A PRACTICAL SOLUTION FOR LINEAR INFERENCE COMPUTATIONS

by

MICHAEL GARFIELD SCHLAX

B.Sc. (Engineering Geoscience)


A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES
Department of Geophysics and Astronomy

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA
March 1984

© Michael G. Schlax, 1984
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Geophysics and Astronomy

The University of British Columbia
1956 Main Mall
Vancouver, Canada
V6T 1Y3

Date 2/3/84
Abstract

In this thesis, I shall present a matrix form of Backus' theory of linear inference with multiple predictions. The Bayesian approach of Backus allows the treatment of erroneous data and the imposition of the essential a priori bound on the model norm. The $\chi^2$ statistic will be introduced to construct a most likely model and bound the norm of all acceptable models from above and below. This results in more reliable, and possibly more confining, estimates of the predictions than provided by only an upper bound. The algorithm derived is robust and efficient, and estimates comparable to those obtained from Oldenburg's linear programming algorithm have been achieved.

*The work done in this thesis was carried out with Doug Oldenburg and the body of this thesis is essentially that of a paper entitled "A Practical Solution for Linear Inference Computations" written by Schlax and Oldenburg. As a consequence, the pronoun "we" will be used throughout the remainder of this text.
# Table of Contents

Abstract .......................................................................................................................... i  
List of Figures .................................................................................................................. iii  
Acknowledgements .......................................................................................................... iv  
CHAPTER 1 An Introduction ......................................................................................... 1  
CHAPTER 2 Review of Linear Inference ....................................................................... 4  
CHAPTER 3 Matrix Notation for Inference ................................................................... 10  
CHAPTER 4 A Statistical Formulation ......................................................................... 13  
CHAPTER 5 The Matrix Solution .................................................................................. 18  
CHAPTER 6 A Priori Bounds on $\|m^w\|$ ................................................................... 21  
CHAPTER 7 Bounds on $\|m^L\|$ .................................................................................... 25  
CHAPTER 8 A Numerical Example ............................................................................... 27  
CHAPTER 9 Discussion ................................................................................................. 34  
CHAPTER 10 Conclusion ............................................................................................... 39  
References ....................................................................................................................... 40  
Appendix 1 ...................................................................................................................... 41  
Appendix 2 ...................................................................................................................... 44  
Appendix 3 ...................................................................................................................... 47  
Appendix 4 ...................................................................................................................... 49  
Appendix 5 ...................................................................................................................... 52
List of Figures

Figure 1  A Geometrical Example of the Inference Problem ..  7
Figure 2  Numerical Example I ................................. 30
Figure 3  Numerical Example II ................................. 32
ACKNOWLEDGEMENTS

This research owes a great deal to Doug Oldenburg. Besides offering me the opportunity to pursue my research interests, he provided guidance, friendship, and contributed many hours of thought and discussion to this thesis. His interest in my studies, and perpetually available help have made my M.Sc. program a fruitful and enjoyable experience.

Jose Julian Cabrera Gomez kept las maquinas horribles at bay and saved me much grief by finding my infinite loops.

Diane Donnelly did a remarkable job of interpreting my "handwriting" and turning out fine manuscripts.

And, finally, thanks to Bob Krider, for making me feel normal.
CHAPTER 1 AN INTRODUCTION

For many physical problems, there is a linear functional relationship between observational data and the model. This relationship can be written as a Fredholm equation of the first kind

$$\gamma_i = \int_0^L m_i(x) g_i(x) \, dx \quad i = 1, \ldots, D \quad (1)$$

In equation (1), $\gamma_i$ is the $i$'th datum, $m_i(x)$ is a model describing that property of the earth about which information is sought, and $g_i(x)$ is the data kernel corresponding to the $i$'th datum. The functions $g_i(x)$ are determined from the geometry and physical equations relevant to the problem and are assumed to be known.

A particular experiment will yield $D$ numerical values, $\gamma_i$ which are contaminated with error. These data provide a set of linear constraints upon $m_i(x)$. Linear inverse theory uses these constraints to extract information about the model $m_i(x)$ in three possible ways: (1) to construct a model which reproduces the data; (2) to take linear combinations of the data to generate unique averages of the model; (3) to calculate the values of other linear functionals of the model. The first two approaches, which are respectively referred to as model construction and appraisal, are routinely used and do not require further explanation. However, the third approach, linear inference, has had limited exposure.

The goal of linear inference is to use the (inaccurate) observations to predict values of other linear functionals of $m_i(x)$, that is, to estimate other values $\tilde{\gamma}_i$ given by
\[ \tilde{\gamma}_i = \int \int_0^1 m_F(x) p_L(x) \, dx \quad i=1, \ldots, P \quad (2) \]

These predictions may represent averages of \( m_F(x) \) in the sense of Backus and Gilbert (1968, 1970) or Oldenburg (1983), or they may be associated with some physical parameters which depend upon \( m_F(x) \). The functions \( p_L(x) \), which we refer to as the prediction kernels, will be determined by these mathematical or physical considerations and are assumed to be known.

The mathematical foundations of linear inference theory were presented in a series of elegant and formidable papers by Backus (1970a,b,c, 1972). These have received little attention in the geophysical literature but some attempts have been made to use this work. Parker (1971) applied inference theory to the calculation of seamount magnetism, and Johnson and Gilbert (1972) used the results of Backus (1970a) to assess the non-uniqueness inherent in the inversion of teleseismic data. Parker (1977a) presented a matrix form of Backus' (1970a) solution of the inference problem with \( P > 1 \) and error-free data. In that paper, Parker showed how more confined estimates of the predictions could be had when they were considered as the coefficients of some expansion of \( m_F(x) \). Jackson (1979) also presented a solution to the inference problem after assuming that \( m_F(x) \) could be represented by a finite number of parameters. More recently, Oldenburg (1983) has applied linear programming to the solution of a linear inference problem. He calculated upper and lower bounds to boxcar averages of \( m_F(x) \) in order to assess non-uniqueness.

In this paper, we present a matrix form of Backus' (1970a) solution to the inference problem with multiple predictions and
inaccurate data. We shall give a brief review of Backus' theory of linear inference, and present the appropriate matrix notation. Using probability density functions, we shall show how the data and other a priori information about \( m_F(x) \) may be combined through Bayes' theorem to define a set of models which contains \( m_F(x) \). The range of values of the predictions arising from this set of models will be calculated to provide probabilistic estimates for \( \tilde{\gamma}_i \).

We wish to emphasize that we are presenting a practical form of Backus' work. With minor exceptions, the theory of our presentation is his, and we hope that this treatment of linear inference, presented at an elementary level, will demonstrate the power of Backus' ideas.
CHAPTER 2 REVIEW OF LINEAR INERENCE

We shall first present a brief review of Backus' (1970a) treatment of the linear inference problem. From the data set (equation (1)) we seek to estimate the predictions in equation (2). We also suppose that the error associated with each $\gamma_i$ is Gaussian distributed with zero-mean and known standard deviation, and that the covariance matrix for the errors is known.

We assume that $m_\xi(x)$, $g_i(x)$, and $p_i(x)$ are all elements of a real Hilbert space $\mathcal{H}$ with a well-defined inner product

$$<u,v> = \int u(x) v(x) \, dx$$

and norm $||v|| = <u,u>^{\frac{1}{2}}$

With this notation, equations (1) and (2) can be written as

$$\gamma_i = <m_\xi,q_i> \quad \text{i=1, ..., D} \quad (3)$$

$$\tilde{\gamma}_i = <m_\xi,p_i> \quad \text{i=1, ..., P} \quad (4)$$

In light of the geometric structure of $\mathcal{H}$, equations (3) and (4) may be characterized as the projections of $m_\xi$ onto the vectors $q_i$ and $p_i$. We shall exploit this geometry to find a solution.

The vectors $q_i$ and $p_i$ allow a convenient decomposition of $\mathcal{H}$. Assuming, without loss of generality (Backus, 1970a), that the set \{g_1,...,g_p, p_1,...,p_p\} is linearly independent, we define three finite-dimensional subspaces of $\mathcal{H}$. Let asp \{\} denote the algebraic span, or the set of all linear combinations, of a set of vectors in
\( \mathcal{H} \) and define \( \mathcal{R} = \text{asp} \{ g_1 \ldots g_p, p_1 \ldots p_p \} \) and \( \mathcal{D} = \text{asp} \{ g_1 \ldots g_b \} \). The third vector space, \( \mathcal{P} = \mathcal{R} \oplus \mathcal{D} \), is the orthogonal complement of \( \mathcal{D} \) in \( \mathcal{R} \). Thus \( \mathcal{R} = \mathcal{D} \oplus \mathcal{P} \), where \( \oplus \) denotes the orthogonal direct sum of two subspaces.

Let \( m^\circ \), \( m^\circ \) and \( m^\circ \) be the orthogonal projections of \( m^\circ \) onto \( \mathcal{D} \), \( \mathcal{R} \) and the complement of \( \mathcal{R} \) respectively. Then \( m^\circ = m^\circ + m^\circ + m^\circ \). From the definition of the subspaces it follows that the data equations may be rewritten as

\[
\gamma_i = \langle m^\circ, g_i \rangle 
\]

Similarly, the predictions take the form

\[
\tilde{\gamma}_i = \langle m^\circ, p_i \rangle + \langle m^\circ, p_i \rangle 
\]

\[
= \langle m^\circ, p_i \rangle + \langle m^\circ, p_i \rangle 
\]

\[
= \langle m^\circ, p''_i \rangle + \langle m^\circ, p''_i \rangle 
\]

which we write as

\[
\tilde{\gamma}_i = \tilde{\gamma}''_i + \tilde{\gamma}^\dagger_i 
\]

Equations (5) - (7) demonstrate a fundamental fact about the inference problem; the data constrain only \( m^\circ \) (equation 5). In fact, the data are equivalent to knowledge of \( m^\circ \). Although \( \tilde{\gamma}''_i \) may always be found from the data, the data provide no information about \( \tilde{\gamma}^\dagger_\). It therefore follows from equations (6) and (7) that \( \tilde{\gamma}_i \) cannot be determined unless further constraints are placed upon \( m^\circ \) or unless \( p^\dagger_i = 0 \). We shall attempt to estimate \( \tilde{\gamma}_i \) by placing a priori
constraints upon $m_E$.

On the basis of physical reasoning there will always exist a positive number $M$ such that $|<m_E, p^\perp>| \leq M$. The Cauchy-Schwartz inequality assures us that

$$|\gamma^\perp| = |<m_E^\perp, p^\perp>| \leq M |p^\perp|$$

from which we can write $\gamma^\perp = \frac{\gamma^\perp}{|\gamma^\perp|}$. The value of $M$ in equation (8) may be replaced by $M_\perp$ if it is hypothesized that $|<m_E^\perp, p^\perp>| \leq M_\perp$.

We would like to demonstrate the concepts outlined above by considering a very simple analogy using vectors in the 2-dimensional Euclidean space $E_2$, illustrated in Figure 1. We consider $E_2$ with the usual basis $\{\hat{x}, \hat{y}\}$, inner product, and norm. Suppose there is some vector $m_E$ in $E_2$, and that the only information available about $m_E$ is the datum, $\gamma = <m_E, \hat{x}>$. It is desired to have an estimate $\tilde{\gamma} = <m_E, p>$ where the vector $p = ax^\perp + by^\perp$ is the prediction kernel. For this example the subspace $D$ is the $x$-axis, and $P$ is the $y$-axis ($p^\perp = b\hat{y}$). Our datum $\gamma$ provides information about $m_E^x$, that is $m_E^x = \gamma \hat{x}$, but provides no information about $m_E^y$, the $y$-component of $m_E$. Equation (6) shows that our prediction is $\tilde{\gamma} = \gamma a + <m_E^x, p^\perp>$ and consequently is undetermined unless $m_E^x$ is bounded. Hypothesizing that $|<m_E^\perp, p^\perp>| \leq M_\perp$ will confine $m_E$ to the solid line in Figure 1. Corresponding estimates of $\tilde{\gamma}$ are then given by

$$\gamma a - bM_\perp \leq \tilde{\gamma} \leq \gamma a + bM_\perp$$

It is clear from equation (9) that the usefulness of this approach
A Geometrical Example of the Inference Problem

From the datum \( \gamma = \langle m_L, \hat{\gamma} \rangle \) we may find \( m_L = \gamma \hat{\gamma} \). The prediction vector is \( p = a\hat{\gamma} + b\hat{\gamma} \). The bound \( ||m_L^*|| \leq M_\perp \) confines \( m_L \) to the solid line, from which we estimate \( \tilde{\gamma} = \langle p, m_L^* \rangle \) as

\[
\gamma a - bM_\perp \leq \tilde{\gamma} \leq \gamma a + bM_\perp.
\]
Figure 1
will depend heavily on the magnitude of the hypothesized bound $M_L$ and therefore some effort should be made on the part of the investigator to make $M_L$ as small as possible.

There are two extensions to the inference problem outlined thus far which we would like to consider. The first is to include the effects of the data errors; the second is to allow the bounds on $||m_e^h||$ or $||m_e^1||$ to be probabilistic. These two extensions will permit us to define regions in $D$ and $P$ which contain $m_e^h$ and $m_e^1$ at some level of confidence. The evaluation of $<m_e^h, p_i^h>$ and $<m_e^1, p_i^1>$ for all models in those regions will provide the desired estimates for $\bar{r}_i$.

In order to make these extensions and to set up a formalism within which the computations are easily carried out, we devote the next two sections to setting up a matrix notation and introducing Bayesian subjective probabilities.
CHAPTER 3  MATRIX NOTATION FOR INFERENCE

Equation (6) demonstrates that the calculations required to find the $\gamma_i$ involve only elements of the finite-dimensional vector spaces $D$ and $P$. Vectors on those subspaces of $H$ may be represented as $D$- and $P$- tuples of real numbers and inner products of elements of $D$ or $P$ may be computed using matrix operations. We shall exploit this finite-dimensional representation and develop the suitable matrix equations for the inference problem. Details may be found in Appendix 1.

First, let us define some notation. Consider the subspace $D$, with basis $\{g_1, \ldots, g_D\}$. If $x$ is any element of $D$ we may write

$$x = \sum_{i=1}^{D} a_i g_i$$

(10)

Equivalently $x$ may be represented as the $D$-tuple consisting of the coefficients $a_i$, that is, as the column vector $[x]^T = (a_1, a_2, \ldots, a_D)$. Thus $[x]$ is the matrix of coefficients.

If $x$ and $y$ are vectors in $D$, their inner product appears in matrix form as

$$\langle x, y \rangle = [x]^T \Gamma [y]$$

(11)

where $\Gamma_{ij} = \langle g_i, g_j \rangle$. $\Gamma$ is called the inner product matrix, and it is positive definite and symmetric. With this notation, the norm of $x$ is given by

$$||x||^2 = \langle x, x \rangle = [x]^T \Gamma [x]$$

(12)
We now turn our attention to the calculation of the predictions. Equation (7) can be written as a P-tuple in $E^p$, that is, as

$$\vec{\gamma} = \vec{\gamma}^n + \vec{\gamma}^\perp = (\gamma_1^n \ldots \gamma_p^n)^T + (\gamma_1^\perp \ldots \gamma_p^\perp)^T$$

(13)

Our goal is to find a convenient way to evaluate this equation. For any model $m^n_\varepsilon$ in $\mathcal{D}$, the P-tuple $\vec{\gamma}^n$ may be evaluated by

$$\vec{\gamma}^n = G^T[m^n_\varepsilon]$$

(14)

where $G_{ij} = \langle g_i, p_j \rangle$. To evaluate the second component of the prediction, $\vec{\gamma}^\perp$, we introduce, as a basis for $\mathcal{D}$, the set $\{p_1^\perp, \ldots, p_p^\perp\}$. Let the inner product matrix with respect to this basis be $\Lambda_{ij} = \langle p_i^\perp, p_j^\perp \rangle$. Since $p_i^\perp = p_i - p_i^n$, it is easily shown that

$$\Lambda^\perp = \Lambda - G^T \Gamma^{-1} G$$

(15)

where $\Lambda_{ij} = \langle p_i, p_j \rangle$. Thus, for any model $m_\varepsilon^\perp$ in $\mathcal{D}$, we have

$$\vec{\gamma}^\perp = \Lambda^\perp[m_\varepsilon^\perp]$$

(16)

Equations (14) and (16) permit the computation of $\vec{\gamma}$ for any model $m_\varepsilon = m^n_\varepsilon + m_\varepsilon^\perp$.

The data will be used to define a region in $\mathcal{D}$ containing $m^n_\varepsilon$ at a certain level of confidence. For this purpose, it is necessary to understand the geometrical significance of the data. This will follow from consideration of the dual basis for $\mathcal{D}$, which we write
as \{g'\ldots g^p\}. The data are the coefficients of \(m''_E\) with respect to the dual basis (Backus, 1970a). If \([m''_E]_d\) is the coefficient matrix of \(m''_E\) with respect to the dual basis, it can be shown (Appendix 2) that

\[
[m''_E] = \Gamma^{-1}[m''_E]_d
\]

\[
= \Gamma^{-1}\tilde{y}
\]

where \(\tilde{y}^T = (\gamma_1\ldots\gamma_D)\) are the data.

Equation (17) may be recognized as the equation defining the smallest model of Backus and Gilbert. Clearly, \(m''_E\) is the model with smallest norm that satisfies the data equations (5); this is the distinguishing property of the orthogonal projection.
CHAPTER 4 A STATISTICAL FORMULATION

The previous chapter showed how the predictions $\tilde{y}$ could be calculated if $m^*$ and $m^+$ were provided. The practical difficulty is that neither of these functions can be determined exactly. Let us first consider the difficulties involved with computing $m^*$. The data $\tilde{y}$ are inaccurate and may be written as $\tilde{y} = \tilde{y}_t + \tilde{e}$ where $\tilde{y}_t$ are the true data and $\tilde{e}$ is an error vector. We shall suppose that each element of $\tilde{e}$ has a Gaussian distribution with zero-mean. Let $V$ be the covariance matrix of the errors. If $\tilde{y}_t$ are the true data, then a joint pdf for $\tilde{y}$ on the vector space $E^2$ is given by

$$p(\tilde{y} | \tilde{y}_t) = \text{const} \times \exp \left[ -\frac{1}{2} (\tilde{y} - \tilde{y}_t)^T V^{-1} (\tilde{y} - \tilde{y}_t) \right]$$

Let

$$m^* = \sum_{i} \beta_i g_i$$

Equation (17) shows that for any given $\tilde{y}$, the coefficients $\beta$ in equation (19) are $\tilde{\beta} = \Gamma^{-1} \tilde{y}$; that is, the coefficients $\tilde{\beta}$ are linear combinations of the data $\tilde{y}$. It follows that statistical uncertainties in $\tilde{\beta}$ can be related directly to the statistics of the data errors. If $\tilde{\beta}_t = \Gamma^{-1} \tilde{y}_t$, the pdf for $\tilde{\beta}$ is

$$p(\tilde{\beta} | \tilde{\beta}_t) = \text{const} \times \exp \left[ -\frac{1}{2} (\tilde{\beta} - \tilde{\beta}_t)^T V^{-1} (\tilde{\beta} - \tilde{\beta}_t) \right]$$

Equation (20) establishes the statistical relationship between data inaccuracies and statistical variability of the coefficients $\tilde{\beta}$. Variability of the constructed portion of the model on $\Omega$ can
therefore be determined.

Geometrically, equation (20) has the following interpretation. At the 1-σ level of confidence, \( \tilde{\beta} \) is contained within the hyperellipsoidal region centered on \( \tilde{\beta}_t \) defined by the positive definite quadratic form

\[
(\tilde{\beta} - \tilde{\beta}_t)^T \Gamma \Gamma^{-1} \Gamma (\tilde{\beta} - \tilde{\beta}_t) \leq 1 \tag{21}
\]

We shall be concerned only with this geometrical interpretation of pdfs. For our purposes it is necessary only to write the argument of the pdf, that is, equation (20) will be written as

\[
p(\beta|\tilde{\beta}_t) \sim (\tilde{\beta} - \tilde{\beta}_t)^T \Gamma \Gamma^{-1} \Gamma (\tilde{\beta} - \tilde{\beta}_t)
\]

We can also define a pdf for \( \tilde{\beta}_t \), which expresses our belief that \( ||m^\mu|| \leq M_\mu \). The appropriate pdf is

\[
p(\tilde{\beta}_t) \sim \tilde{\beta}_t^T M_\mu^{-2} \tilde{\beta}_t \tag{22}
\]

and this states that, at the 1-σ level of confidence, \( m^\mu \) is contained within a sphere of radius \( M_\mu \) centered on \( \tilde{\beta}_t \). In a similar manner, a pdf for \( \tilde{\eta}_t \), the coefficients of \( m_t^\mu \) with respect to the basis \( \{p^1_t, \ldots, p^r_t\} \), may be defined if \( ||m_t^\mu|| \leq M_\perp \). That pdf is

\[
p(\tilde{\eta}_t) \sim \tilde{\eta}_t^T M_\perp^{-2} \tilde{\eta}_t \tag{23}
\]

Equations (20) and (22) represent two pdfs for the coefficients \( \tilde{\beta} \) and \( \tilde{\beta}_t \). We would like to combine them into one pdf for \( \tilde{\beta}_t \) and to do so we shall use the approach pursued by Bayesian
Bayesians maintain that before the experiment is performed, the investigator has some rational conviction regarding the result \( \tilde{\gamma} \) of an experiment. This a priori conviction may be expressed as a pdf, \( p_{\text{prior}}(\tilde{\gamma}_t) \), where \( \tilde{\gamma}_t \) is the true (error free) outcome of the experiment. After the experiment is performed and the data \( \tilde{\gamma} \) are obtained, the investigator will alter his subjective notions in the light of the result and arrive at an a posteriori pdf for \( \tilde{\gamma}_t \). We write the pdf for \( \tilde{\gamma}_t \), given the datum \( \tilde{\gamma} \) as \( p_{\text{post}}(\tilde{\gamma}_t|\tilde{\gamma}) \). A straightforward extension of Bayes' theorem (Backus (1970a), Savage (1962)) leads to the following formula for \( p_{\text{post}}(\tilde{\gamma}_t|\tilde{\gamma}) \)

\[
p_{\text{post}}(\tilde{\gamma}_t|\tilde{\gamma}) = \kappa \cdot p_{\text{data}}(\tilde{\gamma}|\tilde{\gamma}_t) p_{\text{prior}}(\tilde{\gamma}_t)
\]

where \( \kappa \) is a suitable normalizing constant. \( p_{\text{post}}(\tilde{\gamma}_t|\tilde{\gamma}) \) will have some expected value \( \tilde{\gamma}_t \) and covariance.

In our problem, our a priori conviction that \( ||m_\gamma^\alpha|| \leq M_\eta \) led to an a priori pdf \( p_{\text{prior}}(\tilde{\beta}_t) \) given by equation (22). The outcome of the experiment yielded the pdf \( p_{\text{data}}(\tilde{\beta}|\tilde{\beta}_t) \) given by equation (20). Thus, through Bayes' theorem, we have

\[
p_{\text{post}}(\tilde{\beta}_t|\tilde{\beta}) \rightarrow (\tilde{\beta}-\tilde{\beta}_t)^T \Gamma \nu^{-1} \Gamma (\tilde{\beta}-\tilde{\beta}_t) + \tilde{\beta}_t^T \Gamma M^{-2} \tilde{\beta}_t
\]

It should be noted that while we shall use pdfs defined on the finite-dimensional spaces \( E_\beta \) and \( E_\beta \), the formal development of Backus' theory requires the use of probability measures defined on the infinite-dimensional space \( \mathcal{H} \). This is a difficult topic, and is covered in detail by Backus (1970c, 1972). Also, our use of pdfs to define regions containing \( m_\gamma^\alpha \) and \( m_\gamma^\beta \) in a probabilistic sense does
not imply that these models are the outcome of some random process; we are simply using the pdfs to express our uncertainties about \( m_{\mathbf{k}} \) and \( \tilde{m}_{\mathbf{F}} \) in a quantitative manner. This means of expressing uncertainty (or degree of knowledge) is fundamental to Bayesian analysis.

There has been considerable controversy over the application of Bayes' theorem. Rather than focusing on the mathematical validity of the theorem, which is unquestioned, objectivists question our ability to define \( p_{\text{prior}}(\mathbf{r}_t) \), and the use of \( p_{\text{data}}(\mathbf{r}_t|\mathbf{r}_a) \) to define a "pdf for \( \mathbf{r}_t \)". When applying Bayes' theorem to geophysical inference problems, we note: (1) in many, if not all geophysical problems, the norm of \( m_{\mathbf{F}} \) is generally constrained by widely accepted physical intuition; (2) the geometrical significance of the data makes its use in defining a region containing \( m_{\mathbf{F}} \) by probabilistic means reasonable, and (3) as Backus (1970a, 1970c, 1972) points out, the solution obtained by using Bayes' theorem reduces to the correct one when the data are error-free, and also in the case with one prediction as determined without recourse to Bayesian techniques. In light of this, and given the simple, efficient and robust solution obtained, we feel, along with Backus, that the Bayesian approach is acceptable and advantageous.

Often, linear constraints on \( m_{\mathbf{F}} \) are available prior to obtaining data from an experiment. If an a priori constraint may be expressed in terms of a linear functional then we are justified in formally treating this information as data and including it in the data set. For example, we may be convinced that the average of \( m_{\mathbf{F}}(x) \) over some subinterval \([a,b]\) of \([0,1]\) lies near a certain
value $\mu$. This conviction may be expressed as

$$\mu \pm \delta\mu = \int_\text{[}a, b\text{]} m(x) \, dx$$

(26)

where $B([a,b])$ is a suitable averaging function (e.g., a boxcar over $[a,b]$) and $\delta\mu$ is an "error" assigned in accordance with the strength of our conviction. Jackson (1979) coined the term "a priori data" for this type of constraint and argued for its inclusion in any data set. Henceforth, the term data will apply to both observational data and any a priori linear constraints that we feel justified in including.
CHAPTER 5 THE MATRIX SOLUTION

We shall now apply the concepts and notation of the previous chapters and give the matrix form for the solution to the inference problem. Expressing the data as a D-tuple \( \vec{\gamma} = (\gamma_1, \ldots, \gamma_D)^T \) and assuming that all of the data are contaminated with zero-mean Gaussian errors we write \( \vec{\gamma} = \vec{\gamma}_t + \delta \vec{\gamma} \) (where \( \vec{\gamma}_t \) is the true value) and express this as the pdf

\[
p_{data}(\vec{\gamma} | \vec{\gamma}_t) \rightarrow (\vec{\gamma} - \vec{\gamma}_t)^T \mathbf{V}^{-1} (\vec{\gamma} - \vec{\gamma}_t)
\]

Proceeding to construct a pdf for the coefficients \( \vec{\beta}_t = [m_c^c] \) in \( E_D \), we have \( p_{data}(\vec{\beta} | \vec{\beta}_c) \) in equation (20) and \( p_{prior}(\vec{\beta}_t) \) (equation (22)) resulting from our conviction that \( ||m_c^c|| \leq M_h \). Combining these with Bayes' theorem produces equation (25). Writing \( p_{post}(\vec{\beta}_t | \vec{\beta}) \) as \( p(\vec{\beta}_t) \), and completing the square (Appendix 3), we obtain

\[
p(\vec{\beta}_t) \rightarrow (\vec{\beta}_t - \vec{\beta}_c)^T (\Gamma \mathbf{V}^{-1} \Gamma^* + M_h^2 \Gamma) (\vec{\beta}_t - \vec{\beta}_c)
\]

with

\[
\vec{\beta}_c = (\Gamma \mathbf{V}^{-1} \Gamma + M_h^2 \Gamma)^{-1} \Gamma \mathbf{V}^{-1} \vec{\gamma}
\]

Equations (27) and (28) define a region in \( E_D \) which will contain the coefficients of \( m_c^c \) at a given level of confidence. At the 1-\( \sigma \) level, that will be the hyperellipsoid defined by the positive definite quadratic form

\[
(\vec{\beta}_t - \vec{\beta}_c)^T (\Gamma \mathbf{V}^{-1} \Gamma + M_h^2 \Gamma) (\vec{\beta}_t - \vec{\beta}_c) \leq 1
\]

The model \( m_c \), given by \( \vec{\beta}_c = [m_c] \) is the most likely candidate for \( m_c^c \),
given the data and our prejudice about the norm of \( m_F \).

Our uncertainty about \( m_F \), expressed by (27) and (28) may be
used to estimate \( \tilde{\varphi}^n \). From (14) we find the corresponding expected
value and covariance matrix for \( \tilde{\varphi}^n \) as

\[
\mathbb{E}[\tilde{\varphi}^n] = \bar{\varphi}_c = G^T \bar{\beta}_c
\]

and

\[
\bar{\varphi}^n = G^T (\Gamma \Gamma^{-1} \Gamma + M_x^{-2} \Gamma)^{-1} G
\]

which leads to the pdf on \( E_P \) for \( \tilde{\varphi}^n \)

\[
p(\tilde{\varphi}^n) \overset{\longrightarrow}{\propto} (\tilde{\varphi}^n - \bar{\varphi}_c)^T \bar{\varphi}^{-1} (\tilde{\varphi}^n - \bar{\varphi}_c)
\]

To calculate \( \tilde{\varphi}^l \), we note that the only information we have
about \( m_F^l \) is the a priori bound \(| |m_F^l|| \leq M_\Lambda \). Thus, if \( \bar{\eta}_c \) is the
P-tuple of coefficients of \( m_F^l \) with respect to the basis for \( \mathcal{D} \), we
find the pdf defined in equation (23). Consequently \( \mathbb{E}[\bar{\eta}_c] = 0 \) and
the covariance matrix of \( \tilde{\varphi}^l \) is

\[
\bar{\varphi}^l = M_\Lambda^2 \Lambda_\Lambda
\]

Our desired values of \( \tilde{\varphi} \) are found by combining \( \tilde{\varphi}^n \) and \( \tilde{\varphi}^l \) as
independent random variables. The expected value and the covariance
matrix are respectively:

\[
\mathbb{E}[\tilde{\varphi}] = \bar{\varphi}_c = G^T \bar{\beta}_c
\]
and

\[ \vec{\mathcal{V}}^2 = \vec{\mathcal{V}}''^2 + \vec{\mathcal{V}}'^2 \]  

(35)

At the 1-\(\sigma\) level of confidence the P-tuple \(\vec{y}\) will be contained in the region defined by

\[ (\vec{\mathcal{V}} - \vec{\mathcal{V}}_c)^T \mathcal{V}^{-1} (\vec{\mathcal{V}} - \vec{\mathcal{V}}_c) \leq 1 \]  

(36)

In practice, it is more convenient to write

\[ \vec{\gamma}_i = \vec{\gamma}_c + \vec{\gamma}^{\text{nu}}_i \]  

(37)

if the covariances may be ignored.

It is convenient to use data whose errors are independent with unit variance. It is always possible to find linear combinations of the data which yield an equivalent data set with those properties (Gilbert, 1971). Once this normalization has been done, the solution requires the inversion of the matrices \(\mathcal{V}\) and \((\mathcal{V} + \mathcal{M}^2\mathcal{I})\). In many geophysical problems, \(\mathcal{V}\) will be poorly conditioned. The addition of \(\mathcal{M}^2\mathcal{I}\) to \(\mathcal{V}\) will generally allow the calculation of a stable inverse for \((\mathcal{V} + \mathcal{M}^2\mathcal{I})\); \(\mathcal{V}^{-1}\), however, must be approximated by discarding eigenvalues and eigenvectors of \(\mathcal{V}\). Since \(\mathcal{V}^{-1}\) appears only as a premultiplier of \(\mathcal{G}\), the effect of approximating \(\mathcal{V}^{-1}\) is to approximate \(\mathcal{G}^{-1}\). The resulting error in \(\vec{\gamma}_i\) may be quantified and held well beneath data error. The details of this stabilization procedure may be found in Appendix 4.
In this section we shall present arguments leading to a more reasonable and restrictive pdf for $m^\varepsilon$. The problem with the pdf for $m^\varepsilon$ derived in the last section was the nature of the a priori constraint $| | m^\varepsilon | | \leq M\varepsilon$. Such a constraint has the possibility of being unduly weak since it admits functions that the investigator would think unreasonable. In particular, the constraint $| | m^\varepsilon | | \leq M\varepsilon$ easily admits the function $m^\varepsilon = 0$, a result which is highly unlikely for many experiments. In view of this, better results might be expected if numbers $M_L$ and $M_H$ could be found such that $M_L \leq | | m^\varepsilon | | \leq M_H$ and if this constraint could be included (probabilistically) in the manner outlined in the last section. Unfortunately, this cannot be done directly, but it can be accomplished by altering our problem slightly.

Instead of producing a most likely model $m^\varepsilon$, and a range of acceptable models around it, as we did in the last section, we shall try to find a most likely model (in $D$) using model construction techniques. If $\beta_\varepsilon$ again represents the coefficients of $m^\varepsilon$ with respect to the basis of $D$, then the coefficients $\beta_\varepsilon = \Gamma^{-1} \gamma_\varepsilon$ will reproduce the data exactly. But almost all inverse problems are unstable in the sense that small variations in the data produce large fluctuations in the constructed model if the data are to be fit exactly. Therefore, when the data are imprecise, it is important that the constructed model which estimates $m^\varepsilon$ does not reproduce those data exactly. When the data errors are independent, Gaussian, with zero-mean and unit variance, the appropriate measure for the acceptability of a constructed model is the $\chi^2$ misfit. $\chi^2$ is defined by
\[ x^2 = \sum_{i=1}^{D} (\gamma_i^c - \gamma_i)^2 \]  

(38)

where \( \gamma_i^c = \langle m, g_i \rangle \) is the \( i \)'th datum predicted by the constructed model \( m \). If \( D \geq 5 \) the expected value of \( x^2 \) is approximately \( D \) and the standard deviation of \( x^2 \) is \( \sigma_{x^2} = 2\sqrt{2D} \). A model generating a chi-squared misfit much less than \( D \) will fit the data too well and will show structure that is merely an artifact of the noise. Alternatively, if \( x^2 >> D \), the data are fit too poorly and information about the model that is contained in the data will have been lost. We will define as acceptable any model with \( D - \sigma_{x^2} \leq x^2 \leq D + \sigma_{x^2} \). The most likely model \( m_0 \) is that one corresponding to \( x^2 = D \).

The construction of \( m_0 \) takes place by using a standard regularization technique. Recognizing that the instabilities in our problem result from small eigenvalues of \( \Gamma \), we stabilize \( \Gamma \) by introducing the regularization parameter \( b \) and use

\[ [m] = (\Gamma + bI)^{-1} \tilde{\gamma} \]

(39)

rather than (17) to construct acceptable models.

Let \( \lambda_i \), \( i=1...D \) be the eigenvalues of \( \Gamma \) and let \( U \) be the matrix whose columns contain the associated eigenvectors. It is straightforward to show (Appendix 5) that

\[ x^2 = \sum_{i=1}^{D} \gamma_i^2 \left( \frac{b}{\lambda_i + b} \right)^2 \]  

(40)

and
\[ ||m||^2 = \sum_{i=1}^{D} \hat{q}_i^2 \left( \frac{\lambda_i}{(\lambda_i + b)^2} \right) \]  

(41)

where \( \hat{q} = \mathbf{U}^T \mathbf{y} \).

As equations (40) and (41) demonstrate, \( \chi^2 \) is a monotonically increasing function of \( b \), while \( ||m|| \) is a monotonically decreasing function of \( b \). Using (39) we may find those values \( b, b_0, b \) corresponding to \( \chi^2 \) values \( D-\sigma_{X^L}, D, \) and \( D+\sigma_{X^L} \), respectively. Using \( b_0 \) in (39) and (41) we find \( m_0 \) and \( ||m_0|| \). Using \( b_u \) and \( b_h \) in (41), we may find upper and lower bounds \( M_h \) and \( M_u \) for \( ||m|| \), if \( m \) is to be acceptable by our \( \chi^2 \) criterion.

Having constructed \( m_0 \), we may reformulate the inference problem by writing \( m^*_F = m_0 + \delta m^*_F \) and substituting into equation (5). The new (translated) data are

\[ \gamma_i^* = \gamma_i - \langle m_0, g_i \rangle = \langle \delta m^*_F, g_i \rangle \quad i=1,...,D \]  

(42)

Similarly, the new predictions are

\[ \gamma_i^* = \gamma_i^* - \langle m_0, p_i \rangle = \langle \delta m^*_F, p_i \rangle \quad i=1,...,P \]  

(43)

Equations (42) and (43) define a new inference problem depending upon \( \delta m^*_F \) which may be solved by stipulating only an upper bound on \( ||\delta m^*_F|| \). That is, specifying \( ||\delta m^*_F|| < \delta M \) and using the probabilistic approach outlined in the last chapter is now precisely the way to proceed.

A reasonable choice for \( \delta M \) might be \( \delta M = (M_h - M_u)/2 \) but if \( M_h \) and \( M_u \) differ greatly, and \( ||m_0|| \) is much closer to one than the other, choosing \( \delta M \) as some fraction of \( ||m_0|| \) might prove to be a
better choice.

Having defined a pdf in $E_p$ for $\delta m^p$, we may proceed as before and calculate $\tilde{\gamma}''$ via (43). The expected or most likely prediction will be $\tilde{\gamma}'' = G^r[m_0]$ and the variance $\tilde{\nu}''$ will depend upon the more restrictive variance for $\delta m^p$. 
We have shown that the "perpendicular part" of the prediction will depend only on the subjective bound $||m^h|| \leq M_\perp$. Backus suggests that a bound for $||m^h||$ may be found by requiring that the projection of $m^h$ onto any subspace of $\mathcal{H}$ have norm bounded by $M$, the bound for $||m^h||$; that is $M_\perp = M$. Alternatively, given an estimate $m_0$ for $m^h$, we could require that

$$||m^h||^2 \leq ||m^h||^2 - ||m_0||^2$$

If this quantity turns out to be too small, we would probably wish to increase our bound for $||m^h||$.

An alternate approach might be as follows. Suppose that we construct $m_0$ as in the previous chapter. This model was constructed to have minimum norm subject to a data-fitting criterion and is known as the "smallest" model. Suppose we then construct the "flattest" model, $m_f$, whose derivative has minimum norm, subject to a data-fitting criterion. Often the flattest model will represent a more "optimum" or realistic model than the smallest model, and will exhibit less structure. If we feel that $m^h$ has at least as much structure as $m_f$, but less than $m_0$ then

$$M_\perp^2 \leq ||m_0 - m_f||^2$$

might be a reasonable choice for $M_\perp$.

Regardless of these rationales for the choice of $M_\perp$, it should be realized that $\frac{\gamma}{\gamma_\perp}$ will always depend upon the prejudice of the investigator. The question of what constitutes realistic and
reasonable bounds for $||m^+_e||$ deserves careful consideration on a problem-by-problem basis.
CHAPTER 8 A NUMERICAL EXAMPLE

The application of the matrix solution will now be demonstrated by calculating the funnel functions (Oldenburg, 1983) for a numerical example. Let \( m(x) = 1.0 - 0.5 \cos 2\pi x \) on the interval \( [0,1] \) and consider data kernels
\[
g_k(x) = \exp(-(k-1)x) \quad (k=1,...,11)\]
for some \( x_0 \) in \( [0,1] \) we shall attempt to compute upper and lower bounds \( (m^u(x_0,\Delta) \) and \( m^l(x_0,\Delta) \) to the average value

\[
\bar{m}(x_0,\Delta) = <m,B(x_0,\Delta)> \quad (44)
\]

where

\[
B(x_0,\Delta) = \begin{cases} 
1/\Delta \text{ for } |x-x_0| \leq \Delta/2 \\
0 \text{ otherwise}
\end{cases}
\]
is a unimodular boxcar of width \( \Delta \) centered on \( x_0 \). Oldenburg used linear programming techniques to calculate \( m^u(x_0,\Delta) \) and \( m^l(x_0,\Delta) \) and plotted these quantities as functions of \( \Delta \). These functions were called funnel functions and they quantified the uncertainty in averages of \( m(x) \) taken over intervals of width \( \Delta \) about \( x_0 \).

For our case, let \( \{\Delta_1, ..., \Delta_p\} \) be a set of numbers defining the widths of intervals over which we seek to estimate the averages (44). Our prediction kernels are then \( p^l(x) = B(x_0,\Delta_l) \) and the set of \( P \) predictions we seek from the data after specifying an a priori bound on \( ||m|| \) are...
\[ \gamma_{i} = \overline{m}(x_0, \Delta_i) = \langle m, \rho_i \rangle \quad i=1\ldots P \quad (45) \]

Using the matrix solution, we may estimate \( m^u(x_0, \Delta_i) \) and \( m^l(x_0, \Delta_i) \) as \( \gamma_{i} + \gamma_{i}^{\frac{1}{2}} \) and \( \gamma_{i} - \gamma_{i}^{\frac{1}{2}} \), respectively, and present this 1-\( \sigma \) level confinement of the \( \gamma_{i} \) as the funnel functions for \( m(x) \).

The resolving power of the data is further quantified by the uncertainty function

\[ \xi(\Delta) = m^u(x_0, \Delta) - m^l(x_0, \Delta) \]

\( \xi(\Delta) \) is the distance between the upper and lower funnel functions. Its value indicates the uncertainty in the average value of \( m \) over a width \( \Delta \) centered on \( x_0 \).

For the following examples we have chosen an a priori upper bound for \( ||m|| \) as \( M=2.0 \). This is a conservative estimate, given the true value, \( ||m||=1.06 \).

Figure 2a shows the "parallel part" of the funnel functions obtained using the method of Chapter 5, with \( M_M=M \) and \( M_L=0 \). With this choice of \( M_M \) we find \( ||m_c||=2.09 \). Given our assumption that \( M=2.0 \), setting \( M_L=0 \) seemingly provides reasonable estimates for the predictions.

A more conservative calculation, using \( M_L=1.0 \) is shown in Figure 2b. Along with the funnel functions, in Figures 2a and 2b are shown the expected values of the predictions, \( \gamma_{i} \) (solid line) and the true values (dashed line). The predictions are consistently greater than the true values. The reason for this bias is clear from Figure 2c, in which the most likely model, \( m_c \), is plotted along with the true model. The large norm of \( m_c \) and the high values of the prediction are clearly a result of the large amplitude and
oscillation of $m_c$. The uncertainty functions are plotted in Figure 2d. The more conservative estimate using $M_L=1.0$ does not introduce a great deal more uncertainty into the solution.

Figure 3 shows equivalent results obtained using the method of Chapter 6. Again $M=2.0$. The "parallel part" of the predictions give rise to the funnel functions in Figure 3a, with $\delta M=0.5||m_0||$ and $M_L=0$. Since $||m_0||=1.73$, in Figure 3b, we show the funnel functions obtained by setting $M_L^2 = M^2 - ||m_0||^2$. Now the expected values of the predictions are very close to the true value. This is the result of the similarity of $m$ and $m_0$ demonstrated in Figure 3c. The uncertainty functions are plotted in Figure 3d. Now the result of specifying a non-zero $M_L$ is significant; if a more confining estimate of $M_L$ is to be had, there may be much improvement in the resulting uncertainty. The funnel functions of Figures 2a and 2b and Figure 3b display essentially the same uncertainty. The latter, however, provides more reliable estimates of the prediction, and hence superior results. When compared with Oldenburg's (1983) Figure 1a, we note that, at the $1-\sigma$ level of confidence the funnel functions we have calculated are equivalent to his calculated using linear programming.
Funnel functions for \( x_0 = 0.5 \), calculated using the method of Chapter 5, with \( M=2.0 \) are shown for \( M_J=0 \) in (a), and for \( M_J=1.0 \) in (b). Also shown in (a) and (b) are the expected values of the predictions (solid line) and the true model averages (dotted line). The most likely model, \( m_x \) is plotted with the true model in (c). The uncertainty functions associated with (a) and (b) are the lower and upper curves, respectively, in (d).
Figure 2
Figure 3
Numerical Example II

Funnel functions for $x_0=5$, calculated using the method of Chapter 6, with $M=2.0$ and $\delta M=.5 ||m_0||$, are shown for $M=0$ in (a) and $M^2=M^2-||m_0||^2$ in (b). Also shown are the expected values of the predictions (solid line) and the true model averages (dotted line). The most likely model, $m_0$ is plotted with the true model in (c). The uncertainty functions associated with (a) and (b) are the lower and upper curves respectively, in (d).
Figure 3
CHAPTER 9 DISCUSSION

Following Backus (1970a) we have derived a matrix method to solve the linear inference problem. The use of Bayesian probability theory allows a simple treatment of erroneous data.

We have shown how to apply a smoothing constraint on $m_\varepsilon$ using the $\chi^2$ statistic. This may lead to more reasonable and restrictive estimates of the predictions than the a priori constraint on $||m_\varepsilon||$ of Backus (1970a), although he described how the problem could be translated with respect to some particular model $m_0$, as we have done (Backus,(1972)).

Our solution is numerically efficient and robust. The two matrix inversions required may be stabilized and the amount of computation required for the inversions does not increase with the number of predictions desired. Increasing the number of predictions increases only the size of matrices involved in multiplication operations. This characteristic is advantageous in problems of interpolation or appraisal where a large number of predictions may be required. The difficulty of computing the inner product matrices will, of course, vary from problem to problem.

Parker (1977a) derived a matrix solution equivalent to Backus' (1970a) solution for accurate data and showed how to treat error in a non-Bayesian fashion. We shall briefly outline Parker's work using our notation to facilitate comparison.

We have defined the subspace $\mathcal{R}$, spanned by $\{g_1...g_d,p_1...p_p\}$ with dual basis written as $\{g^*_1...g^*_d,p^*_1...p^*_p\}$. Let $\check{\gamma}^* = (\gamma_1...\gamma_d,\gamma_1...\gamma_p)$ be the $D+P$-tuple of data and predictions. $\check{\gamma}^*$ is clearly the coefficient matrix of the orthogonal projection of $m_\varepsilon$ onto $\mathcal{R}$, with respect to the dual basis. That projection has norm
\[ \mathbf{y}^* \mathbf{H}^{-1} \mathbf{y}^* \]  \hspace{1cm} (46)

where

\[ \mathbf{H} = \begin{bmatrix} \Gamma & \mathbf{G} \\ \mathbf{G}^T & \mathbf{A} \end{bmatrix} \]

is the inner product matrix of \( \mathcal{R} \), and the submatrices \( \Gamma \), \( \mathbf{G} \) and \( \mathbf{A} \) are as previously defined. We demand that the projection of \( m_\mathcal{E} \) onto any subspace have norm bounded by \( M \), so that

\[ \mathbf{y}^* \mathbf{H}^{-1} \mathbf{y}^* \leq M \]  \hspace{1cm} (47)

When the data, (assumed error-free) are substituted into (47), we have a positive definite quadratic form defining the range of \( \gamma \) consistent with the data and a priori bound.

Parker shows how a more confining solution may be obtained for a special case of the inference problem. Suppose the predictions represent the coefficients of some expansion of \( m_\mathcal{E} \). Presumably, we have chosen the functions \( p_i(x) \) so that such an expansion represents \( m_\mathcal{E} \) well. Then a bound more confining than (47) may be applied as follows.

Define the subspace \( \mathcal{D} \), with basis \( \{p_1,...,p_\mathcal{D}\} \) and dual basis \( \{p'_1,...,p'_\mathcal{D}\} \). Then the predictions, \( \mathbf{y}^* \) are the coefficients of \( m_\mathcal{E}^* \), the orthogonal projection of \( m_\mathcal{E} \) onto \( \mathcal{D} \), with respect to the dual basis. In approximating \( m_\mathcal{E}(x) \) as an expansion in terms of the functions \( p_i(x) \), we write

\[ m_\mathcal{E}(x) = m_\mathcal{E}^*(x) + m^*(x) \]  \hspace{1cm} (48)
and hope that that part of \( m_e(x) \) not accounted for in our finite expansion, \( m^*(x) \), is small. That is, \( m_e^* \) is a good approximation to \( m_e \). From (48) we find

\[ ||m^*||^2 \leq ||m_e||^2 - ||m_e^*||^2 \]  

(49)

and express our prejudice about \( m^* \) by

\[ ||m^*|| \leq M^* \]  

(50)

where \( M^* \) is small. In matrix form, (50) becomes

\[ \hat{\gamma}^*^T \left[ H^{-1} - \begin{bmatrix} 0 & 0 \\ 0 & \Lambda^{-1} \end{bmatrix} \right] \hat{\gamma}^* \leq M^* \]  

(51)

When the data are substituted into (51), a positive definite quadratic form for \( \hat{\gamma} \) is defined that yields a much more restrictive solution than does (47).

Parker calculated the additional uncertainty imposed on the solution by erroneous data. Given the statistics of the data errors, he calculated the probability that a given prediction is constrained by the data values through arguments associated with the geometry defined in \( \mathbb{R} \) by the quadratic forms (47) or (51). Parker concluded that this procedure is not numerically efficient.

Parker's algorithm was presented in a geometric context. By introducing his parameter-data space, the roles of data, the a priori bounds, and the predictions are made clear. For the problem of model fitting, his bound (50) and associated quadratic form (51)
give superior results. Unfortunately, the use of the dual basis for $R$ requires that both a $D$-dimensional matrix and a $P$-dimensional matrix be inverted, with no method of stabilization proposed for ill-conditioned problems.

The imposition of more restrictive bounds such as (51) is less straightforward with our algorithm. To apply such bounds, the subspace $D$ must be enlarged with suitable functions via the introduction of linear a priori constraints beyond the data, which may become somewhat contrived unless the investigator is careful in controlling his bias.

Jackson (1979) derived matrix equations which may be used to solve the inference problem if $m_f(x)$ is represented by a finite number of parameters. His a priori information was included as linear constraints and via a hypothesized covariance matrix relating the model parameters. It is interesting to note the similarity of his and our results, given the different approaches to the problem.

Oldenburg (1983) has shown how linear programming (LP) may be applied to the solution of linear inference problems. The LP method has several advantages:

(1) The solution is not probabilistic; the bounds on a prediction are the greatest and least consistent with a given data set;

(2) Rather than imposing a bound on $||m_f||$ to find a solution, LP allows the use of non-probabilistic, pointwise constraints.

(3) If applicable, a positivity constraint may be used.
The principle drawback to the use of LP for multiple prediction is that each prediction is done separately; computing time increases rapidly with the number of predictions. For the example presented in this paper, the LP solution required approximately three times as much computing time.
CHAPTER 10 CONCLUSION

We have presented Backus' solution to the linear inference problem with multiple predictions in matrix form. The matrix algorithm is efficient and stable, and with the careful application of linear a priori constraints and smoothing assumptions, it appears as if usefully restrictive solutions may be obtained. We hope that this elementary treatment of Backus' work will help to earn for it the attention it deserves.
REFERENCES


Backus, G.E., 1972, Inference from inadequate and inaccurate data, in Mathematical Problems in the Geophysical Sciences, Amer. Math. Soc., Providence, RI.


APPENDIX 1

Matrix Form of the Predictions

Write the predictions as

\[ \tilde{\gamma}_i = \gamma_i^* + \gamma_i^+ = \langle p_i^*, m_i^* \rangle + \langle p_i^+, m_i^+ \rangle \quad i=1,...,P \]  \hspace{1cm} (A1.1)

We shall express the predictions as elements of \( E_P \), that is as column vectors or \( P \)-tuples

\[ \tilde{\gamma} = \tilde{\gamma}^* + \tilde{\gamma}^+ = (\gamma_1^* \ldots \gamma_P^*)^T + (\gamma_1^+ \ldots \gamma_P^+)^T \]  \hspace{1cm} (A1.2)

To calculate \( \tilde{\gamma}_i^* \), write

\[ m_i^* = \sum_{j=1}^{P} \beta_j g_j \]

so that, for \( i=1,...,P \),

\[ \tilde{\gamma}_i^* = \langle p_i^*, m_i^* \rangle = \langle p_i^*, m_i^* \rangle \]
\[ = \sum_{j=1}^{P} \beta_j \langle p_i^*, g_j \rangle \]
\[ = \sum_{j=1}^{P} \beta_j \langle g_j, p_i \rangle \]

Thus, if \( G_{ij} = \langle g_j, p_i \rangle \),

\[ \tilde{\gamma}_i^* = \sum_{j=1}^{P} G_{ij} \beta_j \]

Since \( \tilde{\beta} = (\beta_1, \ldots, \beta_P)^T = [m_i^*] \), we have

\[ \tilde{\gamma}^* = G^T [m_i^*] \]  \hspace{1cm} (A1.3)
To find $\gamma^T$, we need to know the inner product matrix for $\mathcal{R}$, that is $\Lambda\perp$, where

$$\Lambda_{ij} = \langle p_i^+, p_j^+ \rangle$$

To find a simple expression for $\Lambda\perp$, note that

$$\Lambda_{ij} = \langle p_i^+, p_j^+ \rangle = \langle p_i^-, p_j^-, p_i^+ \rangle$$
$$= \langle p_i^-, p_j \rangle - \langle p_i^-, p_j^+ \rangle$$
$$= \Lambda_{ij} - \Lambda_{ii}$$

Now,

$$\Lambda_{ii} = \langle p_i^-, p_j^+ \rangle = \langle p_i^+ \rangle^T \Gamma [p_j^+]$$

Using a result from Appendix 2, and standard indicial notation,

$$\Lambda_{ij} = (\Gamma^{-1} [p_i^+] d)^T \Gamma (\Gamma^{-1} [p_j] d)$$
$$= [p_i^+]^T \Gamma^{-1} [p_j] d$$
$$= \langle q_k, p_i \rangle \Gamma_{kl}^1 \langle q_l, p_j \rangle$$
$$= G_{ki} \Gamma_{kl}^1 \Gamma_{lj}$$

or

$$\Lambda_{ii} = G^T \Gamma^{-1} G$$

whence

$$\Lambda_{\perp} = \Lambda - G^T \Gamma^{-1} G \quad \text{(A1.4)}$$

Now, on $\mathcal{R}$, for each $i=1\ldots P$, 

$$\gamma_i^+ = \langle p_i^+, m_i^+ \rangle = [p_i^+]^T \Lambda_{\perp} [m_i^+]$$.
Since \{p_1^+, \ldots, p_\rho^+\} is our basis for P, we find

\[ \mathcal{E}_\perp = \Lambda_\perp [m_\xi^+] \]  \tag{A1.5}
APPENDIX 2
Dual Bases, Inner Product Matrices and Orthogonal Projections

Consider the subspace of $\mathcal{H} = \text{span}\{g_1, \ldots, g_D\}$. The vectors $\{g_i\}, \quad i=1, \ldots D$ are assumed to be linearly independent, and thus form a basis for $\mathcal{D}$. The basis dual to this is the set of $D$ vectors $\{g^1, \ldots, g^D\}$ defined by

$$<g_i, g^j> = \delta_{ij} \quad \text{for} \quad i,j=1 \ldots D \quad (A2.1)$$

where $\delta_{ij}$ is the Kronecker delta. The dual basis always exists and is unique.

Now, if one of the elements of the dual basis, $g^j$ is expressed in terms of the original basis,

$$g^j = \sum_{k=1}^{D} B_{jk} g_k$$

By (A2.1),

$$<g_i, g^j> = \sum_{k=1}^{D} B_{jk} <g_i, g_k> = \delta_{ij}$$

Defining the inner product matrix $\Gamma_{ij} = <g_i, g_j>$ leads to $B \Gamma = I$ or $B = \Gamma^{-1}$. (Since $\Gamma$ is positive definite and symmetric, $\Gamma^{-1}$ always exists and is positive, definite and symmetric).

The above relation allows the following characterization of $\Gamma^{-1}$; it is the matrix whose columns contain the coefficients of the vectors $g^i$ with respect to the basis $\{g_1, \ldots, g_D\}$.

Thus if $m$ is in $\mathcal{D}$, and $[m]$ and $[m]_d$ are the matrices of $m$ with respect to the original and dual bases, respectively,
[m] = \Gamma^{-1}[m]_d \quad \text{(A2.2)}

Let m be an element of $\mathcal{R}$. The orthogonal projection of m onto $\mathcal{D}$ is that vector $m''$ in $\mathcal{D}$ such that for all vectors $x$ in $\mathcal{D}$

$$<m-m'',x> = 0 \quad \text{(A2.3)}$$

By the projection theorem $m''$ exists and is unique.

Since $x$ is in $\mathcal{D}$, (A2.3) is equivalent to the D equations

$$<m-m'',g_i> = 0 \quad i=1...D \quad \text{(A2.4)}$$

whence

$$<m,g_i> = \gamma_i = <m'',g_i> \quad \text{(A2.5)}$$

Write $m''$ in terms of the dual basis

$$m'' = \sum_{j=1}^{D} \alpha_j g_j$$

so that (A2.5) becomes

$$\gamma_i = <m,g_i> = \sum_{j=1}^{D} \alpha_j <g_j,g_i> = \sum_{j=1}^{D} \alpha_j \delta_{ij}$$

which implies

$$\gamma_i = a_i \quad i=1...D \quad \text{(A2.6)}$$

Thus, the numbers $\gamma_i = <m,g_i>$ are the coefficients of $m''$ with
respect to the dual basis of $\mathcal{D}$. By (A2.2) then,

$$[m^n] = \Gamma^{-1} [m^n]_d = \Gamma^{-1} \bar{\gamma}$$  \hspace{1cm} (A2.7)

Equation (A2.7) demonstrates that the data are equivalent to knowledge of $m^n$. 
APPENDIX 3

Calculation of $\hat{\beta}_c$

The joint normal distribution for $\vec{\beta}_c = [m_{\vec{\beta}}]$ is given by equation (25)

$$p(\vec{\beta}_c) \rightarrow (\vec{\beta}_c - \vec{\beta}_c)^T \Gamma V^{-1} \Gamma (\vec{\beta}_c - \vec{\beta}_c) + \vec{\beta}_c^T (M_{\vec{\beta}}^2) \Gamma \vec{\beta}_c$$

or

$$p(\vec{\beta}_c) \rightarrow \vec{\beta}_c^T (\Gamma V^{-1} \Gamma + (M_{\vec{\beta}}^2) \Gamma) \vec{\beta}_c - 2\vec{\beta}_c \Gamma V^{-1} \Gamma \vec{\beta} + \text{constant}$$

(A3.1)

where $\vec{\beta} = \Gamma^{-1} \vec{\gamma}$.

Equation (A3.1) is to be rewritten as

$$p(\vec{\beta}_c) \rightarrow (\vec{\beta}_c - \vec{\beta}_c)^T C (\vec{\beta}_c - \vec{\beta}_c) + \text{constant}$$

(A3.2)

which may be expanded to

$$p(\vec{\beta}_c) \rightarrow \vec{\beta}_c^T C \vec{\beta}_c - 2\vec{\beta}_c C \vec{\beta}_c + \text{constant}$$

(A3.3)

For equality between (A3.1) and (A3.2) to hold, up to a constant,

$$C = (\Gamma V^{-1} \Gamma + (M_{\vec{\beta}}^2) \Gamma)$$

(A3.4)

and

$$\vec{\beta}_c C \vec{\beta}_c = \vec{\beta}_c \Gamma V^{-1} \Gamma \vec{\beta}$$
For this to be the case

\[ \hat{\beta}_c = (\Gamma V^{-1} \Gamma + (M_u^2) \Gamma)^{-1} \Gamma V^{-1} \hat{\beta} \]  

(A3.5)

Note that \((\Gamma V^{-1} \Gamma + (M_u^2) \Gamma)\) is positive definite and symmetric, so that it has a positive definite symmetric inverse.

Recalling that \(\hat{\beta} = \Gamma^{-1} \gamma\), we find

\[ \hat{\beta}_c = (\Gamma V^{-1} \Gamma + (M_u^2) \Gamma)^{-1} \Gamma V^{-1} \gamma \]  

(A3.6)
APPENDIX 4
Computational Details

If the errors on the data are not independent, linear combinations of the data equations may be found that yield an equivalent data set with independent errors (Gilbert, 1971). Assuming that this has been accomplished it is convenient to normalize each data equation by the standard deviation of the associated error so that the resulting normalized data all have unit variance, that is we use

$$\chi_i = \frac{<m_i^T, g_i>}{\sigma_i} \quad i=1...D \quad (A4.1)$$

as the data equation rather than (5). The normalized data kernels are then $\bar{g}_i = g_i / \sigma_i$ and the inner product matrix $\Gamma$ is found using these normalized kernels. The matrix $V^{-1}$ is then the identity.

To carry out the solution, it is necessary to invert the $DxD$ matrices $(rV^{-1} + M_\parallel^2 I)$ and $\Gamma$. When using normalized kernels, we find

$$(rV^{-1} + M_\parallel^2 I - 1)^{-1} = (\Gamma + M_\parallel^2 I)^{-1} \Gamma^{-1} \quad (A4.2)$$

In many geophysical problems, $\Gamma$ will be poorly conditioned. The addition of $M_\parallel^2 I$ will generally be sufficient to allow the calculation of a stable inverse for $(\Gamma + M_\parallel^2 I)$. $\Gamma$, however, may remain ill-conditioned.

Note that $\Gamma^{-1}$ appears only as a premultiplier of $G$, i.e. as $\Gamma^{-1}G$. By (A2.2) and the definition of $G$, the matrix $\Gamma^{-1}G$ has as columns the coefficients of the vectors $p_i^\prime$. If $\Gamma^{-1}$ is found only
approximately, in a manner described below, the effect is to approximate the \( p_i \). Since the \( p_i \) are found in the expression \( \tilde{\gamma}_i = \langle p_i, m_\varepsilon \rangle \), this approximation will introduce some error into the calculation of the \( \tilde{\gamma}_i \). This error will be quantified, and in cases of practical interest it should be insignificant with respect to data error.

Since \( \Gamma \) is positive definite and symmetric, it possesses the singular value decomposition \( \Gamma = U \Omega U^T \) with inverse \( \Gamma^{-1} = U \Omega^{-1} U^T \), where \( U \) is an orthogonal matrix containing the eigenvectors of \( \Gamma \) as its columns. \( \Omega \) is diagonal, with entries \( \lambda_j \) which are the eigenvalues of \( \Gamma \). The condition number of \( \Gamma \) is defined as \( c = |\lambda_{\min}/\lambda_{\max}| \). If \( c \) is near the precision of the machine used in the calculation of \( \Gamma^{-1} \), it is not possible to find a stable inverse. Suppose the eigenvalues of \( \Gamma \) are ordered with decreasing size and that the \( S \) smallest are discarded so that \( |\lambda_{D-S}/\lambda_1| \) is above machine precision. Then defining \( \tilde{U} \) as the \( (D \times (D-S)) \) matrix of the first \( (D-S) \) eigenvectors of \( \Gamma \) and \( \tilde{\Omega} \) as the \( (D-S) \times (D-S) \) diagonal matrix containing the first \( (D-S) \) eigenvalues of \( \Gamma \), \( \tilde{\Gamma}^{-1} = \tilde{U} \tilde{\Omega}^{-1} \tilde{U}^T \) approximates \( \Gamma^{-1} \).

To calculate the error in \( \tilde{\gamma}_i \) introduced by this approximation write

\[
\tilde{\gamma}_i = \langle \tilde{p}_i, m_\varepsilon \rangle = \langle p_i, m_\varepsilon \rangle + \langle \tilde{p}_i - p_i, m_\varepsilon \rangle
\]

where \( \tilde{p}_i \) is the approximation to \( p_i \) found by using \( \tilde{\Gamma}^{-1} \).

Now \( \langle \tilde{p}_i - p_i, m_\varepsilon \rangle \leq ||\tilde{p}_i - p_i|| ||m_\varepsilon|| \leq \sum_{j=2}^{D-S} ||p_i|| |\lambda_j M| = e \), so if \( U \) and \( \Omega \) are truncated so that \( c \) is above machine precision, and \( e \ll \delta \gamma_i \) for \( i = 1, \ldots, D \), then \( \Gamma^{-1} \) provides a stable approximation to
\( \Gamma^{-1} \) without introducing significant error into the solution.
APPENDIX 5

Regularization of \( \Gamma \)

The regularization procedure used to obtain equation (40) and (41) are most readily understood by using the orthonormal basis for \( D \) given by Parker (1977b).

As in Appendix 4, we consider that the data and associated kernel function have been normalized by the standard deviations of the data errors. The inner product matrix \( \Gamma \) is decomposed into

\[
\Gamma = U \Omega U^T \quad \text{(A5.1)}
\]

We form linear combinations of the normalized kernels to find

\[
\psi_i = \frac{1}{\sqrt{\lambda_i}} \sum_{j=1}^{D} U_{ij} \bar{g}_j \quad \text{(A5.2)}
\]

the vectors \( \psi_i \) are easily shown to be an orthonormal basis for \( D \).

The coefficients of \( m_n^\theta \) with respect to this new basis are

\[
a_i = <m_n^\theta, \psi_i > \\
= \frac{1}{\sqrt{\lambda_i}} \sum_{j=1}^{D} U_{ij} <m_n^\theta, g_j > \\
= \frac{1}{\sqrt{\lambda_i}} \sum_{j=1}^{D} U_{ij} \gamma_j
\]

or

\[
a_i = \frac{\gamma_i}{\sqrt{\lambda_i}} \quad \text{(A5.3)}
\]

if

\[
\gamma = U^T \gamma
\]
Thus

$$\hat{\gamma}_L = \sqrt{\lambda}_L \langle m^\perp, \psi_L \rangle$$  \hspace{1cm} (A5.4)

Now, suppose that $m$ is a model found from

$$\hat{\beta} = [m] = (\Gamma + bI)^{-1} \hat{\gamma}$$  \hspace{1cm} (A5.5)

and $\hat{\gamma}$ is the data we would expect from $m$, that is

$$\gamma^c = \langle m, g_L \rangle$$

Then, as before

$$\hat{\gamma}^c = \sqrt{\lambda}_L \langle m, \psi_L \rangle$$  \hspace{1cm} (A5.6)

where

$$\hat{\gamma}^c = U^T \hat{\gamma}^c$$

Now, to find the coefficients of $m$ with respect to the orthonormal basis, that is $\langle m, \psi_L \rangle$, we note

$$\hat{\gamma} = (\Gamma + bI)\hat{\beta}$$
$$= U \Omega U^T \hat{\beta} + bI \hat{\beta}$$

or
Thus

\[ \hat{\gamma}_i = \sqrt{\lambda_i} \langle m^e_i, \psi_k \rangle \quad (A5.4) \]

Now, suppose that \( m \) is a model found from

\[ \hat{\beta} = [m] = (\Gamma + bI)^{-1} \hat{\gamma} \quad (A5.5) \]

and \( \hat{\gamma} \) is the data we would expect from \( m \), that is

\[ \gamma^e_i = \langle m, g_i \rangle \]

Then, as before

\[ \hat{\gamma}^e_i = \sqrt{\lambda_i} \langle m, \psi_i \rangle \quad (A5.6) \]

where

\[ \hat{\gamma}^e = U^T \hat{\gamma} \]

Now, to find the coefficients of \( m \) with respect to the orthonormal basis, that is \( \langle m, \psi_i \rangle \), we note

\[ \hat{\gamma} = (\Gamma + bI)\hat{\beta} = UU^T \hat{\beta} + bI \hat{\beta} \]

or
\[ \gamma = \hat{\omega}_\beta + b \hat{\theta} \]  

(A5.7)

where \( \hat{\theta} = U^T \hat{\beta} \), since \( UU^T = I \).

From (A5.7),

\[ \hat{\beta}_i = \frac{\gamma_i}{\lambda_i + b} \]  

(A5.8)

and

\[ \hat{\theta} = U \hat{\beta} \]

Now,

\[ m = \sum_{i=1}^{D} \beta_i \bar{g}_i \]

\[ = \sum_{i=1}^{D} \sum_{j=1}^{P} U_{ij} \hat{\beta}_j \bar{g}_i \]

\[ = \sum_{i=1}^{D} \frac{\gamma_i}{\lambda_i + b} \sqrt{\lambda_i} \psi \]

or

\[ <m, \psi_i> = \frac{\hat{\gamma}_i \sqrt{\lambda_i}}{\lambda_i + b} \]  

(A5.9)

With these results we may find \( x^2 \) as

\[ x^2 = \sum_{i=1}^{D} (\gamma_i^2 - \gamma_i) \]

\[ = \sum_{i=1}^{D} (\hat{\gamma}_i^2 - \hat{\gamma}_i^2) \]

\[ = \sum_{i=1}^{D} \lambda_i \left[ <m, \psi_i> - <m, \psi_i> \right]^2 \]

\[ = \sum_{i=1}^{D} \hat{\gamma}_i^2 \left[ \frac{b}{\lambda_i + b} \right]^2 \]  

(A5.10)
(recalling that, for the normalized data, $c_l=1$).

And

$$||m||^2 = <m,m> = \sum_{i=1}^{D} \frac{\gamma_i^2}{\lambda_i + b} \lambda_i$$

using (A5.9) and the fact that the $\psi_l$ are orthonormal.