf{U}_{1p}^T\mathbf{G}_2\mathbf{m}_2$. The inverse problem has been partitioned into two problems, a problem involving only \mathbf{m}_2 (the velocity parameters) that can be solved first, and a problem involving both \mathbf{m}_1 and \mathbf{m}_2 (the velocity and source parameters) that can be solved second. Furthermore, the source parameter data kernel \mathbf{G}_1 is block diagonal, since the source parameters of a given earthquake affect only the travel times associated with that earthquake. The problem of computing the singular-value decomposition of \mathbf{G}_1 can be broken down into the problem of computing the singular-value decomposition of the submatrices. Simultaneous inversions are frequently used, both with the assumption of a vertically stratified velocity structure [Ref. 30] and with three-dimensional heterogeneities [Refs. 22 and 64].

14.2 Velocity Structure from Free Oscillations and Seismic Surface Waves

The earth, being a finite body, can oscillate only with certain characteristic patterns of motion (eigenfunctions) at certain corresponding discrete frequencies ω_{knm} (eigenfrequencies), where the indices (k, n, m) increase with the complexity of the motion. Most commonly, the frequencies of vibration are measured from the spectra of seismograms of very large earthquakes, and the corresponding indices are estimated by comparing the observed frequencies of vibration to those predicted using a reference model of the earth's structure. The inverse problem is to use small differences between the measured and predicted frequencies to improve the reference model.

The most fundamental part of this inverse problem is the relationship between small changes in the earth's material properties (the possibly anisotropic compressional and shear wave velocities and density) and the resultant changes in the frequencies of vibration. These relationships are computed through the variational techniques described in Section 12.10 and are three-dimensional analogs of Eq. (12.41) [Ref. 62]. All involve integrals of the reference earth structure with the eigenfunctions of the reference model. The problem is one in linearized

continuous inverse theory. The complexity of this problem is determined by the assumptions made about the earth model — whether it is isotropic or anisotropic, radially stratified or laterally heterogeneous — and how it is parameterized. A good example of a radially stratified inversion is the Preliminary Reference Earth Model (PREM) of Dziewonski and Anderson [Ref. 34], which also incorporates travel time data, while good examples of laterally heterogeneous inversions are those of Masters *et al.* [Ref. 46] and Giardini *et al.* [Ref. 36].

Seismic surface waves occur when seismic energy is trapped near the earth's surface by the rapid increase in seismic velocities with depth. The phase velocity of these waves is very strongly frequency dependent, since the lower-frequency waves extend deeper into the earth than the higher-frequency waves and are affected by correspondingly higher velocities. A surface wave emitted by an impulsive source such as an earthquake is not itself impulsive, but rather dispersed, since its component frequencies propagate at different phase velocities. The variation of phase velocity $c(\omega)$ with frequency ω can easily be measured from seismograms of earthquakes, as long as the source parameters of the earthquake are known. The inverse problem is to infer the earth's material properties (the possibly anisotropic compressional and shear wave velocities and density) from measurements of the phase velocity function. Since this relationship is very complex, it is usually accomplished through linearization about a reference earth model.

As in the case of free oscillations, the partial derivatives of phase velocity with respect to the material properties of the medium can be calculated using a variational approach, and are similar in form to Eq. (12.41) [e.g., see Ref. 23, Section 7.3]. (In fact, seismic surface waves are just a limiting, high-frequency case of free oscillations, so the two are closely related indeed.) The earth's material properties are usually taken to vary only with depth, so the model functions are all one dimensional.

Sometimes, the surface waves are known to have traveled through several regions of different structure (continents and oceans, for example). The measured phase velocity function $c(\omega)$ is then not characteristic of any one region but is an average. This problem can be handled by assuming that each point on the earth's surface has its own phase velocity function $c(\omega, \mathbf{x})$, related to the observed path-averaged phase velocity $c_a(\omega)$ by

$$c_a^{-1}(\omega) = \frac{1}{S} \int_{\text{path}} \frac{ds}{c(\omega, \mathbf{x})} \quad (14.6)$$

Here the path is taken to be a great circle on the earth's surface, S is its length, and \mathbf{x} is a two-dimensional position vector on the earth's surface. This is a tomography problem for $c(\omega, \mathbf{x})$. Two general approaches are common: either to use a very coarse discretization of the surface of the earth into regions of geological significance [e.g., Ref. 68] or to use a very fine discretization (or parameterization in terms of spherical harmonic expansions) that allows a completely general variation of phase velocity with position [e.g., Refs. 50 and 63]. The local phase velocity functions can then be individually inverted for local, depth-dependent, structure.

14.3 Seismic Attenuation

Internal friction within the earth causes seismic waves to lose amplitude as they propagate. The attenuation can be characterized by the attenuation factor $\alpha(\omega, \mathbf{x})$, which can vary with frequency ω and position x . The fractional loss of amplitude is then

$$\frac{A}{A_0} = \exp\left(-\int_{\text{ray}} \alpha(\omega, \mathbf{x}) ds\right) \quad (14.7)$$

Here A_0 is the initial amplitude. This problem is identical in form to the x-ray imaging problem described in Section 1.3.4. Jacobson *et al.* [Ref. 39] discuss its solution when the attenuation is assumed to vary only with depth. In the general case it can be solved (after linearization) by standard tomographic techniques [e.g., see Ref. 66].

Attenuation also has an effect on the free oscillations of the earth, causing a widening in the spectral peaks associated with the eigenfrequencies ω_{knm} . The amount of widening depends on the relationship between the spatial variation of the attenuation and the spatial dependence of the eigenfunction and can be characterized by a quality factor Q_{knm} (that is, a fractional loss of amplitude per oscillation) for that mode. Variational methods are used to calculate the partial derivative of Q_{knm} with the attenuation factor. One example of such an inversion is that of Masters and Gilbert [Ref. 45].

14.4 Signal Correlation

Geologic processes record signals only imperfectly. For example, while variations in oxygen isotopic ratios $r(t)$ through geologic time t are recorded in oxygen-bearing sediments as they are deposited, the

sedimentation rate is itself a function of time. Measurements of isotopic ratio $r(z)$ as a function of depth z cannot be converted to the variation of $r(t)$ without knowledge of the sedimentation function $t(z)$ [or equivalently $t(z)$].

Under certain circumstances, the function $r(t)$ is known *a priori* (for instance, oxygen isotopic ratio correlates with temperature, which can be estimated independently). In these instances it is possible to use the observed $r^{\text{obs}}(z)$ and the predicted $r^{\text{pre}}(t)$ to invert for the function $t(z)$. This is essentially a problem in signal correlation: distinctive features that can be correlated between $r^{\text{obs}}(z)$ and $r^{\text{pre}}(t)$ establish the function $t(z)$. The inverse problem is

$$r^{\text{obs}}(z) = r^{\text{pre}}[t(z)] \quad (14.8)$$

and is therefore a problem in nonlinear continuous inverse theory. The unknown function $t(z)$ —often called the mapping function—must increase monotonically with z . The solution of this problem is discussed by Martinson *et al.* [Ref. 44] and Shure and Chave [Ref. 59].

14.5 Tectonic Plate Motions

The motion of the earth's rigid tectonic plates can be described by an Euler vector ω , whose orientation gives the pole of the rotation and whose magnitude gives its rate. Euler vector can be used to represent relative rotations, that is, the rotation of one plate relative to another, or absolute motion, that is, motion relative to the earth's mantle. If we denote the relative rotation of plate A with respect to plate B as ω_{AB} , then the Euler vectors of three plates A, B, and C satisfy the relationship $\omega_{AB} + \omega_{BC} + \omega_{CA} = 0$. Once the Euler vectors for plate motion are known, the relative velocity between two plates at any point on their boundary can easily be calculated from trigonometric formulas.

Several geologic features provide information on the relative rotation between plates, including the faulting directions of earthquakes at plate boundaries and the orientation of transform faults, which constrain the direction of the relative velocity vectors, and spreading rate estimates based on magnetic lineations at ridges, which constrain the magnitude of the relative velocity vectors.

These data can be used in an inverse problem to determine the Euler vectors [Refs. 28, 31, 48, and 49]. The main difference between various

authors' approaches is in the manner in which the Euler vectors are parameterized: some authors use their cartesian components; others use their magnitude, azimuth, and inclination. Since the relationship between the magnitude and direction of the relative velocity vectors (the data) and either choice of these model parameters is nonlinear, the two inversions can behave somewhat differently in the presence of noise. Parameterizations in terms of the cartesian components of the Euler vectors seem to produce somewhat more stable inversions.

14.6 Gravity and Geomagnetism

Inverse theory plays an important role in creating representations of the earth's gravity and magnetic fields. Field measurements made at many points about the earth need to be combined into a smooth representation of the field, a problem which is mainly one of interpolation in the presence of noise and incomplete data. Both spherical harmonic expansions and harmonic spline functions [Refs. 58, 60, and 61] have been used in the representations. In either case, the trade-off of resolution and variance is very important. Studies of the earth's core and geodynamo require that measurements of the magnetic field at the earth's surface be extrapolated to the core–mantle boundary. This is an inverse problem of considerable complexity, since the short-wavelength components of the field that are most important at the core–mantle boundary are very poorly measured at the earth's surface. Furthermore, the rate of change of the magnetic field with time (called “secular variation”) is of critical interest. This quantity must be determined from fragmentary historical measurements, including measurements of compass deviation recorded in old ship logs. Consequently, these inversions introduce substantial *a priori* constraints on the behavior of the field near the core, including the assumption that the root-mean-square time rate of change of the field is minimized and the total energy dissipation in the core is minimized [Ref. 27]. Once the magnetic field and its time derivative at the core–mantle boundary have been determined, they can be used in inversions for the fluid velocity near the surface of the outer core.

Fundamentally, the earth's gravity field $\mathbf{g}(\mathbf{x})$ is determined by its density structure $\rho(\mathbf{x})$. In parts of the earth where electric currents and magnetic induction are unimportant, the magnetic field $\mathbf{H}(\mathbf{x})$ is caused by the magnetization $\mathbf{M}(\mathbf{x})$ of the rocks. These quantities are related by

$$\begin{aligned}
 g_i(\mathbf{x}) &= \int_V G_i^g(\mathbf{x}, \mathbf{x}_0) \rho(\mathbf{x}) dV \quad \text{and} \quad H_i(\mathbf{x}) = \int_V G_{ij}^H(\mathbf{x}, \mathbf{x}_0) M_j(\mathbf{x}) dV \\
 G_i^g(\mathbf{x}, \mathbf{x}_0) &= -\gamma \frac{(x_i - x_{0i})}{|\mathbf{x} - \mathbf{x}_0|^3} \\
 G_{ij}^H(\mathbf{x}, \mathbf{x}_0) &= \frac{1}{4\pi\mu_0} \left[\frac{\delta_{ij}}{|\mathbf{x} - \mathbf{x}_0|^3} - \frac{(x_i - x_{0i})(x_j - x_{0j})}{|\mathbf{x} - \mathbf{x}_0|^5} \right]
 \end{aligned} \tag{14.9}$$

Here γ is the gravitational constant and μ_0 is the magnetic permeability of the vacuum. Note that in both cases the fields are linearly related to the sources (density and magnetization), so these are linear, continuous inverse problems. Nevertheless, inverse theory has proved to have little application to these problems, owing to their inherent underdetermined nature. In both cases it is possible to show analytically that the field outside a finite body can be generated by an infinitesimally thick spherical shell of mass or magnetization surrounding the body and below the level of the measurements. The null space of these problems is so large that it is generally impossible to formulate any useful solution, except when an enormous amount of *a priori* information is added to the problem.

Some progress has been made in special cases where *a priori* constraints can be sensibly stated. For instance, the magnetization of sea mounts can be computed from their magnetic anomaly, assuming that their magnetization vector is everywhere parallel (or nearly parallel) to some known direction and that the shape of the magnetized region closely corresponds to the observed bathymetric expression of the sea mount [Refs. 37, 55, and 56].

14.7 Electromagnetic Induction and the Magnetotelluric Method

An electromagnetic field is set up within a conducting medium when it is illuminated by a plane electromagnetic wave. In a homogeneous medium, the field decays exponentially with depth, since energy is dissipated by electric currents induced in the conductive material. In a vertically stratified medium, the behavior of the field is more complicated and depends on the details of the conductivity $\sigma(z)$. The ratio between the horizontal electric field E_x and magnetic field B_y on the

surface of the medium (at $z = 0$) is called the admittance and is defined by

$$Z(\omega) = \frac{E_x(\omega, z = 0)}{i\omega B_y(\omega, z = 0)} \quad (14.10)$$

where ω is frequency. The inverse problem is to determine the conductivity $\sigma(z)$ from measurements of the admittance $Z(\omega)$ at a suite of frequencies ω . Parker [Ref. 54] shows that this problem can be cast into a standard linear continuous inverse theory problem by using the calculus of variations (see Section 12.10) to calculate a linearized data kernel. This method has been used to determine the conductivity of the crust and mantle on a variety of scales [e.g., Refs. 52 and 53]. Oldenborg [Ref. 51] discusses the discretization of the problem and the application of extremal inversion methods (see Chapter 6).

14.8 Ocean Circulation

Ocean circulation models, that is, models of the velocity field of the water and the flux of heat and salt (and other chemical components), are generally based on relatively sparse data, much of it concentrated in the upper part of the water column. Inverse theory provides a means of incorporating *a priori* information, such as the dynamical constraints provided by the Navier–Stokes equations of fluid flow and conservation laws for mass and energy, into the models and thus improving them. One common approach is “box models”—models that are very coarsely parametrized into regions of constant velocity and flux. Box models are discussed in some detail by Wunsch and Minster [Ref. 67].

This page intentionally left blank

APPENDIX A

IMPLEMENTING CONSTRAINTS WITH LAGRANGE MULTIPLIERS

Consider the problem of minimizing a function of two variables, say, $E(x, y)$, with respect to x and y , subject to the constraint that $\phi(x, y) = 0$. One way to solve this problem is to first use $\phi(x, y) = 0$ to write y as a function of x and then substitute this function into $E(x, y)$. The resulting function of a single variable $E(x, y(x))$ can now be minimized by setting $dE/dx = 0$. The constraint equation is used to explicitly reduce the number of independent variables.

One problem with this method is that it is rarely possible to solve $\phi(x, y)$ explicitly for either $y(x)$ or $x(y)$. The method of Lagrange multipliers provides a method of dealing with the constraints in their implicit form.

When the function E is minimized, small changes in x and y lead to

no change in the value of E :

$$dE = \frac{\partial E}{\partial x} dx + \frac{\partial E}{\partial y} dy = 0 \quad (\text{A.1})$$

However, the constraint equations show that the perturbations dx and dy are not independent. Since the constraint equation is everywhere zero, its derivative is also everywhere zero:

$$d\phi = \frac{\partial \phi}{\partial x} dx + \frac{\partial \phi}{\partial y} dy = 0 \quad (\text{A.2})$$

We now consider the weighted sum of these two equations, where the weighting factor λ is called the *Lagrange multiplier*:

$$dE + \lambda d\phi = \left(\frac{\partial E}{\partial x} + \lambda \frac{\partial \phi}{\partial x} \right) dx + \left(\frac{\partial E}{\partial y} + \lambda \frac{\partial \phi}{\partial y} \right) dy = 0 \quad (\text{A.3})$$

The expression $dE + \lambda d\phi$ is equal to zero for any value of λ . If λ is chosen so that the expression within the first set of parentheses of the above equation is equal to zero, then the expression within the other parenthesis must also be zero, since one of the differentials (in this case dy) can be arbitrarily assigned. The problem can now be reinterpreted as one in which $E + \lambda\phi$ is minimized without any constraints. There are now three simultaneous equations for x , y , and the Lagrange multiplier λ :

$$\left[\frac{\partial E}{\partial x} + \lambda \frac{\partial \phi}{\partial x} \right] = 0, \quad \left[\frac{\partial E}{\partial y} + \lambda \frac{\partial \phi}{\partial y} \right] = 0, \quad \phi(x, y) = 0 \quad (\text{A.4})$$

The generalization of this technique to M unknowns \mathbf{m} and q constraints $\phi(\mathbf{m}) = 0$ is straightforward. One Lagrange multiplier is added for each constraint. The $M + q$ simultaneous equations for $M + q$ unknowns and Lagrange multipliers are then

$$\frac{\partial E}{\partial m_i} + \sum_{j=1}^q \lambda_j \frac{\partial \phi_j}{\partial m_i} = 0 \quad \text{and} \quad \phi_i(\mathbf{m}) = 0 \quad (\text{A.5})$$

APPENDIX B

L_2 INVERSE THEORY WITH COMPLEX QUANTITIES

Some inverse problems (especially those that deal with data that has been operated on by Fourier or other transforms) involve complex quantities. Many of these problems can readily be solved with a simple modification of the theory (the exception is the part involving inequality constraints, which we shall not discuss). The definition of the L_2 norm must be changed to accommodate the complex nature of the quantities. The appropriate change is to define the squared length of a vector \mathbf{v} to be $\|\mathbf{v}\|_2^2 = \mathbf{v}^H \mathbf{v}$, where \mathbf{v}^H is the *Hermetian transpose*, the transpose of the complex conjugate of the vector \mathbf{v} . This choice ensures that the norm is a nonnegative real number. When the results of L_2 inverse theory are rederived for complex quantities, the results are very similar to those derived previously; the only difference is that all the ordinary transposes are replaced by Hermetian transposes. For instance, the least squares solution is

$$\mathbf{m}^{\text{est}} = [\mathbf{G}^H \mathbf{G}]^{-1} \mathbf{G}^H \mathbf{d} \quad (\text{B.1})$$

Note that all the square symmetric matrices of real inverse theory now become square Hermetian matrices:

$$[\text{cov } \mathbf{x}]_{ij} = \int [x_i - \langle x_i \rangle]^H [x_j - \langle x_j \rangle] P(\mathbf{x}) d\mathbf{x}, [\mathbf{G}^H \mathbf{G}], [\mathbf{G} \mathbf{G}^H], \text{ etc.} \quad (\text{B.2})$$

These matrices have real eigenvalues, so no special problems arise when deriving eigenvalues or singular-value decompositions.

The modification made to Householder transformations is also very simple. The requirement that the Hermetian length of a vector be invariant under transformation implies that a unitary transformation must satisfy $\mathbf{T}^H \mathbf{T} = \mathbf{I}$. The most general unitary transformation is therefore $\mathbf{T} = \mathbf{I} - 2\mathbf{v}\mathbf{v}^H/\|\mathbf{v}\|^2$, where \mathbf{v} is any complex vector. The Householder transformation that annihilates the j th column of \mathbf{G} below its main diagonal then becomes

$$\mathbf{T}_j = \left\{ \mathbf{I} - \frac{1}{|\alpha|(|\alpha| - |G_{jj}|)} \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ G_{j,j} - \alpha \\ G_{j+1,j} \\ \vdots \\ \vdots \\ G_{N,j} \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ [0, \dots, 0, G_{jj}^* - \alpha^*, G_{j+1,j}^*, \dots, G_N^*] \end{bmatrix} \right\}$$

$$\text{where } |\alpha| = \sqrt{\sum_{i=j}^N |G_{ij}|^2} \quad (\text{B.3})$$

* means “complex conjugate,” and the phase of α is chosen to be π away from the phase of G_{jj} . This choice guarantees that the transformation is in fact a unitary transformation and that the denominator is not zero.

REFERENCES

1. Backus, G. E., and Gilbert, J. F. (1967). "Numerical application of a formalism for geophysical inverse problems." *Geophys. J. Roy. Astron. Soc.* **13**, 247–276.
2. Backus, G. E., and Gilbert, J. F. (1968). "The resolving power of gross earth data." *Geophys. J. Roy. Astron. Soc.* **16**, 169–205.
3. Backus, G. E., and Gilbert, J. F. (1970). "Uniqueness in the inversion of gross Earth data." *Phil. Trans. Roy. Soc. London, Ser. A* **266**, 123–192.
4. Claerbout, J. F., and Muir, F. (1973). "Robust modelling with erratic data." *Geophysics* **38**, 826–844.
5. Crosson, R. S. (1976). "Crustal structure modeling of earthquake data: 1. Simultaneous least squares estimation of hypocenter and velocity parameters." *J. Geophys. Res.* **81**, 3036–3046.
6. Cuer, M., and Bayer, R. (1980). "FORTRAN routines for linear inverse problems." *Geophysics* **45**, 1706–1719.
7. Fisher, R. A. (1953). "Dispersion on a sphere." *Phil. Trans. Roy. Soc. London, Ser. A* **217**, 295–305.
8. Franklin, J. N. (1970). "Well-posed stochastic extensions of ill-posed linear problems." *J. Math. Anal. and Appl.* **31**, 682–716.
9. Jackson, D. D. (1972). "Interpretation of inaccurate, insufficient and inconsistent data." *Geophys. J. Roy. Astron. Soc.* **28**, 97–110.
10. Jackson, D. D. (1979). "The use of a priori data to resolve non-uniqueness in linear inversion." *Geophys. J. Roy. Astron. Soc.* **57**, 137–157.

11. Jordan, T. H., and Franklin, J. N. (1971). "Optimal solutions to a linear inverse problem in geophysics." *Proc. Nat. Acad. Sci.* **68**, 291–293.
12. Kaula, W. M. (1966). *Theory of Satellite Geodesy*. Ginn (Blaisdell), Boston, Massachusetts.
13. Lanczos, C. (1961). *Linear differential operators*. Van Nostrand-Reinhold, Princeton, New Jersey.
14. Lawson, C. L., and Hanson, D. J. (1974). *Solving Least Squares Problems*. Prentice-Hall, Englewood Cliffs, New Jersey.
15. Minster, J. F., Jordan, T. J., Molnar, P., and Haines, E. (1974). "Numerical modelling of instantaneous plate tectonics." *Geophys. J. Roy. Astron. Soc.* **36**, 541–576.
16. Penrose, R. A. (1955). "A generalized inverse for matrices." *Proc. Cambridge Phil. Soc.* **51**, 406–413.
17. Rietch, E. (1977). "The maximum entropy approach to inverse problems." *J. Geophys.* **42**, 489–506.
18. Tarantola, A., and Valette, B. (1982). "Generalized non-linear inverse problems solved using the least squares criterion." *Rev. Geophys. Space Phys.* **20**, 219–232.
19. Tarantola, A., and Valette, B. (1982). "Inverse problems = quest for information." *J. Geophys.* **50**, 159–170.
20. Wiggins, R. A. (1972). "The general linear inverse problem: Implication of surface waves and free oscillations for Earth structure." *Rev. Geophys. Space Phys.* **10**, 251–285.
21. Wunsch, C., and Minster, J. F. (1982). "Methods for box models and ocean circulation tracers: Mathematical programming and non-linear inverse theory." *J. Geophys. Res.* **87**, 5647–5662.
22. Aki, K., and Lee, W. H. K. (1977). "Determination of three-dimensional velocity anomalies under a seismic array using first P arrival times from local earthquakes. 1. A homogeneous initial model." *J. Geophys. Res.* **81**, 4381–4399.
23. Aki, K., and Richards, P. G. (1980). *Quantitative Seismology, Theory and Methods*, W. H. Freeman, San Francisco, California.
24. Aki, K., Christoffersson, A., and Huesby, E. S. (1977). "Determination of the three-dimensional structure of the lithosphere." *J. Geophys. Res.* **82**, 277–296.
25. Bée, M., and Jacobson, R. S. (1984). "Linear inversion of body wave data—Part 3: Model parametrization." *Geophysics* **49**, 2088–2093.
26. Bessonova, E. N., Fishman, V. M., Ryaboy, V. Z., and Sitnitova, G. A. (1974). "The tau method for the inversion of travel times, 1. Deep seismic sounding data." *Geophys. J. Roy. Astron. Soc.* **36**, 377–398.
27. Bloxam, J. (1987). "Simultaneous inversion for geomagnetic main field and secular variation, 1. A large scale inversion problem." *J. Geophys. Res.* **92**, 11597–11608.
28. Chase, E. P., and Stuart, G. S. (1972). "The N plate problem of plate tectonics." *Geophys. J. Roy. Astron. Soc.* **29**, 117–122.
29. Creager, K. C., and Jordan, T. H. (1986). "Aspherical structure of the core–mantle boundary from PKP travel times." *Geophys. Res. Lett.* **13**, 1497–1500.
30. Crosson, R. S. (1976). "Crustal structure modeling of earthquake data. 1. Simultaneous least squares estimation of hypocenter and velocity parameters, *J. Geophys. Res.* **81**, 3036–3046, 1976.
31. De Mets, C., Gordon, G. G., Stein, S., Argus, D. F., Engeln, J., Lundger, P., Quible, D. G., Stein, C., Weisstein, S. A., Weins, D. A., and Woods, D. F. (1985). "NUVEL-1: A new global plate motion data set and model." *Eos Trans. AGU* **66**, 368–369.

32. Dorman, L. M., and Jacobson, R. S. (1981). "Linear inversion of body wave data—Part 1: Velocity structure from travel times and ranges." *Geophysics* **46**, 135–151.
33. Dziewonski, A. M. (1984). "Mapping the lower mantle; determination of lateral heterogeneity in P velocity to degree and order six." *J. Geophys. Res.* **89**, 5925–5952.
34. Dziewonski, A. M., and Anderson, D. L. (1981). "Preliminary Reference Earth Model." *Phys. Earth Planet. Interiors* **25**, 297–358.
35. Ellsworth, W. L., and Koyanagi, R. Y. (1977). "Three dimensional crust and mantle structure of Kilauea volcano, Hawaii." *J. Geophys. Res.* **82**, 5379–5394.
36. Giardini, D., Li, X., and Woodhouse, J. H. (1987). "Three-dimensional structure of the earth from splitting in free-oscillation spectra." *Nature* **325**, 405–411.
37. Grossling, B. F. (1970). "Seamount magnetism," in *The Sea*, Vol. 4. Wiley, New York.
38. Hager, B. H., Clayton, R. W., Richards, M. A., Comer, R. P., and Dziewonski, A. M. (1985). "Lower mantle heterogeneity, dynamic topography, and the geoid." *Nature* **313**, 541–545.
39. Jacobson, R. S., Shor, G. G., and Dorman, L. M. (1981). "Linear inversion of body wave data—Part 2: Attenuation vs. depth using spectral ratios." *Geophysics* **46**, 152–162.
40. Johnson, L. E., and Gilbert, F. (1972). "Inversion and inference for teleseismic array data," in *Methods of Computational Physics 12*, B. A. Bolt, ed. Academic Press, New York.
41. Jordan, T. H., and Sverdrup, K. A. (1981). "Teleseismic location techniques and their application to earthquake clusters in the south-central Pacific." *Bull. Seismol. Soc. Am.* **71**, 1105–1130.
42. Kennett, B. N. L., and Orcutt, J. A. (1976). "A comparison of travel time inversions for marine refraction profiles." *J. Geophys. Res.* **81**, 4061–4070.
43. Klein, F. W. (1985). "User's guide to HYPOINVERSE, a program for VAX and PC350 computers to solve for earthquake locations." U.S. Geologic Survey Open File Report 85-515, 24 pp.
44. Martinson, D. G., Menke, W., and Stoffa, P. (1982). "An inverse approach to signal correlation." *J. Geophys. Res.* **87**, 4807–4818.
45. Masters, G., and Gilbert, F. (1983). "Attenuation in the earth at low frequencies." *Phil. Trans. Roy. Soc. London Ser. A* **388**, 479–522.
46. Masters, G., Jordan, T. H., Silver, P. G., and Gilbert, F. (1982). "Aspherical earth structure from fundamental spheroidal mode data." *Nature* **298**, 609–613.
47. Menke, W. (1977). "Lateral heterogeneities in P velocity under the Tarbella array of the lesser Himalayas of Pakistan," *Bull. Seismol. Soc. Am.* **67**, 725–734.
48. Minster, J. B., and Jordan, T. H. (1978). "Present day plate motions." *J. Geophys. Res.* **83**, 5331–5354.
49. Minster, J. B., Jordan, T. H., Molnar, P., and Haines, E. (1974). "Numerical modeling of instantaneous plate tectonics." *Geophys. J. Roy. Astron. Soc.* **36**, 541–576.
50. Nataf, H.-C., Nakanishi, I., and Anderson, D. L. (1986). "Measurement of mantle wave velocities and inversion for lateral heterogeneity and anisotropy. 3. Inversion." *J. Geophys. Res.* **91**, 7261–7308.
51. Oldenburg, D. W. (1983). "Funnel functions in linear and nonlinear appraisal." *J. Geophys. Res.* **88**, 7387–7398.
52. Oldenburg, D. W., Whittall, K. P., and Parker, R. L. (1984). "Inversion of ocean bottom magnetotelluric data revisited." *J. Geophys. Res.* **89**, 1829–1833.

53. Parker, R. L. (1970). "The inverse problem of the electrical conductivity of the mantle." *Geophys. J. Roy. Astron. Soc.* **22**, 121–138.
54. Parker, R. L. (1977). "The Frechet derivative for the one dimensional electromagnetic induction problem." *Geophys. J. Roy. Astron. Soc.* **39**, 543–547.
55. Parker, R. L. (1988). "A statistical theory of seamount magnetism." *J. Geophys. Res.* **93**, 3105–3115.
56. Parker, R. L., Shure, L., and Hildebrand J. (1987). "An application of inverse theory to seamount magnetism." *Rev. Geophys.* **25**, 17–40.
57. Pavlis, G. L., and Booker, J. R. (1980). "The mixed discrete–continuous inverse problem: Application to the simultaneous determination of earthquake hypocenters and velocity structure." *J. Geophys. Res.* **85**, 4801–4809.
58. Sandwell, D. T. (1987). "Biharmonic spline interpolation of GEOS-3 and SEASAT altimeter data." *Geophys. Res. Lett.* **14**, 139–142.
59. Shure, L., and Chave, A. D. (1984). "Comments on 'An inverse approach to signal correlation'." *J. Geophys. Res.* **89**, 2497–2500.
60. Shure, L., Parker, R. L., and Backus, G. E. (1982). "Harmonic splines for geomagnetic modeling." *Phys. Earth Planet. Interiors* **28**, 215–229.
61. Shure, L., Parker, R. L., and Langel, R. A. (1985). "A preliminary harmonic spline model for MAGSAT data." *J. Geophys. Res.* **90**, 11505–11512.
62. Takeuchi, H., and Saito, M. (1972). "Seismic surface waves," in *Methods of Computational Physics 11*, B. A. Bolt, ed. Academic Press, New York.
63. Tanimoto, T., and Anderson, D. L. (1985). "Lateral heterogeneity and azimuthal anisotropy of the upper mantle, Love and Rayleigh waves 100–250 sec." *J. Geophys. Res.* **90**, 1842–1858.
64. Thurber, C. H. (1983). "Seismic detection of the summit magma complex of Kilauea volcano, Hawaii." *Science* **223**, 165–167.
65. Vasco, D. W. (1986). "Extremal inversion of travel time residuals." *Bull. Seismol. Soc. Am.* **76**, 1323–1345.
66. Wong, J., Hurley, P., and West, G. F. (1983). "Crosshole tomography and seismic imaging in crystalline rocks." *Geophys. Res. Lett.* **10**, 686–689.
67. Wunsch, C., and Minster, J. F. (1982). "Methods for box models and ocean circulation tracers: Mathematical programming and nonlinear inverse theory." *J. Geophys. Res.* **87**, 5647–5662.
68. Yu, G. K., and Mitchell, B. J. (1979). "Regionalized shear wave velocity models of the Pacific upper mantle from observed Love and Rayleigh wave dispersion." *Geophys. J. Roy. Astron. Soc.* **57**, 311–341.

INDEX

A

- Absorption coefficient, 14
- Admittance, 171
- Airgun, 189
- A priori distribution, 83, 147
- A priori information, 48
- Attenuation, 267
- Autocorrelation, 189
- Auxiliary variable, 10, 30
- Averages
 - localized, 19, 103
 - non-unique, 106, 123
 - unique, 104
 - weighted, 19, 101

B

- Backprojection, 179
- Backsolving, 112, 222
- Basis
 - in Simplex method, 257
 - of vector space, 110
- Bias, 146
- Bordering method, 74
- Bounding values
 - of localized averages, 106
 - of model parameters, 17

C

- CAT scanner, 13
- Calculus of variations, 160, 217
- Camera, 183
- Central limit theorem, 29
- Chasing algorithm, 255
- Chi-squared distribution, 32, 60
- Chi-squared test, 32
- Circulation, ocean, 271
- Combining distributions, 158
- Complex number, 275
- Conditional distribution, 87, 159
- Conductivity, electrical, 270
- Confidence interval, 33, 134
- Convergence
 - of nonlinear algorithm, 153
 - of nonnegative least squares algorithm, 130
- Convolution, 187
- Correlated data, 24
- Covariance, 26
- Covariance matrix, 27
- Covariance size function, 68
- Crosscorrelation, 189
- Crossover error, 190
- Curve fitting
 - Gaussian, 15, 210
 - plane, 43

quadratic, 11, 42
straight line, 10, 41, 202

D

Damped least squares, 52, 71, 93
Data, 7
Data kernel, 9
Data resolution matrix, 62
Density, 269
Dike, 186
Discretization of continuous problems, 174
Distribution, 18, 22

E

Earthquake location, 213, 261
Echo sounding, 189
Eigenfrequencies, 217
Eigenvalues
 of real symmetric matrix, 251
 transformations using, 117
Empirical orthogonal function analysis, 167
Equality constraints, linear
 definition, 55
 examples
 crossover error problem, 190
 Gaussian curve fitting, 213
 straight line problem, 56
Householder transformations, 114, 236
implementation using
Lagrange multipliers, 56
probability distributions, 85
singular-value decomposition, 125
weight matrices, 56
Estimate, 17, 27, 105
Euler angle, 268
Even-determined problem, 47
Existence criterion, 45
Expectation, 23
Explicit equation, 9
Exponential distribution, 133

F

Factor analysis, 15, 161
Factor loading, 163
F distribution, 97
Feasibility of inequality constraints, 127
 determination using
 graphical interpretation, 127, 256
 L₂ algorithm, 130
 Simplex method, 249
Filter design, 187
Finite element derivative, 53, 214
Fisher distribution, 207
Flatness, measure of, 53
FORTRAN 77, 221
 subroutines, 226, 230, 233, 234, 236, 242

Forward theory, 2

Fourier slice theorem, 178
Fourier transform, 178
F ratio, 97
Free oscillations, 265
F test, 97

G

Gaussian distribution, 29
Gauss-Jordan reduction, 222
Generalized inverse, 61
Geomagnetism, 269
Givens-Householder method, 253
Givens transformation, 255
Golub's method, 231
Gravity, 269

H

Hermetian transpose, 275
Householder-Reinch method, 254
Householder transformation, 111, 231, 276
Hyperplane, 110

I

Image enhancement, 183
Implicit equation, 8

Importance, 64
Induction, electromagnetic, 270
Inequality constraint, 57, 86, 126, 240
Inexact model, 89, 159
Information in distribution, 158
Integral equation, 9
Invariance, 145
Inverse of matrix, 229
Inverse theory, 2
 continuous, 3, 171
 discrete, 3

J

Jacobian determinant, 144, 178

K

Kuhn-Tucker theorem, 127

L

Lagrange multiplier
applications
 Backus-Gilbert generalized inverse, 74
 constrained least squares problem, 56
 Fisher distribution, 209
 minimum-length solution, 50
 nonlinear least squares problem, 150
derived, 273
Least squares error solution, 40, 67, 94
Least squares principle, 36
Length, 36
 generalized, 52
Levinson recursion, 189
Likelihood, 80
Linear independence, 109
Linear inverse problem, 9
Linearization
 by Taylor series, 14, 152
 by transformation of variables, 147
Linearly close problems, 155
Linear programming
 definition, 138
 solution by Simplex method, 256

M

Magnetic field, 269
Magnetization, 269
Magnetotelluric method, 270
Mapping, 109
Mapping function, 268
Matrix inversion, 229
Maximum entropy method, 160
Maximum likelihood method, 80

 estimates of means

 exponential, 135

 Fisher, 207

 Gaussian, 79

solution to inverse problems

 general case, 159

 linear exponential, 135

 linear Gaussian, 92

 nonlinear Gaussian, 149

Maximum likelihood point, 23

Mean, 23

Median, 136

Minimum-length solution, 50, 67, 94

Mixed-determined problems, 50

 partitioning, 50

 solution, 121

 vector space representation, 118

Model, 7

N

Natural generalized inverse, 121
Natural solution, 119
Navier-Stokes equations, 271
Noise, 21
Nonlinear equation, 8, 147, 156
Norm, 36
Null distribution, 157
Null space, 118
Null vector, 101

O

Objective function, 256
Operator, 61
Orbit, 192
Overdetermined problems, 47, 231

P

Parameterization, 143
 Parameter separation, 264
 Partial pivoting, 223
 Phase velocity, 266
 Physicality constraints, 165
 Pivoting, 223
 Pixel, 183
 Plate tectonics, 268
 Precision parameter, 208
 Prediction error, 32
 L_1 , 135
 L_2 , 37
 L_∞ , 141
 weighted, 54
 Projected distribution, 90

Q

Q-mode factor analysis, 167
 Quartimax factor, 166

R

Radon transform, 177
 Random variable, 21
 Ray parameter, 262
 Ray tracing, 214, 262
 Realization, 21
 Resolution
 of data, 62
 of model, 64, 172
 relationship to localized average, 104
 trade-off with variance, 76
 R-mode factor analysis, 167
 Robustness, 39
 Roughness, measure of, 53

S

Sea mount, 270
 SEASAT, 190
 Shift parameter, 254
 Sidelobes, 71
 Signal correlation, 267
 Significance, test of, 96

Simplex method, 256

Singular values, 120

Singular-value decomposition
 definition, 119
 derivation, 124

Solution, type of, 16

Spherical harmonic expansion, 267

Spread function

 Backus-Gilbert, 71, 172
 Dirichlet, 68

Sturm sequence, 242

Surface waves, 265

T

Tau function, 262
 Teleseisms, 264
 Time series, 187
 Tomography
 acoustic, 11, 194
 relationship to continuous inverse
 theory, 176
 using earthquakes, 217
 x-ray, 13
 Toeplitz matrix, 189
 Trade-off curve, 78, 174
 Transformation
 definition, 110
 of equality constraints, 114
 of inequality constraints, 125, 130
 that diagonalizes weight matrices, 117
 Travel time curves, 262

U

Underdetermined problem, 48, 231
 Uniqueness, 102
 Unitary transformation, 111, 276
 Unit covariance matrix, 65

V

Variance
 definition, 23
 relationship to confidence interval, 33,
 134

- | | |
|---|---|
| relationship to error curvature, 59
Varimax factor, 166
Vector space, 109
Vibrational problem, 217 | W
Weighted average, 101
Weighted least squares solution, 54
Weight matrices, 53, 117, 165 |
|---|---|

This page intentionally left blank

INTERNATIONAL GEOPHYSICS SERIES

Edited by

J. VAN MIEGHEM
(1959–1976)

ANTON L. HALES
(1972–1979)

WILLIAM L. DONN
Lamont-Doherty
Geological Observatory
Columbia University
Palisades, New York
(1980–1986)

Current Editors

RENATA DMOWSKA
Division of Applied Science
Harvard University

JAMES R. HOLTON
Department of Atmospheric Sciences
University of Washington
Seattle, Washington

- Volume 1 BENO GUTENBERG. Physics of the Earth's Interior. 1959*
- Volume 2 JOSEPH W. CHAMBERLAIN. Physics of the Aurora and Airglow. 1961*
- Volume 3 S. K. RUNCORN (ed.). Continental Drift. 1962
- Volume 4 C. E. JUNGE. Air Chemistry and Radioactivity. 1963
- Volume 5 ROBERT G. FLEAGLE and JOOST A. BUSINGER. An Introduction to Atmospheric Physics. 1963*
- Volume 6 L. DUFOUR AND R. DEFAY. Thermodynamics of Clouds. 1963
- Volume 7 H. U. ROLL. Physics of the Marine Atmosphere. 1965
- Volume 8 RICHARD A. CRAIG. The Upper Atmosphere: Meterology and Physics. 1965

* Out of print.

- Volume 9 WILLIS L. WEBB. Structure of the Stratosphere and Mesosphere. 1966*
- Volume 10 MICHELE CAPUTO. The Gravity Field of the Earth from Classical and Modern Methods. 1967
- Volume 11 S. MATSUSHITA and WALLACE H. CAMPBELL (eds.). Physics of Geomagnetic Phenomena. (In two volumes.) 1967
- Volume 12 K. YA. KONDRATYEV. Radiation in the Atmosphere. 1969
- Volume 13 E. PALMÉN and C. W. NEWTON. Atmospheric Circulation Systems: Their Structure and Physical Interpretation. 1969
- Volume 14 HENRY RISHBETH and OWEN K. GARRIOTT. Introduction to Ionospheric Physics. 1969
- Volume 15 C. S. RAMAGE. Monsoon meteorology. 1971*
- Volume 16 JAMES R. HOLTON. An Introduction to Dynamic Meteorology. 1972*
- Volume 17 K. C. YEH and C. H. LIU. Theory of Ionospheric Waves. 1972
- Volume 18 M. I. BUDYKO. Climate and Life. 1974
- Volume 19 MELVIN E. STERN. Ocean Circulation Physics. 1975
- Volume 20 J. A. JACOBS. The Earth's Core. 1975*
- Volume 21 DAVID H. MILLER. Water at the Surface of the Earth: An Introduction to Ecosystem Hydrodynamics. 1977
- Volume 22 JOSEPH W. CHAMBERLAIN. Theory of Planetary Atmospheres: An Introduction to Their Physics and Chemistry. 1978*
- Volume 23 JAMES R. HOLTON. Introduction to Dynamic Meterology, Second Edition. 1979
- Volume 24 ARNETT S. DENNIS. Weather Modification by Cloud Seeding. 1980
- Volume 25 ROBERT G. FLEAGLE and JOOST A. BUSINGER. An Introduction to Atmospheric Physics, Second Edition. 1980
- Volume 26 KUO-NAN LIOU. An Introduction to Atmospheric Radiation. 1980
- Volume 27 DAVID H. MILLER. Energy at the Surface of the Earth: An Introduction to the Energetics of Ecosystems. 1981
- Volume 28 HELMUT E. LANDSBERG. The Urban Climate. 1981
- Volume 29 M. I. BUDYKO. The Earth's Climate: Past and Future. 1982
- Volume 30 ADRIAN E. GILL. Atmosphere–Ocean Dynamics. 1982
- Volume 31 PAOLO LANZANO. Deformations of an Elastic Earth. 1982
- Volume 32 RONALD T. MERRILL and MICHAEL W. McELHINNY. The Earth's Magnetic Field: Its History, Origin and Planetary Perspective. 1983
- Volume 33 JOHN S. LEWIS and RONALD G. PRINN. Planets and Their Atmospheres: Origin and Evolution. 1983
- Volume 34 ROLF MEISSNER. The Continental Crust: A Geophysical Approach. 1986
- Volume 35 M. U. SAGITOV, B. BODRI, V. S. NAZARENKO, and KH. G. TADZHIDINOV. Lunar Gravimetry. 1986
- Volume 36 JOSEPH W. CHAMBERLAIN and DONALD M. HUNTER. Theory of Planetary Atmospheres: An Introduction to Their Physics and Chemistry, Second Edition. 1987

* Out of Print.

- Volume 37 J. A. JACOBS. *The Earth's Core*, Second Edition. 1987
- Volume 38 J. R. APEL. *Principles of Ocean Physics*. 1987
- Volume 39 MARTIN A. UMAN. *The Lightning Discharge*. 1987
- Volume 40 DAVID G. ANDREWS, JAMES R. HOLTON, and CONWAY B. LEOVY. *Middle Atmosphere Dynamics*. 1987
- Volume 41 PETER WARNECK. *Chemistry of the Natural Atmosphere*. 1988
- Volume 42 S. PAL ARYA. *Introduction to Micrometeorology*. 1988
- Volume 43 MICHAEL C. KELLEY. *The Earth's Ionosphere: Plasma Physics and Electrodynamics*. 1989
- Volume 44 WILLIAM R. COTTON and RICHARD A. ANTHES. *Storm and Cloud Dynamics*. 1989
- Volume 45 WILLIAM MENKE. *Geophysical Data Analysis: Discrete Inverse Theory*. 1989
- Volume 46 S. GEORGE PHILANDER. *El Niño, La Niña, and the Southern Oscillation*. 1990

This page intentionally left blank