Linear Inference and Underparameterized Models

ROBERT L. PARKER

Institute of Geophysics and Planetary Physics, Scripps Institution of Oceanography, University of California, San Diego La Jolla, California 92093

A version of Backus's theory of linear inference is developed by using a new finite-dimensional space. This approach affords a clear geometric interpretation of the essential role played by a priori model smoothing assumptions and also facilitates the construction of a theory for the treatment of random data errors that is quite different from the treatment of Backus. When the unknown parameters form a (necessarily incomplete) description of the model, it is possible to formulate a special smoothing assumption that is particularly appropriate; in practical examples this strategy often leads to tighter bounds on the model uncertainty than those obtained with previous assumptions. An analysis of the numerical aspects of the problem forces one to the conclusion that the theory is not competitive numerically with conventional least squares parameter estimation, unless one of the large submatrices in the problem possesses a simple inverse. An example of this kind is discussed briefly.

INTRODUCTION

The classical statistical theory of parameter estimation rests on the assumption that a small number of parameters control the deterministic behavior of the observations but that the actual measurements are influenced by an additional random disturbance. Then the theory of least squares approximation leads directly to estimates of model parameters and their confidence limits; for a lucid summary of the theory, see *Mathews and Walker* [1964], and for an exhaustive treatment, consult *Rao* [1974]. Quite often in geophysics, however, it is not legitimate to maintain that only a few variables determine the observations: for example, the external gravitational field of the earth requires in principle infinitely many parameters for its full description. Usually, the geophysicist is content with estimates of a few of the relevant quantities, in this case, the first few spherical harmonic coefficients. The customary procedure is to set up a model containing the desired parameters as unknowns and to find them by the least squares fitting method. Now the formal statistical basis for least squares fitting is absent, because the misfit between model prediction and actual observation stems from model deficiency, not noise in the measurements. Furthermore, it is obvious that the standard recipe for calculating parameter uncertainty is inappropriate.

A formal way of handling this problem is to define the earth as the realization of a statistical process [e.g., *Franklin*, 1970]. One disadvantage of this in my opinion is that in order to make progress, quite strong (and sometimes unwarranted) assumptions are required but that even then, there may be no reduction in the number of unknowns: for example, if it is assumed that a parameter in the earth is a particular sample of a stationary, second-order Gaussian process (a very strong requirement), then the autocorrelation function is still an unknown continuous function, quite as difficult to determine as the original parameter distribution itself from a finite number of observations. Furthermore, the attribution of statistical properties to the physical model is sometimes very artificial, as it is in the case in which density or seismic velocity varies with depth.

This paper describes a theory for determining parameters and their uncertainties in the case of a linear model defined upon an infinite-dimensional space without introducing statistical elements into the definition of the model. It is therefore another formulation of the linear inverse problem encountered so often in geophysics [*Parker*, 1977]. The approach adopted here differs from most previous work [e.g., *Backus and Gilbert*, 1970; *Wiggins*, 1972; *Jackson*, 1972] by emphasizing the role of assumptions that must be introduced to limit indeterminacy in the required predictions; in fact, the present paper follows closely the development of *Backus* [1970*a*, *b*, *c*, 1972] but proceeds on a more elementary level. Following Backus, we assume that the system can be modeled as an element in an infinite-dimensional, separable Hilbert space. The results of the observations and the parameters to be found are taken to be continuous linear functionals of the model; more general linear functionals can be reduced to this form [*Backus*, 1970*b*, 1972]. *Backus* [1970*a*] pointed out that apart from certain special cases the parameters are entirely undetermined by the data unless something is assumed about the model. Note parenthetically that this unhappy situation does not always pertain to nonlinear processes: nonlinear predictors [*Parker*, 1975] or nonlinear model equations [*Parker*, 1972] sometimes yield concrete information about the earth without the need for further assumptions. The nonlinearity often arises through a physical constraint like positivity of the unknown model [*Sabatier*, 1977], but it should be stressed that it is the nonlinearity of such a constraint that allows it to be effective in reducing ambiguity; if the physical constraint were linear, it would look just like an additional datum and would therefore leave the model as undetermined as before. In the linear problem the difficulty of unbounded parameter uncertainty can be avoided if we reject models that are physically unreasonable by restricting the class of possible solutions in some way. There is obviously more than one way of introducing such smoothing assumptions. Backus chose to assume a maximum value on the norm of the model [*Backus*, 1970*a*] or of the model's projection onto some subspace [*Backus*, 1970*c*, 1972]. In our application the unknowns are coefficients in a representation of the model; we suggest that a reasonable restriction is on the maximum deviation of the true structure from our representation of it. We can then express our confidence that the chosen representation is a good one by using rather small values for this bound.

The mathematical machinery used by Backus was the geometry of linear functionals in a Hilbert space. New insight is gained, however, by the introduction of a finite-dimensional space, one in which every element combines the unknown parameters and the data. In this space the smoothing assump-

Paper number 7R0682.

tion works by confining possible elements to a restricted region, for example, with Backus's assumption, to a hyperellipsoid centered on the origin. Fixing the data values passes a hyperplane through the parameter-data space; the region common to the hyperplane and the volume defined by the smoothing assumption is the one in which all possible values of the unknown parameters lie. The mathematical details will be filled out in the next few sections. Perhaps the greatest benefit proceeding from the use of parameter-data space is the very clear intuitive picture it gives of the relation between measurement, assumption, and determined parameters. This enables us to give an entirely elementary derivation of the effects of statistical noise in the observations without the need of a Bayesian framework. Furthermore, matrix notation naturally gives an indication of the numerical labor involved. A careful study of this aspect reveals, unfortunately, that really large data sets are quite intractable with our approach, unless there exists a particular, felicitous formulation of the underlying problem. We are able to give an example having this useful property.

BACKUS REVISITED

The basic ideas are as follows: we are given observations in the form of D real numbers $\gamma_1, \gamma_2, \cdots, \gamma_i, \cdots, \gamma_D$, which are exact for the moment. These D data are to be used in a linear inverse problem to investigate an earth model m_E; instead of attempting to describe m_E exactly we shall find estimates of P parameters $\tilde{\gamma}_1, \tilde{\gamma}_2, \cdots, \tilde{\gamma}_k, \cdots, \tilde{\gamma}_P$, which are also linear functionals of the model m_E. These parameters could be local averages of the structure or, as we shall emphasize later, coefficients in an expansion of m_E in terms of a convenient basis set of functions (like spherical harmonics). *Backus* [1970*a*] shows that without additional information about m_E we can say nothing whatever about $\tilde{\gamma}_k$; he proposes that we may often know an upper limit on the size of m_E, expressed in terms of the norm m_E. Even with a knowledge of this bound the parameters $\tilde{\gamma}_k$ are not uniquely specified; instead, they can assume only a limited range of values, which is described by the Backus equation (4). In this section we review *Backus*'s [1970*a*] work and obtain his equation (4) with the aid of parameter-data space. The notation used here will be similar to his, but we establish some additional quantities necessary for the new viewpoint. For the moment, uncertainty in the data is ignored. Backus writes for each of the D data

$$\gamma_i = \langle g_i, m_E \rangle \qquad i = 1, 2, \cdots, D \tag{1}$$

where m_E is the element of a Hilbert space $\mathcal{K}$ representing the actual earth and $g_i \in \mathcal{K}$ is an element associated with γ_i, the ith datum; angle brackets denote the inner product in $\mathcal{K}$. Equation (1) represents the theoretical connection between the model m_E and each observable γ_i, and it is central to the development that D is finite, a point beyond dispute for any practical situation. We wish to determine the values of P numbers $\tilde{\gamma}_k$ (the predictions), and these too are continuous linear functionals on $\mathcal{K}$:

$$\tilde{\gamma}_k = \langle \tilde{g}_k, m_E \rangle \quad k = 1, 2, \cdots, P \tag{2}$$

This relationship expresses the fact that if we actually knew m_E, we could find all the quantities $\tilde{\gamma}_k$. In least squares fitting, it is unusual to give a procedure for calculating the parameters from a known model, but one can usually be derived if it is not obvious. For example, if P coefficients in a polynomial fit to a profile are sought, they can in principle be discovered from P integrals over the true profile (e.g., by expanding each term in orthogonal Legendre polynomials); or when a layered model is considered, a given layer value must be interpreted as the mean value within that layer, and this interpretation is now clearly in the form of (2). From this approach to least squares fitting, it will be seen that orthogonal $\tilde{g}_k$ are often natural, as is true in the case of nonoverlapping layers.

A smoothing assumption is introduced by claiming that we can find an upper bound on the norm of the model, i.e.,

$$\|m_E\|^2 = \langle m_E, m_E \rangle \leq M_0^2 \tag{3}$$

In our application, other assumptions are more useful, but we shall pursue the analysis with Backus's constraint for a while.

We now introduce $\mathcal{K}$, the parameter-data space mentioned earlier: $\mathcal{K}$ is simply an N-dimensional Euclidean coordinate space R^N, where $N = P + D$. An element $\Gamma \in \mathcal{K}$ is an ordered N-tuple of real numbers $\{\tilde{\gamma}_1, \tilde{\gamma}_2, \cdots, \tilde{\gamma}_P, \gamma_1, \gamma_2, \cdots, \gamma_D\}$; i.e., the unknown parameters followed by the data comprise the components of Γ; we shall write the jth component as Γ_j. We define two orthogonal subspaces of $\mathcal{K}$: $\mathcal{P}$ and $\mathcal{D}$. These are the parameter and data subspaces, so that an element $\tilde{\gamma} \in \mathcal{P}$ has zero for its last D components, while $\gamma \in \mathcal{D}$ has zero for its first P components. Operations in $\mathcal{K}$ will be carried out mainly with conventional matrix algebra notation. In particular, a quadratic form will soon be encountered written $\Gamma^T\mathbf{Q}\Gamma$; this is to be interpreted as $\sum_{i,j}\Gamma_i Q_{ij}\Gamma_j$, the $N \times N$ matrix $\mathbf{Q}$ consisting of components Q_{ij}. Finally, for convenience, N elements $G_j \in \mathcal{K}$ are defined as follows: $G_j = \tilde{g}_j$ for $1 \leq j \leq P$, $G_j = g_{j-P}$ for $P < j \leq N$.

Having disposed of these formalities, let us now ask how the inequality (3) expresses itself in $\mathcal{K}$. We argue as follows: fix an arbitrary element $\Gamma \in \mathcal{K}$, which then fixes the parameters $\tilde{\gamma}_k$ and data values γ_i. Then if m is any element consistent with Γ,

$$\Gamma_i = \langle G_i, m \rangle \qquad i = 1, 2, \cdots, N$$

Now find the smallest value of $\|m\|^2$ for an m satisfying the above relation; this minimum value will be called $\|m_0\|^2$. If $\|m_0\|^2$ violates (3), Γ is not consistent with the smoothing assumption, because if the smallest value of $\|m\|^2$ exceeds M_0^2, every other value will also. Conversely, if $\|m_0\|^2 \leq M_0^2$, there is at least one model consistent with (1), (2), and (3), so that Γ is then an acceptable element, consistent with our assumption. In Appendix A we derive the value of $\|m_0\|^2$:

$$\|m_0\|^2 = \Gamma^T\hat{\mathbf{G}}^{-1}\Gamma \tag{4}$$

where the matrix $\hat{\mathbf{G}}^{-1}$ is the inverse of $\hat{\mathbf{G}}$, whose elements are $\hat{G}_{ij} = \langle G_i, G_j \rangle$. To exist at all, $\hat{\mathbf{G}}^{-1}$ must be derived from the $G_i \in \mathcal{K}$ that are linearly independent. *Backus* [1970*a*] argues that any linear inference problem can be reduced to one having this property; we shall not repeat this discussion here. From (3) and (4),

$$\Gamma_a^T\hat{\mathbf{G}}^{-1}\Gamma_a \leq M_0^2 \tag{5}$$

where the $\Gamma_a \in \mathcal{K}$ are the acceptable elements consistent with the smoothing assumption (3). It is easily seen that $\hat{\mathbf{G}}$ is positive definite, and therefore so is $\hat{\mathbf{G}}^{-1}$; hence (5) defines the interior and boundary of a hyperellipsoid in $\mathcal{K}$. We denote this region by Q_0 so that $\Gamma_a \in Q_0 \subset \mathcal{K}$.

When the D data values are assigned their actual numerical values, this defines a P-dimensional hyperplane $\Delta \subset \mathcal{K}$. If the hyperplane misses Q_0 completely, these data values contradict (3). More important, when Δ cuts the hyperellipsoid, the common region $Q_0 \cap \Delta$ contains the acceptable values of the

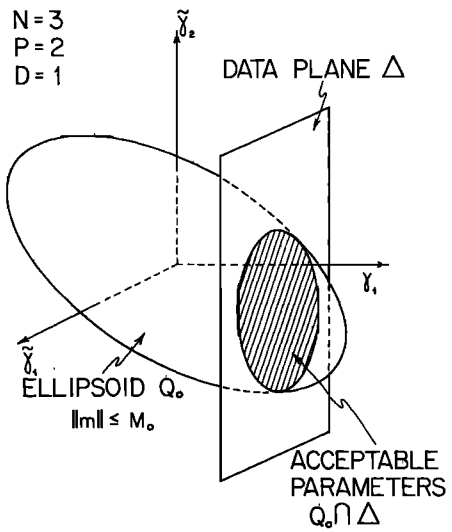

Fig. 1. Parameter-data space when $D = 1$ and $P = 2$; i.e., only one datum has been measured, but two parameters are required. With the assumption that $\|m\| \leq M_0$, all acceptable points lie inside the ellipsoid Q_0. Given a specific value for the datum, the predicted parameters both lie in the shaded ellipse $Q_0 \cap \Delta$.

unknown parameters. A hyperplane section of a hyperellipsoid is, of course, another hyperellipsoid (see Figure 1). Because the data subspace is orthogonal to the parameter subspace, it is a trivial matter to discover the equation of the parameter hyperellipsoid. It should be pointed out that the parameter ellipsoid obtained here must be the same one defined by (4) in *Backus*'s [1970*a*] paper, although the numerical correspondence is not immediately apparent. To clarify this, we partition the matrix $\hat{\mathbf{G}}$ and the vector Γ_a into parts corresponding to the parameter subspace and data subspace: then (5) is written as

$$\begin{pmatrix} \tilde{\gamma} \\ \cdots \\ \gamma \end{pmatrix}^T \begin{bmatrix} \tilde{G} & \vdots & H^T \\ \cdots & \cdots & \cdots \\ H & \vdots & G \end{bmatrix}^{-1} \begin{pmatrix} \tilde{\gamma} \\ \cdots \\ \gamma \end{pmatrix} \leq M_0^2 \tag{6}$$

where $\tilde{G}_{kl} = \langle \tilde{g}_k, \tilde{g}_l \rangle$, $H_{il} = \langle g_i, \tilde{g}_l \rangle$, $G_{ij} = \langle g_i, g_j \rangle$, and $\tilde{\gamma}$ and γ are the projections of Γ_a onto $\mathcal{P}$ and $\mathcal{D}$. When the matrix $\hat{\mathbf{G}}^{-1}$ is written out as an explicit partition also, the detailed correspondence to Backus's result emerges: these manipulations appear in Appendix B. Numerically, it is often simpler to invert $\hat{\mathbf{G}}$ and partition it afterward than to construct the coefficients of the parameter subspace quadratic form by geometric considerations, as Backus did.

If we look again at (6), we can verify that fixing γ makes (6) into an equation for a quadratic form for $\tilde{\gamma}$, exactly as we said earlier.

From a mathematical point of view the space $\mathcal{K}$ is isomorphic to the subspace $\mathcal{R} \subset \mathcal{K}$ (defined by Backus), which is the one spanned by the elements $g_1, g_2, \cdots, g_D, \tilde{g}_1, \tilde{g}_2, \cdots, \tilde{g}_P$. Any vector $g \in \mathcal{R}$ can be represented by the linear combination

$$g = \sum_{k,l} \tilde{\gamma}_k \tilde{g}^k + \gamma_l g^l$$

and it is easy to verify that the coefficients $\tilde{\gamma}_k$ and γ_l so defined are in fact the components of $\Gamma \in \mathcal{K}$. In going from $\mathcal{R}$ to $\mathcal{K}$ we emphasize the measured numbers γ_l and the quantities that we wish to know rather than their corresponding elements in $\mathcal{K}$; by making the measurements into a subspace orthogonal to that of the desired parameters we stress the fact that they are unrelated in a linear theory without some additional structure.

UNDERPARAMETERIZED LEAST SQUARES MODELS

We have covered a small part of the general theory of linear inference set out by Backus. Our particular problem is model fitting, and in this application the unknown parameters $\tilde{\gamma}_k$ assume a special role: they are coefficients in some sort of representation of m_E. The appropriate representation in a linear theory is an expansion of the form

$$m_E = \tilde{\gamma}_k \tilde{g}^k + m_* \tag{7}$$

The P elements $\tilde{g}^k \in \mathcal{K}$ are the dual basis for $\tilde{g}_k$; they are linear combinations of $\tilde{g}_k$ having the property $\langle \tilde{g}_k, \tilde{g}^l \rangle = \delta_k{}^l$. Note that the sum is over repeated indices when one is up and the other down. The element $m_* \in \mathcal{K}$ contains the information in m_E that is not in our incomplete, finite expansion. We have chosen an expansion in $\tilde{g}^k$ rather than in $\tilde{g}_k$, because when (2) is applied to (7), we find

$$\tilde{\gamma}_k = \tilde{\gamma}_k + \langle \tilde{g}_k, m_* \rangle$$

or

$$0 = \langle \tilde{g}_k, m_* \rangle \qquad k = 1, 2, \cdots, P$$

Furthermore, this implies

$$0 = \langle \tilde{g}^k, m_* \rangle \qquad k = 1, 2, \cdots, P \tag{8}$$

because the $\tilde{g}^k$ are merely linear combinations of $\tilde{g}_k$. Equation (8) assures us that no part of the expansion $\tilde{\gamma}_k \tilde{g}^k$ lies in m_*, as we require.

Because our expansion accounts for m_E in some degree, it is rather unnatural to apply a bound on $\|m_E\|$ as Backus suggests. Indeed, it has been found in several practical applications that (3) leads to extremely pessimistic limits on the uncertainties in the solution; the discouraging results of work on one case [*Parker*, 1971] led to the developments described here. It has been found that alternative smoothing assumptions give distinctly better results in most cases; the reason for this is that we chose the expansion $\tilde{\gamma}_k \tilde{g}^k$ to contain functions that are likely to represent m_E well, whereas the equivalent representation $m_0{}^{\prime\prime} = \gamma_i g^i$ in Backus's paper may be very poor. We express our prejudice (as Backus puts it) in terms of the difference between m_E and $\tilde{\gamma}_k \tilde{g}^k$. Thus we assume

$$\|m_*\|^2 \leq M_1^2 \tag{9}$$

It is intuitive that if we have what we believe is a fairly complete description of m_E in (7), we are justified in choosing a rather small value of M_1, and thereby we hope to arrive at a more confined distribution of parameters than that obtained with (3). Obviously, making M_0^2 near to $\|m_0\|^2$ (see Appendix B) has a similar effect, but as we shall see, the two assumptions are different in their essential character.

To put (9) into our geometrical picture, use (7) to find $\|m_E\|^2$:

$$\|m_E\|^2 = \tilde{\gamma}_k \tilde{\gamma}_l \langle \tilde{g}^k, \tilde{g}^l \rangle + 2\tilde{\gamma}_k \langle \tilde{g}^k, m_* \rangle + \|m_*\|$$

and in view of (8),

$$\|m_E\|^2 = \|m_*\|^2 + \tilde{\gamma}_k \tilde{\gamma}_l \langle \tilde{g}^k, \tilde{g}^l \rangle$$

so that from (9),

$$\|m\|^2 - \tilde{\gamma}_k \tilde{\gamma}_l \langle \tilde{g}^k, \tilde{g}^l \rangle \leq M_1^2 \tag{10}$$

where m is any model satisfying (9). But we already know the smallest possible value of $\|m\|^2$, since it is given by (4); and $\langle \tilde{g}^k, \tilde{g}^l \rangle$ can be written in a matrix as $\hat{G}^{-1}$ (see Appendix B). Now (10) can be recast into the following form analogous to (6):

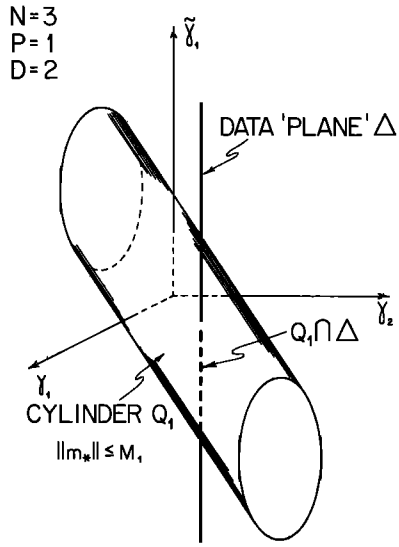

Fig. 2. Parameter-data space when $D = 2$ and $P = 1$. Here the assumption $\|m_*\| \leq M_1$ is illustrated: this requires all acceptable points to lie inside the cylinder Q_1. When particular values are given to the two data, the associated acceptable interval for the parameter is the dashed line $Q_1 \cap \Delta$.

$$\Gamma_\alpha{}^T\left\{\begin{bmatrix} \tilde{G} & H^T \\ H & G \end{bmatrix}^{-1} - \begin{bmatrix} \tilde{G}^{-1} & 0 \\ 0 & 0 \end{bmatrix}\right\}\Gamma_\alpha \leq M_1^2$$

or

$$\Gamma_\alpha{}^T \mathbf{G}_*\Gamma_\alpha \leq M_1^2 \tag{11}$$

Here the zeros represent submatrices of the appropriate size all filled with zeros. This is evidently somewhat different from (6), as we should expect; what is not apparent is whether G_* is associated with a hyperellipsoid or an unbounded quadratic region. Somewhat surprisingly, the quadratic form (11) is a cylinder: it is a degenerate hyperellipsoid with precisely P 'infinitely long' principal axes (see Figure 2).

To show this, we derive (11) in another way. We work with the element m_*, seeking to minimize $\|m_*\|^2$ directly; now the linear constraints are (8) and (1) expanded with (7):

$$\gamma_i - \tilde{\gamma}_k\langle \tilde{g}^k, g_i\rangle = \langle g_i, m_*\rangle \quad i = 1, 2, \cdots, D \tag{12}$$

When Γ is fixed, the left sides of (8) and (12) constitute a fixed vector $\Phi \in \mathcal{K}$, so that the theorem of Appendix A applies immediately:

$$\|m_*\|_{\min}{}^2 = \Phi^T\tilde{\mathbf{G}}^{-1}\Phi \tag{13}$$

where $\tilde{\mathbf{G}}$ is formed with inner products between all the g_i and $\tilde{g}^k$ exactly as $\mathbf{G}$ was found. But Φ can be written explicitly in terms of Γ; we have

$$\Phi = \mathbf{L}\Gamma = \begin{bmatrix} 0 & \vdots & 0 \\ -H\tilde{G}^{-1} & \vdots & I \end{bmatrix}\begin{pmatrix} \tilde{\gamma} \\ \cdots \\ \gamma \end{pmatrix} \tag{14}$$

where the symbol I stands for a unit matrix of the correct size (here $D \times D$). Thus after (14) is substituted into (13) and the usual arguments are gone through, we have

$$\Gamma_\alpha{}^T\{\mathbf{L}^T\tilde{\mathbf{G}}^{-1}\mathbf{L}\}\Gamma_\alpha \leq M_1^2 \tag{15}$$

From its partition in (14) it is clear that the matrix $\mathbf{L}$ has a P-dimensional subspace in which $\mathbf{L}\Gamma = 0$; that is; there is a P-fold degenerate, zero eigenvalue. Elements Γ_α lying in this null space can be arbitrarily large in magnitude without violating (15). Since $\tilde{\mathbf{G}}$ is positive definite, the remaining eigenvalues of $\mathbf{L}^T\tilde{\mathbf{G}}^{-1}\mathbf{L}$ are positive, thus proving the assertion that assumption (9) defines a cylinder Q_1 in $\mathcal{K}$. It remains to be shown that (11) and (15) are really the same: they are, but a direct algebraic verification is tedious and will not be inflicted upon the reader or the compositor.

The region of overlap of Δ, the data hyperplane, and the cylinder Q_1 is still a quadratic region containing the acceptable solutions $\tilde{\gamma}_k$. Whether this region is bounded or not depends on the rank of the upper left matrix in (11); equivalently, we can inquire whether the parameter subspace $\mathcal{P}$ intersects the null space of $\mathbf{L}$ anywhere apart from $\Gamma = 0$. Using (14), we find that this is equivalent to

$$H\tilde{G}^{-1}\gamma = 0 \qquad \tilde{\gamma} \neq 0 \tag{16}$$

where $\tilde{\gamma} \in \mathcal{P}$, and it will be recalled that $\mathbf{H}$ is a matrix of D rows and P columns. If $P > D$, it is obviously possible to satisfy (16), since there are more free components in $\tilde{\gamma}$ (there are P of them) than constraints (only D). The more usual case with $P \leq D$ is slightly more complicated: (16) can be satisfied only if $\mathbf{H}$ has less than P linearly independent rows. This is rather unlikely but is possible in principle; for example, it happens when all the g_i and $\tilde{g}_k$ are orthogonal to each other. When (16) has a solution with $P \leq D$, the conventional least squares fitting procedure also breaks down because $\mathbf{H}^T\mathbf{H}$ is singular (see the section below on comparison with least squares fitting).

The consequences of these results for practical analysis deserve discussion. As a concrete case, consider this: suppose that we wish to discover the dipole moment of the geomagnetic field from measurements of the vertical flux density on the earth's surface. The expansion $\tilde{\gamma}_k\tilde{g}^k$ in (7) is a spherical harmonic representation of the field, and the first three terms contain the information that we are seeking. If we stop at $P = 3$ in (7), we should expect $\|m_*\|^2$ to be rather large, and thus (11) may not confine the derived coefficients very well; however, if we increase P, we know that $\|m_*\|^2$ must decrease, and perhaps a more confining region for $\tilde{\gamma}_1$, $\tilde{\gamma}_2$, and $\tilde{\gamma}_3$ will ensue. The uncertainty region for these first three parameters alone is found by projecting the region $Q_1 \cap \Delta$ onto the subspace of interest, the algebraic span of the elements $\{\tilde{\gamma}_1, \tilde{\gamma}_2, \tilde{\gamma}_3\}$ (see Appendix C for details). In other words, the introduction of additional parameters, of no immediate interest in themselves, may be able to improve the accuracy of the desired quantities. This is probably correct initially, but we can see that increasing P indefinitely does not work, for when $P > D$, we know that $Q_1 \cap \Delta$ is unbounded, and it is almost certain that the uncertainty in $\tilde{\gamma}_1$, $\tilde{\gamma}_2$, $\tilde{\gamma}_3$ becomes infinite too. In practical problems, then, there is an optimum value of P, probably larger than the number of parameters of immediate concern but not more than the number of data; there appears to be no way of discovering the best value other than the obvious one of trying successively larger systems until improvement ceases.

There are other forms that the smoothing assumption can take; we shall briefly mention another. Instead of (9) let

$$\|m_*\|/\|m\| \leq M_2 \tag{17}$$

Here in place of the absolute size of $\|m_*\|$ we estimate a bound on its ratio with $\|m\|$. Clearly, $0 \leq M_2 \leq 1$. This bound is useful when we are seeking a quantity whose magnitude is unknown but whose variability can be guessed. We simply state the region confined by (17):

$$\Gamma_a{}^T\left\{\begin{bmatrix} \tilde{G} & H^T \\ H & \tilde{G} \end{bmatrix}^{-1} - \frac{1}{1-M_2{}^2}\begin{bmatrix} \tilde{G}^{-1} & 0 \\ 0 & 0 \end{bmatrix}\right\}\Gamma_a \leq 0 \qquad (18)$$

A few properties will be asserted without proof. The matrix in braces normally has some negative and positive eigenvalues: it is possible that they could all be negative, in which case (18) implies the whole of $\mathcal{K}$; however, they can never all be positive (which would, of course, isolate $\Gamma_a = 0$ as the only permitted element). Thus the region defined in (18) is always unbounded: it is a hypercone which will be called Q_2. The intersection with the data hyperplane may or may not be bounded, but whereas with (9), unbounded regions were unlikely when $P < D$, here they are not uncommon. Assumption (17) is a more delicate one to use than either (9) or (3).

UNCERTAINTY IN THE DATA

When the data are not precisely known, further uncertainty is introduced into the determined parameters. First let us consider the situation in which it is possible to define a region in $\mathfrak{D}$, the data subspace, within which the observations must lie. Now every element of Γ obeying this restriction lies inside a cylinder C in $\mathcal{K}$. We saw that the additional constraint of a smoothing assumption confines Γ to the interior and surface of a quadratic region Q_0, Q_1, Q_2, depending on the choice of criterion. Thus acceptable values Γ_a lie in the common region of C and Q_n, i.e., $\Gamma_a \in C \cap Q_n$. Unfortunately, even when C has a really simple section in $\mathfrak{D}$ (such as a hyperellipse), its intersection with Q_n is not a simple figure, and the computation of the allowed parameters by projecting $C \cap Q_n$ onto $\mathcal{P}$ will in general be exceedingly complex. This also appears to be a difficulty with statistical uncertainty in the data.

Rather than known upper and lower bounds on a measured variable we normally specify its statistical distribution or, in the case of many observations, a joint probability distribution of the data vector $\gamma \in \mathfrak{D}$. To make the situation clearer, we begin by studying the simplest possible case with $P = 1$ and $D = 1$; thus $\mathcal{K}$ is a plane, and its subspaces $\mathcal{P}$ and $\mathfrak{D}$ are the $\tilde{\gamma}_1$ and γ_1 axes. For definiteness we consider the criterion (3) so that the acceptable vectors Γ_a lie in an elliptical region Q_0, as is shown in Figure 3. For any exact value of γ_1 there would be an interval in $\tilde{\gamma}_1$; therefore a correct statistical statement about uncertainty in $\tilde{\gamma}$ must concern intervals. We ask for the probability that a particular value, say, $\tilde{\gamma}_1'$, lies inside any of the possible intervals associated with the data. The set of all data values whose intervals include $\tilde{\gamma}_1'$ can easily be seen to be the projection onto the γ_1 axis of that part of the line $\tilde{\gamma}_1 = \tilde{\gamma}_1'$ which lies inside Q_0. The required probability is then simply found by integrating the probability density function for the observed value of γ_1 over this set, which is of course a finite segment of the γ_1 axis (see Figure 3). Hence we compute directly a confidence value for a given parameter value; a 90% confidence region would be the one inside which the probability is never less than 0.9.

These ideas generalize without difficulty to $\mathcal{K}$ spaces of higher dimension. A particular parameter vector $\tilde{\gamma}' \in \mathcal{P}$ is associated with a hyperplane Π, and the intersection $Q_n \cap \Pi$ is the region containing all possible Γ associated with $\tilde{\gamma}'$, exactly as $Q_n \cap \Delta$ contained all Γ associated with a given data vector. When $Q_n \cap \Pi$ is projected onto $\mathfrak{D}$, we have the region over which the probability density function must be integrated to discover the probability that $\tilde{\gamma}'$ is associated with some $\gamma \in \mathfrak{D}$. Despite simplicity in principle, actual calculations are not easy because of the practical problems of performing integrations in spaces of high dimension. Simplifications are possible in some circumstances. For example, if the statistical contribution to the uncertainty is small and the error law Gaussian, we can give a relatively simple approximation. The part of the surface of $Q_n \cap \Pi$ nearest the center of the error distribution is treated as a hyperplane; then the pdf (probability density function) can be integrated over an entire half space, a procedure which is easily done if the variables are first transformed to an uncorrelated set with unit variance.

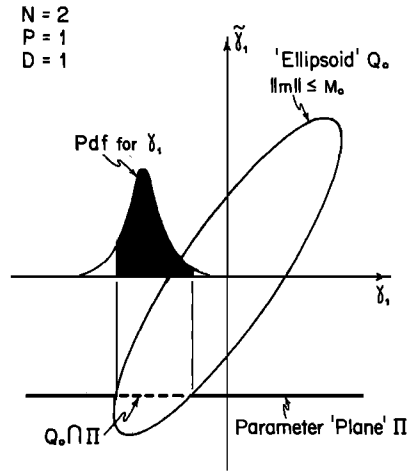

Fig. 3. Estimation of the probability that a certain parameter value is associated with data whose statistical behavior is known and with the assumption $\|m\| \leq M_0$. The set of all possible data values compatible with the value of the parameter is the dashed line $Q_0 \cap \Pi$. The probability is calculated by integrating the probability density function for γ_1 over that interval.

In the next few sections, in which numerical matters will be discussed, we shall not attempt to include the computational work of these statistical considerations because, in all but the smallest problems, they impose an impractically large, additional burden.

COMPARISON WITH LEAST SQUARES FITTING

Conventional least squares modeling depends upon the assumption that disagreement between the model predictions and actual observation arises from zero-mean, random noise; therefore the difference between prediction and observation is made as small as possible by varying the unknown parameters. The residuals (observed minus predicted values) in our linear modeling problem are given by

$$r_i = \gamma_i - \langle g_i, \tilde{g}^k \rangle \tilde{\gamma}_k \qquad i = 1, 2, \cdots, D$$

or

$$\mathbf{r} = \gamma - H\tilde{G}^{-1}\tilde{\gamma} \qquad (19)$$

where $\mathbf{r}, \gamma \in \mathfrak{D}$, $\tilde{\gamma} \in \mathcal{P}$, and H and $\tilde{G}$ are defined in (6). When the residuals are minimized in an rms sense, $\tilde{\gamma}$ is found so as to make the length $(\mathbf{r}^T\mathbf{r})^{1/2}$ smallest: the usual treatment yields

$$\tilde{\gamma}_L = \tilde{G}(H^TH)^{-1}H^T\gamma \qquad (20)$$

Whereas this solution is obtained by minimizing a distance in $\mathfrak{D}$, the center of the hyperellipse $Q_0 \cap \Delta$ is the result of minimizing a distance in $\mathcal{K}$: suppose that we ask for the parameter vector $\tilde{\gamma}_0$ associated with the model fitting the data with the smallest norm $\|m\|$. It is easy to show that (see Appendix A)

$$\tilde{\gamma}_0 = H^TG^{-1}\gamma \qquad (21)$$

This is also the expression for the position of the center of the region Q_0, derived in Appendix B (B5). There is no obvious connection between the two solutions $\tilde{\gamma}_L$ and $\tilde{\gamma}_0$, and they are usually far apart in numerical examples.

Something more interesting occurs when we seek the model with the smallest $\|m_*\|$ in (9). This solution gives rise to a parameter vector $\tilde{\gamma}_*$ that is at the center of the quadratic region $Q_1 \cap \Delta$; a little algebra gives

$$\tilde{\gamma}_* = \tilde{G}(H^T G^{-1} H)^{-1} H^T G^{-1} \gamma \tag{22}$$

The solutions $\tilde{\gamma}_L$ and $\tilde{\gamma}_*$ are more similar than a superficial comparison of (20) and (22) suggests. It is not hard to show that $\tilde{\gamma}_*$ can be obtained directly from (19) by seeking the solution that minimizes $\mathbf{r}^T G^{-1} \mathbf{r}$; since G^{-1} is positive definite, $(\mathbf{r}^T G^{-1} \mathbf{r})^{1/2}$ is a valid norm of $\mathbf{r}$, and therefore $\tilde{\gamma}_L$ and $\tilde{\gamma}_*$ are simply minimizers of two different norms of the misfit vector. In some practical calculations, $\tilde{\gamma}_L$ and $\tilde{\gamma}_*$ have been found to be surprisingly close. We can be more precise and show that

$$\|\tilde{\gamma}_* - \tilde{\gamma}_L\|_2 \leq \tfrac{1}{2}(\rho_G^{1/2} - \rho_G^{-1/2}) \cdot \|(H^T H)^{-1}\|_2^{1/2} \cdot \|\tilde{G}\|_2 \cdot \|\mathbf{r}_L\|_2 \tag{23}$$

where $\rho_G = \lambda_{\max}/\lambda_{\min}$, the ratio of the largest and smallest eigenvalues of G; $\mathbf{r}_L$ is the value of $\mathbf{r}$ in (19) when $\tilde{\gamma} = \tilde{\gamma}_L$ (the least squares solution); and the two norms of the matrices are the norms subordinate upon the Euclidean norm of the vectors [see *Wilkinson*, 1965, p. 56]. The proof of (23) is found from a consideration of the right-angled triangle $H\tilde{\gamma}_*$, $H\tilde{\gamma}_L$, and $\mathbf{r}_L$ and the calculation of the minimum $\|H(\tilde{\gamma}_* - \tilde{\gamma}_L)\|_2$. This equation shows that if the rms misfit is small, $\tilde{\gamma}_*$ and $\tilde{\gamma}_L$ are likely to be similar.

Let us now discuss the computational aspects of the solutions. First we need to calculate the necessary matrix elements associated with the various formulations. In the linear inference approach in (5) and (11) there are about $\frac{1}{2}(D + P)^2$ independent matrix elements for $\hat{\mathbf{G}}$ (recall that $\hat{\mathbf{G}}$ is symmetric); with the least squares approach, only H and $\tilde{G}$ are required and consist of DP and about $\frac{1}{2}P^2$ elements, respectively. Furthermore, the calculation of H is frequently much easier than that of the submatrix G of $\hat{\mathbf{G}}$. When $D \gg P$, a relationship which is common in least squares analysis, there are roughly $\frac{1}{2}D/P$ more calculations involved in setting up the required matrices for the linear inference method than there are in the corresponding least squares problem.

The actual calculation of (20) is not by a direct construction of $H^T H$ and its inversion. It is not widely appreciated in the geophysical community that this direct attack leads to matrices that are unnecessarily poorly conditioned and that numerically stable alternatives are available. A modern technique for solving (20) is by means of the Q-R factorization of H [*Wilkinson and Reinsch*, 1971, contribution I/8]. The number of multiplications needed here is somewhat less than DP^2. To invert $\hat{\mathbf{G}}$ for (5) or (11) requires $\frac{1}{3}(D + P)^3$ multiplications, although this figure can be reduced by a factor of 2 if $\hat{\mathbf{G}}$ is partitioned and the submatrix G factorized rather than inverted (more about this in a moment). Thus the methods of this paper are at a considerable disadvantage computationally in comparison to least squares modeling, by a factor of about $\frac{1}{6}D^2/P^2$ in computing time.

In the approach advocated in this paper, there are two special circumstances in which direct inversion of $\hat{\mathbf{G}}$ should be avoided. The first is the one in which the inner products $\langle G_i, G_j \rangle$ are calculated by numerical quadrature: this is in fact quite a common situation. Then it will be possible to write out the quadrature explicitly:

$$\langle G_i, G_j \rangle = \sum_k w_k f_i(x_k) f_j(x_k)$$

where the x_k are the sampling coordinates of the quadrature scheme and the w_k the associated weights. If the weights are positive (as they are for the more stable methods, like the Gauss-Legendre formulas), this can be rewritten as

$$\langle G_i, G_j \rangle = \sum_k [w_k^{1/2} f_i(x_k)][w_k^{1/2} f_j(x_k)]$$

or

$$\hat{\mathbf{G}} = \mathbf{K}^T \mathbf{K}$$

where

$$K_{kj} = w_k^{1/2} f_j(x_k)$$

Now the inverse of $\hat{\mathbf{G}}$ can be found rather precisely by using the singular value decomposition of $\mathbf{K}$ [*Wilkinson and Reinsch*, 1971, contribution I/10]. While this method is slower than the direct inversion of $\hat{\mathbf{G}}$, it yields much more precise answers, because the condition number of $\mathbf{K}$ is exactly equal to the square root of that of $\hat{\mathbf{G}}$.

The other situation in which inversion of $\hat{\mathbf{G}}$ is not recommended is this: it is sometimes possible to set up the problem so that the submatrix G can be inverted or factored in far fewer operations than the customary $\frac{1}{3}D^3$. Then the number of calculations for the investigation of (5) or (11) also can be reduced to this order, provided that $D \gg P$. We work with the partition (B1) of $\hat{\mathbf{G}}$ in Appendix B, which is the form originally derived geometrically by Backus. In (B1) it is the submatrix C that constitutes the major computational task, because it is of the order of $D \times D$: from (B1),

$$C = [G - H\tilde{G}^{-1}H^T]^{-1} \tag{24}$$

If $\tilde{G}$ is factored by the Cholesky decomposition [*Wilkinson and Reinsch*, 1971, contribution I/1]

$$\tilde{G} = \tilde{L}\tilde{L}^T$$

where $\tilde{L}$ is a lower-triangular matrix, then (24) can be written as

$$C = [G - \tilde{H}\tilde{H}^T]^{-1} \tag{25}$$

with $\tilde{L}\tilde{H}^T = H^T$. To create $\tilde{H}$ requires only $\frac{1}{2}P^2D$ operations plus those needed to factor $\tilde{G}$. Now (25) is ready for the application of the so-called Sherman-Morrison-Woodbury identity, which reads

$$(A + UV^T)^{-1} \equiv A^{-1} - A^{-1}U(I + V^T A^{-1} U)^{-1} V^T A^{-1}$$

From this we see

$$C = G^{-1} + G^{-1}\tilde{H}(I - \tilde{H}^T G^{-1} \tilde{H})^{-1} \tilde{H}^T G^{-1} \tag{26}$$

Here the matrix $I - \tilde{H}^T G^{-1} \tilde{H}$ is of the order of $P \times P$, so that its inversion represents only $\frac{1}{3}P^3$ operations. When G^{-1} can be obtained cheaply, (26) is obviously an efficient way to find C. Normally, the actual components of C are not required (this is untrue of the submatrix A), only such products as $C\gamma$; therefore a factored form of G^{-1} suffices rather than the explicit inverse, and that can create economies of computer time and sometimes storage.

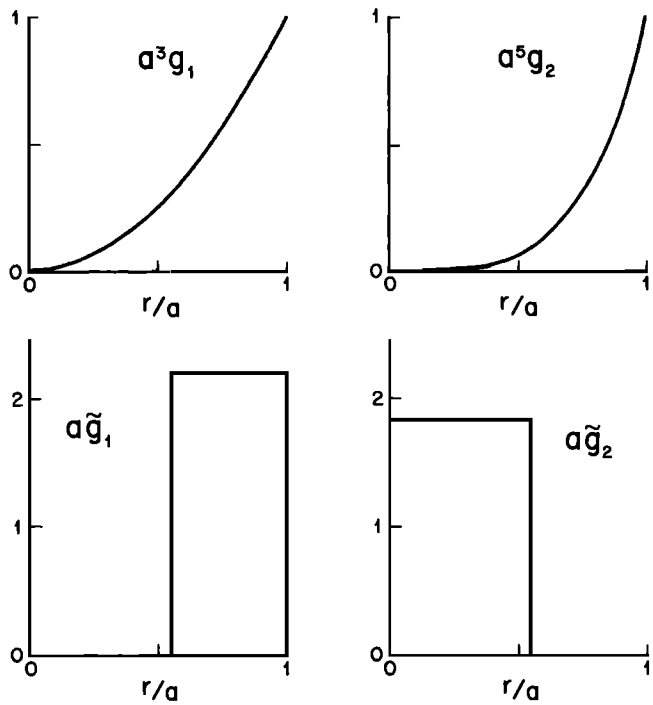

Fig. 4. The four functions g_1, g_2, $\tilde{g}_1$, and $\tilde{g}_2$. Because $\tilde{g}_1$ and $\tilde{g}_2$ are orthogonal, the dual basis elements $\tilde{g}^1$ and $\tilde{g}^2$ are simply scaled versions of $\tilde{g}_1$ and $\tilde{g}_2$.

Unless the particular form of G allows rapid inversion or factorization, the considerable computational effort involved relegates the linear inference approach to problems with relatively few data (not more than several hundred). In contrast, least squares modeling is extraordinarily cheap numerically and has been applied to very large data sets ($D > 10^6$). It is unfortunate that such an unwanted increase in computing time usually accompanies the attempt to assess the uncertainty in the solutions to underparameterized problems.

A GEOPHYSICAL ILLUSTRATION

We assume values for the earth's mass and moment of inertia, together with the radius of the core, and we ask for the mean densities of the mantle and core. The unknown model will be $\rho(r)$, the density in a spherically symmetric earth of radius a; the natural function space is $L^2(0, a)$, the space of square-integrable functions on $(0, a)$. With the inner product

$$\langle u, v\rangle = \int_0^a u(r)v(r)\,dr/a$$

this space is a well-known example of a Hilbert space. The two data are connected to the model by

$$\frac{\bar{\rho}}{3} = \gamma_1 = \int_0^a \rho(r)r^2\,dr/a^3$$

$$\tfrac{2}{3}\bar{\rho}C/M_E a^2 = \gamma_2 = \int_0^a \rho(r)r^4\,dr/a^5$$

where $\bar{\rho}$ is the mean density (5517 kg m^{-3}), C the polar moment of inertia, and M_E the mass of the earth ($C/M_E a^2 = 0.33078$). Thus in (1), $g_1 = r^2/a^2$, and $g_2 = r^4/a^4$.

We wish to predict the mean densities of the mantle and core (without the r^2 weighting in the definition of $\bar{\rho}$): for (2) we have

$$\gamma_1 = \int_b^a \rho(r)\,dr/(a-b)$$

$$\gamma_2 = \int_0^b \rho(r)\,dr/b$$

where b is the core radius. The functions g_1, g_2, $\tilde{g}_1$, and $\tilde{g}_2$ are shown in Figure 4.

Setting $a = 1$ and $b = 0.547$, we find

$$\hat{G}^{-1} = \left[\begin{array}{cc:cc} 2.2075 & 0 & 0.6154 & 0.4199 \\ & 1.828 & 0.0997 & 0.0179 \\ \hdashline & & 0.2 & 0.1429 \\ & & & 0.1111 \end{array}\right]$$

$$= \left[\begin{array}{cc:cc} 6.7037 & 1.9345 & -40.114 & 25.930 \\ & 1.2409 & -14.785 & 11.497 \\ \hdashline & & 316.35 & -252.77 \\ & & & 234.15 \end{array}\right]$$

where only the upper triangle of the symmetric array is shown. After substitution of numerical values for γ into Γ_a (we use $\gamma_1 = 1.839$ Mg m^{-3} and $\gamma_2 = 0.9125$ Mg m^{-3}) the following equation for $Q_1 \cap \Delta$ results:

$$\tilde{\gamma}^T \begin{bmatrix} 6.7037 & 1.9345 \\ 1.9345 & 1.2409 \end{bmatrix} \tilde{\gamma} + 2\tilde{\gamma}^T \begin{bmatrix} -50.11 \\ -16.70 \end{bmatrix} + 41615 \le M_0^2$$

For a fixed value of M_0 this is an elliptical region containing all possible solutions obeying (5); we show these regions for several values of M_0 in Figure 5. The smallest possible value of M_0, the one that causes the four-dimensional hyperellipsoid to touch the data hyperplane, is 5.89 Mg m^{-3}. M_0 is of course the rms density of the whole earth; using the model 1066B of

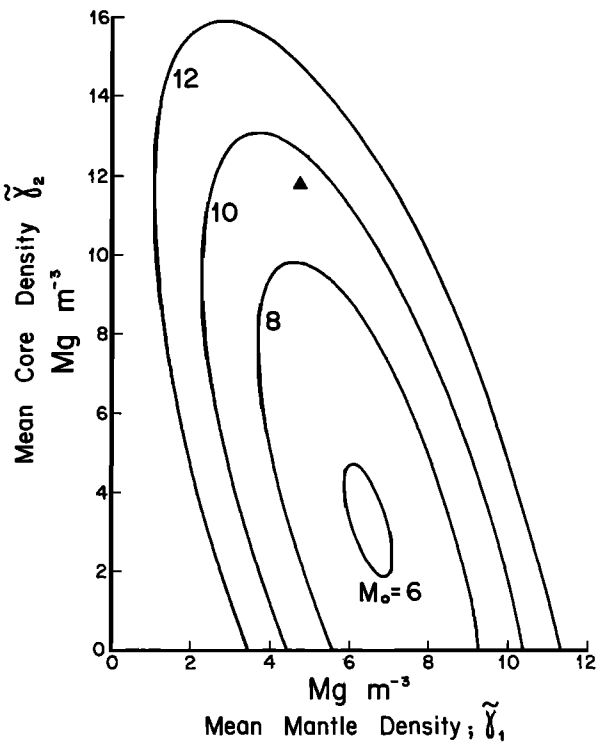

Fig. 5. Mean mantle and core densities derived only from the values of the mass and moment of inertia of the earth, together with a bound of the rms density M_0. The uncertainty regions lie inside the ellipses labeled with values of M_0.

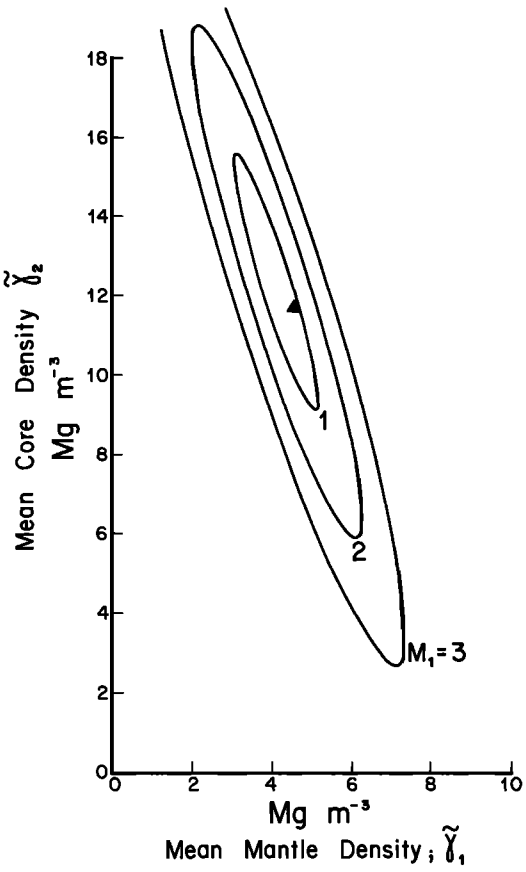

Fig. 6. Same as Figure 5, but here the rms deviation from a two-shell model is bounded above by M_1.

Gilbert and Dziewonski [1975], we find $\|\rho\| = 9.36$ Mg m^{-3}, so that with the Backus criterion the class of all possible solutions lies slightly inside the ellipse labeled 10 in Figure 5. This region contains negative densities for the core; clearly, the solutions are very poorly constrained. Good estimates of the true mean densities of the mantle and core (made from 1066B) are 4.71 and 11.9 Mg m^{-3}, and the point is plotted as a triangle in the figure.

If we carry out the equivalent calculations for (11), we arrive at Figure 6, in which the values of M_1 are varied. Suppose that we believe that the value of $\|\rho_*\|$ is less than 1 Mg m^{-3}. Then $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ are confined inside the ellipse labeled 1 in that figure; $\|\rho_*\|$ is the rms difference between the true density and a model made up of two uniform shells. For 1066B, $\|\rho_*\| = 0.86$ Mg m^{-3}, indicating that the two-shell model is in fact remarkably good and that in all probability, $\|\rho_*\| < 1$ Mg m^{-3} for the real earth. The confinement of our unknown parameters is considerably greater here than in the former case.

FOURIER SERIES

It is frequently useful in geophysical studies to obtain a Fourier series representation of a digitally sampled record; for example, it is the normal method in determining the tidal response of the earth or the amplitude of the diurnal geomagnetic fluctuation. We can apply our method to this problem, the particular interest being that the matrix $\mathbf{G}$ has an especially simple inverse if the equations are set up in a certain way. A signal $y(t)$ is recorded at specific times $0, t_1, t_2, \cdots, t_D$ during an interval $0 \le t \le T$; we shall assume that $y(t)$ is continuously differentiable. Then we may represent the data as

$$\gamma_i = y(t_i) - y(0) = \int_0^T H(t_i - t)m(t)\,dt \qquad (27)$$

$$i = 1, 2, \cdots, D$$

where $m(t)$ is the derivative of y and $H(t)$ is the Heaviside function. This representation is in fact an example of 'quelling' by integration [*Backus*, 1970*b*] of the expression

$$\gamma_i = \int_0^T \delta(t_i - t)y(t)\,dt$$

We shall assume that $m \in L^2(0, T)$ and the g_i implied by (27) are obviously also square integrable; the same inner product is taken as was used in the first example. The parameters to be determined are the coefficients in a Fourier series (which will exist because $\|m\|$ is finite); the expression for the cosine coefficient at radian frequency ω_k is

$$\tilde{\gamma}_k = \int_0^T -\frac{\sin \omega_k t}{\omega_k T} m(t)\,dt$$

there being a similar equation for the sine term. These then are the equivalents of (2).

To reach our estimate of uncertainty in the parameters, we need a smoothing assumption. Here $\|m\|$ is a rather good intuitive measure of signal roughness, and we easily see how restricting its magnitude suppresses unwanted wiggliness between the sampled data: the choice of (9) will be natural when most of the energy is confined to relatively few spectral lines, whose amplitudes are required; the statement that M_1 is small is equivalent to the more familiar idea that the digital sampling is not 'aliased.'

Our main purpose, however, is to point out that even when the samples are not equally spaced in t, the submatrix G in (6) or (11) has a very simple inverse. The elements of G are given by

$$\langle g_i, g_j \rangle = \int_0^T H(t_i - t)H(t_j - t)\,dt$$
$$= \min(t_i, t_j)$$

When the samples are arranged in natural order $0 < t_1 < t_2 \cdots t_D$, the inverse of G is tridiagonal: explicitly,

$$(G^{-1})_{i,i-1} = (t_{i-1} - t_i)^{-1} \qquad 1 < i \le D$$
$$(G^{-1})_{i,i+1} = (t_i - t_{i+1})^{-1} \qquad 1 \le i < D$$
$$(G^{-1})_{i,i} = (G^{-1})_{i,i-1}(G^{-1})_{i,i+1}(t_{i+1} - t_{i-1}) \qquad 1 < i < D$$
$$(G^{-1})_{1,1} = -(G^{-1})_{1,2}t_2/t_1$$
$$(G^{-1})_{D,D} = -(G^{-1})_{D,D-1}$$

All other elements of G^{-1} vanish.

This remarkably simple inverse allows the procedures described above in the section on comparison with least squares fitting to be brought into play. Submatrices G arising in some other problems possess tridiagonal inverses, e.g., the class of matrices

$$R_{ij} = \begin{cases} p_i q_j & 1 \le i \le j \le D \\ q_i p_j & 1 \le j < i \le D \end{cases}$$

(of which our present example is a special case). Such matrices occur when the model is represented as an expansion of orthogonal functions on a line (like Legendre polynomials) and a particular quelling is invoked to induce convergence of the series defining the elements of G.

DISCUSSION

A careful evaluation of the uncertainty in a geophysical model is something that few geophysicists undertake. Even in the class of linear problems, much current practice gives results based on unstated and unquantified assumptions. Backus's work, while very complete, has failed to enjoy much application. It is hoped that this somewhat discursive and more elementary treatment of the problem will be more easily accessible to the practicing geophysicist.

APPENDIX A

This appendix sets out the proof of the theorem used to delineate the regions associated with the various smoothing assumptions. Whereas the result can be obtained at once by calculus of variations, that method does not show us that the value obtained is indeed the smallest possible one.

Theorem. Given an element $m \in \mathcal{H}$ satisfying the N linear constraints,

$$\phi_i = \langle f_i, m\rangle \qquad i = 1, 2, \cdots, N \tag{A1}$$

where the N elements $f_i \in \mathcal{H}$ are linearly independent, the least possible value of $\langle m, m\rangle$ is given by

$$\|m_0\|^2 = \sum_{i,j} \phi_i (f^{-1})_{ij} \phi_j$$

where f^{-1} is the inverse of the matrix f whose elements are $f_{ij} = \langle f_i, f_j\rangle$.

Proof. Define an element $m_0 \in \mathcal{H}$ as

$$m_0 = \sum_{i,j} (f^{-1})_{ij} \phi_i f_i \tag{A2}$$

The matrix f^{-1} exists because of the linear independence of f_i. We verify that m_0 satisfies (A1) and achieves the value $\|m_0\|^2$. First calculate

$$\langle f_k, m_0\rangle = \left\langle f_k, \sum_{i,k} (f^{-1})_{ij} \phi_i f_j \right\rangle$$

$$= \sum_{i,k} \langle f_k, f_j\rangle (f^{-1})_{ij} \phi_i = \sum_i \delta_{ik} \phi_i = \phi_k$$

Thus m_0 satisfies (A1). Next

$$\langle m_0, m_0\rangle = \left\langle \sum_{i,j} (f^{-1})_{ij} \phi_i f_j, \sum_{l,k} (f^{-1})_{lk} \phi_l f_k \right\rangle$$

$$= \sum_{i,j,k,l} \phi_i (f^{-1})_{ij} \langle f_j, f_k\rangle (f^{-1})_{lk} \phi_l$$

$$= \sum_{i,j,l} \phi_i (f^{-1})_{ij} \delta_{jl} \phi_l = \sum_{i,l} \phi_i (f^{-1})_{ij} \phi_j$$

Now consider another element $n \in \mathcal{H}$ satisfying (A1). We compute $\|n\|^2$:

$$\|n\|^2 = \|m_0 + (n - m_0)\|^2$$

$$= \|m_0\|^2 + \|n - m_0\|^2 + 2\langle m_0, n - m_0\rangle$$

$$= \|m_0\|^2 + \|n - m_0\|^2 + 2 \sum_{j,k} \langle f_j, n - m_0\rangle (f^{-1})_{jk} \phi_k$$

But

$$\phi_i = \langle f_i, n\rangle$$

and

$$\phi_i = \langle f_i, m_0\rangle$$

Therefore by subtracting,

$$0 = \langle f_i, n - m_0\rangle$$

Thus

$$\|n\|^2 = \|m_0\|^2 + \|n - m_0\|^2 \geq \|m_0\|^2$$

APPENDIX B

Here we establish the equivalence of Backus's result (4) and our (6). In principle, there is no need to do this because from their definitions they must be identical; however, in going through a direct verification we arrive at intermediate results useful elsewhere in the paper. First we need the explicit partition of $\tilde{\mathbf{G}}^{-1}$. This is straightforward matrix algebra:

$$\begin{bmatrix} \tilde{G} & \vdots & H^T \\ \cdots & \vdots & \cdots \\ H & \vdots & G \end{bmatrix}^{-1} = \begin{bmatrix} A & \vdots & B^T \\ \cdots & \vdots & \cdots \\ B & \vdots & C \end{bmatrix} \tag{B1}$$

where

$$A = (\tilde{G} - H^T G^{-1} H)^{-1}$$

$$C = (G - H\tilde{G}^{-1} H^T)^{-1}$$

$$B = CH\tilde{G}^{-1} = -(AH^T G^{-1})^T$$

Now break (6) down into terms in A, B, and C:

$$\tilde{\gamma}^T A \tilde{\gamma} + 2\tilde{\gamma}^T B \gamma + \gamma^T C \gamma \leq M_0^2$$

or

$$(\tilde{\gamma} - \tilde{\gamma}_0)^T A (\tilde{\gamma} - \tilde{\gamma}_0) \leq M_0^2 - \gamma^T C \gamma + \tilde{\gamma}_0^T A \tilde{\gamma}_0 \tag{B2}$$

where

$$\gamma_0 = -A^{-1} B^T \gamma$$

The form of (B2) is exactly like Backus's (4):

$$(\tilde{\gamma}_k - \tilde{\gamma}_k^{\parallel})(\tilde{\gamma}_l - \tilde{\gamma}_l^{\parallel})\langle \tilde{g}_\perp^k, \tilde{g}_\perp^l\rangle \leq M^2 - \|m_0^{\parallel}\|^2$$

The notation of Backus will be used extensively here. First we need an explicit form for f^i, the elements of the dual basis to the set $\{f_1, f_2, \cdots, f_N\}$ with $f_i \in \mathcal{H}$. By definition,

$$\langle f^i, f_j\rangle = \delta_j^i$$

It will be immediately apparent that

$$f^i = \langle f^i, f^j\rangle f_j = (f^{-1})^{ij} f_j \tag{B3}$$

where f^{-1} is the inverse of the matrix f with elements $f_{ij} = \langle f_i, f_j\rangle$.

Let us first unscramble $\tilde{g}_k^{\perp}$: this denotes the part of $\tilde{g}_k$ not in the subspace spanned by $\{g_1, g_2, \cdots, g_D\}$. Therefore we have

$$\langle \tilde{g}_k^{\perp}, g_i\rangle = 0 \tag{B4}$$

Again it can be verified that the following definition fits (B4):

$$\tilde{g}_k^{\perp} = \tilde{g}_k - \langle \tilde{g}_k, g_i\rangle g^i$$

Let us now compute $\langle \tilde{g}_k^{\perp}, \tilde{g}_l^{\perp}\rangle$, since from (B3), $\langle \tilde{g}_\perp^k, \tilde{g}_\perp^l\rangle$ is the inverse matrix. By direct substitution,

$$\langle \tilde{g}_k^{\perp}, \tilde{g}_l^{\perp}\rangle = \langle \tilde{g}_k, \tilde{g}_l\rangle - 2\langle \tilde{g}_k, g_i\rangle\langle g^i, \tilde{g}_l\rangle + \langle \tilde{g}_k, g_i\rangle\langle g^i, g^j\rangle\langle g_j, \tilde{g}_l\rangle$$

Using (B3) liberally in this equation gives us

$$\langle \tilde{g}_k^{\perp}, \tilde{g}_l^{\perp}\rangle = \langle \tilde{g}_k, \tilde{g}_l\rangle - \langle g_k, g_i\rangle\langle g^i, g^j\rangle\langle g_j, \tilde{g}_l\rangle$$

and upon close scrutiny of the definitions of A, G, H, and $\tilde{G}$ we see that $\langle \tilde{g}_\perp^k, \tilde{g}_\perp^l\rangle$ is the same as A.

Next consider $\gamma_k^{\parallel}$: Backus defines this vector in two steps,

$$\gamma_k^{\parallel} = \langle \tilde{g}_k, m_0^{\parallel} \rangle$$

and

$$m_0^{\parallel} = \gamma_i g^i$$

Therefore

$$\gamma_k^{\parallel} = \langle \tilde{g}_k, g^i \rangle \gamma_i = \langle \tilde{g}_k, g_j \rangle \langle g^j, g^i \rangle \gamma_i$$

But in (B2) we have

$$\gamma_0 = -A^{-1}B^T\gamma = H^T G^{-1}\gamma \tag{B5}$$

again proving them to be the same things in different guises.

The equality of $\|m_0^{\parallel}\|^2$ with $\gamma^T C\gamma - \tilde{\gamma}^T A\gamma_0$ is more easily shown by fiddling with the matrix notation. First, however, note that

$$\|m_0\|^2 = \gamma_i \gamma_j \langle g^i, g^j \rangle$$

Directly from the definitions and (B5) we have

$$X = \gamma^T C\gamma - \tilde{\gamma}_0^T A\gamma_0 = \gamma^T [C - G^{-1}HAH^T G^{-1}]\gamma$$

But from the definition of B we have

$$AH^T G^{-1} = (CH\tilde{G}^{-1})^T = \tilde{G}^{-1}H^T C$$

Thus

$$\begin{aligned} X &= \gamma^T [I - G^{-1}HG^{-1}H^T]C\gamma \\ &= \gamma^T [I - G^{-1}(G - C^{-1})]C\gamma \\ &= \gamma^T G^{-1}\gamma = \|m_0\|^2 \end{aligned}$$

APPENDIX C

It has been suggested that we may improve the accuracy of a set of parameters by the introduction of ancillary coefficients in the expansion (7). The region in which acceptable solutions lie will be a hyperellipsoid in the P unknowns, but if we are really interested in only $P' < P$ of them, we need to obtain the boundary of the region projected onto the subspace $\mathcal{P}'$ of the interesting parameters. Let us work in terms of new variables that vanish at the center of $Q_t \cap \Delta$:

$$\sigma = \tilde{\gamma} - \tilde{\gamma}_*$$

$\tilde{\gamma}_*$ is given in (22). Then the acceptable region of solutions is the quadratic form

$$\sigma^T A\sigma \le R^2 \tag{C1}$$

where A is the $P \times P$ upper left submatrix of $\mathbf{G}_*$ in (11) and

$$R^2 = M_1^2 - \|m_*\|_{\min}^2$$

Here $\|m_*\|_{\min}^2$ is the smallest value of $\|m_*\|^2$; it is the one associated with the central model $\tilde{\gamma}_*$ and is found by putting $\tilde{\gamma}_*$ and γ into (11).

If $\tau \in \mathcal{P}$, then there is a projection matrix T such that

$$\tau = T\sigma \tag{C2}$$

where T obeys $T^2 = T^T = T$. When σ obeys (C1), τ covers the interior of the projection of $Q_t \cap \Delta$ onto $\mathcal{P}'$. Let us find the stationary values of $(\tau^T\tau)^{1/2} = |\tau|$: these will be the principal axes of the boundary of the projected region, which is itself a hyperellipsoid. Fix σ to be in the surface (force equality in (C1)), and introduce a Lagrange multiplier in the usual way:

$$|\tau|^2 = \tau^T\tau - \lambda(\sigma^T A\sigma - R^2)$$

From (C2) and the properties of the projection matrix T,

$$|\tau|^2 = \sigma^T T\sigma - \lambda(\sigma^T A\sigma - R^2)$$

Stationary values of $|\tau|^2$ occur when

$$(T - \lambda A)\sigma_0 = \mathbf{0} \tag{C3}$$

which is a two-matrix eigenvalue problem. It is now easy to show that the stationary values of $|\tau|^2$ are given by $\lambda_n R^2$, where the λ_n are the eigenvalues of (C3); the orientation information is contained in the eigenvectors of (C3).

If the P' relevant unknowns are arranged to be the first P' elements of Γ, then T is simply

$$T = \begin{bmatrix} I & \vdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \vdots & 0 \end{bmatrix}$$

where I is a $P' \times P'$ unit submatrix. When A in (C1) is positive definite, as it normally is, the numerical solution of (C3) is relatively simple [see *Wilkinson*, 1965, p. 34]. It is perhaps obvious that all but P' of the eigenvalues of (C3) are exactly zero. The case in which $P' = 1$ is very simple: the projection of the uncertainty region onto the $\tilde{\gamma}_1$ axis is given by

$$\Delta\tilde{\gamma}_1 = R(\sum a_{1m}^2/\lambda_m)^{1/2}$$

where λ_m is the mth eigenvalue of A and a_{1m} is component 1 of the eigenvector associated with λ_m.

Acknowledgments. This work was carried out in part at the Australian National University while the author held a Guggenheim Fellowship. The author wishes to thank the John Simon Guggenheim Memorial Foundation for making this opportunity available and the Research School of Earth Sciences at the Australian National University for making their facilities available. The work was completed under National Science Foundation grant EAR 75-2618, 'Studies of Geophysical Inference Problems: Theory and Practice.'

REFERENCES

Backus, G. E., Inference from inadequate and inaccurate data, I, *Proc. Nat. Acad. Sci. U.S.*, *65*, 1–7, 1970*a*.

Backus, G. E., Inference from inadequate and inaccurate data, II, *Proc. Nat. Acad. Sci. U.S.*, *65*, 281–287, 1970*b*.

Backus, G. E., Inference from inadequate and inaccurate data, III, *Proc. Nat. Acad. Sci. U.S.*, *67*, 282–289, 1970*c*.

Backus, G. E., Inference from inadequate and inaccurate data, in *Mathematical Problems in the Geophysical Sciences*, American Mathematical Society, Providence, R. I., 1972.

Backus, G. E., and F. Gilbert, Uniqueness in the inversion of inaccurate gross earth data, *Phil. Trans. Roy. Soc. London, Ser. A*, *266*, 123–192, 1970.

Franklin, J. N., Well-posed stochastic extension of ill-posed problems, *J. Math. Anal. Appl.*, *31*, 682–716, 1970.

Gilbert, F., and A. M. Dziewonski, An application of normal mode theory to the retrieval of structural parameters and source mechanisms from seismic spectra, *Phil. Trans. Roy. Soc. London, Ser. A*, *278*, 187–269, 1975.

Jackson, D. D., Interpretation of inaccurate, insufficient and inconsistent data, *Geophys. J. Roy. Astron. Soc.*, *28*, 97–110, 1972.

Mathews, J., and R. L. Walker, *Mathematical Methods of Physics*, W. A. Benjamin, New York, 1964.

Parker, R. L., The determination of seamount magnetism, *Geophys. J. Roy. Astron. Soc.*, *24*, 321–324, 1971.

Parker, R. L., Inverse theory with grossly inadequate data, *Geophys. J. Roy. Astron. Soc.*, *29*, 123–138, 1972.

Parker, R. L., The theory of ideal bodies for gravity interpretation, *Geophys. J. Roy. Astron. Soc.*, *42*, 315–334, 1975.

Parker, R. L., Understanding inverse theory, *Annu. Rev. Earth Planet. Sci.*, *5*, 35–64, 1977.

Rao, C. R., *Linear Statistical Inference and Its Applications*, John Wiley, New York, 1974.
Sabatier, P. C., Positivity constraints in linear inverse problems, *Geophys. J. Roy. Astron. Soc.*, *48*, 415–442, 1977.
Wiggins, R. A., The general linear inverse problem: Implication of surface waves and free oscillations for earth structure, *Rev. Geophys. Space Phys.*, *10*, 251–285, 1972.
Wilkinson, J. H., *The Algebraic Eigenvalue Problem*, Oxford at the Clarendon Press, London, 1965.
Wilkinson, J. H., and C. Reinsch, *Handbook for Automatic Computation*, vol. 2, *Linear Algebra*, Springer, New York, 1971.

(Received November 1, 1976;
accepted July 22, 1977.)