Geophysical Journal (1988) **92,** 125–142

Bayesian inference in geomagnetism

George E. Backus

Institute of Geophysics and Planetary Physics, University of California, San Diego, A-025, *La Jolla,* CA 92093, *USA*

Accepted 1987 July 1. Received 1987 June 23; in original form 1987 March 2

SUMMARY

In the existence half of a geophysical inverse problem (finding a model to fit the data), any method of regularization is acceptable, and the damping parameter λ should be made as large as still permits a reasonable model to fit the data adequately. In the uniqueness half of the inverse problem (finding all reasonable models that fit the data) two common methods for regularizing are stochastic inversion (SI) and Bayesian inference (BI). In both methods λ is determined by the observer's prior beliefs. If the errors and the prior model distribution are both gaussian, SI and BI lead to the same calculations, but are interpreted differently. In Gubbins and Bloxham's (G & B's) recent use of surface and satellite magnetic data to find the radial magnetic field at the core–mantle boundary (CMB), their choice of BI seems appropriate. However, their method of choosing λ is suited to the existence problem rather than the uniqueness problem and overestimates the resolution which the data provide on the CMB. As a prior belief, the heat flow bound at the CMB would call for a λ 6000 times smaller than G & B's smallest λ. How this change would affect G & B's conclusions cannot be ascertained without repeating their calculations with the smaller λ, but recent work by Shure, Parker & Langel (1985) suggests that the data cannot determine the Gauss coefficients of the core for degrees above 10.

Key words: Bayesian inference, core field, Gauss coefficient errors, geomagnetic downward extrapolation, stochastic inversion

1 INTRODUCTION

Every geophysical inverse problem has two halves: finding at least one plausible model which fits the data (the existence problem), and finding all plausible models which fit the data (the uniqueness problem). If no physically reasonable model fits the data within experimental error, the theory must be revised. If many reasonable models fit the data, then only the properties common to all these models are certain to be properties of the Earth.

Inverse theory supposes that the direct problem has been solved. The direct theory has produced a function F which tells the observer what data $\mathbf{y}$ he will obtain if the correct model for the Earth is $\mathbf{x}$. If the errors $\boldsymbol{\varepsilon}$ of observation are included, the data actually observed are

$$\mathbf{y} = F(\mathbf{x}) + \boldsymbol{\varepsilon}. \tag{1.1}$$

In the inverse problem the observer tries to estimate $\mathbf{x}$ from (1.1), given F, $\mathbf{y}$ and the statistics of $\boldsymbol{\varepsilon}$. Usually the inverse function F^{-1} fails to exist or is discontinuous, making it impossible to solve (1.1) for $\mathbf{x}$ directly or to interpret the solution. This difficulty is usually met by 'regularizing' F, that is, by finding a one-parameter family of functions F_λ such that for $\lambda > 0$ the function F_λ has a continuous inverse, while for all sufficiently small λ the values of $F_\lambda(\mathbf{x})$ and $F(\mathbf{x})$ differ by less than the expected errors of measurement. The parameter λ is usually called the damping parameter.

In the existence half of the inverse problem, regularization can be introduced *ad hoc* as a way of stabilizing ill-conditioned numerical calculations and producing one model which does fit the data (Tikhonov 1963; Backus & Gilbert 1967; Tikhonov & Arsenin 1972; Shure, Parker & Backus 1982). For this purpose, the functional form of F_λ and the value of λ are irrelevant, and can be chosen for convenience. Usually λ is taken to be the largest value which permits an acceptable fit to the data, since as λ decreases the calculations usually become more ill-conditioned (the modulus of continuity of F_λ^{-1} increases).

Regularization enters the uniqueness problem as a way to introduce into the data analysis the observer's prior beliefs about which earth model $\mathbf{x}$ is likely to be correct (Franklin 1970; Backus 1970b, 1971; Jackson 1979). Those beliefs determine the functional form of F_λ and the appropriate value for λ. To justify to himself his choice of λ, the observer must privately understand and examine his prior opinions, what he believed about the correct model $\mathbf{x}$ before he obtained the data $\mathbf{y}$. To convince others that his choice of λ is correct the observer must publicly describe and defend the prior beliefs on which he bases that choice of λ.

The present paper compares two regularization techniques widely used in inverse problems, namely stochastic inversion (SI) and Bayesian inference (BI). The general aim is not to obtain the formal results, which are well-known, but to explicate the foundations of the two techniques, so as to make clear when each is appropriate, and to contribute to the discussion of how to quantify prior beliefs so as to make them useful in practical calculations.

The present paper also has the specific aim of discussing the regularizations in the recent very important work of Gubbins and Bloxham (G & B) on using surface and satellite measurements of the geomagnetic field **B** to estimate the radial component, B_r, at the core–mantle boundary (CMB) (Gubbins 1983, 1984; Gubbins & Bloxham 1985; Bloxham & Gubbins 1986; Bloxham 1986). G & B conclude that the data indicate that frozen flux and geostrophy are poor approximations in the fluid upper core and that B_r on the CMB shows conspicuous small-scale features whose locations have remained fixed relative to the mantle for 250 yr (Bloxham 1986), perhaps serving as markers for CMB topography, mantle thermal structure, or the geometry of the core dynamo. If these conclusions are sustained, they will profoundly affect studies of main-field geomagnetism.

The analysis reported here suggests that the prior beliefs imposed by G & B on B_r at the CMB may not be shared by many workers in geomagnetism. Workers who do not share those beliefs will want to adopt a very much smaller value of the damping parameter λ than that used by G & B. They will regard G & B's published work as showing that G & B's conclusions are permitted by the data but not required by it; that G & B have solved the existence problem but not the uniqueness problem.

G & B use intensity data from satellites and directional data from old explorers, so their F in (1.1) is non-linear, but in solving the uniqueness problem they find a local solution and linearize F by Frechet differentiation. The problems posed by a non-linear F are technical rather than conceptual, so we will restrict our attention to linear Fs. The plan of this paper is to describe geometrically the pathologies in F which call for regularization, and which make the use of prior beliefs indispensable in the uniqueness half of almost every geophysical inverse problem. Next we compare SI and BI as techniques for invoking prior beliefs to deal with those pathologies, and conclude with G & B (1985) that BI is more appropriate to their problem than SI. Then we discuss how to implement prior beliefs in BI. Finally we use this implementation to examine the prior beliefs adopted by G & B. We conclude that if they were to use a damping parameter smaller than their smallest by a factor of about 6000 then they would have a solution of the uniqueness problem for B_r on the CMB as well as of the existence problem. Whether such a large decrease in G & B's damping parameter will seriously alter their conclusions cannot be ascertained until they repeat their calculation, but published evidence suggests that some of the conclusions might be questioned.

2 THE GEOMETRY OF NON-UNIQUENESS

Our data are real numbers, $y_1, \ldots, y_d$, with errors $\varepsilon_1, \ldots, \varepsilon_d$, and the finiteness of human resources assures that $d < \infty$. Most Earth models require for their complete specification an infinite sequence of real parameters, $x_1, x_2, x_3, \ldots$. In the geomagnetic work of G & B, the data are cartesian components or other magnetic elements of **B** measured at finitely many places on and above the Earth's surface. The model parameters are the internal Gauss coefficients g_l^m at the CMB, since magnetic sources outside the core are being treated as errors of measurement. The column vectors $\mathbf{x} = (x_1, x_2, x_3, \ldots)^T$, $\mathbf{y} = (y_1, \ldots, y_d)^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_d)^T$ are the model vector, the data vector, and the error vector; T means matrix transpose. The linearized form of (1.1) is

$$\mathbf{y} - \boldsymbol{\eta} = \mathbf{F}(\mathbf{x} - \boldsymbol{\xi}) + \boldsymbol{\varepsilon} \tag{2.1}$$

where $\boldsymbol{\xi}$ is a known model vector, $\mathbf{F}$ is the known $d \times \infty$ matrix of partial derivatives of $F(\mathbf{x})$ at $\mathbf{x} = \boldsymbol{\xi}$, $\boldsymbol{\eta} = F(\boldsymbol{\xi})$, and the juxtaposition $\mathbf{F}(\mathbf{x} - \boldsymbol{\xi})$ denotes matrix multiplication. The linear space of all $d \times 1$ column vectors will be denoted by D and called the data space. Both $\mathbf{y}$ and $\boldsymbol{\varepsilon}$ belong to D. The linear space of all $\infty \times 1$ column vectors $\mathbf{x}$ for which $\mathbf{Fx}$ converges will be denoted by M and called the model space. The correct Earth model is in M if F is linear or if the correct model is close to $\boldsymbol{\xi}$.

Let $\langle \varepsilon_i \rangle$ denote the statistical expected value of the random variable ε_i, and let $\langle \boldsymbol{\varepsilon} \rangle$ be the column vector whose ith component is $\langle \varepsilon_i \rangle$. A non-zero $\langle \boldsymbol{\varepsilon} \rangle$ can be absorbed into $\boldsymbol{\eta}$ in (2.1), so no generality is lost by assuming that

$$\langle \boldsymbol{\varepsilon} \rangle = \mathbf{0}.$$

Then $\mathbf{V}_\varepsilon$, the $d \times d$ variance matrix of $\boldsymbol{\varepsilon}$, is given by

$$\mathbf{V}_\varepsilon = \langle \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^T \rangle.$$

The ij entry of $\mathbf{V}_\varepsilon$ is $\langle \varepsilon_i \varepsilon_j \rangle$.

Computers are finite, so we can never compute all the model parameters x_i of an infinite model vector $\mathbf{x}$. We must restrict our computations to truncated vectors

$$\mathbf{x}_m = (x_1, \ldots, x_m)^T$$

in a truncated model space M_m of dimension m. This is not a difficulty of principle when, as in the geomagnetic case, we can choose m so large that we are sure that the x_i with $i > m$ are too small to affect the data by more than the error of measurement. Our confidence in such an m for the geomagnetic case comes from two bounds on the Gauss coefficients g_l^m at the CMB, both of the form

$$\sum_{l=1}^{\infty} C_l \sum_{m=-l}^{l} |g_l^m|^2 < Q \tag{2.2}$$

with suitably chosen positive coefficients C_l and Q. One such bound comes from the fact that the rest mass of the energy in the geomagnetic field can be no more than the mass of the Earth, 6×10^{24} kg. The other bound comes from an estimate of the electrical conductivity κ in the core, and our belief that the total ohmic heating rate in the core is probably less than the geothermal heat flow at the Earth's surface, 3×10^{13} W. Given any positive error ε in measuring **B** on and above the surface of the Earth, and any inequality (2.2), Appendix A provides an $l^*(\varepsilon)$ such that the total effect of all Gauss coefficients g_l^m with $l > l^*(\varepsilon)$ is less than ε at every measuring station, even if the g_l^m add co-operatively rather than randomly. For example, Appendix A shows that if $\varepsilon = 1$ nT (somewhat less than the MAGSAT errors) then the mass bound (2.2) permits $l^* = 70$ and $m = l^*(l^* + 2) = 5040$, while the heat bound (2.2) permits $l^* = 31$ and $m = 1023$. If $\varepsilon = 2 \times 10^{-6}$ nT (one flux quantum through a superconducting loop of area $1m^2$) then the mass bound permits $l^* = 92$ and $m = 8648$, while the heat bound permits $l^* = 52$ and $m = 2808$. In fitting the data vector $\mathbf{y}$ with a model vector $\mathbf{x}$, we can safely set $g_l^m = 0$ for $l > l^*$. If $l > l^*$, changing g_l^m from 0 to any physically acceptable value will change the data by less than the errors of measurement.

In real data analysis, the foregoing values of m are uncomfortably large. In practice, the procedure is to find the smallest m such that M_m contains a model $\mathbf{x}$ which fits the data acceptably, and then to repeat the computations for slightly larger m and see whether the results depend on m. Unless we increase m to the values justified by (2.2), this is a logically unsatisfactory situation in the uniqueness problem. Fortunately, increases in computer power seem likely to make the values of m obtained from (2.2) accessible in the near future, and, indeed, the heat flow values of l^* can already be used if we have enough money to pay for the computer time.

In any case, we must work with a finite-dimensional truncated model space M_m, and we must replace (2.1) with

$$\mathbf{y} - \boldsymbol{\eta} = \mathbf{F}_m(\mathbf{x} - \boldsymbol{\xi})_m \tag{2.3}$$

where $\mathbf{F}_m$ is the $d \times m$ matrix consisting of the first m columns of $\mathbf{F}$. This fact leads some workers to argue that we need never consider the true infinite-dimensional model space M. For the existence problem, these workers are right. Indeed, economy dictates choosing the smallest m for which we can find an $\mathbf{x}_m$ that fits the data acceptably. For the uniqueness problem, the situation is different. If we are interested in finding *all* models which fit the data, we must begin with the true M and use (2.2) to justify a choice of M_m. Then we know that we have no information except (2.2) about the $\mathbf{x}$ with $i > m$. But there may be a further non-uniqueness within M_m, especially if $m > d$, but occasionally even when $m \le d$. To discuss this question calls for careful notation.

We write the space spanned by the columns of $\mathbf{F}^T$ as $M_\mathbf{F}$, and the orthogonal complement of $M_\mathbf{F}$ as $M_\mathbf{F}^\perp$. Thus $M_\mathbf{F}^\perp$ is the linear subspace of all model vectors $\mathbf{x}_\mathbf{F}^\perp$ such that $\mathbf{F}\mathbf{x}_\mathbf{F}^\perp = \mathbf{0}$; $M_\mathbf{F}^\perp$ is also the linear subspace of all model vectors $\mathbf{x}_\mathbf{F}^\perp$ such that $\mathbf{x}_\mathbf{F}^T\mathbf{x}_\mathbf{F}^\perp = \mathbf{0}$ for every $\mathbf{x}_\mathbf{F}$ in $M_\mathbf{F}$. For every model vector $\mathbf{x}$ in M, there are unique vectors $\mathbf{x}_\mathbf{F}$ and $\mathbf{x}_\mathbf{F}^\perp$ such that $\mathbf{x}_\mathbf{F}$ is in $M_\mathbf{F}$, $\mathbf{x}_\mathbf{F}^\perp$ is in $M_\mathbf{F}^\perp$, and

$$\mathbf{x} = \mathbf{x}_\mathbf{F} + \mathbf{x}_\mathbf{F}^\perp. \tag{2.4}$$

Since $\dim M_\mathbf{F}^\perp \ge \dim M - \dim D$, if $\dim M > \dim D$ then $M_\mathbf{F}^\perp$ will contain non-zero vectors. They are the source of the non-uniqueness.

On D we introduce the inner product (Halmos 1950)

$$(\mathbf{y}, \mathbf{y}')_\varepsilon = \mathbf{y}^T\mathbf{V}_\varepsilon^{-1}\mathbf{y}' \tag{2.5}$$

and the corresponding length

$$\|\mathbf{y}\|_\varepsilon = (\mathbf{y}, \mathbf{y})_\varepsilon^{1/2}. \tag{2.6}$$

The scheme (2.6) amounts to measuring data vectors in units equal to the error of measurement, and permits combining measurements with different physical dimensions, such as angles and magnetic intensities. If the data $\mathbf{y}$ are used to estimate v model parameters (v is the emp of (5.8)) and $v \ll d$, then the expected value of $\|\boldsymbol{\varepsilon}\|_\varepsilon^2$ will be $d - v$ (Theil 1963; G & B 1985), so a crude way to state the implications for $\mathbf{x}$ of measuring a particular $\mathbf{y}$ in (2.1) is that $\mathbf{x}$ satisfies

$$\|\mathbf{y} - \boldsymbol{\eta} - \mathbf{F}(\mathbf{x} - \boldsymbol{\xi})\|_\varepsilon^2 \le d - v. \tag{2.7}$$

If $\mathbf{x}$ is any model which satisfies (2.7), and $\mathbf{x}_\mathbf{F}^\perp$ is any model in $M_\mathbf{F}^\perp$, and α is any real number, then $\mathbf{x} + \alpha\mathbf{x}_\mathbf{F}^\perp$ is another model which satisfies (2.7). The set of models $\mathbf{x}$ which satisfy (2.7) is a cylinder in M which extends to infinity in both directions along every non-zero member of $M_\mathbf{F}^\perp$. All the models in this infinitely long cylinder satisfy the data within experimental error.

Suppose g is a continuous real-valued linear function on M, and that we would like to know $g(\mathbf{x})$ for the true model $\mathbf{x}$. What can the data tell us? For such a g there is a unique model vector $\mathbf{g}$ in M such that, for all $\mathbf{x}$ in M, $g(\mathbf{x}) = \mathbf{g}^T\mathbf{x}$. Then there are unique vectors $\mathbf{g}_\mathbf{F}$ in $M_\mathbf{F}$ and $\mathbf{g}_\mathbf{F}^\perp$ in $M_\mathbf{F}^\perp$ such that

$$\mathbf{g} = \mathbf{g}_\mathbf{F} + \mathbf{g}_\mathbf{F}^\perp. \tag{2.8}$$

The expressions (2.4) and (2.8) imply

$$\mathbf{g}^T\mathbf{x} = \mathbf{g}_\mathbf{F}^T\mathbf{x}_\mathbf{F} + (\mathbf{g}_\mathbf{F}^\perp)^T\mathbf{x}_\mathbf{F}^\perp \tag{2.9}$$

and $\mathbf{x}_\mathbf{F}^\perp$ is completely unconstrained by the data vector $\mathbf{y}$. It is possible to take $\mathbf{x}_\mathbf{F}^\perp = \alpha\mathbf{g}_\mathbf{F}^\perp$ for any real α whatever, and if $\mathbf{g}_\mathbf{F}^\perp \ne \mathbf{0}$ then varying α allows $\mathbf{g}^T\mathbf{x}$ to take any value; $\mathbf{g}^T\mathbf{x}$ is completely unconstrained by the data if $\mathbf{g}_\mathbf{F}^\perp \ne \mathbf{0}$.

We will reject many of the models in the infinite cylinder (2.7) as 'unreasonable' because we believe we know more about the Earth than that it satisfies the data $\mathbf{y}$. The goal of the uniqueness half of inverse theory is to quantify our prior beliefs and to use them in the inversion. The effect of *a priori* beliefs concerning the true $\mathbf{x}$ is to truncate the infinite cylinder (2.7). The two simplest kinds of *a priori* beliefs are linear and quadratic. Linear beliefs take the form

$$a < \mathbf{c}^T\mathbf{x} < b \tag{2.10}$$

where a and b are known real numbers and $\mathbf{c}$ is a known model vector. A limiting case of (2.10) is $0 < \mathbf{c}^T\mathbf{x} < \infty$. This can arise, for example, if $x_1, x_2, \ldots,$ are the coefficients in a cubic spline fit to the density and $\mathbf{c}^T\mathbf{x}$ is the value of density at one location.

Quadratic beliefs take a form generalized from (2.2);

$$(\mathbf{x} - \boldsymbol{\mu})^T\mathbf{C}(\mathbf{x} - \boldsymbol{\mu}) \le Q \tag{2.11}$$

where Q is real and positive, $\boldsymbol{\mu}$ is a model, $\mathbf{C}$ is an $m \times m$ symmetric positive semi-definite matrix, and Q, $\boldsymbol{\mu}$ and $\mathbf{C}$ are known. An example of (2.11) in geomagnetism is the belief that the rate of ohmic heat production in the core is less than the rate of heat flow out of the Earth's surface. If the maximum electrical conductivity of the core is $3 \times 10^5\ \mathrm{S\,m^{-1}}$ then this comparison implies (Parker 1972; Gubbins 1975)

$$\sum_{l=1}^{\infty} \frac{(l+1)(2l+1)(2l+3)}{l} \sum_{m=-l}^{l} |g_l^m|^2 < 3 \times 10^{17}\ \mathrm{nT}^2 \tag{2.12}$$

where the g_l^m are the Schmidt quasi-normalized Gauss coefficients for B_r at the CMB. Another geomagnetic example of (2.11) is

$$\sum_{l=1}^{\infty} \frac{(l+1)}{(2l+1)} \sum_{m=-l}^{l} |g_l^m|^2 < 2 \times 10^{33}\ \mathrm{nT}^2,$$

which follows from the demand that the rest mass of the energy in the geomagnetic field be less than the mass of the Earth.

If $\mathbf{C}^{-1}$ exists, then for every $\mathbf{g}$ in M, (2.7) and (2.11) will confine $\mathbf{g}^T\mathbf{x}$ to a finite interval; (2.7) and (2.10) will do so only if there are enough linear bounds (2.10) that their $\mathbf{c}$'s span $M_{\mathbf{F}}^{\perp}$. Whether the bounds obtained on $\mathbf{g}^T\mathbf{x}$ from (2.7) and (2.10) or (2.11) are tight enough to be useful can be decided only by calculation with the given $\mathbf{F}, \mathbf{g}$ and bounds (Backus 1970a; Parker 1977). The present paper focuses on quadratic bounds because of their occurrence in geomagnetism and because when a and b are both finite the linear bound (2.10) can be written as a quadratic bound (2.11), with $\mathbf{C} = \mathbf{c}\mathbf{c}^T$, $Q = \frac{1}{4}(b-a)^2$, and $\boldsymbol{\mu}$ any solution of $\mathbf{c}^T\boldsymbol{\mu} = \frac{1}{2}(a+b)$.

If $\dim M \leq \dim D$ and $M_{\mathbf{F}}^{\perp}$ contains only the zero vector, then the set of $\mathbf{x}$ in M which satisfy (2.7) is a finite ellipsoid. This may not eliminate the non-uniqueness problem, because $\mathbf{F}$ is often very ill-conditioned (the largest eigenvalue of $\mathbf{F}^T\mathbf{F}$ is very much larger than the smallest). Then the ellipsoid (2.7) will be very long in some directions, and including prior beliefs like (2.10) or (2.11) in the data analysis may provide tight bounds on $\mathbf{g}^T\mathbf{x}$ where (2.7) alone does not.

3 A COMPARISON OF STOCHASTIC INVERSION AND BAYESIAN INFERENCE

In Jackson's (1979) graphic language, the bounds (2.7), (2.10) and (2.11) are 'hard'. All of them overstate our certainty. For example, we do not really believe the claim (2.7) that the error vector $\boldsymbol{\varepsilon}$ lies inside the sphere in D whose radius is $(d-v)^{1/2}$. What we know about $\boldsymbol{\varepsilon}$ is its statistical distribution in data space, which makes (2.7) probable but not certain. The statistical distribution of $\boldsymbol{\varepsilon}$ is the 'soft' analogue of (2.7) (Jackson 1979).

Similarly, it is difficult to defend the precise value $Q = 3 \times 10^{17}\,\mathrm{nT}^2$ in (2.12). The electrical conductivity of the core is poorly known, the ohmic heat production in the core may not be in steady equilibrium with mantle heat flow, and some surface heat flow is supplied by radioactivity in the mantle. Why not replace the 'hard' bound (2.10) or (2.11) by a 'soft' bound, a probabilistic description of where we think the true model $\mathbf{x}$ is located in the model space M? For example, (2.10) could be 'softened' to a probability distribution for $\mathbf{c}^T\mathbf{x}$, with mean $\frac{1}{2}(a+b)$ and standard deviation $\frac{1}{2}(b-a)$. This makes no sense for the bounds $0 < \mathbf{c}\cdot\mathbf{x} < \infty$, but those bounds usually come from such trustworthy physics (e.g. positive densities) that we are willing to believe their hard forms. The hard quadratic bound (2.11) could be replaced by a probability distribution on M whose mean is $\boldsymbol{\mu}$ and whose variance matrix is $Qm^{-1}\mathbf{C}^{-1}$. There are problems about how well these soft bounds represent the original hard bounds, but we defer those to the next section. The question pursued here is, what does such 'soft' information about a model really mean? How can we justify using a probability distribution to describe the location of the one true model $\mathbf{x}$ in the model space M? How do we work with such distributions, and how do we defend them?

3.1 Stochastic inversion (SI) and Bayesian inference (BI)

SI and BI answer these questions in different ways. Both procedures assume that the probability distribution of the error vector $\boldsymbol{\varepsilon}$ in the data space D is known, except perhaps for a few parameters to be estimated from the data vector $\mathbf{y}$ (see equation 21 of G & B 1985).

3.1.1 Stochastic inversion

SI arose from Wiener's (1949) work. Its use in geophysical inversion was suggested by Franklin (1970) and amplified by Jackson (1979) and Gubbins (1983). In SI one imagines a series of experiments in each of which a model vector $\mathbf{x}$ is drawn at random from M and the data vector $\mathbf{y}$ is measured, where $\mathbf{y} = \mathbf{F}(\mathbf{x} - \boldsymbol{\xi}) + \boldsymbol{\eta} + \boldsymbol{\varepsilon}$. The $d \times m$ matrix $\mathbf{F}$, the model vector $\boldsymbol{\xi}$ and the data vectors $\boldsymbol{\eta}$ and $\mathbf{y}$ are known, but the model vector $\mathbf{x}$ which was drawn from M and the random error $\boldsymbol{\varepsilon}$ are unknown. The probability distribution on D of the errors $\boldsymbol{\varepsilon}$ is supposed to be known, as is the probability distribution on M which governs the random process of drawing models from M. The two random processes for $\boldsymbol{\varepsilon}$ and $\mathbf{x}$ are supposed to be uncorrelated, so, because $\langle \boldsymbol{\epsilon} \rangle = \mathbf{0}$,

$$\langle \mathbf{x}\boldsymbol{\varepsilon}^T \rangle = \mathbf{0}, \tag{3.1}$$

i.e. $\langle x_i \varepsilon_j \rangle = 0$.

The conceptual framework of SI has been used, for example, to direct anti-aircraft fire. The gunner observes the enemy aircraft's path before firing, knows the aircraft's aerodynamic characteristics, and has observed enough enemy pilots to have some idea of their evasive tactics under fire. He aims so as to minimize the expected value of the square of his miss distance. Another problem treated by SI is decipherment of microwave transmissions when they can be viewed as members of a large population with known statistics.

We slightly generalize Jackson's (1979) very clear derivation of the formalism for SI, partly to illuminate the underlying assumptions, and partly to establish a notation. Let $\mathbf{x}_{PR}$ and $\mathbf{V}_{PR}$ be the mean and the variance matrix of the probability distribution on M which describes the statistics of randomly drawing a model $\mathbf{x}$ from M. Let

$$\mathbf{y}_{PR} = \mathbf{F}(\mathbf{x}_{PR} - \boldsymbol{\xi}) + \boldsymbol{\eta} \tag{3.2a}$$

so that

$$\mathbf{y} - \mathbf{y}_{PR} = \mathbf{F}(\mathbf{x} - \mathbf{x}_{PR}) + \boldsymbol{\varepsilon}. \tag{3.2b}$$

The experiment of drawing $\mathbf{x}$ from M, measuring its $\mathbf{y}$, and using $\mathbf{y}$ to estimate $\mathbf{x}$ is repeated many times, at least in imagination. We seek a linear method which estimates $\mathbf{x}$ from $\mathbf{y}$ in every repetition. Whatever the method, it will be described by a fixed $m \times d$ matrix $\mathbf{H}$ such that if $\mathbf{y}$ is the observed data vector then our estimate for the $\mathbf{x}$ which produced it is $\mathbf{H}(\mathbf{y} - \mathbf{y}_{PR}) + \mathbf{x}_{PR}$. We denote this estimate by $\mathbf{x}_{PO}$, so

$$\mathbf{x}_{PO} - \mathbf{x}_{PR} = \mathbf{H}(\mathbf{y} - \mathbf{y}_{PR}). \tag{3.3}$$

The error committed through accepting $\mathbf{x}_{PO}$ for the true $\mathbf{x}$ is $\mathbf{x}_{PO} - \mathbf{x}$. Its expected value vanishes because $\langle \mathbf{x} \rangle = \mathbf{x}_{PR}$, $\langle \boldsymbol{\varepsilon} \rangle = \mathbf{0}$, and hence $\langle \mathbf{y} \rangle = \mathbf{y}_{PR}$. Let $\mathbf{V}_{PO}$ be the variance matrix of the error vector $\mathbf{x}_{PO} - \mathbf{x}$:

$$\mathbf{V}_{PO} = \langle (\mathbf{x}_{PO} - \mathbf{x})(\mathbf{x}_{PO} - \mathbf{x})^T \rangle. \tag{3.4}$$

Then from (2.1), (3.1) and (3.3)

$$\mathbf{V}_{PO} = (\mathbf{HF} - \mathbf{I})\mathbf{V}_{PR}(\mathbf{HF} - \mathbf{I})^T + \mathbf{HV}_\varepsilon\mathbf{H}^T, \tag{3.5}$$

where $\mathbf{I}$ is the $m \times m$ identity matrix. The trace of (3.4) is $\langle |\mathbf{x}_{PO} - \mathbf{x}|^2 \rangle$, which is a scalar estimate of the average error committed in many trials in which $\mathbf{x}_{PO}$ is always accepted for $\mathbf{x}$.

SI consists in observing that it is desirable to minimize this average error by a shrewd choice of $\mathbf{H}$, assuming that the parameters x_i have been normalized to be of equal and independent interest. The desired $\mathbf{H}$ will minimize the trace of (3.5), so it is

$$\mathbf{H} = \mathbf{V}_{PR}\mathbf{F}^T(\mathbf{V}_\varepsilon + \mathbf{FV}_{PR}\mathbf{F}^T)^{-1}. \tag{3.6}$$

Jackson (1979) rewrites (3.6) as

$$\mathbf{H} = (\mathbf{F}^T\mathbf{V}_\varepsilon^{-1}\mathbf{F} + \mathbf{V}_{PR}^{-1})^{-1}\mathbf{F}^T\mathbf{V}_\varepsilon^{-1}. \tag{3.7}$$

Substituting (3.7) in (3.5) gives

$$\mathbf{V}_{PO} = (\mathbf{F}^T\mathbf{V}_\varepsilon^{-1}\mathbf{F} + \mathbf{V}_{PR}^{-1})^{-1}; \tag{3.8}$$

this is the variance matrix of the estimation error $\mathbf{x}_{PO} - \mathbf{x}$ when $\mathbf{H}$ is optimally chosen as (3.7). For this optimal $\mathbf{H}$, the estimate of $\mathbf{x}$ is the $\mathbf{x}_{PO}$ given by

$$\mathbf{x}_{PO} - \mathbf{x}_{PR} = \mathbf{V}_{PO}\mathbf{F}^T\mathbf{V}_\varepsilon^{-1}(\mathbf{y} - \mathbf{y}_{PR}), \tag{3.9}$$

where $\mathbf{y}_{PR} = \mathbf{F}(\mathbf{x}_{PR} - \boldsymbol{\xi}) + \boldsymbol{\eta}$.

SI provides a rational way of estimating a random vector $\mathbf{x}$ from noisy measurements of linear data. The estimate is the x_{PO} of (3.9). The expected value of the error in this estimate vanishes, and the variance matrix of the error is provided by the theory as $\mathbf{V}_{PO}$ in (3.8).

3.1.2 Bayesian inference

BI is much older than SI, having been introduced in 1764 by the Reverend Thomas Bayes. It is widely used in business and economics and has elicited an extensive literature; see, e.g., Jeffreys (1961), Lindley (1965), Savage (1972), Box & Tiao (1973) and Berger (1985). Advocates of the use of BI in geophysical inverse problems include Backus (1970b, 1971), Tarantola & Valette (1982) and Gubbins & Bloxham (1985).

BI is based on an idealized description of how rational observers quantify their beliefs in the face of uncertainty, and how they modify those beliefs in the light of new information. BI supposes that a rational observer can describe his uncertainty about the location of the true model vector $\mathbf{x}$ in its model space M by assigning to $\mathbf{x}$ a probability distribution on M, called the observer's subjective or personal probability distribution. We simplify the discussion by assuming that this distribution has a continuous density function f. Therefore we must assume $\dim M < \infty$, since there is no sensible volume element in infinite-dimensional Hilbert spaces (Loewner 1939). The case $\dim M = \infty$ is treated in Appendix B.

One way to test the rationality of the observer and to learn his personal probability density f is to examine the bets he is willing to make about the location of the true $\mathbf{x}$ in the model space M. Suppose that for some subset X of M, odds of $a(X)$ to $b(X)$ are proposed for the bet that the true $\mathbf{x}$ lies in X. Suppose the odds are adjusted so that the observer is willing to take either side of the bet. Then for that observer's f

$$\int_X d\mathbf{x}^m f(\mathbf{x}) = a(X)/[a(X) + b(X)]. \tag{3.10}$$

A rational observer's bets will be described by a single density function f, which can be determined from (3.10) if his odds $a(X)$: $b(X)$ are known for every rectangular subset X of M.

Different observers may have different fs, reflecting different opinions about where $\mathbf{x}$ is likely to be in M. Furthermore, most observers will alter their fs if they obtain new information. An observer's fs before and after measuring the data vector $\mathbf{y}$ in (2.1) are written f_{PR} and f_{PO} and are called his prior and posterior personal probability densities for $\mathbf{x}$ in M.

To those who accept the existence of personal probabilities, calculating f_{PO} from f_{PR} and the data vector $\mathbf{y}$ is an exercise in probability theory. Consider the space $M \times D$ consisting of all pairs $(\mathbf{x}, \mathbf{y})$ with $\mathbf{x}$ in M and $\mathbf{y}$ in D. The point $(\mathbf{x}, \mathbf{y})$ in $M \times D$ represents the event that the true model is $\mathbf{x}$ and that measurements produce the data vector $\mathbf{y}$. Before measuring $\mathbf{y}$, the observer has opinions about the probable location of $(\mathbf{x}, \mathbf{y})$ in $M \times D$, and these are described by a prior personal probability density $f_{PR}(\mathbf{x}, \mathbf{y})$ on $M \times D$. Since the observer is rational, the laws of probability connect his personal probabilities on M, D and $M \times D$. If he knew nothing about $\mathbf{y}$ he would assign to $\mathbf{x}$ the marginal probability density on M:

$$f_{PR}(\mathbf{x}, D) = \int_D d\mathbf{y}^d f_{PR}(\mathbf{x}, \mathbf{y}). \tag{3.11a}$$

If he knew nothing about $\mathbf{x}$, he would assign to $\mathbf{y}$ the marginal probability density on D:

$$f_{PR}(M, \mathbf{y}) = \int_M d\mathbf{x}^m f_{PR}(\mathbf{x}, \mathbf{y}). \tag{3.11b}$$

If he knew that the data vector was $\mathbf{y}$ he would assign to $\mathbf{x}$ the conditional probability for $\mathbf{x}$ given $\mathbf{y}$:

$$f_{PR}(\mathbf{x} \mid \mathbf{y}) = f_{PR}(\mathbf{x}, \mathbf{y})/f_{PR}(M, \mathbf{y}). \tag{3.12a}$$

If he knew that the model was $\mathbf{x}$, he would assign to $\mathbf{y}$ the conditional probability for $\mathbf{y}$ given $\mathbf{x}$:

$$f_{PR}(\mathbf{y} \mid \mathbf{x}) = f_{PR}(\mathbf{x}, \mathbf{y})/f_{PR}(\mathbf{x}, D). \tag{3.12b}$$

From (3.12)

$$f_{PR}(\mathbf{x} \mid \mathbf{y}) = f_{PR}(\mathbf{y} \mid \mathbf{x}) f_{PR}(\mathbf{x}, D)/f_{PR}(M, \mathbf{y}). \tag{3.13}$$

But from the definitions, $f_{PR}(\mathbf{x}, D) = f_{PR}(\mathbf{x})$ and $f_{PR}(\mathbf{x} \mid \mathbf{y}) = f_{PO}(\mathbf{x})$. Writing $C(\mathbf{y})$ for $f_{PR}(M, \mathbf{y})$ then converts (3.13) into Bayes' theorem:

$$f_{PO}(\mathbf{x}) = C(\mathbf{y})^{-1} f_{PR}(\mathbf{y} \mid \mathbf{x}) f_{PR}(\mathbf{x}). \tag{3.14a}$$

Since

$$\int_M d\mathbf{x}^m f_{PO}(\mathbf{x}) = 1, \tag{3.14b}$$

therefore in (3.14*a*)

$$C(\mathbf{y}) = \int_M d\mathbf{x}^m f_{PR}(\mathbf{y} \mid \mathbf{x}) f_{PR}(\mathbf{x}). \tag{3.14c}$$

The observer can calculate $f_{PR}(\mathbf{y} \mid \mathbf{x})$ from his knowledge of F and the statistics of $\boldsymbol{\varepsilon}$ in (1.1), since $f_{PR}(\mathbf{y} \mid \mathbf{x})$ is just the probability density in D that the data vector $\mathbf{y}$ will be observed when the true model is $\mathbf{x}$. If $h(\boldsymbol{\varepsilon})$ is the probability density for the error vector $\boldsymbol{\varepsilon}$ in (1.1), then

$$f_{PR}(\mathbf{y} \mid \mathbf{x}) = h(\mathbf{y} - F(\mathbf{x})). \tag{3.14d}$$

Equations (3.14) are the calculations undertaken by the observer to reassess his opinion about the likely location of the true model in the model space, once he obtains the data described by the vector $\mathbf{y}$.

The foregoing calculations can be carried out explicitly when F is linear and $f_{PR}(\mathbf{x})$ and $h(\boldsymbol{\varepsilon})$ are gaussian. As with SI, we accept (2.1) and (3.1), and write the variance matrix of $h(\boldsymbol{\varepsilon})$ as $\mathbf{V}_\varepsilon$. We denote the mean and the variance matrix of $f_{PR}(\mathbf{x})$ by $\mathbf{x}_{PR}$ and $\mathbf{V}_{PR}$, and we define $\mathbf{y}_{PR}$ by (3.2a). Then (3.14d) becomes

$$f_{PR}(\mathbf{y} \mid \mathbf{x}) = h(\tilde{\mathbf{y}} - \mathbf{F}\tilde{\mathbf{x}}), \tag{3.15a}$$

where

$$\tilde{\mathbf{x}} = \mathbf{x} - \mathbf{x}_{PR}, \; \tilde{\mathbf{y}} = \mathbf{y} - \mathbf{y}_{PR}. \tag{3.15b}$$

When $h(\boldsymbol{\varepsilon})$ and $f_{PR}(\mathbf{x})$ are gaussian, (3.14a) and (3.15a) imply

$$\begin{aligned} -2 \ln f_{PO}(\mathbf{x}) = {} & (\tilde{\mathbf{y}} - \mathbf{F}\tilde{\mathbf{x}})^T \mathbf{V}_\varepsilon^{-1} (\tilde{\mathbf{y}} - \mathbf{F}\tilde{\mathbf{x}}) \\ & + \tilde{\mathbf{x}}^T \mathbf{V}_{PR}^{-1} \tilde{\mathbf{x}} + \psi(\mathbf{y}), \end{aligned} \tag{3.16}$$

where $\boldsymbol{\psi}(\mathbf{y})$ involves $C(\mathbf{y})$ and the determinants of $\mathbf{V}_\varepsilon$ and $\boldsymbol{V}_{PR}$ but does not involve $\mathbf{x}$. From (3.16) it follows that $f_{PO}(\mathbf{x})$ is also gaussian in $\mathbf{x}$. Let $\mathbf{V}_{PO}$ and $\mathbf{x}_{PO}$ be the variance matrix and the mean of f_{PO}. They can be found immediately by rearranging (3.16), and they turn out to be given exactly by (3.8) and (3.9).

To summarize, suppose the errors of observation have a gaussian distribution in the data space D, with variance matrix $\mathbf{V}_\varepsilon$ and mean $\mathbf{0}$. Suppose that before he observes the data vector $\mathbf{y}$ the observer's prior personal probability distribution for $\mathbf{x}$ in the model space M is gaussian with variance matrix $\mathbf{V}_{PR}$ and mean $\mathbf{x}_{PR}$. Then after observing $\mathbf{y}$ the observer's posterior personal probability distribution for $\mathbf{x}$ in M is gaussian, with variance matrix $\mathbf{V}_{PO}$ given by (3.8) and mean $\mathbf{x}_{PO}$ given by (3.9). It follows that calculations done for SI can be interpreted in terms of BI and vice versa. Appendix B extends these assertions to infinite-dimensional model spaces.

3.1.3 A comparison of SI and BI

SI has over BI the advantage of objectivity; the results of SI are the same for all observers. Moreover, the fundamental equations (3.8) and (3.9) are valid in SI for all distributions, not merely for gaussians. The only objective part of BI is the method used to compute f_{PO} from f_{PR}; different observers will often have different f_{PR}s and f_{PO}s and will agree only on f_{PO}/f_{PR} (and perhaps not even on that if the error distribution includes a subjective element; see below). Moreover, with BI equations (3.8) and (3.9) are valid only for gaussian distributions, although they can be used for distributions which are approximately gaussian (Gubbins & Bloxham 1985, equation (18) ff.). A secondary advantage of BI over SI is that BI is easily extended to nonlinear Fs. The main advantage of BI over SI is that in some inverse problems SI is not a natural way to incorporate prior subjective beliefs into the processing of the data. In SI the variance matrix $\mathbf{V}_{PR}$ describes the random process of drawing, at least in imagination, many models $\mathbf{x}$ out of the model space M. The conceptual framework of SI makes difficult the treatment problems where there can be only one correct model and there is no natural way to regard that model as a member of a large population with known statistics. Such problems are the province of BI, which enables different observers to make objective comparisons of their subjective beliefs when there is only one correct model.

Illustrations will clarify this point. Suppose an observer obtains magnetic anomaly profiles along surface ship tracks and wants to extrapolate them down to the sea-bed. If there is a large library of deep tow lines, it is sensible to view any one of them as a member of a statistical population and to choose a method of inference which minimizes the rms error in the sea-bed estimates, averaged over the whole population. Similarly, over several eons a chaotic core dynamo will produce a large population of B_rs at the CMB whose statistics might be accessible through paleomagnetic data (Constable & Parker 1988). In both these examples, SI is the method of choice if the population statistics of the models have been estimated from adequately large samples. On the other hand, perhaps such sampling has not yet been done. Even if it has, the observer may feel that he knows more about his particular sea-bed profile than is conveyed by membership of that model in a large class of observed models. Then the observer wants to bring to bear all of his relevant knowledge and judgement. This leads him to BI as the natural technique for quantifying a 'soft' form of his beliefs as a probability distribution on M.

Another advantage of BI over SI has to do with the error $\boldsymbol{\varepsilon}$ in (1.1) and (2.1). So far, $\boldsymbol{\varepsilon}$ has been attributed entirely to errors of measurement. But no model space ever gives a complete description of everything which might contribute to the data vector $\mathbf{y}$, so $\boldsymbol{\varepsilon}$ will contain errors due to inadequacies of the model space. One example is truncation of an infinite-dimensional model space to one whose finite dimensional permits numerical computation. The errors $\boldsymbol{\varepsilon}$ produced by model space inadequacies can often be computed in a hard form like (2.11), in which the value of Q is somewhat uncertain. Such errors do not fit naturally into SI. With BI, however, there is no difficulty in viewing the distribution of errors as another personal probability distribution. The errors due to model inadequacies are estimated not by collecting data but by the introspection required to understand a theory. Tarantola & Valette (1982) discuss these questions at greater length, and such questions also appear in Section 4 below.

When the probability distributions for $\boldsymbol{\varepsilon}$ in D and $\mathbf{x}$ in M are gaussian, BI and SI call for the same calculations. Therefore it is not urgent that the reader be persuaded to adopt BI in studying the work of Gubbins & Bloxham. However, the present section has established that BI is an appropriate conceptual framework for interpreting all of

G & B's calculations, even those done with SI in mind. We undertake that interpretation, in order to clarify exactly what prior beliefs G & B accept and to ascertain whether the reader will accept them.

4 BOUND-SOFTENING PROBLEMS IN GAUSSIAN BAYESIAN INFERENCE

In BI with gaussian probability distributions, the mathematical behaviours of soft and hard bounds can differ, so an observer who softens beliefs (inequalities) into a gaussian personal probability distribution must examine carefully whether his mathematical implementation of the soft bounds really reflects what he believes.

To discuss such questions, it is necessary first to consider the general theory of how an observer who accepts gaussian Bayesian inference will digest new information by altering his personal probability distribution for the model $\mathbf{x}$ in the model space M of m-dimensional column vectors. First, some notation is needed. Let $\boldsymbol{\mu}$ be any vector in M. Let $\mathbf{C}$ be any $m \times m$ symmetric positive semi-definite matrix. Let $r = r(\mathbf{C})$ be the rank of $\mathbf{C}$, and let $c_1, \ldots, c_r$ be the positive eigenvalues of $\mathbf{C}$. Let

$$K(\mathbf{C}) = (2\pi)^{-r/2}(c_1 c_2 \cdots c_r)^{1/2} \tag{4.1a}$$

and

$$P_{\mathbf{C},\boldsymbol{\mu}}(\mathbf{x}) = (\mathbf{x}-\boldsymbol{\mu})^T\mathbf{C}(\mathbf{x}-\boldsymbol{\mu}) \tag{4.1b}$$

and

$$f_{\mathbf{C},\boldsymbol{\mu}}(\mathbf{x}) = K(\mathbf{C})\exp\left[-P_{\mathbf{C},\boldsymbol{\mu}}(\mathbf{x})/2\right]. \tag{4.1c}$$

The function $f_{\mathbf{C},\boldsymbol{\mu}}$ will be called the gaussian $(\mathbf{C}, \boldsymbol{\mu})$. If $r = m$, then $\mathbf{C}^{-1}$ exists and $f_{\mathbf{C},\boldsymbol{\mu}}$ is a gaussian probability density on M. If $r < m$, then $\mathbf{C}^{-1}$ does not exist and

$$\int_M d\mathbf{x}^m f_{\mathbf{C},\boldsymbol{\mu}}(\mathbf{x}) = \infty$$

so $f_{\mathbf{C},\boldsymbol{\mu}}$ cannot be normalized to a probability density on M. The gaussian $(\mathbf{C}, \boldsymbol{\mu})$ will be called regular or singular according as $\mathbf{C}^{-1}$ exists or does not. Singular gaussians are cylinder measures on M (Backus (1971) gives a review with references to the original literature). It is important to note that the function $f_{\mathbf{C},\boldsymbol{\mu}}$ is completely determined if those parts of the inhomogeneous quadratic polynomial $P_{\mathbf{C},\boldsymbol{\mu}}(\mathbf{x})$ are known which are homogenous of degrees 2 and 1 in $\mathbf{x}$. This is the same as knowing the gradient of $P_{\mathbf{C},\boldsymbol{\mu}}(\mathbf{x})$ with respect to $\mathbf{x}$.

Suppose the observer's personal probability distribution for $\mathbf{x}$ in M is gaussian $(\mathbf{C}_1, \boldsymbol{\mu})$. Since this is a probability distribution, it must be regular, and $\mathbf{C}_1^{-1}$ exists. Now suppose the observer attends a lecture which provides him with good evidence, independent of his prior opinions, that the true model $\mathbf{x}$ satisfies an inequality (2.11). He can rewrite that inequality as

$$P_{\mathbf{C}_2,\boldsymbol{\mu}_2}(\mathbf{x}) \le r(\mathbf{C}_2) \tag{4.2}$$

if he defines $\mathbf{C}_2 = \mathbf{C}r(\mathbf{C})/Q$ and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}$.

How does this new information alter the observer's personal probability distribution? Let d be the rank of $\mathbf{C}_2$, let $c_1, c_2, \ldots, c_d$ be the positive eigenvalues of $\mathbf{C}_2$, and in M let $\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2, \ldots, \hat{\mathbf{c}}_d$ be an orthonormal set of eigenvectors of $\mathbf{C}_2$ with those eigenvalues; i.e. $\mathbf{C}_2\hat{\mathbf{c}}_i = c_i\hat{\mathbf{c}}_i$ (no sum on i). Let D denote the linear space of d-dimensional column vectors, and let $\hat{\mathbf{y}}_i$ be the vector in D whose ith component is 1, all other components being 0. Define the $d \times m$ matrix $\mathbf{F}$ as

$$\mathbf{F} = \sum_{i=1}^{d} \hat{\mathbf{y}}_i\hat{\mathbf{c}}_i^T. \tag{4.3a}$$

Define the vector $\mathbf{y}$ in D as

$$\mathbf{y} = \mathbf{F}\boldsymbol{\mu}_2. \tag{4.3b}$$

Define the $d \times d$ matrix $\mathbf{V}_\varepsilon$ by

$$\mathbf{V}_\varepsilon^{-1} = \sum_{i=1}^{d} c_i\hat{\mathbf{y}}_i\hat{\mathbf{y}}_i^T. \tag{4.3c}$$

Then

$$P_{\mathbf{C}_2,\boldsymbol{\mu}_2}(\mathbf{x}) = (\mathbf{y}-\mathbf{F}\mathbf{x})^T\mathbf{V}_\varepsilon^{-1}(\mathbf{y}-\mathbf{F}\mathbf{x}). \tag{4.4}$$

Now the observer can imagine experimental attempts to measure the data vector $\mathbf{F}\mathbf{x}$ of the true model $\mathbf{x}$. The measured data vector $\mathbf{y}$ includes an experimental error $\boldsymbol{\varepsilon}$, so $\mathbf{y} = \mathbf{F}\mathbf{x} + \boldsymbol{\varepsilon}$. Because of (4.4), the observer can regard (4.2) as the hard version of the belief that the error vector $\boldsymbol{\varepsilon}$ has a gaussian distribution in D, with mean $\mathbf{0}$ and variance matrix V_ε. And the observer can regard the lecture as an experiment which produced for the data vector $\mathbf{y}$ the value (4.3b). Then Section 3 shows how the observer will alter his personal probability distribution for $\mathbf{x}$ on M from a prior distribution f_{PR} to a posterior distribution f_{PO} in response to obtaining the particular value $\mathbf{y}$ for the data vector. We have already assumed

$$f_{PR}(\mathbf{x}) = f_{\mathbf{C}_1,\boldsymbol{\mu}_1}(\mathbf{x}) \tag{4.5a}$$

and the errors are gaussian, so

$$f_{PO}(\mathbf{x}) = f_{\mathbf{C},\boldsymbol{\mu}}(\mathbf{x}), \tag{4.5b}$$

where $\mathbf{C}^{-1} = \mathbf{V}_{PO}$ and $\boldsymbol{\mu} = \mathbf{x}_{PO}$ are given by (3.8) and (3.9). Some algebra can be avoided by observing that (3.16) implies

$$P_{\mathbf{C},\boldsymbol{\mu}}(\mathbf{x}) = P_{\mathbf{C}_1,\boldsymbol{\mu}_1}(\mathbf{x}) + P_{\mathbf{C}_2,\boldsymbol{\mu}_2}(\mathbf{x}) + \text{constant}, \tag{4.6a}$$

where the constant involves $\mathbf{C}_i$ and $\boldsymbol{\mu}_i$ but not $\mathbf{x}$. Therefore

$$\mathbf{C} = \mathbf{C}_1 + \mathbf{C}_2 \tag{4.6b}$$

and

$$\mathbf{C}\boldsymbol{\mu} = \mathbf{C}_1\boldsymbol{\mu}_1 + \mathbf{C}_2\boldsymbol{\mu}_2. \tag{4.6c}$$

Equation (4.6c) can always be solved for $\boldsymbol{\mu}$, because $\mathbf{C}^{-1}$ exists. To see this, note that for any non-zero model vector $\mathbf{x}$, $\mathbf{x}^T\mathbf{C}\mathbf{x} \ge \mathbf{x}^T\mathbf{C}_1\mathbf{x} > 0$ by hypothesis, because $\mathbf{x}^T\mathbf{C}_2\mathbf{x} \ge 0$. It follows also that the observer's posterior probability distribution is regular even if the new information $f_{\mathbf{C}_2,\boldsymbol{\mu}_2}$ is not.

The foregoing discussion shows that the observer can incorporate the hard belief (4.2) into his opinion about where $\mathbf{x}$ is in M by softening (4.2) to the gaussian $(\mathbf{C}_2, \boldsymbol{\mu}_2)$ and replacing his prior personal probability, the gaussian $(\mathbf{C}_1, \boldsymbol{\mu}_1)$, with a posterior personal probability which is gaussian $(\mathbf{C}, \boldsymbol{\mu})$, the $\mathbf{C}$ and $\boldsymbol{\mu}$ being given by (4.6). In short, the softened form of a hard quadratic belief is a possibly singular gaussian on the model space M which operates through (4.6) to convert the observer's regular gaussian prior to a regular gaussian posterior personal probability distribution on M.

When there is a solution of $\boldsymbol{\mu}_2$ of

$$\mathbf{F}\boldsymbol{\mu}_2 = \mathbf{y}. \tag{4.7a}$$

then in fact (4.6) is the only rule the observer needs for digesting new information. Digesting an inaccurately measured value $\mathbf{y}$ for a data vector which depends linearly on $\mathbf{x}$ can be viewed as a special case of (4.6). Because of (2.6) and (4.4), in (4.6) the observer can simply choose

$$\mathbf{C}_2 = \mathbf{F}^T\mathbf{V}_\varepsilon^{-1}\mathbf{F}. \tag{4.7b}$$

Having started with the belief $f_{\mathbf{C}_1,\boldsymbol{\mu}_1}$ the observer was persuaded by the new information $f_{\mathbf{C}_2,\boldsymbol{\mu}_2}$ to change his belief to $f_{\mathbf{C},\boldsymbol{\mu}}$ as in (4.6). If he obtains still more new information in the form $f_{\mathbf{C}_3,\boldsymbol{\mu}_3}$, he will modifiy the new prior $f_{\mathbf{C},\boldsymbol{\mu}}$ to a posterior $f_{\mathbf{C}'\boldsymbol{\mu}'}$ with $\mathbf{C}' = \mathbf{C} + \mathbf{C}_3 = \mathbf{C}_1 + \mathbf{C}_2 + \mathbf{C}_3$ and $\mathbf{C}'\boldsymbol{\mu}' = C\boldsymbol{\mu} + \mathbf{C}_3\boldsymbol{\mu}_3$. In general, if the observer starts with the prior personal belief $f_{\mathbf{C}_1,\boldsymbol{\mu}_1}$ and obtains $N-1$ new pieces of quadratic information about $\mathbf{x}$, whose soft forms are $f_{\mathbf{C}_i,\boldsymbol{\mu}_i}$ with $i = 2, \dots, N$, then the observer will modify his personal probability distribution from $f_{\mathbf{C}_1,\boldsymbol{\mu}_1}$ to $f_{\mathbf{C},\boldsymbol{\mu}}$ where

$$P_{\mathbf{C},\boldsymbol{\mu}}(\mathbf{x}) = \sum_{i=1}^{N} P_{\mathbf{C}_i,\boldsymbol{\mu}_i}(\mathbf{x}) + \text{const.} \tag{4.8a}$$

so

$$\mathbf{C} = \sum_{i=1}^{N} \mathbf{C}_i \tag{4.8b}$$

and

$$\mathbf{C}\boldsymbol{\mu} = \sum_{i=1}^{N} \mathbf{C}_i\boldsymbol{\mu}_i. \tag{4.8c}$$

Clearly the order in which the observer acquires the new pieces of information will make no difference, and any further information is added without inconsistency. This situation makes it unnecessary to single out $f_{\mathbf{C}_1,\boldsymbol{\mu}_1}$ as the observer's prior, or to assume that it is regular. If the observer holds all the beliefs $f_{\mathbf{C}_i,\boldsymbol{\mu}_i}$ with $i = 1, \dots, N$, he can consistently adopt $f_{\mathbf{C},\boldsymbol{\mu}}$ in (4.8) as his personal probability distribution for $\mathbf{x}$ in M even if each $f_{\mathbf{C}_i,\boldsymbol{\mu}_i}$ is singular, as long as the $\mathbf{C}$ of (4.8b) has an inverse. Thus separate pieces of quadratic information, none adequate alone to produce a personal probability density on M, can combine to do so. Jackson's (1979) equation (44) is our (4.8b), derived in the context of SI and with the assumption that $\mathbf{C}_i\mathbf{C}_j = \mathbf{0}$ if $i \neq j$. No such assumption is needed in (4.8b).

Jeffreys (1961) has advocated accepting singular gaussians as personal probability distributions. To do so produces major theoretical complications (Backus 1971), and it seems preferable to use only regular gaussians for personal probabilities, although quadratic information will often arrive in the form of a singular gaussian. The analytical machinery of probability theory is too valuable to be given up lightly.

One pitfall which awaits the Bayesian observer is now visible. Suppose his prior is $f_{\mathbf{C},\mathbf{0}}$ where $\mathbf{C} = \sigma^{-2}\mathbf{I}$. Then his expected value for x_1^2 is $\langle x_1^2 \rangle = \sigma^2$. Suppose now that d messengers arrive, each reporting a measurement of x_1 whose outcome could be described in hard terms as $-\sigma < x_1 < \sigma$, or in soft terms by the one-dimensional singular gaussian $(2\pi)^{-1/2}\sigma^{-1}\exp(-x_1^2/2\sigma^2)$. According to (4.8), the Bayesian observer should now adopt for his personal probability the gaussian $f_{\mathbf{C}',\mathbf{0}}$ where $\mathbf{C}'$ is diagonal with σ^{-2} in all diagonal positions but the first, where $(d+1)\sigma^{-2}$ appears. Does the observer really want to accept an opinion whose hard form is $-\sigma/(d+1)^{1/2} < x_1 < \sigma/(d+1)^{1/2}$? To decide, he must judge whether he really believes that the d measurements of x_1 were independent of each other and of the information which led him to his own prior opinion. If they are, he does adopt $f_{\mathbf{C}',\mathbf{0}}$. The new variance of x_1 is just what ordinary probability theory would give for the variance of the mean of $d+1$ independent measurement of x_1.

A second pitfall concerns model spaces of high (or infinite) dimension. An assumption is hidden in replacing the hard belief (4.2) by the soft belief $f_{\mathbf{C}_2,\boldsymbol{\mu}_2}$. We can confine the discussion to (2.7) when $\mathbf{C}$ is diagonal and non-singular. Suppose, then, that there are positive constants $\Delta_1, \dots, \Delta_m$, Q such that the observer believes

$$\sum_{i=1}^{m} (x_i/\Delta_i)^2 \leq Q. \tag{4.9}$$

Softening this hard inequality as in Section 4 leads him to adopt for his prior personal probability distribution the gaussian $f_{\mathbf{C},\mathbf{0}}$ where $\mathbf{C}$ is diagonal, its ith diagonal entry being $mQ^{-1}\Delta_i^{-2}$. Then the expected value which the observer assigns to $(x_i/\Delta_i)^2$ is

$$\langle (x_i/\Delta_i)^2 \rangle = Q/m. \tag{4.10}$$

Nothing in (4.9) justifies the belief (4.10) that all the terms $(x_i/\Delta_i)^2$ are likely to be about the same size. This belief is one the observer holds independently of (4.9). It is a belief which he may want to reconsider if (4.9) is a finite-dimensional approximation to an infinite sum,

$$\sum_{i=1}^{\infty} (x_i/\Delta_i)^2 \leq Q, \tag{4.11}$$

as in (2.12), for example. When $m \to \infty$, the right side of (4.10) approaches 0, so when $\dim M = \infty$ the observer is claiming exact *a priori* knowledge that all x_i vanish.

Bold and cautious observers will adopt different resolutions of this difficulty. The bold observer may convince himself that he can estimate the relative importance of the various terms in (4.11). He may believe that he knows numbers γ_i such that for each i his prior leads to the expected value

$$\langle (x_i/\Delta_i)^2 \rangle_{\text{bold}} = K\gamma_i, \tag{4.12a}$$

where K is a proportionality constant. If he also believes (4.11), then taking expected values to soften that equation yields

$$K\sum_{i=1}^{\infty} \gamma_i = Q. \tag{4.12b}$$

Therefore, consistency requires him to verify that the $\gamma_1, \gamma_2, \dots$, which he believes appropriate in (4.12a) do sum to a finite value. Absorbing that sum in K permits the normalization

$$\sum_{i=1}^{\infty} \gamma_i = 1, \tag{4.13a}$$

and then (4.12b) yields $K = Q$, so that

$$\langle (x_i/\Delta_i)^2 \rangle_{\text{bold}} = \gamma_i Q. \tag{4.13b}$$

Even in the absence of an *a priori* belief (4.12a) based on the physics, the bold observer may be willing to adopt such a belief arbitrarily, simply to make the problem tractable. If he thinks that at any rate the $\langle (x_i/\Delta_i)^2 \rangle$ are likely to be smaller for larger i, one reasonable if arbitrarily choice for γ_i is

$$\gamma_i = 6(\pi i)^{-2}. \tag{4.14}$$

This series converges rather slowly, and (4.13b) does not commit the observer to a high degree of certainty about x_i/Δ_i until i is very large. The bold observer will adopt for his prior personal probability distribution the gaussian $(\mathbf{C}^{\text{bold}}, \mathbf{0})$, where $\mathbf{C}^{\text{bold}}$ is diagonal, and the ith diagonal entry is $Q^{-1}\Delta_i^{-2}\gamma_i^{-1}$.

The cautious observer, unwilling to commit himself *a priori* about the ratios of the terms in (4.11), will resign himself to the fact that from (4.11) all he knows with certainty about any individual term $(x_i/\Delta_i)^2$ is that it is not larger than Q. The cautious observer will soften these certainties to the gaussian $(\mathbf{C}^{\text{safe}}, \mathbf{0})$, where $\mathbf{C}^{\text{safe}}$ is diagonal, and the ith diagonal entry is $Q^{-1}\Delta_i^{-2}$. For the cautious observer, the prior expected value of the ith term in (4.11) is

$$\langle (x_i/\Delta_i)^2 \rangle_{\text{safe}} = Q. \tag{4.15}$$

Equation (4.15) understates what the observer thinks he knows *a priori*, because (4.15) makes the expected value of the sum in (4.11) infinite, and makes no use of the knowledge that the sum converges. If, nevertheless, (4.15) leads to useful conclusions, those conclusions will inspire a high level of confidence.

There are intermediate courses between the bold and the conservative in which $1 \geq \gamma_1 \geq \gamma_2 \geq \gamma_3 \geq \cdots$ and the sum of the γ_i is larger than 1, perhaps even infinite. A real observer would probably want to start with the conservative stance (4.15). If it did not lead to useful error bounds for the $\mathbf{g}^T\mathbf{x}$ in which he was interested, he might be emboldened to try (4.13b) with γ_i decreasing but summing to infinity. Failure here might lead him to use γ_i whose sum was finite but larger than 1. Finally, he might be driven to adopt (4.13) *in toto*, recognizing that it would be difficult to convince cautious observers of his conclusions.

5 RESOLUTION IN LINEAR GAUSSIAN BAYESIAN INFERENCE

In linear inverse problems, the data vector $\mathbf{y}$ will determine the value of $\mathbf{g}^T\mathbf{x}$ for some linear functionals $\mathbf{g}$ with a very small uncertainty, while contributing almost nothing to the determination of $\mathbf{g}^T\mathbf{x}$ for other $\mathbf{g}$. For example, $\mathbf{y}$ will permit accurate estimation of some x_i but not of others. If $\mathbf{g}^T\mathbf{x}$ is accurately determined by the data, then $\mathbf{g}^T\mathbf{x}$ is said to be 'resolved' by the data. The present section discusses how to find and count the linear functionals $\mathbf{g}$ on M such that $\mathbf{g}^T\mathbf{x}$ is resolved by the data. Other characterizations of these resolved functionals are also given.

In the absence of prior information, the data resolve $\mathbf{g}^T\mathbf{x}$ if and only if $\mathbf{g}$ is in $M_{\mathbf{F}}$ (Backus & Gilbert 1968, 1970). Backus (1970b) discusses resolution when hard prior information (inequalities) is available. For soft prior information (probability distributions on M), Jackson (1979) discusses resolution in SI. The present section considers resolution in linear, gaussian BI.

We use the notation of Section 3. We suppose that the observer's data $\mathbf{y}$ are related to the model $\mathbf{x}$ by (2.1), that the errors $\boldsymbol{\varepsilon}$ are gaussian $(\mathbf{V}_\varepsilon^{-1}, \mathbf{0})$, and that the observer's prior personal probability distribution on M is gaussian $(\mathbf{V}_{PR}^{-1}, \mathbf{x}_{PR})$. Then his posterior personal probability distribution is gaussian $(\mathbf{V}_{PO}^{-1}, \mathbf{x}_{PO})$, where $\mathbf{V}_{PO}$ and $\mathbf{x}_{PO}$ are given by (3.8) and (3.9). If z is any random variable (function) on M, we denote by $\langle z \rangle_{PR}$ and $\langle z \rangle_{PO}$ the statistical expected values of z under the observer's prior and posterior personal probability distributions. Thus, for example, if P stands for either PR or PO,

$$\langle \mathbf{x} \rangle_P = \mathbf{x}_P \tag{5.1a}$$

and

$$\langle (\mathbf{x}-\mathbf{x}_P)(\mathbf{x}-\mathbf{x}_P)^T \rangle_P = \mathbf{V}_P. \tag{5.1b}$$

If $\mathbf{g}$ is a linear functional on M, then in BI $\mathbf{g}^T\mathbf{x}$ is resolved by the data if those data cause the variance of $\mathbf{g}^T\mathbf{x}$ to be significantly smaller under the observer's posterior personal probability distribution than under his prior. To see which $\mathbf{g}$ are so resolved, it is useful to introduce on M a new parametrization, $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_m)$ chosen so that if $i \neq j$ then ζ_i and ζ_j are subjectively uncorrelated both before and after the observer obtains the data $\mathbf{y}$, i.e. $\langle (\zeta_i - \langle \zeta_i \rangle_P)(\zeta_j - \langle \zeta_j \rangle_P) \rangle_P = 0$ if $i \neq j$ and P stands for either PR or PO. It will be convenient to normalize $\boldsymbol{\zeta}$ so that the prior variance of each ζ_i is 1. Then its posterior variance ≤ 1, and if $\ll 1$ then that particular ζ_i has been constrained or resolved by the data.

To find the parameters ζ_i, we define

$$\mathbf{A} = \mathbf{F}^T\mathbf{V}_\varepsilon^{-1}\mathbf{F} \tag{5.2a}$$

$$\mathbf{B} = \mathbf{V}_{PR}^{-1} \tag{5.2b}$$

$$\mathbf{C} = \mathbf{V}_{PO}^{-1} \tag{5.2c}$$

so that (3.8) becomes

$$\mathbf{C} = \mathbf{A} + \mathbf{B}. \tag{5.2d}$$

The matrices $\mathbf{A}$ and $\mathbf{B}$ are symmetric, $\mathbf{B}$ is positive definite, and $\mathbf{A}$ is positive semi-definite. Then $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$ is symmetric and positive semi-definite. Let $\mathbf{B}^{1/2}\boldsymbol{\mu}_1, \ldots, \mathbf{B}^{1/2}\boldsymbol{\mu}_m$ be its eigenvectors and $\alpha_1, \ldots, \alpha_m$ its eigenvalues in decreasing order. Then the α_i are real and

$$\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m \text{ is a basis for } M \tag{5.3a}$$

$$\boldsymbol{\mu}_i^T\mathbf{B}\boldsymbol{\mu}_j = \delta_{ij} \qquad (i, j = 1, \ldots, m) \tag{5.3b}$$

$$\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_m \geq 0 \tag{5.3c}$$

$$\mathbf{A}\boldsymbol{\mu}_j = \alpha_j\mathbf{B}\boldsymbol{\mu}_j \qquad (j = 1, \ldots, m). \tag{5.3d}$$

Here δ_{ij} is the Kronecker delta ($\delta_{ij} = 1$ if $i = j$, $\delta_{ij} = 0$ if $i \neq j$). For any $\mathbf{x}$ in M, there are real numbers $\zeta_1, \ldots, \zeta_m$ such that

$$\mathbf{x} = \sum_{i=1}^{m} \zeta_i\boldsymbol{\mu}_i \tag{5.4a}$$

and (5.3b) implies that

$$\zeta_i = \boldsymbol{\mu}_i^T \mathbf{B}\mathbf{x}. \tag{5.4b}$$

Since $\mathbf{B} = \mathbf{B}^T$, (5.4b) can be written

$$\zeta_i = \mathbf{x}^T \mathbf{B}\boldsymbol{\mu}_i, \tag{5.4c}$$

therefore

$$\zeta_i \zeta_j = \boldsymbol{\mu}_i^T \mathbf{B}\mathbf{x}\mathbf{x}^T \mathbf{B}\boldsymbol{\mu}_j. \tag{5.4d}$$

The parameters $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_m)$ depend linearly on $\boldsymbol{x} = (x_1, \ldots, x_m)$. Since the observer's prior and posterior distributions for $\mathbf{x}$ are gaussian, the same is true for $\boldsymbol{\zeta}$. To find the prior and posterior variance matrices for $\boldsymbol{\zeta}$, take expected values in (5.4b, c, d). Taking P to be either PR or PO, the result is

$$\langle \zeta_i \rangle_P = \boldsymbol{\mu}_i^T \mathbf{B}\mathbf{x}_P$$
$$\langle \zeta \rangle_P = \mathbf{x}_P^T \mathbf{B}\boldsymbol{\mu}_j$$
$$\langle \zeta_i \zeta_j \rangle_P = \boldsymbol{\mu}_i^T \mathbf{B} \langle \mathbf{x}\mathbf{x}^T \rangle_P \mathbf{B}\boldsymbol{\mu}_j.$$

Therefore, with P either PR or PO,

$$\langle \zeta_i \zeta_j \rangle_P = \boldsymbol{\mu}_i^T \mathbf{B}\mathbf{V}_P \mathbf{B}\boldsymbol{\mu}_j + \langle \zeta_i \rangle_P \langle \zeta_j \rangle_P.$$

It follows that for $P = PR$ or $P = PO$, the $m \times m$ variance matrix of $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_m)$ has as its ij entry

$$\langle (\zeta_i - \langle \zeta_i \rangle_P)(\zeta_j - \langle \zeta_j \rangle_P) \rangle_P = \boldsymbol{\mu}_i^T \mathbf{B}\mathbf{V}_P \mathbf{B}\boldsymbol{\mu}_j. \tag{5.5a}$$

Since $\mathbf{V}_{PR} = \mathbf{B}^{-1}$, (5.5a) and (5.3b) imply

$$\langle (\zeta_i - \langle \zeta_i \rangle_{PR})(\zeta_j - \langle \zeta_j \rangle_{PR}) \rangle_{PR} = \delta_{ij}. \tag{5.5b}$$

To evaluate (5.5a) for $P = PO$, we must set $\mathbf{V}_P = \mathbf{V}_{PO} = (\mathbf{A} + \mathbf{B})^{-1}$ in (5.5a). To proceed, we note from (5.3d) that

$$(\mathbf{A} + \mathbf{B})\boldsymbol{\mu}_j = (1 + \alpha_j)\mathbf{B}\boldsymbol{\mu}_j$$

so that

$$(\mathbf{A} + \mathbf{B})^{-1}\mathbf{B}\boldsymbol{\mu}_j = (1 + \alpha_j)^{-1}\boldsymbol{\mu}_j. \tag{5.5c}$$

Inserting this expression on the right in (5.5a) and using (5.3b) we obtain

$$\langle (\zeta_i - \langle \zeta_i \rangle_{PO})(\zeta_j - \langle \zeta_j \rangle_{PO}) \rangle_{PO} = (1 + \alpha_j)^{-1}\delta_{ij}. \tag{5.5d}$$

Before the observer learns of the data $\mathbf{y}$, his prior personal probability distribution on M makes $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_{\mathrm{m}})$ a gaussian, with ζ_i and ζ_j independent when $i \neq j$, and with each ζ_i having variance 1. After learning of $\mathbf{y}$, the observer alters his prior to his posterior distribution. His new, posterior, opinion assigns to $\boldsymbol{\zeta}$ a gaussian distribution, and ζ_i and ζ_j are still independent if $i \neq j$, but now ζ_i has the variance $(1 + \alpha_i)^{-1}$.

The parameters $\zeta_1, \ldots, \zeta_m$ are linear combinations of the original parameters $x_1, \ldots, x_m$. Those linear combinations ζ_i for which $(1 + \alpha_i)^{-1} \ll 1$ are constrained or resolved by the data; the others are not. The unconstrained or unresolved ζ_i are those for which

$$\alpha_i \ll 1. \tag{5.6}$$

If (5.6) holds, the data produce almost no improvement in the accuracy with which the observer believes he knows ζ_i. Such an improvement occurs only when (5.6) fails.

The α_i can also be used to determine which ζ_i are really useful in fitting the data. In equation (2.1), if the model is changed by $\delta\mathbf{x}$, the data will change by

$$\delta\mathbf{y} = \mathbf{F}\delta\mathbf{x}.$$

A dimensionless way of estimating the size of $\delta\mathbf{y}$ is to compare it with the errors in the data by calculating

$$\delta\mathbf{y}^T \mathbf{V}_\varepsilon^{-1} \delta\mathbf{y} = \delta\mathbf{x}^T \mathbf{A}\delta\mathbf{x}. \tag{5.7a}$$

But if $\delta\mathbf{x} = \sum_{i=1}^m (\delta\zeta_i)\,\boldsymbol{\mu}_i$ then (5.3d, b) give

$$\delta\mathbf{y}^T \mathbf{V}_\varepsilon^{-1} \delta\mathbf{y} = \textstyle\sum_{i=1}^m \alpha_i (\delta\zeta_i)^2. \tag{5.7b}$$

The *a priori* variance of ζ_i is 1, so $|\delta\zeta_i| \gg 1$ is very improbable *a priori*. Therefore, if $\alpha_i \ll 1$, only improbably large changes in ζ_i will change the data vector by as much as its experimental error of measurement.

In the existence problem, an effective computational technique is to use only those ζ_i for which (5.6) fails. The other ζ_i do not help to fit the model to the data, so nothing is lost by setting them equal to their values (5.4b) for $\mathbf{x} = \mathbf{x}_{PR}$. In the uniqueness problem, it must be kept in mind that the ζ_i which satisfy (5.6) have unit posterior variance, and may contribute to errors in estimating $\mathbf{g}^T\mathbf{x}$ even if they contribute nothing to fitting the data.

The model parameters ζ_i for which (5.6) fails are 'effective' in the sense that varying them by reasonable amounts can improve the fit to the data. They are 'resolved' in the sense that the data shrink their subjective personal variances appreciably. A rough estimate of the number of effective or resolved model parameters is

$$\mathrm{emp} = \sum_{i=1}^{m} \alpha_i (1 + \alpha_i)^{-1}. \tag{5.8}$$

To justify this estimate, note that the ith term in (5.8) is nearly 0 if $\alpha_i \ll 1$ and nearly 1 if $\alpha_i \gg 1$. If m is large and α_i decreases rapidly with i when $\alpha_i \leq 1$ then (5.8) is approximately the number of ζ_i for which (5.6) fails.

The estimate (5.8) may be rough, but has the advantage that it can be calculated without solving (5.3) for the α_i. To see this, note that

$$(\mathbf{A} + \mathbf{B})^{-1}(\mathbf{A} + \mathbf{B})\boldsymbol{\mu}_j = \boldsymbol{\mu}_j.$$

Subtracting (5.5c) from this equation gives

$$(\mathbf{A} + \mathbf{B})^{-1}\mathbf{A}\boldsymbol{\mu}_j = \alpha_j (1 + \alpha_j)^{-1}\boldsymbol{\mu}_j.$$

Multiplying on the left by $\boldsymbol{\mu}_i^T\mathbf{B}$ and using (5.3b) gives

$$\boldsymbol{\mu}_i^T \mathbf{B}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{A}\boldsymbol{\mu}_j = \alpha_j (1 + \alpha_j)^{-1}\delta_{ij}. \tag{5.9a}$$

Let $\mathbf{S}$ be the $m \times m$ matrix whose jth column is $\boldsymbol{\mu}_j$. Then (5.3b) asserts that $\mathbf{S}^{-1}$ is the $m \times m$ matrix whose ith row is $\boldsymbol{\mu}_i^T\mathbf{B}$. Therefore (5.9a) implies

$$\mathbf{S}^{-1}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{A}\mathbf{S} = \mathrm{diag}\,[\alpha_j (1 + \alpha_j)^{-1}] \tag{5.9b}$$

where $\mathrm{diag}\,[\beta_j]$ is the $m \times m$ diagonal matrix whose jth diagonal entry is β_j. For any matrix $\mathbf{Q}$, let $\mathrm{tr}\,\mathbf{Q}$ be its trace. Then $\mathrm{tr}\,\mathbf{S}^{-1}\mathbf{Q}\mathbf{S} = \mathrm{tr}\,\mathbf{Q}$, so (5.9b) and (5.8) imply

$$\mathrm{emp} = \mathrm{tr}\,\mathbf{R}, \tag{5.9c}$$

where

$$\mathbf{R} = (\mathbf{A} + \mathbf{B})^{-1}\mathbf{A} \tag{5.9d}$$

or

$$\mathbf{R} = \mathbf{V}_{PO}\mathbf{F}^T\mathbf{V}_\varepsilon^{-1}\mathbf{F}. \tag{5.9e}$$

In SI, the matrix **R** of (5.9e) is called the resolution matrix, and (5.9c) is arrived at by different arguments (Jackson 1979).

6 GEOMAGNETIC PRIORS ON THE CMB

Drawing inferences about B_r at the CMB from data obtained at and above Earth's surface requires *a priori* bounds on the spherical harmonics of high degree in B_r at the CMB. We consider three sources of such information: the results of geomagnetic studies done before the work of Gubbins & Bloxham, the heat flow bound, and the prior beliefs used by Gubbins & Bloxham. The last are deferred to the next section.

To establish a notation, let c be the radius of the core, let b be the radius of the Earth, and let $g_l^m(a)$ be a Schmidt quasi-normalized internal Gauss coefficient radius $r = a$. These coefficients parametrize the geomagnetic field **B** in the region $r > a$ as $\mathbf{B} = \sum_{l=1}^{\infty} \mathbf{B}_l$ where $\mathbf{B}_l = -\nabla\psi_l$ and

$$\psi_l(\mathbf{r}) = a(a/r)^{l+1} \sum_{m=-l}^{l} g_l^m(a) Y_l^m(\hat{\mathbf{r}}). \qquad (6.1)$$

In (6.1), **r** is the position vector relative to the centre of the Earth, $r = |\mathbf{r}|$, $\hat{\mathbf{r}} = \mathbf{r}/r$, and $Y_l^m(\hat{\mathbf{r}})$ is a surface spherical harmonic of degree l and longitudinal order m. The Y_l^m are Schmidt quasi-normalized and orthogonal; i.e. the average value of $Y_l^m(\hat{\mathbf{r}})^* Y_{l'}^{m'}(\hat{\mathbf{r}})$ on the unit sphere is $(2l+1)^{-1}$ if $l = l'$ and $m = m'$, and vanishes otherwise. The asterisk denotes complex conjugation, and is omitted if real harmonics are used. Let $R_l(a)$ denote the average of $|\mathbf{B}_l|^2$ on the spherical surface $S(a)$ of radius a. Then (Mauersberger 1956; Lucke 1957; Lowes 1966)

$$R_l(a) = (l+1) \sum_{m=-l}^{l} |g_l^m(a)|^2 \qquad (6.2)$$

Langel & Estes (1982) used satellite and observatory data to estimate $R_l(b)$ for $l \leq 23$, and Cain *et al.* (1986) refined their analysis and carried to $l \leq 45$. Both observers found that except for an anomalously large dipole term ($l = 1$), the dependence of $R_l(b)$ on l is well described by

$$R_l(b) = R_l^K(b) + R_l^C(b), \qquad (6.3a)$$

where, for $i = K$ or C,

$$R_l^i(b) = P_i(p_i)^l. \qquad (6.3b)$$

For Cain *et al.* $P_K = 1.365 \times 10^9\,\mathrm{nT}^2$, $p_K = 0.269$, $P_C = 29.6\,\mathrm{nT}^2$, $p_C = 0.977$; Langel and Estes found roughly the same values. With these values $R_l^K(b) \gg R_l^C(b)$ for $l \leq 12$ and $R_l^C(b) \gg R_l^K(b)$ for $l \geq 16$, while $R_l^K(b) = R_l^C(b)$ at $l = 13.6$. The now usual interpretation of this situation (Lowes 1974) is that $P_i(p_i)^l$ is produced by sources inside a sphere whose radius is approximately $a_i = bp_i^{1/2}$. Then $a_K = 3300$ km and $a_C = 6300$ km, so $R_l^K(b)$ is attributed to the core and $R_l^C(b)$ to the crust.

If this interpretation is accepted, then for $l \geq 16$ the magnetic signal at and above the surface of the Earth comes entirely from the crust. That signal for $l \geq 14$ contains no information about the core except that $R_l^K(b) \leq R_l^C(b)$. Since $R_l^K(c) = R_l^K(b)(b/c)^{2l+4}$, it follows from the P_C and p_C of Cain *et al.* (1986) that

$$R_l^K(c) \leq 330 \times (3.34)^l\,\mathrm{nT}^2 \text{ if } l \geq 14. \qquad (6.4)$$

The expression $R_l^K(b) = P_K(p_K)^l$ is verified from the data only for $2 \leq l \leq 12$. A bold observer might be willing to extrapolate this purely empirical relationship to all l. Then the P_K and p_K of Cain *et al.* (1986) give

$$R_l^K(c) \leq 1.52 \times 10^{10} \times (0.899)^l\,\mathrm{nT}^2 \text{ for } l \geq 2. \qquad (6.5)$$

The bound (6.5) is stricter than (6.4) at all l for which (6.4) is available. Both sets of bounds have the form (4.2) with $\boldsymbol{\mu}_2 = \mathbf{0}$, and can be softened as described in Section 4.

A bold observer, willing to adopt the strong bounds (6.5), would be able to extract from further data much more 'information' about B_r at the CMB than a cautious observer who confined himself to (6.4). Agreement of the bold observer's conclusions with observations in other areas or with theory might act on the cautious observer's personal probability distribution so as to incline him to something stronger than (6.4). For example, Busse's (1978) dynamo does have the symmetrical flux spots which G & B see at high northern and southern latitudes, adding weight to G & B's conclusions from a source not just derived from the data or their own prior beliefs. The weight would be even greater if it were established that the core dynamo must be a Busse dynamo.

The separation of core and crustal magnetic fields has been approached differently by Meyer *et al.* (1983), who use crustal geology to estimate crustal susceptibility on a global scale. The accuracy of the model is somewhat uncertain, both because of the limits on geological information and because permanent magnetization is neglected. As an alternative, Langel (Shure *et al.* 1985) has proposed that (6.3) can be used to estimate the total rms error produced by the crustal field in satellite measurements of the core field. Langel supposes that the crustal signal $R_l^C(b)$ is given by (6.3b) for all $l \leq 45$. Then $|\mathbf{B}^C|_a^2$, the square of the crustal field at radius $a > b$, averaged over the sphere of radius a, satisfies $|\mathbf{B}^C|_a^2 \geq 3\,\varepsilon_C(a)^2$ where

$$3\,\varepsilon_C(a)^2 = \sum_{l=1}^{45} P_C(p_C)^l (b/a)^{2l+4}.$$

The mean squared intensity of one cartesian component of $\mathbf{B}^C$ is $\frac{1}{3}|\mathbf{B}^C|_a^2$, or $\varepsilon_C(a)^2$. From the values of P_C and p_C found by Cain *et al.* (1986)

$$\varepsilon_C(a) = 3.2\zeta^{3/2}(1 - \zeta^{45})^{1/2}(1 - \zeta)^{-1/2}\,\mathrm{nT}, \qquad (6.6a)$$

where

$$\zeta = (0.977)(b/a)^2. \qquad (6.6b)$$

Table 1 gives $\varepsilon_C(a)$ at various altitudes $a - b$. Table 1 grossly underestimates $\varepsilon_C(b)$, but gives about the right value for the

Table 1. A lower bound for the global rms of one cartesian component of the crustal field at various altitudes above the Earth's surface.

altitude km	$\varepsilon_C(a)$ nT
0	16.4
100	12.2
200	9.8
300	8.1
400	7.00
500	6.16
600	5.50
800	4.53
1000	3.84

error remaining in observatory data if the station corrections obtained by satellite are included (Langel *et al.* 1982). If Langel's rather than Meyer's approach to crustal fields is adopted, then the probability distribution of the errors of measurement $\boldsymbol{\varepsilon}$ for the core field should include an error of the order of the entry in Table 1 corresponding to the altitude where **B** is measured. This is at least as large as the instrumental and navigational errors, so Langel's approach to crustal fields suggest treating the probability distribution of ε by subjective, Bayesian techniques. Section 4 shows how to do so.

Another source of prior information about B_r on the CMB is the heat-flow bound. From the pre-Maxwell equations (which omit displacement current and charge advection) and Ohm's law, Gubbins (1975) showed that if Φ is the rate of ohmic heat production in the core, if κ is the maximum electrical conductivity in the core, and if μ_0 is the magnetic permeability of the vacuum ($4\pi \times 10^{-7}\,\mathrm{N\,A^{-2}}$) then

$$\sum_{l=1}^{\infty} \frac{(l+1)(2l+1)(2l+3)}{l} \sum_{m=-l}^{l} |g_l^m(c)|^2 < Q \tag{6.7a}$$

where

$$Q = (4\pi c)^{-1} \Phi \kappa \mu_0^2. \tag{6.7b}$$

Taking $\kappa = 3 \times 10^5\,\mathrm{S\,m^{-1}}$, $c = 3.48 \times 10^6\,\mathrm{m}$, and $\Phi = 3 \times 10^{13}\,\mathrm{W}$ (the observed heat flow out of the Earth's surface) gives $Q = 0.3\,\mathrm{T}^2$ or $3 \times 10^{17}\,\mathrm{nT}^2$, as in (2.12). The series on the left in (6.7a) certainly converges, because the rate of ohmic heat production in the core is finite. The value of Q is, however, quite uncertain, and values ranging from 3×10^{16} to $3 \times 10^{18}\,\mathrm{nT}^2$ could probably be accepted, and ought to be explored by any observer who uses (6.7a) to construct a prior personal probability distribution for processing the observations of **B**. A gaussian prior personal probability distribution on any finite-dimensional space of Gauss coefficients is easy to construct from (6.7a), the procedure being that discussed in connection with (4.2), but it involves the second pitfall mentioned in Section 4.

If the observer is bold enough to adopt for (6.7a) convergence factors γ_i as in (4.13), then the analogue of (4.14) for the double sum (6.7a) would be

$$\gamma_l^m = 6(\pi l)^{-2}(2l+1)^{-1}. \tag{6.8}$$

The extra factor $(2l+1)^{-1}$ in (6.8) takes care of the sum over m in (6.7a). Such a bold observer will adopt on the model space M of sequences of Gauss coefficients a prior personal probability distribution which assigns to $g_l^m(c)$ the mean value zero and the variance

$$\langle |g_l^m(c)|^2 \rangle_{PR} = 6Q\pi^{-2} l^{-1}(l+1)^{-1}(2l+1)^{-2}(2l+3)^{-1}, \tag{6.9}$$

correlations between different g_l^m being zero because there are no cross terms in (6.7a). An observer too cautious to adopt convergence factors will accept only the safer claim

$$\langle |g_l^m(c)|^2 \rangle_{PR} = Ql(l+1)^{-1}(2l+1)^{-1}(2l+3)^{-1}, \tag{6.10}$$

which limits each Gauss coefficient only by requiring it not to produce by itself more ohmic heat in the core than is observed in the surface heat flow. When (6.10) is extrapolated to the Earth's surface, it is less than the crustal field limit (6.3b) for $l \geq 29$, so the heat flow constraint is stronger than one might expect. At present, the best a conservative observer could do would be to use (6.10) for $1 \leq l \leq 13$ and $29 \leq l$, and to use (6.4) for $14 \leq l \leq 28$. However, (6.3) was itself obtained by a regularization (truncation) aimed at solving the existence half of the geomagnetic inverse problem, not the uniqueness half. A purist would object to this circularity, and would accept for his prior only (6.10).

7 THE PRIORS USED BY GUBBINS & BLOXHAM

Gubbins (1983) and Gubbins & Bloxham (1985) argue that we do not know $\mathbf{V}_\varepsilon$ but only its shape, a $d \times d$ matrix $\mathbf{V}_0$ such that

$$\mathbf{V}_\varepsilon = \sigma^2 \mathbf{V}_0. \tag{7.1}$$

They estimate σ by fitting a model $\mathbf{x}$ to the data and regarding the vector of residuals as $\boldsymbol{\varepsilon}$. Gubbins (1983) and G & B (1985) take $\mathbf{V}_0$ to be diagonal and use various evidences of internal consistency in the data to estimate the ratios of the diagonal entries. The normalization of $\mathbf{V}_0$ is such that when $\sigma = 1$ the rms misfit of the model to the satellite data is 6 nT. In most of G & B's work, σ is between 1.0 and 1.3.

The prior personal probability distributions adopted by G & B (1985) for B_r on the CMB (where $r = c$) have uncorrelated Gauss coefficients with two different candidates for the variances:

$$\langle |g_l^m(c)|^2 \rangle_{PR} = (\sigma^2/4\pi\lambda)(b/c)^4(2l+1)l^{-p} \tag{7.2}$$

with $p = 5$ or 6 and the damping parameter λ to be determined. G & B use the heat flow bound to justify (7.2), and say that it is rigorous for $p = 5$ but not for $p = 6$. Indeed, for $p = 6$, (7.2) at large l is the same as (6.9) and requires the introduction of the arbitrary convergence factor (6.8). For $p = 5$, (7.2) is also not a rigorous consequence of the heat flow bound. The rigorous version of that bound, insofar as one can talk about rigor in BI, is (6.10), which corresponds to (7.2) with $p = 3$. The values $p = 4$ and $p = 5$ correspond to the intermediate γ_i mentioned in the last paragraph of Section 4.

In order not to be accused of timidity, we accept the bold heat flow bound (7.2) with $p = 6$. G & B (1985) give no details about how they choose λ, but Gubbins (1983) says that it is chosen 'to give a satisfactory fit to the data'. As they observe, if λ is too large, a posteriori models are determined mainly by the *a priori* beliefs that went into them rather than by the data, and then σ becomes unrealistically large. G & B regard $\sigma \approx 1$ as providing a satisfactory fit to the data. This leads to adopt values of λ between 5×10^{-13} and $10^{-11}\,\mathrm{nT}^{-2}$, but they consider all values of λ valid if the solutions are used only in conjunction with their error estimates (G & B 1985, p. 708). This position is crucial and seems inappropriate to the uniqueness half of Bayesian inversion, in which λ describes an *a priori* belief.

Equation (7.2) with $p = 6$ is essentially the same as (6.9) for large l if we make the identification

$$\lambda = 2\pi(3Q)^{-1}(b/c)^4\sigma^2\,\mathrm{nT}^{-2}. \tag{7.3a}$$

If $\sigma \approx 1$ then (7.3a) is

$$\lambda = 23/Q \text{ nT}^{-2}. \tag{7.3b}$$

As remarked in Section 5, equating ohmic heat production in the core with the heat flow out of the Earth's surface gives $Q = 3 \times 10^{17}$ nT^2, but a value $3 \times 10^{16} < Q < 3 \times 10^{18}$ is probably acceptable to most geophysicists. In the spirit of BI, uncertainties should never shrink the variance of a personal probability distribution, so we probably ought to work with $Q = 3 \times 10^{18}$ nT^2. Again giving G & B the benefit of a doubt, we adopt $Q = 3 \times 10^{17}$ nT^2. Then (7.3b) gives

$$\lambda = 8 \times 10^{-17} \text{ nT}^{-2}. \tag{7.4}$$

This is 6000 times smaller than G & B's smallest λ, and 1.2×10^5 times smaller than their heavily damped models. Their choice of λ cannot be justified from the heat flow bound.

To state the question in another way, G & B (1985) appear to have approached the data with a prior personal probability distribution for B_r on the CMB which implies, from the hardened form of (7.2) with $p = 6$ and $\lambda = 5 \times 10^{-13}$, that

$$R_l^K(c) \leq 3.5 \times 10^{13}(l + 1/2)^2 l^{-6} \text{ nT}^2. \tag{7.5}$$

The belief (7.5) cannot be justified from the heat flow bound at any l. That belief amounts to assuming that the ohmic heating rate in the core is no more than 1/6000 of the geothermal heat flow at the Earth's surface. The belief (6.4) will justify (7.5) only for $l \leq 14$, where (6.4) is unavailable (this appears to be a numerical coincidence; for $\lambda = 10^{-11}$, (6.4) justifies (7.5) only for $l \leq 12.5$). The bold belief (6.5) will justify (7.5) for $1 \leq l \leq 24$, but the evidence for (6.5) is only in $1 \leq l \leq 12$ or possibly (Shure *et al.* 1985) $1 \leq l \leq 10$. Whether to accept the belief (7.5) on such evidence is a personal matter, but (7.5) is certainly not the heat flow bound.

The real basis for G & B's choice of λ seems to be that they increase it until the fit to the data becomes unacceptable. But this means that they are solving the existence half of the inverse problem, not the uniqueness half. They have found models which do fit the data within 10 nT, but they can give no information about the other physically reasonably models which do so. If they are right that the errors of observation of the core field are of the order of 10 nT, then using the heat flow bound as a solution to the uniqueness problem requires that their calculations be repeated with a damping parameter no larger than (7.4)–6000 times smaller than their smallest. Table 1 suggests that G & B's 10 nT is a reasonable estimate of the errors of observation of the core field.

Some observers will be reluctant to go beyond the heat flow bound in constructing prior beliefs about B_r at the CMB. Those cautious observers will be unable to accept that G & B have established their conclusions until they repeat their calculations with a much smaller damping parameter. Some suggestions about the possible outcome of such recalculation can be found in Shure *et al.* (1985). They adopt an observational error for the core field of 10 nT, and approach the existence half of the core field modelling problem in a new way which produces as good a fit to the data as any other models except G & B's. Their model agrees with the models of Langel & Estes (1985) and Cain *et al.* (1986) for $1 \leq l \leq 10$ but not for higher l. This suggests that for $l > 10$ the core field may not be resolvable from the surface. If so, the circle of confusion (averaging disc) imposed on the CMB has a diameter of about 35° (Booker 1968) and it seems unlikely that the small-scale features reported by G & B will be resolvable.

Among the workers active in main field geomagnetism, there has been a reluctance to use the small damping parameters proposed here, partly for reasons of numerical stability and partly out of concern that such small λs would produce no useful error bounds even for the dipole. The former concern is amenable to modern numerical techniques, and the latter has not been tested. My own conjecture is that the safe heat flow bound (6.10) will give accurate values for the Gauss coefficients up to $l = 10$. The possibility of invoking other prior information besides heat flow should be energetically explored.

To many users of SI, the fact that weaker prior information calls for a decreased damping parameter λ will seem paradoxical, as it increases the roughness and thus the apparent resolution of the posterior model $\mathbf{x}_{PO}$ and increases the number of 'resolved' parameters, (5.8). The point is that a parameter is 'resolved' if its *prior* variance is significantly decreased by the data. As shown in Appendix C, decreasing λ will increase the posterior variance of the model parameters whose mean values produce the roughness, and some reasonable models which fit the data will not be rough. The roughness in $\mathbf{x}_{PO}$ will not really be resolved by the data, and in fact the resolution will be poorer for smaller λ. In the existence half of the inverse problem, the observer avoids such small λs because they complicate the model and destabilize the computation without much improving the fit to the data. In the uniqueness half of the inverse problem, if no prior information is available, probably the observer will also avoid such small λ; they usually lie on a part of the trade-off curve where the variances of the averages calculable from the data alone are too large to be useful (Backus & Gilbert 1970). It is only the observer using quantified prior information to solve the uniqueness problem who may be forced to use such small λ when he estimates a $\mathbf{g}^T\mathbf{x}$ not calculable from the data alone (i.e. when $\mathbf{g}$ is not in $M_{\mathbf{F}}$).

8 CONCLUSIONS

In most linear geophysical inverse problems, the uniqueness half appears to be insoluble without the aid of 'infinite-dimensional beliefs', beliefs of a character designed to truncate the infinitely long cylinder of models capable of fitting all the data with acceptable accuracy. Such beliefs can take the 'hard' form of inequalities or the 'soft' form of probability distributions in the model space M. The 'soft' beliefs are analytically quite tractable, and the arithmetic is the same for stochastic inversion (SI) as for Bayesian inference (BI), if the probability distributions which describe prior beliefs and errors are both gaussian. If an inverse problem can be viewed as a sampling problem, in which at least in imagination there are many drawings of sample models $\mathbf{x}$ from M, each followed by an inversion of the data, and if there is objective quantitative evidence about the sampling distribution in M, then SI is the preferred interpretation of the calculations. If there is only one true model, and no sources of prior information about it except the observer's whole body of knowledge, then BI gives a more accurate description of how an observer inverts data. Therefore, in such 'one-shot' inversions, BI gives the observer clearer guidelines than SI for quantifying his subjective prior beliefs. BI has the disadvantage of being subjective, but the advantage of great generality. It generalizes easily to non-linear problems and BI via (4.8) incorporates a wider class of data and beliefs than SI; in particular, the $\mathbf{C}_i$ can overlap.

When the general guidelines for comparing SI with BI are applied to the recent work of Gubbins & Bloxham (G & B) on modelling the geomagnetic field at the core–mantle boundary, it appears that they are well-advised to abandon SI for BI in their work. However, their method of choosing their damping parameter λ is appropriate to the existence problem rather than the uniqueness problem. They choose λ to permit a reasonable model to fit the data; they do not defend their λ on the basis of prior beliefs. Therefore, G & B have found a model of B_r at the CMB which is physically reasonable and fits the data, but they have not settled the question of how many such models exist, or whether all these models share in common the properties which they attribute to B_r at the CMB. To do so would require of an observer whose prior belief is the heat flow bound that he use a λ at least 6000 times smaller than the smallest λ used by G & B. The resulting core field would be more poorly resolved than G & B's at the CMB, and some published work suggests that the diameter of the resolving circle (averaging disc) for B_r on the CMB might be as large as 35°.

It will be of great interest to repeat G & B's calculations with a much smaller damping parameter, and to explore other prior beliefs besides the heat flow bound which are not circular (i.e. not based on the outcome of accepting them).

ACKNOWLEDGMENTS

So many colleagues have shared discussion of these topics that it is impractical to thank them all. J. Bloxham, C. Constable, D. Gubbins, V. Keilis-Borok, R. Langel, R. Parker, A. Tarantola and C. Voorhies have made especially incisive comments. The work was supported in part by NSF grants EAR 85-21453 and EAR 86-07014 and in part by NASA grant NAG 5-818.

REFERENCES

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

Backus, G., 1970b. Inference from inadequate and inaccurate data, III, *Proc. Natl. Acad. Sci.*, **67,** 282–289.

Backus, G., 1971. Inference from inadequate and inaccurate data, *Lectures in Applied Mathematics,* **14,** 1–105, Amer. Math. Soc., Providence, R.I.

Backus, G. & Gilbert, J. F., 1967. Numerical applications of a formalism for geophysical inverse problems, *Geophys. J. R. astr. Soc.*, **13,** 247–276.

Backus, G. & Gilbert, J. F., 1968. The resolving power of gross earth data, *Geophys. J. R. astr. Soc.* **16,** 169–205.

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

Bayes, T., 1764. An essay towards solving a problem in the doctrine of chances, *Phil. Trans. Roy. Soc.*, **53,** 370–418.

Berger, J. O., 1985. *Statistical Decision Theory and Bayesian Analysis,* Springer-Verlag, New York.

Bloxham, J., 1986. Models of the magnetic field at the core–mantle boundary for 1715, 1777 and 1842, *J. geophys. Res.*, **91,** 13954–13966.

Bloxham, J. & Gubbins, D., 1986. Geomagnetic field analysis-IV. Testing the frozen flux hypothesis, *Geophys. J. R. astr. Soc.*, **84,** 139–152.

Booker, J. R., 1969. Geomagnetic data and core motions, *Proc. R. Soc. A,* **309,** 27–40.

Box, G. & Tiao, G., 1973. *Bayesian Inference in Statistical Analysis,* Addison-Wesley, Reading, MA.

Busse, F., 1978. Magnetohydrodynamics of the Earth's dynamo, *Ann. Rev. Fluid Mech.*, **10,** 435–462.

Cain, J., Wang, Z. & Schmitz, D., 1986. The geomagnetic modal spectrum for 1980 and core-crustal separation, submitted to *PEPI.*

Constable, C. & Parker, R., 1988. Submitted to *J. geophys. Res.*

Franklin, J., 1970. Well-posed stochastic extensions of ill-posed linear problems, *J. Math. Analysis Applic.*, **31,** 682–716.

Gubbins, D., 1975. Can the Earth's magnetic field be sustained by core oscillations? *Geophys. Res. Lett.*, **2,** 409–412.

Gubbins, D., 1983. Geomagnetic field analysis-I. Stochastic inversion, *Geophys. J. R. astr. Soc.*, **73,** 641–652.

Gubbins, D., 1984. Geomagnetic field analysis-II. Secular variation consistent with a perfectly conducting core, *Geophys. J. R. astr. Soc.,* **77,** 753–766.

Gubbins, D. & Bloxham, J., 1985. Geomagnetic field analysis-III. Magnetic fields on the core–mantle boundary, *Geophys. J. R. astr. Soc.*, **80,** 695–714.

Halmos, P. R., 1950. *Measure Theory,* Van Nostrand, New York.

Halmos, P. R., 1958. *Finite-Dimensional Vector Spaces,* Van Nostrand, New York.

Jackson, D., 1979. The use of *a priori* data to resolve non-uniqueness in linear inversion, *Geophys. J. R. astr. Soc.*, **57,** 137–158.

Jeffreys, H., 1961. *Theory of Probability,* 3rd edition. Clarendon Press, Oxford.

Kolmogorov, A. N., 1950. *Foundations of Probability,* Chelsea, New York.

Langel, R. & Estes, R., 1982. A geomagnetic field spectrum, *Geophys. Res. Lett.*, **9,** 250–253.

Langel, R. & Estes, R., 1985. The near-earth magnetic field at 1980 determined from Magsat data, *J. Geophys. R.*, **90,** 2495–2509.

Langel, R., Estes, R. & Mead, G., 1982. Some new methods in geomagnetic field modelling applied to the 1960–1980 epoch, *J. Geomagn. Geoelect.*, **34,** 327–349.

Lindley, D., 1965. *Introduction to Probability and Statistics,* Cambridge University Press, Vol. 1, 259, pp; Vol. 2, 292 pp. New York.

Loewner, K., 1939. Grundzüge einer Inhaltslehre im Hilbertschen Raume, *Ann. Math.* (2), **40,** 816–833.

Lowes, F. J., 1966. Mean-square values on sphere of spherical harmonic vector fields, *J. geophys. Res.*, **71,** 2179.

Lowes, F. J., 1974. Spatial power spectrum of the main geomagnetic field, and extrapolation to the core, *Geophys. J. R. astr. Soc.*, **36,** 717–730.

Lucke, O., 1957. Uber Mittelwerte von Energiedichten der Kraftfelder, Wiss, Z. Päd. Hochschule Potsdam, *Math.-Nat. Reihe* **3,** 39–46.

Mauersberger, P., 1956. Das Mittel der Energiedichte des geomagnetischen Hauptfeldes an der Erdoberfläche und seine säkulare Änderung, *Gerlands Beitr. Geophys.*, **65,** 207–215.

Meyer, J., Hufen, J.-H., Siebert, M. & Hahn, A., 1983. Investigations of the internal geomagnetic field by means of a global model of the Earth's crust, *J. Geophys.*, **52,** 71–84.

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

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

Savage, L. J., 1972. *The Foundations of Statistics,* Dover, New York.

Shure, L., Parker, R. & Backus, G., 1982. Harmonic splines for geomagnetic modelling, *Phys. Earth Planet. Int.*, **28,** 215–229.

Shure, L., Parker, R. & Langel, R., 1985. A preliminary harmonic spline model from magsat data, *J. geophys. Res.*, **90,** 11505–11512.

Tarantola, A. & Valette, B., 1982. Generalized nonlinear inverse problems solved using the least squares criterion, *Rev. Geophys. Space Phys.*, **20,** 219–232.

Theil, H., 1963. On the use of incomplete prior information in regression analysis, *J. Am. statist. Ass.*, **58,** 401–414.

Tikhonov, A. N., 1963. Regularization of ill-posed problems, *Doklady Akad. Nauk SSSR*, **153,** 1–6.

Tikhonov, A. N. & Arsenin, V. Y., 1972. *Solutions of Ill-Posed Problems* (English translation 1977, Winston, New York.

Wiener, N., 1949. *The Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*, John Wiley, New York.

APPENDIX A

How a quadratic bound truncates the geomagnetic model space to finite dimensionality

The Schmidt quasi-normalized Gauss coefficients $g_l^m(a)$ at the CMB ($r=a$) must satisfy (2.12) if the core generates ohmic heat at a rate less than the geothermal heat flux at the Earth's surface. Those Gauss coefficients must satisfy (2.13) because the rest mass of the geomagnetic field energy is less than that of the Earth. Both (2.12) and (2.13) are of the form (2.2). We will show here that for all $\varepsilon > 0$ (2.2) provides a function $l^*(\varepsilon)$ such that the $g_l^m(a)$ with $l > l^*(\varepsilon)$ cannot together produce a magnetic field as large as ε anywhere on or above the surface of the Earth ($r \geq b$).

In the notation of (6.1), for any integer l^* write

$$\mathbf{B}^{(l^*)} = \sum_{l=l^*+1}^{\infty} \mathbf{B}_l. \tag{A1}$$

It will suffice to produce from (2.2) a decreasing function $\varepsilon(l^*)$ such that $\varepsilon(l^*) \rightarrow 0$ as $l^* \rightarrow \infty$ and

$$|\mathbf{B}^{(l^*)}(r\hat{\mathbf{r}})| < \varepsilon(l^*) \tag{A2}$$

for all $r \geq b$. Then $l^*(\varepsilon)$ is the inverse function.

Define $\boldsymbol{\nabla}_1$ by the equation

$$\boldsymbol{\nabla} = \hat{\mathbf{r}}\, \partial_r + r^{-1} \boldsymbol{\nabla}_1. \tag{A3}$$

Then in any system r, θ, λ of spherical polar coordinates,

$$\boldsymbol{\nabla}_1 = \hat{\boldsymbol{\theta}}\, \partial_\theta + \hat{\boldsymbol{\lambda}} \operatorname{cosec} \theta\, \partial_\lambda.$$

Define $\mathbf{b}_l^m(\hat{\mathbf{r}}) = a\boldsymbol{\nabla}(a/r)^{l+1} Y_l^m(\hat{\mathbf{r}})$ at $r = a$, so

$$\mathbf{b}_l^m(\hat{\mathbf{r}}) = \hat{\mathbf{r}}(l+1) Y_l^m(\hat{\mathbf{r}}) - \boldsymbol{\nabla}_1 Y_l(\hat{\mathbf{r}}). \tag{A4}$$

Then from (6.1) it follows that

$$\mathbf{B}_l(r\hat{\mathbf{r}}) = (a/r)^{l+2} \sum_{m=-l}^{l} g_l^m(a) \mathbf{b}_l^m(\hat{\mathbf{r}}). \tag{A5}$$

By Schwarz's inequality,

$$|\mathbf{B}_l(a\hat{\mathbf{r}})|^2 \leq \left[\sum_{m=-l}^{l} |g_l^m(a)|^2\right]\left[\sum_{m=-l}^{l} |\mathbf{b}_l^m(\hat{\mathbf{r}})|^2\right] \tag{A6}$$

The key observation which continues the proof is that

$$\sum_{m=-l}^{l} |\mathbf{b}_l^m(\hat{\mathbf{r}})|^2 = (l+1)(2l+1). \tag{A7}$$

for all unit vectors $\hat{\mathbf{r}}$. It is easier to prove a generalization of (A7), namely

$$\sum_{m=-l}^{l} \mathbf{b}_l^m(\hat{\mathbf{r}}) \cdot \mathbf{b}_l^m(\hat{\mathbf{s}})^* = (2l+1) \times [l+1)\mu P_l(\mu) - (1-\mu^2)\, \partial_\mu P_l(\mu)] \tag{A8}$$

where $\hat{\mathbf{r}}$ and $\hat{\mathbf{s}}$ are any unit vectors, $\mu = \hat{\mathbf{r}} \cdot \hat{\mathbf{s}}$, and P_l is the lth Legendre polynomial. The proof of (A8) begins with the addition theorem for spherical harmonics,

$$\sum_{m=-l}^{l} Y_l^m(\hat{\mathbf{r}}) Y_l^m(\hat{\mathbf{s}})^* = P_l(\hat{\mathbf{r}} \cdot \hat{\mathbf{s}}). \tag{A9}$$

Writing the gradients with respect to the two independent position variables $\mathbf{r}$ and $\mathbf{s}$ as $\boldsymbol{\nabla}_r$ and $\boldsymbol{\nabla}_s$, we define both $\boldsymbol{\nabla}_{1r}$ and $\boldsymbol{\nabla}_{1s}$ from (A3) as

$$\boldsymbol{\nabla}_r = \hat{\mathbf{r}}\, \partial_r + r^{-1} \boldsymbol{\nabla}_{1r} \tag{A10a}$$

$$\boldsymbol{\nabla}_s = \hat{\mathbf{s}}\, \partial_s + s^{-1} \boldsymbol{\nabla}_{1s}. \tag{A10b}$$

Clearly $\boldsymbol{\nabla}_r \mathbf{r} = \mathbf{I}$, the 3-D identity tensor, so (A10a) implies

$$\boldsymbol{\nabla}_{1r} \hat{\mathbf{r}} = \mathbf{I} - \hat{\mathbf{r}}\hat{\mathbf{r}}.$$

Similarly

$$\boldsymbol{\nabla}_{1s} \hat{\mathbf{s}} = \mathbf{I} - \hat{\mathbf{s}}\hat{\mathbf{s}}.$$

Therefore, if $\mu = \hat{\mathbf{r}} \cdot \hat{\mathbf{s}}$, then

$$\boldsymbol{\nabla}_{1r} \mu = \hat{\mathbf{s}} - \hat{\mathbf{r}}\mu$$

and

$$\boldsymbol{\nabla}_{1s} \mu = \hat{\mathbf{r}} - \hat{\mathbf{s}}\mu.$$

Then applying $\boldsymbol{\nabla}_{1r}$ to (A9) gives

$$\sum_{m=-l}^{l} [\boldsymbol{\nabla}_1 Y_l^m(\hat{\mathbf{r}})] Y_l^m(\hat{\mathbf{s}})^* = (\hat{\mathbf{s}} - \mu\hat{\mathbf{r}})\, \partial_\mu P_1(\mu), \tag{A11a}$$

and applying $\boldsymbol{\nabla}_{1s}$ to (A9) gives

$$\sum_{m=-l}^{l} Y_l^m(\hat{\mathbf{r}}) [\boldsymbol{\nabla}_1 Y_l^m(\hat{\mathbf{s}})]^* = (\hat{\mathbf{r}} - \mu\hat{\mathbf{s}})\, \partial_\mu P_l(\mu). \tag{A11b}$$

Applying $\boldsymbol{\nabla}_{1r}$ to (A11b) gives

$$\sum_{m=-l}^{l} [\boldsymbol{\nabla}_1 Y_l^m(\hat{\mathbf{r}})][\boldsymbol{\nabla}_1 Y_l^m(\hat{\mathbf{s}})]^* = [\mathbf{I} - \hat{\mathbf{r}}\hat{\mathbf{r}} - \hat{\mathbf{s}}\hat{\mathbf{s}} + \hat{\mathbf{r}}\hat{\mathbf{s}}\mu]\, \partial_\mu P_l + (\hat{\mathbf{s}} - \hat{\mathbf{r}}\mu)(\hat{\mathbf{r}} - \hat{\mathbf{s}}\mu)\, \partial_\mu^2 P_l. \tag{A12}$$

From (A4), (A9), (A11) and (A12) it follows that

$$\sum_{m=-l}^{l} \mathbf{b}_l^m(\hat{\mathbf{r}}) \mathbf{b}_l^m(\hat{\mathbf{s}})^* = \hat{\mathbf{r}}\hat{\mathbf{s}}(l+1)^2 P_l(\mu) + [I - (l+2)(\hat{\mathbf{r}}\hat{\mathbf{r}} + \hat{\mathbf{s}}\hat{\mathbf{s}}) + (2l+3)\mu\hat{\mathbf{r}}\hat{\mathbf{s}}]\, \partial_\mu P_l(\mu) + (\hat{\mathbf{s}} - \hat{\mathbf{r}}\mu)(\hat{\mathbf{r}} - \hat{\mathbf{s}}\mu)\, \partial_\mu^2 P_i(\mu). \tag{A13}$$

Taking traces on both sides of (A13), and using Legendre's equation to replace $(1-\mu^2)\, \partial_\mu^2 P_l$ by $2\mu\, \partial P_l - l\,(l+1) P_l$ gives (A8). Parenthetically, one can set $\hat{\mathbf{r}} = \hat{\mathbf{s}}$ in (A13), to obtain

$$\sum_{m=-l}^{l} \mathbf{b}_l^m(\hat{\mathbf{r}}) \mathbf{b}_l^m(\hat{\mathbf{r}})^* = \tfrac{1}{2} l(l+1)\mathbf{I} + \tfrac{1}{2}(l+1)(l+2)\hat{\mathbf{r}}\hat{\mathbf{r}}, \tag{A14}$$

since $P_l(1) = 1$ and $\partial_\mu P_l(1) = l(l+1)/2$. Taking traces on both sides of (A14) gives (A7) directly.

Having established (A7), we can write (A6) as

$$|\mathbf{B}_l(a\hat{\mathbf{r}})|^2 \leq (l+1)(2l+1) \sum_{m=-l}^{l} |g_l^m(a)|^2.$$

spline model from magsat data, *J. geophys. Res.*, **90,** 11505–11512.

Tarantola, A. & Valette, B., 1982. Generalized nonlinear inverse problems solved using the least squares criterion, *Rev. Geophys. Space Phys.*, **20,** 219–232.

Theil, H., 1963. On the use of incomplete prior information in regression analysis, *J. Am. statist. Ass.*, **58,** 401–414.

Tikhonov, A. N., 1963. Regularization of ill-posed problems, *Doklady Akad. Nauk SSSR*, **153,** 1–6.

Tikhonov, A. N. & Arsenin, V. Y., 1972. *Solutions of Ill-Posed Problems* (English translation 1977, Winston, New York.

Wiener, N., 1949. *The Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications,* John Wiley, New York.

APPENDIX A

How a quadratic bound truncates the geomagnetic model space to finite dimensionality

The Schmidt quasi-normalized Gauss coefficients $g_l^m(a)$ at the CMB ($r = a$) must satisfy (2.12) if the core generates ohmic heat at a rate less than the geothermal heat flux at the Earth's surface. Those Gauss coefficients must satisfy (2.13) because the rest mass of the geomagnetic field energy is less than that of the Earth. Both (2.12) and (2.13) are of the form (2.2). We will show here that for all $\varepsilon > 0$ (2.2) provides a function $l^*(\varepsilon)$ such that the $g_l^m(a)$ with $l > l^*(\varepsilon)$ cannot together produce a magnetic field as large as ε anywhere on or above the surface of the Earth ($r \geq b$).

In the notation of (6.1), for any integer l^* write

$$\mathbf{B}^{(l^*)} = \sum_{l=l^*+1}^{\infty} \mathbf{B}_l. \tag{A1}$$

It will suffice to produce from (2.2) a decreasing function $\varepsilon(l^*)$ such that $\varepsilon(l^*) \to 0$ as $l^* \to \infty$ and

$$|\mathbf{B}^{(l^*)}(r\hat{\mathbf{r}})| < \varepsilon(l^*) \tag{A2}$$

for all $r \geq b$. Then $l^*(\varepsilon)$ is the inverse function.

Define $\boldsymbol{\nabla}_1$ by the equation

$$\boldsymbol{\nabla} = \hat{\mathbf{r}}\, \partial_r + r^{-1} \boldsymbol{\nabla}_1. \tag{A3}$$

Then in any system r, θ, λ of spherical polar coordinates,

$$\boldsymbol{\nabla}_1 = \hat{\boldsymbol{\theta}}\, \partial_\theta + \hat{\boldsymbol{\lambda}} \operatorname{cosec} \theta\, \partial_\lambda.$$

Define $\mathbf{b}_l^m(\hat{\mathbf{r}}) = a \boldsymbol{\nabla}(a/r)^{l+1} Y_l^m(\hat{\mathbf{r}})$ at $r = a$, so

$$\mathbf{b}_l^m(\hat{\mathbf{r}}) = \hat{\mathbf{r}}(l+1) Y_l^m(\hat{\mathbf{r}}) - \boldsymbol{\nabla}_1 Y_l(\hat{\mathbf{r}}). \tag{A4}$$

Then from (6.1) it follows that

$$\mathbf{B}_l(r\hat{\mathbf{r}}) = (a/r)^{l+2} \sum_{m=-l}^{l} g_l^m(a) \mathbf{b}_l^m(\hat{\mathbf{r}}). \tag{A5}$$

By Schwarz's inequality,

$$|\mathbf{B}_l(a\hat{\mathbf{r}})|^2 \leq \left[\sum_{m=-l}^{l} |g_l^m(a)|^2 \right] \left[\sum_{m=-l}^{l} |\mathbf{b}_l^m(\hat{\mathbf{r}})|^2 \right] \tag{A6}$$

The key observation which continues the proof is that

$$\sum_{m=-l}^{l} |\mathbf{b}_l^m(\hat{\mathbf{r}})|^2 = (l+1)(2l+1). \tag{A7}$$

for all unit vectors $\hat{\mathbf{r}}$. It is easier to prove a generalization of (A7), namely

$$\sum_{m=-l}^{l} \mathbf{b}_l^m(\hat{\mathbf{r}}) \cdot \mathbf{b}_l^m(\hat{\mathbf{s}})^* = (2l+1) \\ \times [l+1)\mu P_l(\mu) - (1-\mu^2)\, \partial_\mu P_l(\mu)] \tag{A8}$$

where $\hat{\mathbf{r}}$ and $\hat{\mathbf{s}}$ are any unit vectors, $\mu = \hat{\mathbf{r}} \cdot \hat{\mathbf{s}}$, and P_l is the lth Legendre polynomial. The proof of (A8) begins with the addition theorem for spherical harmonics,

$$\sum_{m=-l}^{l} Y_l^m(\hat{\mathbf{r}}) Y_l^m(\hat{\mathbf{s}})^* = P_l(\hat{\mathbf{r}} \cdot \hat{\mathbf{s}}). \tag{A9}$$

Writing the gradients with respect to the two independent position variables $\mathbf{r}$ and $\mathbf{s}$ as $\boldsymbol{\nabla}_r$ and $\boldsymbol{\nabla}_s$, we define both $\boldsymbol{\nabla}_{1r}$ and $\boldsymbol{\nabla}_{1s}$ from (A3) as

$$\boldsymbol{\nabla}_r = \hat{\mathbf{r}}\, \partial_r + r^{-1} \boldsymbol{\nabla}_{1r} \tag{A10a}$$

$$\boldsymbol{\nabla}_s = \hat{\mathbf{s}}\, \partial_s + s^{-1} \boldsymbol{\nabla}_{1s}. \tag{A10b}$$

Clearly $\boldsymbol{\nabla}_r \mathbf{r} = \mathbf{I}$, the 3-D identity tensor, so (A10a) implies

$$\boldsymbol{\nabla}_{1r} \hat{\mathbf{r}} = \mathbf{I} - \hat{\mathbf{r}}\hat{\mathbf{r}}.$$

Similarly

$$\boldsymbol{\nabla}_{1s} \hat{\mathbf{s}} = \mathbf{I} - \hat{\mathbf{s}}\hat{\mathbf{s}}.$$

Therefore, if $\mu = \hat{\mathbf{r}} \cdot \hat{\mathbf{s}}$, then

$$\boldsymbol{\nabla}_{1r} \mu = \hat{\mathbf{s}} - \hat{\mathbf{r}}\mu$$

and

$$\boldsymbol{\nabla}_{1s} \mu = \hat{\mathbf{r}} - \hat{\mathbf{s}}\mu.$$

Then applying $\boldsymbol{\nabla}_{1r}$ to (A9) gives

$$\sum_{m=-l}^{l} [\boldsymbol{\nabla}_1 Y_l^m(\hat{\mathbf{r}})] Y_l^m(\hat{\mathbf{s}})^* = (\hat{\mathbf{s}} - \mu\hat{\mathbf{r}})\, \partial_\mu P_1(\mu), \tag{A11a}$$

and applying $\boldsymbol{\nabla}_{1s}$ to (A9) gives

$$\sum_{m=-l}^{l} Y_l^m(\hat{\mathbf{r}}) [\boldsymbol{\nabla}_1 Y_l^m(\hat{\mathbf{s}})]^* = (\hat{\mathbf{r}} - \mu\hat{\mathbf{s}})\, \partial_\mu P_l(\mu). \tag{A11b}$$

Applying $\boldsymbol{\nabla}_{1r}$ to (A11b) gives

$$\sum_{m=-l}^{l} [\boldsymbol{\nabla}_1 Y_l^m(\hat{\mathbf{r}})][\boldsymbol{\nabla}_1 Y_l^m(\hat{\mathbf{s}})]^* \\ = [\mathbf{I} - \hat{\mathbf{r}}\hat{\mathbf{r}} - \hat{\mathbf{s}}\hat{\mathbf{s}} + \hat{\mathbf{r}}\hat{\mathbf{s}}\mu]\, \partial_\mu P_l + (\hat{\mathbf{s}} - \hat{\mathbf{r}}\mu)(\hat{\mathbf{r}} - \hat{\mathbf{s}}\mu)\, \partial_\mu^2 P_l. \tag{A12}$$

From (A4), (A9), (A11) and (A12) it follows that

$$\sum_{m=-l}^{l} \mathbf{b}_l^m(\hat{\mathbf{r}}) \mathbf{b}_l^m(\hat{\mathbf{s}})^* = \hat{\mathbf{r}}\hat{\mathbf{s}}(l+1)^2 P_l(\mu) \\ + [\boldsymbol{I} - (l+2)(\hat{\mathbf{r}}\hat{\mathbf{r}} + \hat{\mathbf{s}}\hat{\mathbf{s}}) + (2l+3)\mu\hat{\mathbf{r}}\hat{\mathbf{s}}]\, \partial_\mu P_l(\mu) \\ + (\hat{\mathbf{s}} - \hat{\mathbf{r}}\mu)(\hat{\mathbf{r}} - \hat{\mathbf{s}}\mu)\, \partial_\mu^2 P_i(\mu). \tag{A13}$$

Taking traces on both sides of (A13), and using Legendre's equation to replace $(1-\mu^2)\, \partial_\mu^2 P_l$ by $2\mu\, \partial P_l - l\,(l+1)P_l$ gives (A8). Parenthetically, one can set $\hat{\mathbf{r}} = \hat{\mathbf{s}}$ in (A13), to obtain

$$\sum_{m=-l}^{l} \mathbf{b}_l^m(\hat{\mathbf{r}}) \mathbf{b}_l^m(\hat{\mathbf{r}})^* = \tfrac{1}{2} l(l+1)\mathbf{I} + \tfrac{1}{2}(l+1)(l+2)\hat{\mathbf{r}}\hat{\mathbf{r}}, \tag{A14}$$

since $P_l(1) = 1$ and $\partial_\mu P_l(1) = l(l+1)/2$. Taking traces on both sides of (A14) gives (A7) directly.

Having established (A7), we can write (A6) as

$$|\mathbf{B}_l(a\hat{\mathbf{r}})|^2 \leq (l+1)(2l+1) \sum_{m=-l}^{l} |g_l^m(a)|^2.$$

But from (A5) and (A1),

$$\mathbf{B}^{(l^*)}(r\hat{\mathbf{r}}) = \sum_{l=l^*+1}^{\infty} (a/r)^{l+2}\mathbf{B}_l(a\hat{\mathbf{r}})$$

$$= \sum_{l=l^*+1}^{\infty} K_l^{-1}(a/r)^{l+2}K_l\mathbf{B}_l(a\hat{\mathbf{r}})$$

where the K_l are any non-zero constants. Hence, by Schwarz's inequality,

$$|\mathbf{B}^{(l^*)}(r\hat{\mathbf{r}})|^2 \le \left[\sum_{l=l^*+1}^{\infty} K_l^{-2}(a/r)^{2l+4}\right]\left[\sum_{l=l^*+1}^{\infty} K_l^2\,|\mathbf{B}_l(a\hat{\mathbf{r}})|^2\right]$$

$$\le \left[\sum_{l=l^*+1}^{\infty} K_l^{-2}(a/r)^{2l+4}\right]$$

$$\times\left[\sum_{l=l^*+1}^{\infty} K_l^2(l+1)(2l+1)\sum_{m=-l}^{l}|g_l^m(a)|^2\right].$$

If the Gauss coefficients satisfy (2.2) and we choose K_l so that

$$K_l^2(l+1)(2l+1) = C_l$$

then (2.2) implies

$$|\mathbf{B}^{(l^*)}(r\hat{\mathbf{r}})|^2 \le Q \sum_{l=l^*+1}^{\infty} \frac{(l+1)(l+2)}{C_l}(a/r)^{2l+4}.$$

Thus, finally, if $r \ge b$, then

$$|\mathbf{B}^{(l^*)}(r\hat{\mathbf{r}})| \le \varepsilon(l^*) \tag{A15}$$

where

$$\varepsilon(l^*)^2 = Q \sum_{l=l^*+1}^{\infty} \frac{(l+1)(l+2)}{C_l}(a/b)^{2l+4}. \tag{A16}$$

For the mass bound, (2.13) gives $C_l = (l+1)/(2l+1)$, and (A16) can be summed in closed form to give

$$\varepsilon(l^*)^2 = QF[l^*, (a/b)^2]$$

where

$$F[L, x] = x^{L+3}\left[\frac{(2L+1)}{(1-x)} + \frac{8L}{(1-x)^2} + \frac{8}{(1-x)^3}\right].$$

For the heat-flow bound, $C_l = l^{-1}(l+1)(2l+1)(2l+3)$, so

$$\varepsilon(l^*)^2 = Q \sum_{l=l^*+1}^{\infty} \frac{l}{2l+3}(a/r)^{2l+4}.$$

The series can be summed in terms of incomplete B functions, but it is easier to observe that $\varepsilon(l^*) < \tilde{\varepsilon}(l^*)$ where

$$\tilde{\varepsilon}(l^*)^2 = \frac{Q}{2} \sum_{l=l^*+1}^{\infty} (a/b)^{2l+4}.$$

Then (A15) remains true if ε is replaced by $\tilde{\varepsilon}$.

APPENDIX B

Bayesian inference on infinite-dimensional model spaces

Most statistical theory, such as computing moments and characteristic functions, involves integration with respect to a probability distribution. The existence and decent limiting behaviour of the integrals depends on the fact that the probability distribution is a finite, positive, countably additive measure (Halmos 1950) on the model space M. Such probability distributions cannot have density functions when $\dim M = \infty$ because then there is no sensible volume element in M. When $\dim M = \infty$ the probability distributions closest to being gaussian are those introduced by Kolmogorov (1950). His infinite-dimensional gaussians are the obvious candidates to soften hard inequalities when $\dim M = \infty$.

To discuss Kolmogorov measures, some notation is needed. If X is a subset of M, let $P\,(\mathbf{x} \in X)$, or simply $P(X)$, denote the probability that models lie in the set X. Even when $\dim M < \infty$, it is not always possible to assign a probability for every subset of M (Halmos 1950). The subsets X for which $P(X)$ has a well-defined value are called P-measurable, or simply measurable.

If $\dim M = \infty$, and $\mathbf{x} = (x^1, x^2, \ldots)^T$ is a column vector in M, and k is a positive integer, we define

$$\mathbf{x}_k = (x^1, \ldots, x^k)^T. \tag{B1}$$

Let M_k denote the set of all k-dimensional column vectors, so $M_\infty = M$. If X_k is a subset of M_k and X_l is a subset of M_l, let $X_k \times X_l$ denote the subset of M_{k+l} consisting of all $(k+l)$-dimensional column vectors $(\mathbf{x}_k^T, \mathbf{x}_l^T)^T$ with $\mathbf{x}_k$ in X_k and $\mathbf{x}_l$ in X_l. If $\mathbf{V}$ is an infinite matrix whose entries are V_{ij} with $1 \le i, j < \infty$, let $\mathbf{V}_k$ denote the $k \times k$ matrix whose entries are V_{ij} with $1 \le i, j \le k$.

To construct a Kolomogorov measure on M, it is necessary to start with an ordinary probability measure P_k on M_k for each positive integer k. Furthermore these distributions must be consistent with one another in the sense that for any positive integers k and l, P_k is the marginal distribution on M_k of the distribution P_{k+l} on M_{k+l}. That is, if X_k is a measurable subset of M_k, then $X_k \times M_l$ is a measurable subset of M_{k+l}, and

$$P_k(X_k) = P_{k+l}(X_k \times M_l). \tag{B2}$$

Kolmogorov showed that when this consistency condition is satisfied, then there is exactly one probability distribution (positive, finite, countably additive measure) P on M such that for every positive integer k and every measurable subset X_k of M_k, $X_k \times M_\infty$ is a measurable subset of M, and

$$P_k(X_k) = P(X_k \times M_\infty). \tag{B3}$$

Furthermore, if all the P_k have continuous densities g_k, then for each $\mathbf{x}_k$ in M_k there is a probability measure $P_{\mathbf{x}_k}$ on M with the following property: let ΔX_k be the infinitesimal rectangular box in M_k with one corner at $\mathbf{x}_k$ and one at $\mathbf{x}_k + d\mathbf{x}_k$. Let X be any measurable subset of M. Then $\Delta X_k \times X$ is a measurable subset of M, and

$$P(\Delta X_k \times X) = g_k(\mathbf{x}_k)\,dx_k^1 \cdots dx_k^k P_{\mathbf{x}_k}(X). \tag{B4}$$

Equation (B4) is as close as P comes to having density.

The gaussian Kolmogorov measures are constructed as follows: let $\boldsymbol{\mu}$ be any member of M and let $\mathbf{V}$ be any infinite symmetric matrix. Suppose that for each positive integer k, the $k \times k$ matrix $\mathbf{V}_k$ in the upper left corner of $\mathbf{V}$ is positive definite. Let P_k be the gaussian probability distribution on M_k whose mean is $\boldsymbol{\mu}_k$ and whose variance matrix is $\mathbf{V}_k$. It is a well-known result of the theory of finite-dimensional gaussians that these P_k satisfy the consistency conditions (B2). The Kolmogorov measure P which they generate on

M is called the gaussian probability distribution on M with mean $\boldsymbol{\mu}$ and variance matrix $\mathbf{V}$.

To carry out Bayesian inference in this situation, we start with a prior personal probability distribution on M which is gaussian with mean $\mathbf{x}_{PR}$ and variance matrix $\mathbf{V}_{PR}$. We must be able to calculate $g(\mathbf{y} \mid \mathbf{x})$, the probability density of the data vector $\mathbf{y}$, given that the correct model is $\mathbf{x}$. Then for each positive integer k we calculate

$$C(\mathbf{x}_k) = \int_M dP_{\mathbf{x}_k}(\mathbf{x}) \tag{B5a}$$

and

$$g(\mathbf{y} \mid \mathbf{x}_k) = C(\mathbf{x}_k)^{-1} \int_M g(\mathbf{y} \mid (\mathbf{x}_k^T, \mathbf{x}^T)^T)\, dP_{\mathbf{x}_k}(\mathbf{x}). \tag{B5b}$$

Here $g(\mathbf{y} \mid \mathbf{x}_k)$ is the probability density for the data vector $\mathbf{y}$, given that the first k components of the correct model vector $\mathbf{x}$ are $\mathbf{x}_k$. Equations (B5) permit finite-dimensional Bayesian inference on each M_k. The consistency conditions (B2) hold for the various P_k^{PO}, so they generate a posterior Kolmogorov distribution P_{PO}. If $g(\mathbf{y} \mid \mathbf{x})$ is gaussian in $\mathbf{y}$ for every $\mathbf{x}$, and P_{PR} is gaussian, then P_{PO} is gaussian.

In the infinite-dimensional gaussian case, (3.8) and (3.9) are still true if interpreted carefully. In equation (3.8), nothing guarantees the existence of $\mathbf{V}_{PR}^{-1}$ when it is infinite, so (3.8) must be rewritten as

$$\mathbf{V}_{PO} = (\mathbf{I} + \mathbf{V}_{PR}\mathbf{F}^T\mathbf{V}_\varepsilon^{-1}\mathbf{F})^{-1}\mathbf{V}_{PR}. \tag{B6}$$

The facts that $\mathbf{V}_k^{PR}$ and $\mathbf{V}_\varepsilon$ are positive definite and that $\dim M_{\mathbf{F}} < \infty$ assure that the inverses in (B6) exist as long as $V_{PR}\mathbf{F}^T$ converges. This last requirement is not a mere mathematical artifact, because the direct problem is not solved correctly unless $\mathbf{F}(\mathbf{x} - \mathbf{x}_{PR})$ converges for almost every $\mathbf{x}$ in M (i.e. the set of $\mathbf{x}$ for which $\mathbf{F}(\mathbf{x} - \mathbf{x}_{PR})$ does not converge has probability zero).

APPENDIX C

Pervasive effects of a broader prior in BI

It seems intuitively obvious that, for a given data set, the smaller is the observer's prior uncertainty about the location of the true model $\mathbf{x}$ in model space M, the smaller will be the posterior variance which the observer assigns to any function of $\mathbf{x}$, say $g(\mathbf{x})$. That the formalism of gaussian BI leads to this obvious result is useful not as a proof of the result but as a verification that BI does describe (at least ideally) how we think.

The formal Bayesian result to be proved is as follows: suppose two different observers have gaussian priors with variance matrices $\mathbf{V}_{PR}$ and $\mathbf{V}'_{PR}$. Suppose they both learn of the same observations $\mathbf{y}$, with a given $\mathbf{F}$ and $\mathbf{V}_\varepsilon$, so that they calculate their posterior variance matrices, $\mathbf{V}_{PO}$ and $\mathbf{V}'_{PO}$, from (3.8). Suppose that $\mathbf{V}'_{PR} \geq \mathbf{V}_{PR}$ (i.e. $\mathbf{x}^T\mathbf{V}'_{PR}\mathbf{x} \geq \mathbf{x}^T\mathbf{V}_{PR}\mathbf{x}$ for every $\mathbf{x}$ in M). Then for any real function g on M, the primed observer assigns a larger posterior variance to $g(\mathbf{x})$ than does the unprimed observer. We will prove this result only for linear g, with $g(\mathbf{x}) = \mathbf{g}^T\mathbf{x}$ for some $\mathbf{g}$ in M. The posterior variance of $\mathbf{g}^T\mathbf{x}$ is $\mathbf{g}^T\mathbf{V}_{PO}\mathbf{g}$ for the unprimed observer, and $\mathbf{g}^T\mathbf{V}'_{PO}\mathbf{g}$ for the primed observer. Therefore we want to show that $\mathbf{g}^T\mathbf{V}'_{PO}\mathbf{g} \geq g^T\mathbf{V}_{PO}\mathbf{g}$ for every $\mathbf{g}$ in M. That is, we want to show that $\mathbf{V}'_{PR} \geq \mathbf{V}_{PR}$ implies $\mathbf{V}'_{PO} \geq \mathbf{V}_{PO}$.

The proof requires two lemmas about $m \times m$ symmetric matrices $\mathbf{P}$ and $\mathbf{Q}$.

Lemma 1

Suppose $\mathbf{P} \geq \mathbf{0}$. Then $\mathbf{Q}^T\mathbf{P}\mathbf{Q} \geq \mathbf{0}$.

Proof. By hypothesis, $\mathbf{x}^T\mathbf{P}\mathbf{x} \geq 0$ for every $\mathbf{x}$ in M. Since $\mathbf{Q}\mathbf{x}$ is in M, $(\mathbf{Q}\mathbf{x})^T\mathbf{P}(\mathbf{Q}\mathbf{x}) \geq 0$, so $\mathbf{x}^T\mathbf{Q}^T\mathbf{P}\mathbf{Q}\mathbf{x} \geq 0$ for every $\mathbf{x}$ in M.

Lemma 2

Suppose $\mathbf{0} < \mathbf{P} \leq \mathbf{Q}$. Then $\mathbf{0} < \mathbf{Q}^{-1} \leq \mathbf{P}^{-1}$.

Proof. For any t in $0 \leq t \leq 1$ define $\mathbf{R}(t) = (1-t)\mathbf{P} + t\mathbf{Q}$. Then for each t, $\mathbf{R}(t) > \mathbf{0}$, so $\mathbf{R}(t)^{-1}$ exists and is symmetric and $\mathbf{R}(t)^{-1} > \mathbf{0}$. Moreover, $\mathbf{I} = \mathbf{R}(t)\mathbf{R}(t)^{-1}$ so $\mathbf{0} = \partial_t\mathbf{I} = \partial_t\mathbf{R}\mathbf{R}^{-1} + \mathbf{R}\,\partial_t(\mathbf{R}^{-1})$. Therefore $\partial_t(\mathbf{R}^{-1}) = -\mathbf{R}^{-1}\,\partial_t\mathbf{R}\mathbf{R}^{-1}$, or

$$\partial_t(\mathbf{R}^{-1}) = -\mathbf{R}^{-1}(\mathbf{Q} - \mathbf{P})\mathbf{R}^{-1}.$$

Then by Lemma 1, $-\partial_t(\mathbf{R}^{-1}) \geq 0$, so for any fixed $\mathbf{x}$ in M, $\partial_t[\mathbf{x}^T\mathbf{R}(t)^{-1}\mathbf{x}] \leq 0$. Therefore $\mathbf{x}^T\mathbf{R}(0)^{-1}\mathbf{x} \geq \mathbf{x}^T\mathbf{R}(1)^{-1}\mathbf{x}$, or $\mathbf{x}^T\mathbf{P}^{-1}\mathbf{x}^T \geq \mathbf{x}^T\mathbf{Q}^{-1}\mathbf{x}$, which was to be proved.

We return to the proof that $\mathbf{V}'_{PR} \geq \mathbf{V}_{PR}$ implies $\mathbf{V}'_{PO} \geq \mathbf{V}_{PO}$. Let $\mathbf{A} = \mathbf{F}^T\mathbf{V}_\varepsilon^{-1}\mathbf{F}$, $\mathbf{B} = \mathbf{V}_{PR}^{-1}$ and $\mathbf{B}' = (\mathbf{V}'_{PR})^{-1}$. In BI, $\mathbf{0} < \mathbf{V}_{PR}$, $\mathbf{0} < \mathbf{V}'_{PR}$, and $\mathbf{0} \leq \mathbf{A}$, so by Lemma 2 the hypothesis $\mathbf{V}_{PR} \leq \mathbf{V}'_{PR}$ implies $\mathbf{0} < \mathbf{B}' \leq \mathbf{B}$. Then $\mathbf{0} < \mathbf{A} + \mathbf{B}' \leq \mathbf{A} + \mathbf{B}$, and another application of Lemma 2 gives $\mathbf{0} < (\mathbf{A} + \mathbf{B})^{-1} \leq (\mathbf{A} + \mathbf{B}')^{-1}$. This finishes the proof, because $(\mathbf{A} + \mathbf{B})^{-1} = \mathbf{V}_{PO}$ and $(\mathbf{A} + \mathbf{B}')^{-1} = \mathbf{V}'_{PO}$.

APPENDIX D

Improbable data

In defence of Bayesian inference it is necessary to answer a question raised by C. Voorhies (1987). What does an observer do if his *a priori* personal probability distribution on the model space M makes it extremely improbable that the observed data vector $\mathbf{y}$ could come from the true model $\mathbf{x}$? Observers who ignore the problem will not convince their co-workers to accept their conclusions. To treat the problem, observers very confident of their priors will revise the model space M and the modelling function F in (1.1). Observers more confident of their models than their priors will want to modify their priors so as to make the observed $\mathbf{y}$ less improbable. The modified priors would then be the priors used in Section 3.

The last approach will probably be the first one tried by most observers, so it deserves brief quantitative discussion. For more detail, see Berger (1985, p. 94). The test for whether $\mathbf{y}$ is improbable *a priori* is based on the observer's marginal *a priori* probability for $\mathbf{y}$ in the data space D. The density for this probability is (3.11b), which is the same as the $C(\mathbf{y})$ of (3.14c). It can be computed from the observer's prior, $f_{PR}(\mathbf{x})$, and the probability distribution of the errors of observation.

In linear gaussian BI all the calculations can be done explicitly. The density $C(\mathbf{y})$ is guassian. Its mean is the $\mathbf{y}_{PR}$

of (3.2a), and its variance matrix $\mathbf{V}_\mathbf{y}$ is given by

$$\mathbf{V}_\mathbf{y} = \mathbf{V}_\varepsilon + \mathbf{F}\mathbf{V}_{PR}\mathbf{F}^T. \quad \text{(D1)}$$

To obtain (D1) in BI requires a short (omitted) computation based on the equivalence of (3.6) and (3.7). In SI, (D1) is obvious.

In linear gaussian BI it follows that *a priori* the observer will expect

$$\chi^2 = (\mathbf{y} - \mathbf{y}_{PR})^T \mathbf{V}_\mathbf{y}^{-1} (\mathbf{y} - \mathbf{y}_{PR}) \quad \text{(D2)}$$

to be distributed like chi-squared with d degrees of freedom, where $d = \dim D$. If the observed $\mathbf{y}$ is improbably far from y_{PR} in the χ^2 sense, the observer may want to replace his old prior by a new one with the same $\mathbf{x}_{PR}$ but with the new variance matrix $K\mathbf{V}_{PR}$, where K is chosen so much larger than 1 that there is a probability of 0.5 (for example) that χ^2 will be as large as it is observed to be. Of course, there are other, more complicated ways to increase $\mathbf{V}_{PR}$, and the probability of 0.5 is arbitrary. However, conclusions based on the details of the observer's prior will arouse suspicion, so those details will be unimportant to the conclusions he wants to defend.