Proceedings of the
NATIONAL ACADEMY OF SCIENCES

Volume 65 • Number 1 • January 15, 1970

Inference from Inadequate and Inaccurate Data, I*

George Backus

UNIVERSITY OF CALIFORNIA (LA JOLLA)

Communicated November 7, 1969

Abstract. Having measured D numerical properties of a physical object E which requires many more than D parameters for its complete specification, an observer seeks to estimate P other numerical properties of E. This paper describes how he can proceed when E is adequately described by one member m_E of a Hilbert space $\mathfrak{M}$ of possible models of E, when he believes that the Hilbert norm of m_E is very likely rather smaller than some known number M, and (except for section 6) when all the observed and sought-after properties of E are continuous linear functionals on $\mathfrak{M}$. Section 6 treats Frechet-differentiable nonlinear functionals. A later paper will reduce unbounded functionals on arbitrary topological linear spaces to the present case.

---

1. **Introduction.** An observer believes that, as far as his experiments are concerned, a certain physical object E can be represented by one unknown model m_E in a manifold $\mathfrak{M}$ of possible models m. (For example, $\mathfrak{M}$ might be the set of all real-valued functions square-integrable on the unit interval.) The observer measures D numerical properties of E, the outcome of the measurements being D real numbers γ_i together with an estimate of the joint statistical distribution of the experimental errors $\delta\gamma_i$. From these D data the observer wants to predict the values $\tilde{\gamma}_1, \ldots, \tilde{\gamma}_P$ of P more numerical properties of E and to estimate the errors of his predictions. If dim $\mathfrak{M}$, the dimension of $\mathfrak{M}$, is D or less, the problem can be treated by least squares and presents only computational difficulties. If dim $\mathfrak{M} > D$ there is a conceptual difficulty as well.

The example which stimulated the present paper is a geophysical inverse problem.[1, 2] The object E is the earth. The measured data are the total mass and moment of the earth, the travel times of seismic waves from certain sources to certain receivers, and, for certain seismograms of finite length, those Fourier components which are above the noise level. Such a seismogram yields the periods of a finite number of the earth's elastic-gravitational normal modes of oscillation[3] and the amplitudes and phases to which those modes were excited by the source. If the amplitudes and phases are ignored, then the space $\mathfrak{M}$ of relevant models m consists of all ordered triples $m = (\rho,\kappa,\mu)$ of real valued func-

tions defined and square-integrable inside the earth. Here $\rho(r)$, $\kappa(r)$, and $\mu(r)$ are the density, bulk modulus, and shear modulus at the position r. The properties $\tilde{\gamma}_1, \ldots, \tilde{\gamma}_P$ which the observer wants to predict from the data might be the periods of P not-yet-observed normal modes, or the value of the density or seismic velocity at P different locations in the earth. For example, detection of a low-velocity zone requires values of the velocity at three depths. The manifold just described reflects the observer's expectation that other properties of E, such as anisotropy, will influence neither his data $\gamma_1, \ldots, \gamma_D$ nor his predictions $\tilde{\gamma}_1, \ldots, \tilde{\gamma}_P$. If he finds that no member of $\mathfrak{M}$ will explain his data, the observer may have to admit (among other things) anisotropy[4] in his models, which will then include not two, but 21 elastic coefficients as functions of position in the earth.

In the geophysical example, $\mathfrak{M}$ is infinite-dimensional, while *all* the experimental information about the earth at any one epoch will consist of a finite number of data. This infinite insufficiency of data occurs in other problems of scientific inference, including the interpretation of electron scattering by a nucleus and of spectral emission by the top of a star's atmosphere.

The unknown model in $\mathfrak{M}$ which represents E will be called m_E. It is assumed that if the observer knew m_E he could correctly calculate the outcomes of all the $D + P$ measurements in which he is interested. His methods of calculation are $D + P$ rules which assign real numbers $g_1(m), \ldots, g_D(m), \tilde{g}_1(m), \ldots, \tilde{g}_P(m)$ to every model m in some open subset of $\mathfrak{M}$. These rules are functionals (real-valued functions) on $\mathfrak{M}$. His data are summarized by the statement that to a certain accuracy m_E satisfies the D equations $g_i(m_E) = \gamma_i$, $i = 1, \ldots, D$. From the measured data γ_i and the known functionals $g_i: \mathfrak{M} \to R$ (R is the real line) he wants to learn enough about the unknown m_E to estimate $\tilde{g}_k(m_E)$, $k = 1, \ldots, P$.

In this paper I assume that $\mathfrak{M}$ is a real Hilbert space; I write the inner product of two models m and m' in $\mathfrak{M}$ as $\langle m,m' \rangle$ and the norm $\langle m,m \rangle^{1/2}$ as $\|m\|$. I also assume that the known functionals $g_i: \mathfrak{M} \to R$ and $\tilde{g}_k: \mathfrak{M} \to R$ are linear and continuous. Then there exist members g_i' and $\tilde{g}_k'$ of $\mathfrak{M}$ such that for any m in $\mathfrak{M}$, $g_i(m) = \langle g_i', m \rangle$ and $\tilde{g}_k(m) = \langle \tilde{g}_k', m \rangle$. Henceforth I write g_i' and $\tilde{g}_k'$ without the primes. No generality is lost by assuming that the set $(g_1, \ldots, g_D, \tilde{g}_1, \ldots, \tilde{g}_P)$ is linearly independent. If $(\tilde{g}_1, \ldots, \tilde{g}_D)$ is linearly dependent, some data are redundant or internally inconsistent. If $(g_1, \ldots, g_D)$ is linearly independent but $(g_1, \ldots, g_D, \tilde{g}_1, \ldots, \tilde{g}_P)$ is not, then one of the $\tilde{g}_k$ is a linear combination of its predecessors in the $(D + P)$-tuple $(g_1, \ldots, g_D, \tilde{g}_1, \ldots, \tilde{g}_P)$, and finding $\tilde{g}_k(m_E)$ is trivial if $\tilde{g}_l(m_E)$ can be found for $l < k$. Therefore the set $(\tilde{g}_1, \ldots, \tilde{g}_P)$ can be winnowed until the residue $(g_1, \ldots, g_D, \tilde{g}_{l_1}, \ldots, \tilde{g}_{l_Q})$ is linearly independent and all the discarded $\tilde{g}_k$'s are linear combinations of it.

2. **Notation.** If $\mathfrak{V}$ is any set of vectors in $\mathfrak{M}$, asp$\mathfrak{V}$ (the algebraic span of $\mathfrak{V}$) is the subspace of $\mathfrak{M}$ consisting of all finite linear combinations of vectors in $\mathfrak{V}$. The subspace $\text{asp}(g_1, \ldots, g_D, \tilde{g}_1, \ldots, \tilde{g}_P)$ will be denoted by $\mathfrak{R}$ (for relevant), while $\text{asp}(g_1, \ldots, g_D)$ will be denoted by $\mathfrak{D}$ (for data), and the orthogonal complement of $\mathfrak{D}$ in $\mathfrak{R}$ will be denoted by $\mathfrak{P}$ (for prediction). Thus $\mathfrak{R} = \mathfrak{D} \oplus \mathfrak{P}$. An arbitrary vector m in $\mathfrak{M}$ will be written $m = m^{\|} + m^{\perp} + m^{?}$ where $m^{\|}$ is in $\mathfrak{D}$,

$m^\perp$ is in $\mathfrak{P}$, and $m^?$ is orthogonal to $\mathfrak{R}$. The operators of orthogonal projection from $\mathfrak{M}$ onto $\mathfrak{D}$, $\mathfrak{P}$, and $\mathfrak{R}$ will be written $\mathfrak{D}$, $\mathcal{P}$, and $\mathcal{R}$. Thus $\mathfrak{D}m = m^\parallel$ and $\mathcal{P}m = m^\perp$.

If f and g are any vectors in $\mathfrak{M}$, I denote their tensor product by fg. Thus fg is the bilinear functional on $\mathfrak{M}$ which assigns to any ordered pair (m_1,m_2) of vectors in $\mathfrak{M}$ the real number $\langle f, m_1\rangle \langle g, m_2\rangle$. Then fg can be interpreted as the linear operator on $\mathfrak{M}$ which assigns to an arbitrary vector m in $\mathfrak{M}$ the vector $f\langle g, m\rangle$. Thus, by definition, $fg(m) = f\langle g, \ \rangle$.

An expression a_ib^i or a^ib_i means a sum over all possible values of the repeated index i, say $i = 1, \ldots, N$, but a_ib_i or $a_ib^ic_i$ denotes only a single term, not a sum, and $a_ib_i = c_i$ is interpreted as the N equations $a_1b_1 = c_1, \ldots, a_Nb_N = c_N$.

Given an ordered n-tuple of linearly independent vectors $(f_1, \ldots, f_n)$ in $\mathfrak{M}$, their "dual basis" is the unique ordered n-tuple $(f^1, \ldots, f^n)$ in $\mathcal{S} = \text{asp}(f_1, \ldots, f_n)$ such that $\langle f_i,f^j\rangle = \delta_i^j (= 1$ when $i = j$ and 0 when $i \neq j)$. If m is in $\mathfrak{M}$, then $f_i\langle f^i,m\rangle = f^i\langle f_i,m\rangle$ is $\mathcal{S}m$, the orthogonal projection of m onto $\mathcal{S}$, so the orthogonal projection operator $\mathcal{S}$ can be written $\mathcal{S} = f_if^i = f^if_i$. On $\mathcal{S}$, $\mathcal{S}$ is the identity operator. One easy way to compute f^i is as follows. Define the second order tensor and linear operator $\mathfrak{F} = \sum_{i=1}^{n} f_if_i$. Then $\mathfrak{F}(f^i) = f_i$, so $f^i = \mathfrak{F}^{-1}(f_i)$ if $\mathfrak{F}^{-1}$ refers to the inverse of $\mathfrak{F}$ regarded as an operator on $\mathcal{S}$ only. Since $\mathfrak{F}$ is symmetric and positive definite, $\mathcal{S}$ has an orthonormal basis $\hat{e}_1, \ldots, \hat{e}_n$ consisting of eingenvectors of $\mathfrak{F}$ with positive eingenvalues $\phi_1, \ldots, \phi_n$. Then $\mathfrak{F} = \sum_{i=1}^{n} \phi_i\hat{e}_i\hat{e}_i$ so $\mathfrak{F}^{-1} = \sum_{i=1}^{n} \phi_i^{-1}\hat{e}_i\hat{e}_i$.

3. **Perfectly Accurate Data.** If $\gamma_1, \ldots, \gamma_D$ are measured with perfect accuracy then m_E must satisfy

$$\langle g_i,m\rangle = \gamma_i \quad i = 1, \ldots, D. \tag{1}$$

Let $(g^1, \ldots, g^D)$ be the dual basis for $(g_1, \ldots, g_D)$. If m is in $\mathfrak{M}$ then $m^\parallel = \mathfrak{D}m = g^i\langle g_i,m\rangle$. If m also satisfies equation (1), then $m^\parallel = m_0{}^\parallel$ where, by definition,

$$m_0{}^\parallel = \gamma_ig^i. \tag{2}$$

Thus the fact that m_E satisfies the D equations (1) is equivalent to the single statement $m_E{}^\parallel = m_0{}^\parallel$. Write $\tilde{g}_j = \tilde{g}_j{}^\parallel + \tilde{g}_j{}^\perp$. Then $\tilde{\gamma}_j = \tilde{\gamma}_j{}^\parallel + \tilde{\gamma}_j{}^\perp$ where $\tilde{\gamma}_j{}^\parallel = \langle \tilde{g}_j{}^\parallel,m_E\rangle = \langle \tilde{g}_j{}^\parallel,m_E{}^\parallel\rangle = \langle \tilde{g}_j,m_E{}^\parallel\rangle = \langle \tilde{g}_j,m_0{}^\parallel\rangle$ and $\tilde{\gamma}_j{}^\perp = \langle \tilde{g}_j{}^\perp,m_E\rangle = \langle \tilde{g}_j{}^\perp,m_E{}^\perp\rangle = \langle \tilde{g}_j,m_E{}^\perp\rangle$. The quantity $\tilde{\gamma}_j{}^\parallel$ can be determined from the data via equation (2), but the data contain no information whatever about $m_E{}^\perp$ or $\tilde{\gamma}_j{}^\perp$.

If the observer knows nothing about m_E except equation (1) he can make no predictions about $\langle \tilde{g}_j,m_E\rangle$. In most physical problems however, interest usually centers in a *bounded* region of the model space $\mathfrak{M}$. For example, most geophysicists will be surprised if the earth's central density turns out to be more than 30 gm/cm^3 or if its root-mean-square density turns out to be more than 10 gm/cm^3. A rigorous, nontrivial deduction about the earth's internal density structure which required the assumption that the root-mean-square density was

less than 100 gm/cm^3 would not be absolutely conclusive, but would convince most geophysicists and be interesting to the remainder.

Suppose there is a positive number M such that the observer is willing to hypothecate *a priori* that the unknown norm $||m_E||$ satisfies

$$||m|| \leq M. \tag{3}$$

The question then is whether there are values of M large enough to make equation (3) a convincing hypothesis and yet small enough to provide a useful adjunct to equation (1) in estimating $\tilde{\gamma}_j$. If we assume equation (3), then $\tilde{\gamma}_j{}^\perp$ is bounded by $||\tilde{g}_j{}^\perp||\ M$; and, roughly speakng, $\tilde{\gamma}_j{}^\parallel$ is of the order of $||\tilde{g}_j{}^\parallel||\ ||m_o{}^\parallel||$ in the absence of cancellation. If $||\tilde{g}_j{}^\perp||\ M \ll ||\tilde{g}_j{}^\parallel||\ ||m_o{}^\parallel||$ then equation (3) is useful. Thus a range of M both useful and plausible usually exists if $||\tilde{g}_j{}^\perp|| \ll ||\tilde{g}_j{}^\parallel||$.

To be more precise, let $(\tilde{g}^1{}_\perp, \ldots, \tilde{g}^P{}_\perp)$ be the dual basis to $(\tilde{g}_1{}^\perp, \ldots, \tilde{g}_P{}^\perp)$. Then $m_E{}^\perp = \tilde{g}^k{}_\perp \langle \tilde{g}_k{}^\perp, m_E{}^\perp \rangle = \tilde{\gamma}_k{}^\perp \tilde{g}^k{}_\perp$, so $||m_E{}^\perp||^2 = \tilde{\gamma}_k{}^\perp \tilde{\gamma}_l{}^\perp \langle \tilde{g}^k{}_\perp, \tilde{g}^l{}_\perp \rangle$. Now equation (3) implies $||m_E{}^\perp||^2 \leq M^2 - ||m_0{}^\parallel||^2$ since $m_E = m_o{}^\parallel + m_E{}^\perp + m_E{}^?$. Therefore, equation (3) implies $\tilde{\gamma}_k{}^\perp \tilde{\gamma}_l{}^\perp \langle \tilde{g}^k{}_\perp, \tilde{g}^l{}_\perp \rangle \leq M^2 - ||m_o{}^\parallel||^2$, or

$$(\tilde{\gamma}_k - \tilde{\gamma}_k{}^\parallel)\ (\tilde{\gamma}_l - \tilde{\gamma}_l{}^\parallel)\ \langle \tilde{g}^k{}_\perp, \tilde{g}^l{}_\perp \rangle \leq M^2 - ||m_o{}^\parallel||^2. \tag{4}$$

All the quantities in equation (4) except $\tilde{\gamma}_1, \ldots, \tilde{\gamma}_P$ and M can be computed from the data, so equation (4) gives for each M an ellipsoid to which the data confine the predicted P-tuple $(\tilde{\gamma}_1, \ldots, \tilde{\gamma}_P)$. A conservative observer will pick $M \gg ||m_o{}^\parallel||$, while a bold one may pick $M/||m_o{}^\parallel|| \approx 2$ or 3. An observer who finds he chose $M < ||m_o{}^\parallel||$ will revise his preconceptions about reasonable bounds for $||m_E||$.

4. **Inaccurate Data with Known Variance, P = 1.** Suppose that the expected values of the unknown experimental errors $\delta\gamma_i$ are known to be zero. Then $\langle g_i, m_E \rangle = \gamma_i + \delta\gamma_i$ and γ_i is an unbiased estimate of $\langle g_i, m_E \rangle$. Suppose also that by repeated measurements the observer has estimated the $D \times D$ variance matrix V whose ij'th entry V_{ij} is the expected value of the product $\delta\gamma_i \delta\gamma_j$. Suppose finally that the observer wants to make only one prediction, $\langle \tilde{g}, m_E \rangle$.

For any $\tilde{h}^\parallel$ in $\mathfrak{D}$ there are constants a^i such that $\tilde{g}^\parallel + \tilde{h}^\parallel = a^i g_i$, and moreover $\langle \tilde{g}^\parallel + \tilde{h}^\parallel, m_E \rangle = a^i \langle g_i, m_E \rangle = a^i \gamma_i + a^i \delta\gamma_i$. Write $\tilde{g} = \tilde{g}^\parallel + \tilde{h}^\parallel + (\tilde{g}^\perp - \tilde{h}^\parallel)$ so that $\tilde{\gamma} = \langle \tilde{g}, m_E \rangle = a^i \gamma_i + a^i \delta\gamma_i + \langle \tilde{g}^\perp - \tilde{h}^\parallel, m_E \rangle$. If the observer estimates $\tilde{\gamma}$ as $a^i \gamma_i$, he will commit an error $a^i \delta\gamma_i + \langle \tilde{g}^\perp - \tilde{h}^\parallel, m_E \rangle$. The expected value of the square of this error is $a^i a^j V_{ij} + \langle \tilde{g}^\perp - \tilde{h}^\parallel, m_E \rangle^2$. Granting equation (3) for m_E, this mean squared error can be no larger than $a^i a^j V_{ij} + ||\tilde{g}^\perp - \tilde{h}^\parallel||^2 M^2$, but it can be that large if m_E happens to be $\tilde{g}^\perp - \tilde{h}^\parallel$. Therefore the observer may regard $\epsilon^2(a^1, \ldots, a^D) = a^i a^j V_{ij} + M^2(||\tilde{g}^\perp||^2 + ||\tilde{h}^\parallel||^2)$ as the largest error he is likely to commit when he estimates $\tilde{\gamma}$ as $a^i \gamma_i$. This largest error is an inhomogeneous quadratic polynomial in $(a^1, \ldots, a^D)$, and evidently the observer should choose those constants so as to minimize that polynomial. Only when $V_{ij} = 0$ will he choose the constants so that $\tilde{h}^\parallel = 0$.

Sometimes a decrease in M small enough to leave equation (3) plausible will decrease min $\epsilon^2(a^1, \ldots, a^D)$ from a uselessly large to a usefully small value, permitting choices of a^i which make $a^i \gamma_i$ a very close estimate of $\tilde{\gamma}$. I know no general way to discover when this will happen except numerical calculation of

min $\epsilon^2(a^1, \ldots, a^D)$ for a range of values of M decreasing toward implausibility in equation (3). In 1968, examples[2] of such calculations in geophysical inverse problems were carried out, motivated by an argument different from equation (3) which involved the appearance of a Lagrange multiplier λ as a free parameter. It is clear now that $\lambda = M^2$.

When $P > 1$ and nothing is known about the distribution of the errors $\delta\gamma_i$ except their means and their variance matrix, the foregoing discussion can be applied to each $\tilde{\gamma}_k$ separately. This makes inefficient use of the data, but I have discovered no *linear* computation which leads to a more confining joint distribution for the ordered P-tuple $(\tilde{\gamma}_1, \ldots, \tilde{\gamma}_P)$. However, if $(\delta\gamma_1, \ldots, \delta\gamma_D)$ has a joint normal distribution, there are arguments leading to a joint normal distribution for $(\tilde{\gamma}_1, \ldots, \tilde{\gamma}_P)$.

5. **Normally Distributed Errors in the Data.** Suppose the errors $(\delta\gamma_1, \ldots \delta\gamma_D)$ have a joint normal distribution with mean $(0, \ldots, 0)$ and $D \times D$ variance matrix V. Let $A = V^{-1}$ and write the components of the "accuracy matrix" A as A^{ij}. Then one school of probabilists and most experimentalists are willing to regard the data as giving a probability density p_{data} for the location of $m_E^{\parallel}$ in $\mathfrak{D}$, namely $p_{\text{data}}(m^{\parallel}) = (\text{const.}) \exp(-\frac{1}{2}\langle m^{\parallel} - m_o^{\parallel}, g_i\rangle A^{ij} \langle g_j, m^{\parallel} - m_o^{\parallel}\rangle)$. Define the symmetric operator $\mathfrak{A}$: $\mathfrak{M} \to \mathfrak{M}$ by requiring that $\mathfrak{A}(m) = g_i A^{ij} \langle g_j, m\rangle$. As a second-order tensor, $\mathfrak{A} = A^{ij} g_i g_j$. Then

$$-2 \ln p_{\text{data}}(m^{\parallel}) = (\text{const.}) + \langle m^{\parallel} - m_o^{\parallel}, \mathfrak{A}(m^{\parallel} - m_o^{\parallel})\rangle. \tag{5}$$

If the observer regards the data as imposing a probability distribution on $m_E^{\parallel}$, then it seems appropriate for him to formulate in probabilistic terms the *a priori* judgment that m_E satisfies equation (3). One way to smear equation (3) into an infinite-dimensional analog of a probability distribution[5] is to demand that the orthogonal projection of m onto any finite-dimensional subspace $\mathfrak{S}$ of $\mathfrak{M}$ be a joint normal distribution with mean 0 and variance tensor $M^2\mathcal{S}$, $\mathcal{S}$ being the operator of orthogonal projection from $\mathfrak{M}$ onto $\mathfrak{S}$. Thus the probability density of $\mathcal{S}(m)$ in $\mathfrak{S}$ is p_{prej} where

$$-2 \ln p_{\text{prej}}(\mathcal{S}(m)) = (\text{const.}) + M^{-2}\|\mathcal{S}(m)\|^2. \tag{6}$$

The next question is how to combine the data of equation (5) with the prejudice, equation (6). The prejudice alone would give $-2 \ln p_{\text{prej}}(m^{\parallel}) = \text{const.} + M^{-2}\|m^{\parallel}\|^2$ on $\mathfrak{D}$. If both prejudice and data are available, it seems reasonable to suggest that together they assign to $m_E^{\parallel}$ the probability density $p_{\parallel}$ on $\mathfrak{D}$, where

$$-2 \ln p_{\parallel}(m^{\parallel}) = (\text{const.}) + M^{-2}\|m^{\parallel}\|^2 + \langle m^{\parallel} - m_o^{\parallel}, \mathfrak{A}(m^{\parallel} - m_o^{\parallel})\rangle. \tag{7}$$

If the concentration ellipsoid obtained from the data alone, $\langle m^{\parallel} - m_o^{\parallel}, \mathfrak{A}(m^{\parallel} - m_o^{\parallel})\rangle = 1$, has a principal axis very much longer than M, then in equation (7) the prejudice, equation (3), rather than the data will limit the component of m_E in the direction of that axis.

The centroid of equation (7) is at

$$m_C^{\parallel} = \mathfrak{A}(M^{-2}\mathfrak{D} + \mathfrak{A})^{-1} m_o^{\parallel} \tag{8}$$

where $(M^{-2}\mathfrak{D} + \mathfrak{A})^{-1}$ refers to the inverse of $M^{-2}\mathfrak{D} + \mathfrak{A}$ regarded as an operator on $\mathfrak{D}$ only, where it is, of course, strictly positive definite and hence invertible. The observer's best estimate of $m_E{}^{\|}$, based on both his data of equation (1) and his prejudice of equation (3), is $m_E{}^{\|} = m_C{}^{\|}$. Notice that if $M = \infty$, $m_C{}^{\|} = m_o{}^{\|}$, but that if $M < \infty$ then $\|m_C{}^{\|}\| < \|m_o{}^{\|}\|$. The prejudice, equation (3), leads the observer to assign to $m_E{}^{\|}$ a probability distribution concentrated slightly closer to the origin than the distribution equation (5) obtained from the data alone.

The final question is how to obtain from the data of equation (5) and the prejudice of equation (3) a probability density for $m_E{}^{\|} + m_E{}^{\perp}$ in $\mathfrak{R}$. A conservative probabilist might argue for equation (6) when $\mathfrak{S} = \mathfrak{P}$, but the deterministic form, equation (3), of the prejudice implies $\|m_E{}^{\perp}\|^2 \leq M^2 - \|m_E{}^{\|}\|^2$, and $m_E{}^{\|}$ is estimated from the data as $m_C{}^{\|}$ in equation (8). This suggests that rather than equation (6) the observer can take for the probability density of $m_E{}^{\|}$ on $\mathfrak{P}$ the slightly more concentrated function (const.) $\exp[-\|m^{\perp}\|^2/2(M^2 - \|m_C{}^{\|}\|^2)]$. Since there seems to be no other way in which the data can influence the distribution of $m_E{}^{\perp}$, the author proposes that the observer should regard the data of equation (1) and his prejudice of equation (3) together as establishing for $m_E{}^{\|} + m_E{}^{\perp}$ a joint normal distribution in $\mathfrak{R}$ with probability density $p(\mathfrak{R}m)$ given by

$$-2 \ln p(\mathfrak{R}m) = (\text{const.}) + \|m^{\perp}\|^2(M^2 - \|m_C{}^{\|}\|^2)^{-1} + \|m^{\|}\|^2 M^{-2} + \langle m^{\|} - m_o{}^{\|}, \mathfrak{A}(m^{\|} - m_o{}')\rangle. \quad (9)$$

The centroid of the distribution of equation (9) on $\mathfrak{R}$ is $m_C{}^{\|}$ and its variance tensor is $\mathfrak{B}^{-1}$ where the operator $\mathfrak{B}$: $\mathfrak{R} \to \mathfrak{R}$ is $\mathfrak{B} = \mathcal{P}(M^2 - \|m_C{}^{\|}\|^2)^{-1} + M^{-2}\mathfrak{D} + \mathfrak{A}$. If again $(M^{-2}\mathfrak{D} + \mathfrak{A})^{-1}$ refers to the inverse of $M^{-2}\mathfrak{D} + \mathfrak{A}$ regarded as an operator on $\mathfrak{D}$, then

$$\mathfrak{B}^{-1} = \mathcal{P}(M^2 - \|m_C{}^{\|}\|^2) + (M^{-2}\mathfrak{D} + \mathfrak{A})^{-1}\mathfrak{D}. \quad (10)$$

The best estimate for $m_E{}^{\|} + m_E{}^{\perp}$ in $\mathfrak{R}$ is $m_C{}^{\|}$, and the ellipsoid of concentration of $m_E{}^{\|} + m_E{}^{\perp}$ is $\langle m - m_C{}^{\|}, \mathfrak{B}(m - m_C{}^{\|})\rangle \leq 1$.

The P numbers $\tilde{\gamma}_k$ are values $\langle \tilde{g}_k, m\rangle$ of linear functionals on $\mathfrak{R}$, i.e., linear random variables, so the ordered P-tuple $(\tilde{\gamma}_1, \ldots, \tilde{\gamma}_P)$ has a joint normal distribution in R^P (the space of all ordered P-tuples of real numbers) with centroid $(\tilde{\gamma}_1{}^C, \ldots, \tilde{\gamma}_P{}^C)$ and $P \times P$ variance matrix $\tilde{V}$ where

$$\tilde{\gamma}_k{}^C = \langle \tilde{g}_k, m_C{}^{\|}\rangle = \langle \tilde{g}_k{}^{\|}, m_C{}^{\|}\rangle = \langle (M^{-2}\mathfrak{D} + \mathfrak{A})^{-1}\tilde{g}_k{}^{\|}, \mathfrak{A}\, m_o{}^{\|}\rangle = \gamma_i A^{ij}\langle g_j, (M^{-2}\mathfrak{D} + \mathfrak{A})^{-1}\tilde{g}_k{}^{\|}\rangle \quad (11)$$

and

$$\tilde{V}_{kl} = (M^2 - \|m_C{}^{\|}\|^2)\langle \tilde{g}_k{}^{\perp}, \tilde{g}_l{}^{\perp}\rangle + \langle \tilde{g}_k{}^{\|}, (M^{-2}\mathfrak{D} + \mathfrak{A})^{-1}\tilde{g}_l{}^{\|}\rangle. \quad (12)$$

When the errors in the data are very small, the concentration ellipsoid for the joint normal distribution of $(\tilde{\gamma}_1, \ldots, \tilde{\gamma}_P)$ reduces to the confinement ellipsoid of equation (4). To see this, let ϵ^2 be the smallest eigenvalue of the data error variance matrix V, and write $V = \epsilon^2 V_1, A = \epsilon^{-2}A_1$, $\mathfrak{A} = \epsilon^{-2}\mathfrak{A}_1$. Then hold V_1 and hence A_1 and $\mathfrak{A}_1$ fixed and let ϵ approach 0. From equation (8), $m_C{}^{\|}$ will

approach $m_o^{\|}$, so from equation (11) $\tilde{\gamma}_k^C$ will approach $\tilde{\gamma}_k^{\|}$. From equation (12) $\tilde{V}_{kl}$ will approach $(M^2 - \|m_o^{\|}\|^2)\langle \tilde{g}_k^{\perp}, \tilde{g}_l^{\perp}\rangle$, so $\tilde{V}^{-1}$ will have components $(\tilde{V}^{-1})^{kl}$ which approach $(M^2 - \|m_o^{\|}\|^2)^{-1}\langle \tilde{g}^k{}_{\perp}, \tilde{g}^l{}_{\perp}\rangle$. The concentration ellipsoid for $(\tilde{\gamma}_1, \ldots, \tilde{\gamma}_P)$ is $(\tilde{\gamma}_k - \tilde{\gamma}_k^C)(\tilde{\gamma}_l - \tilde{\gamma}_l^C)(\tilde{V}^{-1})^{kl} \leq 1$, which evidently approaches the confinement ellipsoid of equation (4) appropriate to error free data.

As in section 4, a small decrease in M can sometimes markedly shrink the ellipsoid of concentration for $\tilde{\gamma}_1, \ldots, \tilde{\gamma}_P$. If P is small, this can be discovered by examining the details of the concentration ellipsoid as M decreases toward implausibility in equation (3). If P is large, the observer may prefer to introduce a single figure of merit describing the size of the error ellipsoid. For example, he could choose a positive-definite symmetric $P \times P$ matrix C^{kl} such that he regarded any error P-tuples $(\delta\tilde{\gamma}_1, \ldots, \delta\tilde{\gamma}_P)$ with the same value of $C^{kl}\delta\tilde{\gamma}_k\delta\tilde{\gamma}_l$ as equally serious errors. Then he would call the expected value of $C^{kl}\delta\tilde{\gamma}_k\delta\tilde{\gamma}_l$, "the" error in $(\tilde{\gamma}_1, \ldots, \tilde{\gamma}_P)$, and would study this single error figure, $C^{kl}\tilde{V}_{kl}$, as a function of M, using equation (12).

6. **Nonlinear Functionals.** Suppose now that some of $g_1, \ldots, g_D, \tilde{g}_1, \ldots, \tilde{g}_P$ are nonlinear functionals on the Hilbert space $\mathfrak{M}$. Suppose that somehow[6] the observer has found a model m_o which satisfied $g_i(m_o) = \gamma_i$, $i = 1, \ldots, D$. Suppose also that $g_1, \ldots, g_D, \tilde{g}_1, \ldots, \tilde{g}_P$ are Frechet differentiable at m_o, so that there exist vectors $h_1, \ldots, h_D, \tilde{h}_1, \ldots, \tilde{h}_P$ in $\mathfrak{M}$ and functionals r_i and $\tilde{r}_k$ on $\mathfrak{M}$ such that $\|r_i(m - m_o)\|$ and $\|\tilde{r}_k(m - m_o)\|$ approach 0 as $\|m - m_o\|$ does so, and $g_i(m) = g_i(m_o) + \langle h_i, m - m_o\rangle + \|m - m_o\| r_i(m - m_o)$, with P similar equations for $\tilde{g}_k$.

Let $\mathfrak{R} = \text{asp}(h_1, \ldots, h_D, \tilde{h}_1, \ldots, \tilde{h}_P)$ and call a model m "$\mathfrak{R}$-close to m_o" if $|r_i(m - m_o)| \ll \|h_i\|$, $i = 1, \ldots, D$ and $|\tilde{r}_k(m - m_o)| \ll \|\tilde{h}_k\|, k = 1, \ldots, P$. Then sections 3, 4, and 5 can be applied to the linear functionals h_i, $\tilde{h}_k$. The observed and predicted "data" in this case are all zero, so the result will be an ellipsoid of confinement or concentration for the P-tuple of random variables $\langle \tilde{h}_k, m - m_o\rangle$. If m_E and all the models in this ellipsoid are $\mathfrak{R}$-close to m_o, then the ellipsoid is an approximate confinement or concentration ellipsoid for $(\tilde{\gamma}_1 - g_1(m_o), \ldots, \tilde{\gamma}_P - \tilde{g}_P(m_o))$. However, the system $g_i(m) = \gamma_i$ is now nonlinear, and m_o may not be the root m_E that we want. At present the author knows of no general way to discover whether it is (Monte Carlo methods are not generally adequate[2]). In some problems of geophysical interest, Gerver[7] has shown that if the data functionals $g_1, \ldots, g_D$ are properly chosen then the nonlinear system $g_i(m) = \gamma_i$ admits only solutions $\mathfrak{R}$-close to a given solution.

* Supported in part by the Geophysics section of the National Science Foundation.

[1] Backus, G., and F. Gilbert, *Geophys. J. Roy. Astr. Soc.*, **16**, 169 (1968).

[2] Backus, G., and F. Gilbert, *Phil. Trans. Roy. Soc. London A*, in press.

[3] Benioff, H., F. Press, and S. Smith, *J. Geophys. Res.*, **66**, 605 (1961); Ness, N. F., J. C. Harrison, and L. B. Slichter, *J. Geophys. Res.*, **66**, 621 (1961); Alsop, L. E., G. H. Sutton, and M. Ewing, *J. Geophys. Res.*, **66**, 631 (1961).

[4] Morris, G., R. Raitt, and G. Shor, *J. Geophys. Res.*, **74**, 4300 (1969).

[5] Getoor, R. K., *Pacific J. Math.*, **7**, 885 (1957).

[6] Backus, G., and F. Gilbert, *Geophys. J. Roy. Astr. Soc.*, **13**, 247 (1967).

[7] Gerver, M., Report to the Sixth Symposium of the Working Group on Theory and Computers of the Upper Mantle Committee of the IUGG (abstract to appear in *Geophys. J. Roy. Astr. Soc.*).