Geophysical Journal (1988) **94,** 249–261

Comparing hard and soft prior bounds in geophysical inverse problems

George E. Backus
Institute of Geophysics and Planetary Physics, University of California, San Diego, La Jolla, CA 92093, USA

Accepted 1988 January 4. Received 1988 January 4; in original form 1987 August 12

SUMMARY

In linear inversion of a finite-dimensional data vector **y** to estimate a finite-dimensional prediction vector **z**, prior information about the correct earth model $\mathbf{x}_E$ is essential if **y** is to supply useful limits for **z**. The one exception occurs when all the prediction functionals are linear combinations of the data functionals.

We compare two forms of prior information: a 'soft' bound on $\mathbf{x}_E$ is a probability distribution p_X on the model space X which describes the observer's opinion about where $\mathbf{x}_E$ is likely to be in X; a 'hard' bound on $\mathbf{x}_E$ is an inequality $Q_X(\mathbf{x}_E, \mathbf{x}_E) \leq 1$, where Q_X is a positive definite quadratic form on X. A hard bound Q_X can be 'softened' to many different probability distributions p_X, but all these p_X's carry much new information about $\mathbf{x}_E$ which is absent from Q_X, and some information which contradicts Q_X. For example, all the p_X's give very accurate estimates of several other functions of $\mathbf{x}_E$ besides $Q_X(\mathbf{x}_E, \mathbf{x}_E)$. And all the p_X's which preserve the rotational symmetry of Q_X assign probability 1 to the event $Q_X(\mathbf{x}_E, \mathbf{x}_E) = \infty$. Both stochastic inversion (SI) and Bayesian inference (BI) estimate **z** from **y** and a soft prior bound p_X. If that probability distribution was obtained by softening a hard prior bound Q_X, rather than by objective statistical inference independent of **y**, then p_X contains so much unsupported new 'information' absent from Q_X that conclusions about **z** obtained with SI or BI would seem to be suspect.

Key words: Stochastic inversion, Bayesian inference, prior bounds, isotropic probability distributions.

1 INTRODUCTION

Most geophysical inverse problems require prior information for their solution (Backus 1970a; Franklin 1970), information known to the observer before he obtained the data to be inverted. That prior information is often cast in the form of a probability distribution p_X on the linear space X of possible earth models **x**, but it can also take the form of one or more bounds on the correct earth model $\mathbf{x}_E$. These bounds are usually linear or quadratic. Linear bounds take the form $a \leq \tilde{f}(\mathbf{x}_E) \leq A$, where a and A are known real numbers and $\tilde{f}: X \to R$ is a known real-valued linear function on X (R is the real line). Positivity constraints on the density are an example of linear bounds. Quadratic bounds take the form

$$Q_X(\mathbf{x}_E, \mathbf{x}_E) \leq 1, \tag{1.1}$$

where Q_X is a known positive-definite quadratic form on X. That is, $Q_X(\mathbf{x}_1, \mathbf{x}_2)$ is a real number which depends linearly on each of $\mathbf{x}_1$ and $\mathbf{x}_2$ when the other is fixed; and $Q_X(\mathbf{x}_1, \mathbf{x}_2) = Q_X(\mathbf{x}_2, \mathbf{x}_1)$; and $Q_X(\mathbf{x}, \mathbf{x}) > 0$ unless $\mathbf{x} = \mathbf{0}$. Energy constraints are examples of (1.1). Jackson (1979) calls the probability distributions 'soft' bounds on $\mathbf{x}_E$, and the inequalities 'hard' bounds.

There are two kinds of soft bounds, subjective and objective. A subjective soft bound is a probability distribution p_X on X which represents an observer's subjective personal opinion about where $\mathbf{x}_E$ is likely to be in X. This p_X might be obtained by 'softening' a hard quadratic bound (1.1) when the observer is unwilling to adopt (1.1) with certainty. Then he could replace (1.1) by a Gaussian with mean **0** and variance tensor Q_X^{-1}. Hard linear bounds can also be softened (Jackson 1979). The observer's ability to persuade his colleagues to accept his use of a subjective soft bound to invert the data will depend on his ability to persuade them to share his prior personal probability distribution.

An objective soft bound is a probability distribution p_X on X which models a realizable population of possible models **x**. Such a p_X might be estimated by repeatedly drawing random samples **x** from X, as in the analysis of a stationary time series, or the aiming strategy of an antiaircraft weapon in a protracted war. Alternatively, an objective soft bound might come from a theory of the source of the models: a complete theory of the geodynamo might provide a probability distribution for the Gauss coefficients of the geomagnetic field at the core–mantle boundary. Finally, an objective soft bound might appear as a hypothesis to be tested: perhaps the palaeomagnetic data can be fitted to a statistical model which treats the Gauss coefficients as

uncorrelated gaussian random variables (C. Constable & R. Parker, private communication; they see evidence for some correlations).

Hard linear bounds can be incorporated directly into a geophysical inversion by means of linear programming (Dantzig 1963; Heustis & Parker 1977). Hard quadratic bounds can be incorporated by the method which we will call 'confidence set inference,' CSI (Backus 1970a). Subjective soft bounds are best treated by Bayesian inference, BI, which is directly concerned with how an observer alters his personal probability distribution for $\mathbf{x}_E$ when he learns of new data with known error statistics. Backus (1988) gives references to some of the early work on BI, which goes back to Bayes (1764). Objective soft bounds are best treated by stochastic inversion, SI, which applies a minimum-variance linear estimator to the data vector $\mathbf{y}$ (Franklin 1970; Jackson 1979). The idea is to find the linear mapping $H: Y \to X$ from the data space Y to the model space X which is statistically best for estimating a model $\mathbf{x}$ from its observed data vector $\mathbf{y}$ as $H(\mathbf{y})$. More precisely, H is chosen to minimize the expected value of the squared distance from $\mathbf{x}$ to $H(\mathbf{y})$ in a long series of trials, model vectors $\mathbf{x}$ being drawn at random from X according to p_X, and their data vectors $\mathbf{y}$ being observed and used to estimate $\mathbf{x}$. Both p_X and the error statistics of $\mathbf{y}$ contribute to H.

Some observers (Backus 1988) take the view that stochastic inversion is inappropriate when there is only one correct earth model $\mathbf{x}_E$, and p_X is a prior personal probability distribution, a subjective soft bound. Bayesian inference seems to be the proper procedure here. Others disagree (Jackson 1979). Fortunately, Bayesian inference and stochastic inversion lead to the same result when p_X and the statistics of the errors in the data are gaussian (Backus 1987), so in that case there is no need to choose between SI and BI.

Bayesian inference (and, for some observers, stochastic inversion) can be used with hard prior bounds if those bounds are first softened to subjective soft prior bounds (Backus 1970b; Jackson 1979; Gubbins 1983).

It is the thesis of the present paper that the relationship between hard and soft bounds is not as simple as their names would lead one to expect, and that neither Bayesian inference nor stochastic inversion is appropriate for incorporating a hard quadratic prior bound into a data inversion. If a single inequality (1.1) is really all the prior information the observer wants to use, he must confine himself to confidence set inference. Use of BI and SI introduces new information not contained in (1.1), and this information makes very precise quantitative claims about $\mathbf{x}_E$ which come entirely from the bound-softening process and are independent of the observed data. We illustrate the problem with the example of continuing the geomagnetic field $\mathbf{B}$ down to the core–mantle boundary (CMB). Here, softening the hard heat flow bound (Gubbins 1983; Backus 1987) or the hard energy bound (Backus 1987) makes *a priori* claims about the gauss coefficients of $\mathbf{B}$ at the CMB which many workers in geomagnetism would find preposterous.

We have not investigated the softening of hard linear bounds, and make no comments on that subject. In a later paper we will extend the discussion of confidence set inference begun by Backus (1970a). Preliminary calculations indicate that in many cases estimates of the correct model $\mathbf{x}_E$ will not be very different in CSI from those obtained by correctly executed BI or SI, and that the error estimates in CSI may be larger than those of BI or SI by a factor of the order two. Most of the published Bayesian and stochastic inversions are simply regularizations, the 'prior' information being inferred from the data to be inverted. Therefore, those inversions produce physically acceptable models which fit the data within the expected data errors, but such inversions cannot support the error estimates for $\mathbf{x}_E$ reported by their users (Backus 1988). Defensible error estimates on $\mathbf{x}_E$ are not yet available in these inversions. If the prior information is only a hard quadratic bound (1.1) then those error estimates must come from confidence set inference, not stochastic inversion or Bayesian inference.

If the observer really does have prior information about $\mathbf{x}_E$ which can be described by a probability distribution p_X on the model space X, then he has much valuable information about $\mathbf{x}_E$ not contained in (1.1), and of course he is entitled to use this information in any inversion of the data. If his p_X is objective, then the evidence for it will be objective, and available before the data are obtained. Then he should have no difficulty convincing colleagues to accept the conclusions he draws from the data. However, if his p_X is a subjective prior personal probability distribution, he may have trouble defending it, or accepting it himself, when he realizes how much more he is assuming *a priori* about $\mathbf{x}_E$ than is contained in (1.1).

The plan of this paper is to formulate the linear inverse problem with great generality, so as to make clear why prior information is almost always essential, and so as to exhibit plainly just what prior information goes into the inversion. The substantive content of the paper is the comparison of the prior information about $\mathbf{x}_E$ contained in hard quadratic bounds (1.1) and probability distributions p_X on the model space X (soft bounds).

2 THE NEED FOR PRIOR INFORMATION

Surely data are preferable to opinions. If we have enough good data, why can we not dispense with prior opinions? The reason is that our model spaces are usually infinite-dimensional, and we never have more than finitely many data (Backus & Gilbert 1967). For example, suppose we want to use surface and satellite measurements of the geomagnetic field $\mathbf{B}$ to estimate the radial component B_r at the core–mantle boundary (CMB). The apparently infinite data set from the satellite track can be Fourier analysed, and only finitely many Fourier components will be above the noise. Hence there will be only finitely many data. The model space X can be parametrized by the Schmidt semi-normalized Gauss coefficients g_l^m at the CMB (l is degree, and m is longitudinal order, so $-l \le m \le l$). Clearly $\dim X = \infty$.

It is sometimes supposed that by studying the resolution of the data we can remove the difficulty of having only finitely many equations for infinitely many unknowns. In the example of the geomagnetic field, Gubbins & Bloxham (1985) find that the surface and satellite data do not resolve gauss coefficients at the CMB above degree 20. Backus (1988) shows that no surface and satellite data with a 1 nT error of measurement can resolve Gauss coefficients at the

CMB with degree $l \geq 32$ unless the ohmic heating rate in the core exceeds the total geothermal flow at the Earth's surface. To fit geomagnetic data whose error of measurement is at least 1 nT, we need never use a model space X at the CMB whose dimension exceeds $l(l+2)$ with $l = 31$. That is, $\dim X \leq 1023$. The number of satellite data from MAGSAT exceeds this value by at least a factor of 10 (Langel, Estes & Mead 1982).

In fact, however, arguments based on resolution cannot cut down the model space X to finite dimensionality, because usually the predictions we want to make about the correct model $\mathbf{x}_E$ involve components of $\mathbf{x}_E$ not resolved by the data.

The present section formulates the inverse problem in a way which makes clear when prior information is needed to invert the data. We consider only data and predictions which depend linearly on the model. An approximate discussion of the nonlinear problem appears, for example, in Backus & Gilbert (1968). For the sake of generality, we do not introduce a topology on X, the infinite-dimensional real linear space of models $\mathbf{x}$. In particular, we do not assume that X is a Hilbert space.

Our conventions of notation are as follows: R is the real line. If Y and Z are sets, $y \in Y$ means that y is a member of Y, and $Y \subseteq Z$ means that y is a subset of Z. If a function f assigns to each y in Y a unique value $z = f(y)$ in Z, then we write $f: Y \to Z$. This symbol can also be real as a substantive, 'the function f which maps Y into X.'

A vector or model is simply a member of the linear space X. A dual vector or dual model is a linear function $\tilde{f}: X \to R$, i.e. a linear functional on X. Linearity of $\tilde{f}$ means, of course, that if $b^1, \ldots, b^n \in R$ and $\mathbf{x}_1, \ldots, \mathbf{x}_n \in X$ then

$$\tilde{f}(b^j \mathbf{x}_j) = b^j \tilde{f}(\mathbf{x}_j). \tag{2.1}$$

Here we use the Einstein summation convention: if an index appears once as a subscript and once as a superscript in a single term or a product, the summation over all its possible values is understood. The set of all dual vectors is written $\tilde{X}$. If $a_1, \ldots, a_m \in R$ and $\tilde{f}^1, \ldots, \tilde{f}^m \in \tilde{X}$, then the function $a_i \tilde{f}^i: X \to R$ is defined by requiring that for each vector $\mathbf{x}$ in X

$$(a_i \tilde{f}^i)(\mathbf{x}) = a_i [\tilde{f}^i(\mathbf{x})]. \tag{2.2}$$

From (2.1) it is easy to verify that $a_i \tilde{f}^i \in \tilde{X}$, so $\tilde{X}$ is a real linear space.

By R^n we will mean the real linear space of $1 \times n$ matrices $\mathbf{y} = (y^1, \ldots, y^n)$. In addition to model space X two other real linear spaces enter the inverse problem: $Y = R^d$ and $Z = R^p$. Here Y is the data space and Z is the prediction space. Finally, we know two linear functions, $F: X \to Y$ and $G: X \to Z$. Our inverse problem is summarized as follows: if $\mathbf{x}_E$ is the model in X which best represents the real earth, then

$$\mathbf{y} = F(\mathbf{x}_E) + \delta_R \mathbf{y} + \delta_X \mathbf{y} \tag{2.3a}$$

$$\mathbf{z} = G(\mathbf{x}_E) + \delta_X \mathbf{z}. \tag{2.3b}$$

Here we have collected d data about the Earth, real numbers $y^1, \ldots, y^d$ which make up the data vector $\mathbf{y} = (y^1, \ldots, y^d)$ in Y. We would like to use $\mathbf{y}$ to predict p other data about the Earth, real numbers $z^1, \ldots, z^p$ which make up the prediction vector $\mathbf{z} = (z^1, \ldots, z^p)$ in Z. Some of the z^i may never be directly observable, but we would like to estimate their values. If there were no errors, $\mathbf{y}$ would be $F(\mathbf{x}_E)$ and $\mathbf{z}$ would be $G(\mathbf{x}_E)$. There is a random error $\delta_R y^i$ in each of the observed data y^i, and $\delta_R \mathbf{y} = (\delta_R y^1, \ldots, \delta_R y^d)$ is the random error vector. The failure of the models in X to include all relevant features of the earth produces systematic errors $\delta_X \mathbf{y}$ in $\mathbf{y}$ and $\delta_X \mathbf{z}$ in $\mathbf{z}$. It is crucial in the arguments to follow that dim Y and dim Z are finite.

Of course we do not know the errors $\delta_R \mathbf{y}$, $\delta_X \mathbf{y}$ and $\delta_X \mathbf{z}$, but we must know something about them or the inverse problem is hopeless. We assume that we know a hard quadratic bound like (1.1) on each of the systematic error vectors, $\delta_X \mathbf{y}$ and $\delta_X \mathbf{z}$. We also assume that we know the probability distribution p_R of the random error vector $\delta_R \mathbf{y}$ in the data space Y. Therefore given any $f: T \to R$, we can calculate the expected value of $f(\mathbf{y})$,

$$\langle f(\mathbf{y}) \rangle = \int_Y dp_R(\mathbf{y}) f(\mathbf{y}). \tag{2.4}$$

In particular, we can calculate $\langle \delta_R \mathbf{y} \rangle = (\langle \delta_R y^1 \rangle, \ldots, \langle \delta_R y^d \rangle)$, and redefine $\mathbf{y}$ as $\mathbf{y} - \langle \delta_R \mathbf{y} \rangle$, so we can assume that

$$\langle \delta_R \mathbf{y} \rangle = 0 \tag{2.5}$$

In the geomagnetic example of downward continuation of surface and satellite data to the core–mantle boundary (CMB) we will take for the model space X the space of all magnetic fields $\mathbf{B}$ defined above the CMB, irrotational and solenoidal there, and vanishing at infinity. Obviously $\dim X = \infty$ here. The data $y^1, \ldots, y^d$ are Cartesian components of $\mathbf{B}$ at finitely many sites on and above the surface of the earth. The quantities $z^1, \ldots, z^p$ to be estimated might be p gauss coefficients of $\mathbf{B}$ at the CMB, or the values of the radial component B_r at p sites on the CMB, or the magnetic flux through p null-flux curves on the CMB (Backus 1968; Gubbins & Bloxham, 1985). In the null-flux example, $G: X \to Z$ in (2.3b) is nonlinear and must be linearized (Gubbins & Bloxham, 1985), and our linear calculations will be only approximately correct.

In this geomagnetic example, the total error vector $\delta_R \mathbf{y} + \delta_X \mathbf{y}$ includes instrument errors, site location errors, stray fields, and contributions to $\mathbf{B}$ from crustal magnetization or the electric currents in the mantle, ionosphere and magnetosphere. The errors about which we have statistical information are lumped in $\delta_R \mathbf{y}$, and the remaining errors constitute $\delta_X \mathbf{y}$. It is our thesis that in spaces of high dimension, statistical information is stronger than a quadratic bound, so we claim to know more about $\delta_R \mathbf{y}$ than about $\delta_X \mathbf{y}$. The formulation of the inverse problem can be 'upgraded' by promoting part of $\delta_X \mathbf{y}$ to $\delta_R \mathbf{y}$ or part of either $\delta_X \mathbf{y}$ or $\delta_R \mathbf{y}$ to X. For example, if nothing is known about the magnetic field of the crust at satellite altitudes except a bound on its intensity, it belongs in $\delta_X \mathbf{y}$. If we are willing to treat the crustal field as a 2-D random process on the surface of a sphere, and if we know its statistics, then it belongs in $\delta_R \mathbf{y}$. Finally, we can include the crustal field in X by expanding the model space so that each model includes a jump in B_r at the surface of the Earth (Backus 1986). If we want to include the crustal field in X, so as to model surface as well as satellite observations of $\mathbf{B}$, Langel *et al.* (1982)

point out that we can make station corrections obtained from other satellite data at other times.

The analysis of the linear inverse problem hinges on the following observation: if $\tilde{f}^1, \ldots, \tilde{f}^n$ are linearly independent dual vectors (linear functionals) in $\tilde{X}$ then there are vectors $\mathbf{x}_1, \ldots \mathbf{x}_n$ in X such that for $1 \le i, j \le n$

$$\tilde{f}^i(\mathbf{x}_j) = \delta^i_j, \tag{2.6}$$

where δ^i_j is the Kronecker delta, 1 when $i = j$ and 0 otherwise. We prove this fact by induction of n in imitation of Gram-Schmidt orthogonalization. If $n = 1$, linear independence means simply $\tilde{f}^1 \neq 0$. Then there is an $\mathbf{x}_0$ in X such that $\tilde{f}^1(\mathbf{x}_0) = a \neq 0$. We take $\mathbf{x}_1 = a^{-1}\mathbf{x}_0$. Now suppose we know (2.6) for n, and we want to prove it for $n+1$. We are given linearly independent dual vectors $\tilde{f}^1, \ldots, \tilde{f}^n, \tilde{f}^{n+1}$ in $\tilde{X}$. By induction we can assume that we have found vectors $\xi_1, \ldots, \xi_n$ in X such that for $1 \le i, j \le n$

$$\tilde{f}^i(\xi_j) = \delta^i_j. \tag{2.7a}$$

Then for every vector $\mathbf{x}$ in X we can define a vector $\mathbf{x}^\perp$ by writing

$$\mathbf{x}^\perp = \mathbf{x} - \tilde{f}^i(\mathbf{x})\xi_i, \tag{2.7b}$$

the sum being over $1 \le i \le n$. Suppose that for every $\mathbf{x}$ in X

$$\tilde{f}^{n+1}(\mathbf{x}^\perp) = 0. \tag{2.8a}$$

Then, from (2.7b) and the linearity of $\tilde{f}^{n+1}$,

$$\tilde{f}^{n+1}(\mathbf{x}) = \tilde{f}^i(\mathbf{x})\tilde{f}^{n+1}(\xi_i). \tag{2.8b}$$

Definition (2.2) permits us to write (2.88b) in the form

$$\tilde{f}^{n+1}(\mathbf{x}) = [\tilde{f}^{n+1}(\xi_i)\tilde{f}^i](\mathbf{x}). \tag{2.8c}$$

Since (2.8c) holds for every $\mathbf{x}$ in X, the functions $\tilde{f}^{n+1}: X \to R$ and $[\tilde{f}^{n+1}(\xi_i)\tilde{f}^i]: X \to R$ must be the same. Thus

$$\tilde{f}^{n+1} = \tilde{f}^{n+1}(\xi_i)\tilde{f}^i. \tag{2.8d}$$

But (2.8d) exhibits $\tilde{f}^{n+1}$ as a linear combination of $\tilde{f}^1, \ldots, \tilde{f}^n$, contrary to the assumed linear independence of $\tilde{f}^1, \ldots, \tilde{f}^n, \tilde{f}^{n+1}$. Therefore (2.8a) must fail for some $\mathbf{x}_0$ in X. Then $\tilde{f}^{n+1}(\mathbf{x}_0^\perp) = a \neq 0$. We define

$$\mathbf{x}_{n+1} = a^{-1}\mathbf{x}_0^\perp \tag{2.9a}$$

and, for $1 \le j \le n$,

$$\mathbf{x}_j = \xi_j - \tilde{f}^{n+1}(\xi_j)\mathbf{x}_{n+1}. \tag{2.9b}$$

Then clearly

$$\tilde{f}^{n+1}(\mathbf{x}_{n+1}) = 1; \tag{2.9c}$$

and from (2.9b), for $1 \le j \le n$

$$\tilde{f}^{n+1}(\mathbf{x}_j) = 0;$$

from (2.7a,b), for $1 \le i \le n$

$$\tilde{f}^i(\mathbf{x}_{n+1}) = 0;$$

and then for $1 \le i < j \le n$ equations (2.7a), (2.9b) and (2.9c) imply

$$\tilde{f}^i(\mathbf{x}_j) = \delta^i_j.$$

It follows that $\tilde{f}^i(\mathbf{x}_j) = \delta^i_j$ for $1 \le i, j \le n+1$, and the induction is complete.

The foregoing proof makes clear that the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in (2.6) are not uniquely determined by $\tilde{f}^1, \ldots, \tilde{f}^n$ unless $n = \dim X$. However, $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are linearly independent, for if $a^1, \ldots, a^n \in R$ and $a^j\mathbf{x}_j = \mathbf{0}$ then (2.6) implies $a^j\delta^i_j = 0$, so $a^1 = \cdots = a^n = 0$. The list of dual vectors $\tilde{f}^1, \ldots, \tilde{f}^n$ and the list of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are said to be dual to one another.

Now we return to the inverse problem (2.3) and its need for prior information about $\mathbf{x}_E$. We write the vectors $F(\mathbf{x})$ and $G(\mathbf{x})$ from (2.3) in the forms

$$F(\mathbf{x}) = (\tilde{F}^1(\mathbf{x}), \ldots, \tilde{F}^d(\mathbf{x})) \tag{2.10a}$$

$$G(\mathbf{x}) = (\tilde{G}^1(\mathbf{x}), \ldots, \tilde{G}^p(\mathbf{x})). \tag{2.10b}$$

Here $\tilde{F}^i: X \to R$ and $\tilde{G}^j: X \to R$ are linear functions. They are dual vectors, members of the dual space $\tilde{X}$. For simplicity, we assume there are no errors in the data, that $\tilde{F}_1, \ldots, \tilde{F}^d$ are linearly independent, and that $p = 1$. If $\tilde{G}^1$ is a linear combination of $\tilde{F}^1, \ldots, \tilde{F}^d$, then clearly $\tilde{G}^1(\mathbf{x}_E)$, the prediction, can be calculated directly from $F(\mathbf{x}_E)$, the data vector (Backus & Gilbert 1967, consider this case at length). The uniqueness problem and the need for prior information arise when $\tilde{G}^1$ is not a linear combination of $\tilde{F}^1, \ldots, \tilde{F}^d$. In this case, $\tilde{F}^1, \ldots, \tilde{F}^d, \tilde{G}^1$ are linearly independent. Then we choose model vectors $\mathbf{x}_1, \ldots, \mathbf{x}_d, \boldsymbol{\xi}_1$ dual to $\tilde{F}^1, \ldots, \tilde{F}^d, \tilde{G}^1$. Thus for $1 \le i, j \le d$

$$\tilde{F}^i(\mathbf{x}_j) = \delta^i_j. \tag{2.11a}$$

Also, for $1 \le i \le d$

$$\tilde{F}^i(\boldsymbol{\xi}_1) = \tilde{G}^1(\mathbf{x}_i) = 0, \tag{2.11b}$$

and finally

$$\tilde{G}^1(\boldsymbol{\xi}_1) = 1. \tag{2.11c}$$

Now let b be any real number, and define

$$\mathbf{x} = \mathbf{x}_E + b\boldsymbol{\xi}_1. \tag{2.12}$$

From (2.11), clearly

$$\tilde{F}^i(\mathbf{x}) = \tilde{F}^i(\mathbf{x}_E) = y^i \tag{2.13}$$

and

$$\tilde{G}^1(\mathbf{x}) = \tilde{G}^1(\mathbf{x}_E) + b = z + b. \tag{2.14}$$

From (2.13), $\mathbf{x}$ satisfies the data just as well as does $\mathbf{x}_E$. From (2.14), $\tilde{G}^1(\mathbf{x})$ differs from z by b, which is arbitrary. Thus if all we know about $\mathbf{x}_E$ is that it satisfies the data, we can put no limits whatever on the possible values of the prediction $z = \tilde{G}^1(\mathbf{x}_E)$.

The remedy for this difficulty is apparent. To make $\mathbf{z} = G(\mathbf{x})$ very different from $G(\mathbf{x}_E)$, we must make b very large in (2.12). But if b is too large, the model $\mathbf{x} = \mathbf{x}_E + b\boldsymbol{\xi}$ will be rejected as physically unreasonable. It is the careful examination of what 'physically unreasonable' means which introduces our prior information or beliefs about $\mathbf{x}_E$ into the inverse problem (2.3). To enable the data $\mathbf{y}$ to restrict the prediction vector $\mathbf{z}$, we must have some prior information about the correct earth model $\mathbf{x}_E$. This prior information must confine $\mathbf{x}_E$ to a tractable subset of X, at least with high probability (unless, of course, the prior information is simply the value of $\mathbf{z}$). The manner in which hard or soft prior bounds on $\mathbf{x}_E$ (inequalities or probability distributions

for $\mathbf{x}_E$) reduce the ambiguity in linear inverse problems is discussed elsewhere. See Heustis & Parker (1977) for the use of linear programming with linear inequalities, or Backus (1970a) for confidence set inference with quadratic inequalities. For objective soft bounds (objectively defensible probability distributions), Franklin (1970) and Jackson (1979) discuss stochastic inversion (SI). For subjective soft bounds (personal probability distributions) Backus (1970b) and Tarantola & Valette (1982) discuss Bayesian inference (BI). Backus (1988) gives brief reviews of both SI and BI.

The foregoing conclusions are purely algebraic. They do not require a topology on X, much less a norm or an inner product. The belief that a certain topology on X is relevant to the real earth is itself a prior belief which can be used in the inverse problem. In the subsequent sections, we will see how quadratic hard bounds and probability distributions both lead to physically natural inner products on X. Usually, this prior information is the only natural source of such an inner product on X.

3 INFORMATION LOST AND GAINED IN SOFTENING A HARD BOUND

In the problem of geomagnetic downward continuation, let g_l^m be the Schmidt semi-normalized Gauss coefficient of degree l and longitudinal order m at the core–mantle boundary (CMB), measured in nanoTeslas. Our belief that the energy of the geomagnetic field **B** cannot have a rest mass greater than that of the earth leads us to accept

$$\sum_{l=1}^{\infty} (l+1)(2l+1)^{-1} \sum_{m=-l}^{l} |g_l^m|^2 < 2 \times 10^{33}\ \mathrm{nT}^2 \tag{3.1}$$

(Backus 1988). Our belief that the total rate of heat flow out of the Earth's surface is larger than Gubbins' (1975) expression for the minimum rate of ohmic heating in the core leads to

$$\sum_{l=1}^{\infty} l^{-1}(l+1)(2l+1)(2l+3) \sum_{m=-l}^{l} |g_l^m|^2 < 3 \times 10^{17}\ \mathrm{nT}^2 \tag{3.2}$$

if we think that the electrical conductivity in the core is everywhere less than 3×10^5 mho/meter (Backus 1988). Both prior beliefs, (3.1) and (3.2), are examples of quadratic bounds (1.1). Both beliefs are imprecise. Most geophysicists would confidently reduce the right side of (3.1) by several orders of magnitude. Geophysicists who believe that at least two-thirds of the surface heat flow comes from radioactivity in the crust and mantle would be willing to reduce 3×10^{17} to 10^{17} in (3.2). But others might want to replace 3×10^{17} by 10^{18} in (3.2) because heat pulses from a chaotic core dynamo could produce unsteady heat flow, or because some of the ohmic heat might be produced in the hot source of the core heat engine and recycled into magnetic energy rather than lost to the mantle (Backus 1975). Prior information, although essential to inversion, is almost always imprecise. One way to deal with this imprecision is to verify that the conclusions drawn from an inversion are not sensitive to the bounds on the right of (3.1) or (3.2), as long as those bounds remain in a physically defensible range. Another way is to try to represent the imprecision by replacing (1.1) by a probability distribution p_X on the model space X (Backus 1970b; Jackson 1979; Gubbins 1983). At first sight, this appears reasonable. We show in the present section that, on the contrary, 'softening' (1.1) to a probability distribution adds considerable information about $\mathbf{x}_E$ which is not implied by (1.1).

The natural way to replace (3.2) with a probability distribution p_X is to assume that the g_l^n are independent gaussian random variables with means 0 and variances $3 \times 10^{17} l(l+1)^{-1}(2l+1)^{-1}(2l+3)^{-1}\ \mathrm{nT}^2$. This p_X raises some questions which we want to discuss in general, so we consider the general case (1.1). We are given a model space X and a positive definite quadratic from Q_X on X, and we believe that the real Earth satisfies (1.1).

Then we can introduce on X the dot product $\mathbf{x}_1 \cdot \mathbf{x}_2$, defined by

$$\mathbf{x}_1 \cdot \mathbf{x}_2 = Q_X(\mathbf{x}_1, \mathbf{x}_2) \tag{3.3a}$$

and we can define the length $\|\mathbf{x}\|$ of the model $\mathbf{x}$ as

$$\|\mathbf{x}\| = (\mathbf{x} \cdot \mathbf{x})^{1/2}. \tag{3.3b}$$

If X is not complete in the norm (3.3b), we can complete it, and when we do, X becomes a Hilbert space with inner product (3.3a) (Halmos 1951). In X we can always find an orthonormal basis. We will assume that this basis is denumerable, as is the case in the geomagnetic example and all others where the models are well-behaved scalar or vector fields on finite dimensional domains. Thus there is an infinite sequence of vectors $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots$ in X such that

$$\hat{\mathbf{x}}_i \cdot \hat{\mathbf{x}}_j = \delta_{ij}; \tag{3.4a}$$

and if $\mathbf{x}$ is any vector in X then

$$\mathbf{x} = \sum_{i=1}^{\infty} x_i \hat{\mathbf{x}}_i, \tag{3.4b}$$

where

$$x_i = \hat{\mathbf{x}}_i \cdot \mathbf{x}. \tag{3.4c}$$

The convergence in (3.4b) is with respect to the norm (3.3b). Now the condition (1.1) for physical acceptability of the model $\mathbf{x}$ can be written simply as

$$\|\mathbf{x}\| \le 1 \tag{3.5a}$$

or

$$\sum_{i=1}^{\infty} x_i^2 \le 1. \tag{3.5b}$$

We want to try to 'soften' (3.5) to a probability distribution p_X which injects no new information about $\mathbf{x}$ not already contained in (3.5). A plausible softening procedure begins with the observation that if (3.5) is really all we know, then all we are entitled to claim about any one component x_i is

$$-1 \le x_i \le 1. \tag{3.6}$$

We can inject imprecision into (3.6) by regarding x_i as a random variable with a one-dimensional probability distribution p_i whose mean and variance are 0 and 1. If, as (3.5b) indicates, we really know nothing to distinguish the seaparate x_i's, then all the p_i should be the same as p_1. The $x_1, x_2, \ldots$ should be identically distributed. Furthermore, if p_X distinguishes between x_i and $-x_i$ then p_X includes prior information not present in (3.5). Therefore, using p_X as in

(2.4) to calculate expected values, we should obtain $\langle x_i x_j \rangle = 0$ when $i \neq j$. Hence

$$\langle x_i \rangle = 0 \tag{3.7a}$$

$$\langle x_i x_j \rangle = \delta_{ij}. \tag{3.7b}$$

At this point we encounter a well-known difficulty. From (3.7b) follows

$$\left\langle \sum_{i=1}^{\infty} x_i^2 \right\rangle = \infty. \tag{3.7c}$$

As a probabalistic analogue of (3.5b), (3.7c) is disappointing. Softening (3.5) to a probability distribution appears to have destroyed some information. This accounts for the choice of words: p_X is a 'softened' version of (3.5), being fuzzier than (3.5).

Gubbins (1983) and Backus (1987) discuss a possible remedy for (3.7c). We can introduce convergence factors $\kappa_1, \kappa_2, \ldots$ such that

$$0 < \kappa_i \leq 1 \tag{3.8a}$$

and

$$\sum_{i=1}^{\infty} \kappa_i^2 = 1. \tag{3.8b}$$

Then we can replace (3.7b) by

$$\langle x_i x_j \rangle = \kappa_i^2 \delta_{ij}. \tag{3.8c}$$

Now (3.7c) is replaced by

$$\left\langle \sum_{i=1}^{\infty} x_i^2 \right\rangle = 1, \tag{3.9a}$$

as we might hope from (3.5). However, (3.8c) and (3.7a) are the softened version of the opinion

$$-\kappa_i \leq x_i \leq \kappa_i. \tag{3.9b}$$

Since (3.8b) requires $\kappa_i \to 0$ as $i \to \infty$, clearly (3.8c) goes well beyond (3.6) or even (3.5b) as a statement of what we claim to believe about $\mathbf{x}_E$. The use of convergence factors commits us to accepting much prior information not implied by (3.5). If we really believe this extra prior information, we are certainly entitled to use it in inverting the data. Unfortunately, convergence factors are usually introduced *ad hoc*, simply to avoid (3.7c) and with no other evidence to support them. Conclusions drawn from such prior information are as speculative as the information itself. We are trying to find all physically reasonable models which fit the data with statistically acceptable accuracy, so we prefer to invoke only prior information for which we have good evidence. If (3.5) really represents all the prior information we are willing to accept, then we cannot introduce convergence factors. We must accept (3.7) as properties of any p_X which softens (3.5) and adds no new information. (For a less puritanical view, see Backus 1988.)

To construct p_X, we must arrange that all the one-dimensional marginal distributions, $p_1, p_2, \ldots$, are the same, and have mean 0 and variance 1. Thus, choosing p_1 determines all the p_j. It does not determine p_X. To find p_X from p_1, we note that (3.7b) suggests (but does not require) that $x_1, x_2, x_3, \ldots$ be assumed independent. If $\dim X = n < \infty$, then the density function for p_X in the case of independence becomes simply the product of the densities of $x_1, \ldots, x_n$. If $\dim X = \infty$, then p_X is the Kolmogrov distribution whose projections onto finite-dimensional subspace of X have the product densities (Kolmogorov 1950; see also Backus 1988).

Now suppose we make the very mild assumption that the 1-D distribution p_1 for the separate x_i's has a fourth moment, which we write as $K + 1$. If p_1 is gaussian, this is certainly true, and $K = 2$. But if $\langle x_i^4 \rangle$ exists, then $x_1^2, x_2^2, \ldots$ are identically distributed independent random variables with mean $\langle x_i^2 \rangle = 1$ and variance $\langle x_i^4 \rangle - 1 = K$. For each integer n, define the random variable

$$S_n(\mathbf{x}) = n^{-1} \sum_{i=1}^{n} x_i^2. \tag{3.10}$$

Then S_n has mean 1 and variance K/n. When $n \gg 1$, the central limit theorem (Kendall & Stuart 1977, p. 206) says that $S_n(\mathbf{x})$ is approximately gaussian. The probability that a 1-D gaussian variable is more than three standard deviations from its mean is slightly less than 0.003. Thus, with a probability slightly more than 0.997,

$$|S_n(\mathbf{x}) - 1| < 3(K/n)^{1/2}. \tag{3.11}$$

In short, if n is very large, with high probability we can infer from the softened version of (3.5) a very accurate estimate of the value of $S_n(\mathbf{x}_E)$ for the correct earth model $\mathbf{x}_E$. We obtain this estimate without any data. It was certainly not present in (3.5). It represents new 'information' generated in the process of softening a hard quadratic bound to a probability distribution. The same argument applies, of course, to the data space Y in (2.3). If its dimension is large, we know more about the random error $\delta_R \mathbf{y}$ than about the systematic error $\delta_X \mathbf{y}$, because the former is constrained by a probability distribution p_R on Y, the latter only by a quadratic inequality.

Softening (3.5) to a probability distribution p_X in the manner just described generates still more information about $\mathbf{x}_E$. Let $p_X(U)$ denote the probability that $\mathbf{x}_E$ is a member of the subset U of X. For any positive α and any integer n, let $X_n(\alpha)$ be the set of models $\mathbf{x}$ for which

$$nS_n(\mathbf{x}) \leq \alpha, \tag{3.12a}$$

where S_n is defined by (3.10). Let $X_\infty(\alpha)$ be the set of models $\mathbf{x}$ such that

$$\sum_{i=1}^{\infty} x_i^2 \leq \alpha. \tag{3.12b}$$

Let X_∞ be the set of models $\mathbf{x}$ such that

$$\sum_{i=1}^{\infty} x_i^2 < \infty. \tag{3.12c}$$

Since $nS_n(\mathbf{x})$ is approximately gaussian with mean n and standard deviation $(Kn)^{1/2}$, therefore

$$\lim_{n \to \infty} p_X[X_n(\alpha)] = 0. \tag{3.13a}$$

But for each α, $X_\infty(\alpha) \subseteq X_{n+1}(\alpha) \subseteq X_n(\alpha)$. Therefore (Halmos 1950, p. 38)

$$p_X[X_\infty(\alpha)] = 0. \tag{3.13b}$$

Now for every integer n, $X_\infty(n) \subseteq X_\infty(n+1) \subseteq X_\infty$, while if

$\mathbf{x} \in X_\infty$ then $\mathbf{x} \in X_n$ for some n. Therefore (Halmos 1950, p. 38),

$$p_X(X_\infty) = 0. \tag{3.13c}$$

In short, if we soften (3.5) to a probability distribution in the obvious way, far from losing information, as seems to be suggested by (3.7c), we are led to espouse the belief that with probability 1, $\|\mathbf{x}_E\| = \infty$. Not only does $\mathbf{x}_E$ violate (3.5); it is altogether outside the model space X. The softening process leads a geomagnetician who initially accepts (3.2) to convert to the belief that the ohmic heat production rate in the core is infinite.

Technically, what has happened is that the Kolmogorov distribution p_X obtained by softening (3.5) has as its natural domain the set of all sequences $(x_1, x_2, \ldots)$, square summable or not in the sense of (3.5b). We have just deduced that p_X assigns probability 0 to the set of all those sequences which are square summable.

The argument leading to (3.13c) depends crucially on the assumptions that $x_1, x_2, \ldots$ have the same one-dimensional distribution p_1, and that $x_1, x_2, \ldots$ are independent. The former assumption simply says that our prior information (3.5) does not distinguish between x_i and x_j. This seems a fair view of (3.5). However, the latter assumption, independence, is not obviously contained in (3.5), and it leads to disaster.

Equation (3.7b) does not really entitle us to assume that x_i and x_j are independent, but only that they are uncorrelated. If they are gaussian, lack of correlation implies independence, so if we are to save the idea of bound softening, we must accept model parameters $x_1, x_2, \ldots$ in (3.4b) which are neither gaussian nor independent. We explore this question in the next section.

4 BOUND SOFTENING WITH DEPENDENT, NON-GAUSSIAN MODEL PARAMETERS

The new information in (3.11) arises from the fact that when $x_1, x_2, \ldots$ are identically distributed, independent random variables with mean zero, and

$$r_n^2 = \sum_{i=1}^{n} x_i^2 \tag{4.1a}$$

then the standard deviation of r_n^2,

$$\sigma(r_n^2) = [\langle r_n^4 \rangle - \langle r_n^2 \rangle^2]^{1/2}, \tag{4.1b}$$

grows more slowly with n than the expected value, $\langle r_n^2 \rangle$. Thus the relative error in r_n^2, $\sigma(r_n^2)/\langle r_n^2 \rangle$, tends to zero as n becomes large. We seek a probability distribution p_X on X which avoids this difficulty and represents a fuzzy version of (3.5). Clearly we cannot demand that p_X be gaussian or that $x_1, x_2, \ldots$ be independent. We want to avoid any p_X which introduces new information not contained in (3.5). One obvious property of (3.5) is its failure to distinguish among $x_1, x_2, \ldots$; in fact (3.5) is unchanged by any rotation of the axes $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots$ in X. It seems reasonable to ask that p_X have this property. We will make only the slightly weaker demand that p_X be unaltered when any finite number of axes $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots$ are rotated among themselves, leaving the others fixed. We call such a probability distribution on X 'isotropic.'

We denote by X_n the subspace of X consisting of vectors of the form

$$\mathbf{r}_n = \sum_{i=1}^{n} x_i \hat{\mathbf{x}}_i \tag{4.2a}$$

and by p_n the marginal distribution of p_X on X_n. If each p_n has a density function, f_n, we will call p_X 'regular.' Backus (1987) shows that every isotropic p_X is the weighted sum of a regular isotropic p_X and a p_X which concentrates all its probability at $\mathbf{0}$. For simplicity, here we consider only a regular p_X. Then the isotropy of p_X requires that f_n depend on $x_1, \ldots, x_n$ only through the r_n^2 of (4.1a). Thus

$$dp_n(\mathbf{r}_n) = f_n(x_1^2 + \cdots + x_n^2)\, dx_1 \cdots dx_n. \tag{4.2b}$$

We want to calculate the mean, $\langle r_n^2 \rangle$, and the standard deviation, $\sigma(r_n^2)$, of r_n^2 for an isotropic distribution p_X. The spherical symmetry of p_X implies

$$\langle x_i x_j \rangle = \langle x^2 \rangle \delta_{ij} \tag{4.3a}$$

$$\langle x_i^4 \rangle = \langle x^4 \rangle \tag{4.3b}$$

$$\langle x_i^2 y_i^2 \rangle = \langle x^2 y^2 \rangle \quad \text{if} \quad i \neq j. \tag{4.3c}$$

Here we have written x for x_1 and y for x_2. Furthermore, clearly

$$\langle x^2 \rangle = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy x^2 f_2(x^2 + y^2)$$

$$\langle x^4 \rangle = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy x^4 f_2(x^2 + y^2)$$

$$\langle x^2 y^2 \rangle = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy x^2 y^2 f_2(x^2 + y^2).$$

If we perform these integrations in plane polar coordinates, with $x = \xi^{1/2} \cos\theta$, $y = \xi^{1/2} \sin\theta$, we obtain

$$\langle x^2 \rangle = (\pi/2) \int_0^{\infty} d\xi \xi f_2(\xi) \tag{4.4a}$$

$$\langle x^2 y^2 \rangle = (\pi/8) \int_0^{\infty} d\xi \xi^2 f_2(\xi) \tag{4.4b}$$

$$\langle x^4 \rangle = 3\langle x^2 y^2 \rangle. \tag{4.4c}$$

From (4.3)

$$\langle r_n^2 \rangle = n \langle x^2 \rangle, \tag{4.5a}$$

and

$$\langle r_n^4 \rangle = n \langle x^4 \rangle + n(n-1)\langle x^2 y^2 \rangle,$$

so, from (4.4c)

$$\langle r_n^4 \rangle = n(n+2)\langle x^2 y^2 \rangle. \tag{4.5b}$$

Therefore

$$\sigma(r_n^2)/\langle r_n^2 \rangle = \kappa - 1 + 2\kappa n^{-1}, \tag{4.6a}$$

where

$$\kappa = \langle x^2 y^2 \rangle \langle x^2 \rangle^{-2}. \tag{4.6b}$$

The difficulty (3.11) arises because when $x_1, x_2, \ldots$ are independent, $\kappa = 1$. Can we choose p_X so $\kappa > 1$?

Evidently we must learn to construct isotropic probability distributions. The construction is based on the observation

that since p_n is the marginal distribution of p_X on X_n, it is also the marginal distribution of p_{n+1} on X_n. Therefore one of the Kolmogorov consistency conditions (Kolmogorov 1950) is

$$f_n(\varpi^2) = \int_{-\infty}^{\infty} dz f_{n+1}(\varpi^2 + z^2). \tag{4.7}$$

If we define $\xi = \varpi^2$ and choose the new variable of integration $\eta = \varpi^2 + z^2$, then (4.7) becomes

$$f_n(\xi) = \int_{\xi}^{\infty} d\eta (\eta - \xi)^{-1/2} f_{n+1}(\eta). \tag{4.8}$$

Following Abel (Tricomi 1957, p. 39) we solve (4.3a) for f_{n+1} by appealing to the identity

$$\int_{\zeta}^{\eta} d\xi (\xi - \zeta)^{-1/2} (\eta - \xi)^{-1/2} = \pi. \tag{4.9}$$

We multiply (4.8) by $(\xi - \zeta)^{-1/2}$, integrate over ξ from ζ to ∞, reverse the order of the integrals on the right, and use (4.9) to obtain

$$\int_{\zeta}^{\infty} d\xi (\xi - \zeta)^{-1/2} f_n(\xi) = \pi \int_{\zeta}^{\infty} d\eta f_{n+1}(\eta).$$

Differentiating with respect to ζ and relabeling variables gives

$$f_{n+1}(\xi) = -\pi^{-1} \partial_{\xi} \int_{\xi}^{\infty} d\eta (\eta - \xi)^{-1/2} f_n(\eta). \tag{4.10}$$

Equations (4.8) and (4.10) make very clear that we cannot choose $f_1, f_2, f_3, \ldots$ independently. If we choose one of them, all the others are determined. If we choose f_n for some particular n, how much freedom do we have in this choice?

If we want to choose one of the marginal densities f_n and construct all the others from (4.8) and (4.10), two limitations on our choice are that

$$f_n(\zeta) \geq 0 \tag{4.11a}$$

and that

$$\int_{X_n} dx_1 \cdots dx_n f_n(x_1^2 + \cdots + x_n^2) = 1. \tag{4.11b}$$

Carrying out the integral in n-dimensional spherical polar coordinates gives

$$\pi^{n/2} \Gamma(n/2)^{-1} \int_0^{\infty} d\xi \xi^{n/2-1} f_n(\xi) = 1. \tag{4.11c}$$

A further limitation on our choice of f_n is that any f_m constructed from f_n by (4.8) or (4.10) should also satisfy (4.11). Equations (4.11b,c) cause no trouble. The constructions (4.8) and (4.10) are equivalent to (4.7), which gaurantees (4.11b) for all m. There remains (4.11a). If $m < n$, (4.11a) for f_n and (4.8) imply $f_m(\xi) \geq 0$ for all ξ. To see whether (4.10) ensures $f_m(\xi) \geq 0$ for all ξ when $m > n$ requires some care.

We begin by appealing to Abel's identity (4.9) to simplify (4.1). We iterate (4.8) once, to express f_n in terms of f_{n+2}. Then we reverse the order of the double integral on the right and use (4.9) to obtain

$$f_n(\xi) = \pi \int_{\xi}^{\infty} d\zeta f_{n+2}(\zeta). \tag{4.12a}$$

Differentiating with respect to ζ gives

$$f_{n+2}(\xi) = -\pi^{-1} \partial_{\xi} f_n(\xi). \tag{4.12b}$$

This situation is familiar from seismic travel time inversion. The square of the operator applied to f_{n+1} in (4.8) is π times the integration operator, and the square of the operator applied to f_n in (4.10) is $-\pi$ times the differentiation operator. Given f_n, we find f_{n+1} from (4.10), and then we can find all other f_m with $m \geq n$ by means of (4.12b).

It is clear from (4.12a) that every marginal density function f_n of an isotropic probability distribution p_X has at least one derivative, $\partial_{\xi} f_n(\xi)$. But then induction on (4.12a) shows that each $f_n(\xi)$ is infinitely differentiable for all $\xi \geq 0$. If we are trying to construct an isotropic p_X by choosing one of its marginal densities f_n, we must be careful to choose f_n to be infinitely differentiable.

Furthermore, the derivatives of f_n cannot be arbitrary. If we want $f_{n+2q}(\xi) \geq 0$ for all $\xi \geq 0$ and all integers $q \geq 0$, (4.12b) shows that we must choose an f_n such that

$$(-\partial_{\xi})^q f_n(\xi) \geq 0 \tag{4.13a}$$

for all $\xi \geq 0$ and all integers $q \geq 0$. Moreover, f_n must be such that when f_{n+1} is constructed from it via (4.10), we have

$$(-\partial_{\xi})^q f_{n+1}(\xi) \geq 0. \tag{4.13b}$$

In fact, (4.10) and (4.13a) imply (4.13b). To see this we observe that if f is twice continuously differentiable and dies away at infinity rapidly enough to permit the integrals to converge, then

$$\partial_{\xi} \int_{\xi}^{\infty} d\eta (\eta - \xi)^{-1/2} f(\eta) = \int_{\xi}^{\infty} d\eta (\eta - \xi)^{-1/2} \partial_{\eta} f(\eta).$$

To prove this observation, integrate the integral on the left by parts, and differentiate under the integral sign. Applying this observation $q + 1$ times allows us to infer from (4.10) that

$$(-\partial_{\xi})^q f_{n+1}(\xi) = \pi^{-1} \int_{\xi}^{\infty} d\eta (\eta - \xi)^{-1/2} (-\partial_{\xi})^{q+1} f_n(\xi).$$

Thus if (4.13a) is true for all q, so is (4.13b).

We conclude that $f_n(\xi)$ is the marginal n-dimensional density of an isotropic probability distribution p_X if and only if f_n is continuously differentiable on $\xi \geq 0$ and satisfies (4.13a) for all $\xi \geq 0$ and all integers $q \geq 0$, and also satisfies (4.11c). The choice $f_n(\xi) = (2\pi)^{-n/2} e^{-\xi/2}$ meets these conditions and the resulting p_X is the gaussian considered in section 3. Another isotropic probability distribution can be constructed from

$$f_2(\xi) = \pi^{-1} (\nu + 1) \alpha^{\nu+1} (\alpha + \xi)^{-(\nu+2)}, \tag{4.14}$$

where α is any positive constant and ν is any constant larger than -1.

Now that we have a non-gaussian, dependent isotropic probability distribution, perhaps we can escape from (3.11). It is easy to compute from (4.4) that if $\nu > 1$ the p_X

constructed from (4.14) leads to

$$\langle x^2 \rangle = \alpha/(2\nu)$$

and

$$\langle x^2 y^2 \rangle = \alpha^2/[4\nu(\nu-1)],$$

so in (4.6)

$$\kappa = 1 + (\nu - 1)^{-1}.$$

Therefore $\sigma(r_n^2)/\langle r_n^2 \rangle$, the relative error in r_n^2, does not shrink to zero for large n if we adopt the p_X constructed from (4.14). If we do not want to add information like (3.9b) to (3.5), we should choose $\langle x^2 \rangle = 1$, so in (4.14) we want

$$\alpha = 2\nu.$$

Having escaped from (3.11), can we also escape from (3.13)? Unfortunately, the answer is no. Every isotropic probability distribution p_X will result in (3.13) or will assign probability 1 to the origin, $\mathbf{0} = (0, 0, \ldots)$. The author's proof of this fact was complicated. Gary Egbert has found a simpler proof of a more general result. Let X be as in section 3. For any integer $i > 0$ and any real c and d, the set of all $\mathbf{x} = (x_1, x_2, \ldots)$ in X for which $c < x_i < d$ is called a 'slab.' Let p_X be a probability measure on X which is able to assign probabilities to all slabs (i.e., all slabs are measurable; see Halmos 1950). We call p_X 'symmetric' if it is unchanged when any two coordinates x_i and x_j are interchanged. Clearly every isotropic probability measure on X is symmetric. Egbert proves that if p_X is symmetric then it concentrates all probability at the origin.

Egbert's proof begins by defining

$$\sigma_i(\alpha)^2 = \int_{X_\infty(\alpha)} dp_X(\mathbf{x}) x_i^2, \tag{4.15}$$

where i is any positive integer, α is any positive real number, and $X_\infty(\alpha)$ is as in (3.12b). By the Lebesgue monotone convergence theorem (Halmos 1950),

$$\sum_{i=1}^{\infty} \sigma_i(\alpha)^2 = \int_{X_\infty(\alpha)} dp_X(\mathbf{x}) \left(\sum_{i=1}^{\infty} x_i^2 \right). \tag{4.16}$$

But $\sum_{i=1}^{\infty} x_i^2 \leq \alpha$ in $X_\infty(\alpha)$, so by (4.16)

$$\sum_{i=1}^{\infty} \sigma_i(\alpha)^2 \leq \alpha p_X[X_\infty(\alpha)] \leq \alpha. \tag{4.17}$$

Since p_X is symmetric, $\sigma_i(\alpha)^2$ is the same for all i, so (4.17) implies $\sigma_i(\alpha)^2 = 0$ for all i. Therefore, (4.16) implies

$$\int_{X_\infty(\alpha)} dp_X(\mathbf{x}) \left(\sum_{i=1}^{\infty} x_i^2 \right) = 0. \tag{4.18}$$

For any real ε and α with $0 < \varepsilon < \alpha$, define $X_\infty(\varepsilon, \alpha)$ to be the set of all $\mathbf{x}$ in x for which

$$\varepsilon < \sum_{i=1}^{\infty} x_i^2 \leq \alpha.$$

Then clearly

$$\varepsilon p_X[X_\infty(\varepsilon, \alpha)] \leq \int_{X_\infty(\varepsilon, \alpha)} dp_X(\mathbf{x}) \left(\sum_{i=1}^{\infty} x_i^2 \right) \leq \int_{X_\infty(\alpha)} dp_X(\mathbf{x}) \left(\sum_{i=1}^{\infty} x_i^2 \right).$$

Therefore, by (4.18),

$$p_X[X_\infty(\varepsilon, \alpha)] = 0 \tag{4.19}$$

for every real ε and α with $0 < \varepsilon < \alpha$. Now let $X\backslash\{\mathbf{0}\}$ denote X with $\mathbf{0}$ removed. If $\mathbf{x}$ is in $X\backslash\{\mathbf{0}\}$, then $0 < \|\mathbf{x}\| < \infty$, so there is a positive integer n such that $\mathbf{x}$ is in $X_\infty(n^{-1}, n)$. Moreover, if $m < n$ then $X_\infty(m^{-1}, m) \subseteq X_\infty(n^{-1}, n)$. It follows (Halmos 1950, p. 38) that

$$p_X(X\backslash\{\mathbf{0}\}) = \lim_{n \to \infty} p_X[X_\infty(n^{-1}, n)].$$

Therefore, by (4.19),

$$p_X(X\backslash\{\mathbf{0}\}) = 0. \tag{4.20}$$

Equation (4.20) shows that not only every isotropic probability distribution on X but even every symmetric probability distribution concentrates all probability at the origin. If we do not believe that $\mathbf{x} = \mathbf{0}$ with probability 1, then the situation is as if p_X were gaussian. The natural domain of p_X is the space R^∞ of all infinite sequences $(x_1, x_2, \ldots)$, and p_X assigns probability 0 to the whole subspace of square-summable sequences. Isotropic probability distributions cannot adequately represent the hard bound (3.5) even when they avoid the trap of (3.11).

The argument leading to (4.20) clearly depends on the fact that X is infinite-dimensional. Any particular inverse problem can be studied on a finite-dimensional model space, constructed via (2.6) from a maximal linearly independent subset of the $\tilde{F}^1, \ldots, \tilde{F}^d, \tilde{G}^1, \ldots, \tilde{G}^p$ in (2.10). Perhaps we should permit $\dim X = N < \infty$ and try to soften (3.5) with an isotropic probability distribution p_X on X. The probability densities f_n of the marginal distributions on the subspaces X_n defined by (4.2a) will be related as in (4.8) and (4.10), but the sequence $f_1, f_2, \ldots, f_N$ will terminate at f_N, and from (4.12a) we will be able to claim about f_n only that is has $(N-n)/2$ or $(N-n-1)/2$ derivatives. The obvious candidate for a p_X which softens (3.5) is the one whose probability density on $X = X_N$ is constant when $\|\mathbf{x}\| \leq 1$ and 0 when $\|\mathbf{x}\| > 1$. The constant can be evaluated from (4.11c). Then, from (4.8) and (4.12), if $1 \leq n \leq N$

$$f_n(\xi) = \frac{(N/2)!\,(1-\xi)^{N/2-n/2}}{(N/2-n/2)!\,\pi^{1/2}}.$$

In particular,

$$f_2(\xi) = (2\pi)^{-1} N (1-\xi)^{N/2-1}.$$

Then, from (4.4),

$$\langle x^2 \rangle = (N+2)^{-1}$$
$$\langle X^2 y^2 \rangle = (N+2)^{-1}(N+4)^{-1},$$

so

$$\langle r_n^2 \rangle = n(N+2)^{-1} \tag{4.21a}$$

and

$$\frac{\sigma(r_n^2)}{\langle r_n^2\rangle}=\left[\frac{2(\langle r_n^2\rangle^{-1}-1)}{N+4}\right]^{1/2}. \tag{4.21b}$$

For large N, this distribution does not even avoid the trap (3.11).

It is possible to avoid (3.11) by using (4.14) but terminating the sequence $f_1, f_2, \ldots$ at f_N. However, if N is very large, (4.2b) and (4.12b) show that $p_X[X_N(1)]$ will become very small, and this is the probability of the event (3.5), which our prior beliefs led us to feel sure about. Such a difficulty will arise with any choice of $f_2(\xi)$ which can be continued via (4.10) to arbitrarily high dimension N. If we choose an $f_2(\xi)$ which cannot be continued up beyond a particular X_N, how do we justify our choice of N? An *a priori* restriction on the dimension of the model space is not the sort of prior information that will be very convincing to modern workers. On the other hand, if f_2 is chosen so that (4.10) terminates at $N=d+p$, then the personal probability distribution we accept on X before obtaining the data depends on how many data we are about to obtain.

5 INFORMATION LOST IN HARDENING A SOFT BOUND

To compare further the information content of quadratic inequalities and probability distributions, we will examine the question of trying to represent a probability distribution by a quadratic inequality. Since our starting point is now a probability distribution p_X on a linear space X, (3.13c) leads us to assume that X is finite-dimensional, with $N=\dim X$. The practical application of this section is thus to a discussion of the random errors $\delta_R\mathbf{y}$ in the data space Y, and p_X is the p_R of (2.4). However, to facilitate comparison with sections 3 and 4, and to use their notation, we continue to write X and p_X for the linear space and the probability distribution on it.

The importance of not immediately identifying X with a data space of column vectors is to force us to recognize that X is an abstract linear space without any structure or geometry except what can be built from p_X. Our first task is to construct on X from p_X a positive definite quadratic form Q_X. To do so, we recall the dual space $\tilde{X}$, consisting of all linear functionals $\tilde{f}: X\to R$. Since $\tilde{X}$ is a linear space, it has a dual space $\tilde{\tilde{X}}$. If $\mathbf{x}$ is any fixed vector in X, we can view $\mathbf{x}$ as a member $\tilde{\mathbf{x}}$ of $\tilde{X}$. To so so, for every $\tilde{f}$ in $\tilde{X}$ we define

$$\tilde{\tilde{\mathbf{x}}}(\tilde{f})=\tilde{f}(\mathbf{x}). \tag{5.1a}$$

When $\mathbf{x}$ is fixed, $\tilde{\tilde{\mathbf{x}}}(\tilde{f})$ is a real number which, by (2.2), depends linearly on $\tilde{f}$. Thus $\tilde{\tilde{\mathbf{x}}}$ is indeed a linear functional on $\tilde{X}$, i.e., a member of $\tilde{\tilde{X}}$. Furthermore, since $\dim X<\infty$, every linear functional on $\tilde{X}$ is of the form (5.1a) for exactly one $\mathbf{x}$ in X (Halmos 1958). Thus, in the sense of (5.1a),

$$\tilde{\tilde{X}}=X. \tag{5.1b}$$

First we use $\tilde{X}$ to define the mean value $\langle\mathbf{x}\rangle$ of $\mathbf{x}$ under p_X. If $\tilde{f}\in\tilde{X}$, then $\tilde{f}(\mathbf{x})$ is a random variable on X, with an expected value as defined in (2.4). From (2.4) and (2.2) $\langle\tilde{f}(\mathbf{x})\rangle$ depends linearly on $\tilde{f}$, so by (5.1b) there is a unique vector in X, which we will denote by $\langle\mathbf{x}\rangle$, such that

$$\langle\tilde{f}(\mathbf{x})\rangle=\tilde{f}(\langle\mathbf{x}\rangle) \tag{5.2a}$$

for every $\tilde{f}$ in $\tilde{X}$. Now we can shift the origin of X to $\langle\mathbf{x}\rangle$ so as to achieve the result

$$\langle\mathbf{x}\rangle=\mathbf{0}. \tag{5.2b}$$

Next, we use p_X to define a quadratic form $Q_{\tilde{X}}$ on $\tilde{X}$. For any $\tilde{f}_1$ and $\tilde{f}_2$ in $\tilde{X}$, $\tilde{f}_1(\mathbf{x})\tilde{f}_2(\mathbf{x})$ is a random variable on X, so we can define

$$Q_{\tilde{X}}(\tilde{f}_1,\tilde{f}_2)=\langle\tilde{f}_1(\mathbf{x})\tilde{f}_2(\mathbf{x})\rangle. \tag{5.3a}$$

Clearly, $Q_{\tilde{X}}(\tilde{f}_1,\tilde{f}_2)$ is a real number which depends linearly on each of $\tilde{f}_1$ and $\tilde{f}_2$ when the other is fixed. Also, clearly,

$$Q_{\tilde{X}}(\tilde{f}_1,\tilde{f}_2)=Q_{\tilde{X}}(\tilde{f}_2,\tilde{f}_1). \tag{5.3b}$$

Finally we claim that if $\tilde{f}\neq 0$ then

$$Q_{\tilde{X}}(\tilde{f},\tilde{f})>0, \tag{5.3c}$$

i.e., $Q_{\tilde{X}}$ is positive definite. If $\tilde{f}\in\tilde{X}$, then

$$Q_{\tilde{X}}(\tilde{f},\tilde{f})=\langle\tilde{f}(\mathbf{x})^2\rangle.$$

Thus $Q_{\tilde{X}}(\tilde{f},\tilde{f})\geq 0$. If $Q_{\tilde{X}}(\tilde{f},\tilde{f})=0$, then with probability 1

$$\tilde{f}(\mathbf{x})=0. \tag{5.4}$$

If $\tilde{f}\neq 0$ then (5.4) describes an $(N-1)$ dimensional subspace of X where $\mathbf{x}$ is to be found with probability 1. Obviously we should replace X by that subspace. Continuing in this way be induction, we finally reach (5.3c) unless p_X concentrates all its probability on $\mathbf{0}$, a degenerate case which we ignore.

The positive definite quadratic form $Q_{\tilde{X}}$ on $\tilde{X}$ defines a dot product on $\tilde{X}$, namely

$$\tilde{f}_1\cdot\tilde{f}_2=Q_{\tilde{X}}(\tilde{f}_1,\tilde{f}_2)=\langle\tilde{f}_1(\mathbf{x})\tilde{f}_2(\mathbf{x})\rangle. \tag{5.5}$$

But once we have a dot product on a finite-dimensional vector space X, we can identify X with its dual space $\tilde{X}$ (Halmos 1958). Thus, for every fixed $\tilde{\phi}$ in $\tilde{\tilde{X}}$ there is a unique $\tilde{\phi}$ in $\tilde{X}$ such that for every $\tilde{f}$ in $\tilde{X}$

$$\tilde{\tilde{\phi}}(\tilde{f})=\tilde{\phi}\cdot\tilde{f}. \tag{5.6a}$$

In other words,

$$\tilde{\tilde{X}}=\tilde{X}. \tag{5.6b}$$

Together, (5.1) and (5.6) imply

$$X=\tilde{X}. \tag{5.6c}$$

The details of (5.6c) are as follows. For every fixed $\mathbf{x}$ in X, there is a corresponding $\tilde{\tilde{\mathbf{x}}}\in\tilde{\tilde{X}}$; and vice versa. Setting $\tilde{\tilde{\phi}}=\tilde{\tilde{\mathbf{x}}}$ in (5.6a), we find from $\mathbf{x}$ a unique $\tilde{\mathbf{x}}$ in $\tilde{X}$ such that for every $\tilde{f}$ in $\tilde{X}$,

$$\tilde{\tilde{\mathbf{x}}}(\tilde{f})=\tilde{\mathbf{x}}\cdot\tilde{f}.$$

By (5.1a), this means that if $\mathbf{x}$ is a fixed vector in X, there is a unique linear functional $\tilde{\mathbf{x}}$ in $\tilde{X}$ such that for every $\tilde{f}$ in $\tilde{X}$

$$\tilde{f}(\mathbf{x})=\tilde{f}\cdot\tilde{\mathbf{x}}. \tag{5.6d}$$

Now we can define a quadratic form Q_X on X. For any $\mathbf{x}_1$ and $\mathbf{x}_2$ in X we require simply that

$$Q_X(\mathbf{x}_1,\mathbf{x}_2)=\tilde{\mathbf{x}}_1\cdot\tilde{\mathbf{x}}_2. \tag{5.7a}$$

From (5.6d) it is easy to verify that $\tilde{\mathbf{x}}$ depends linearly on $\mathbf{x}$, so $Q_X(\mathbf{x}_1,\mathbf{x}_2)$ depends linearly on each of $\mathbf{x}_1$ and $\mathbf{x}_2$ when the other is fixed. Also, from (5.5), $\tilde{\mathbf{x}}_1\cdot\tilde{\mathbf{x}}_2=\tilde{\mathbf{x}}_2\cdot\tilde{\mathbf{x}}_1$, so

$Q_X(\mathbf{x}_1, \mathbf{x}_2) = Q_X(\mathbf{x}_2, \mathbf{x}_1)$. Finally, since $Q_{\tilde{X}}$ is positive definite, $Q_X(\mathbf{x}, \mathbf{x}) \geq 0$; and if $Q_X(\mathbf{x}, \mathbf{x}) = 0$ then $\tilde{\mathbf{x}} = 0$. By (5.6c) this implies $\mathbf{x} = \mathbf{0}$, so Q_X is positive definite. From the positive definite quadratic form Q_X on X we can define the obvious dot product,

$$\mathbf{x}_1 \cdot \mathbf{x}_2 = Q_X(\mathbf{x}_1, \mathbf{x}_2), \tag{5.7b}$$

so

$$\mathbf{x}_1 \cdot \mathbf{x}_2 = \tilde{\mathbf{x}}_1 \cdot \tilde{\mathbf{x}}_2. \tag{5.7c}$$

Tracing through the definitions, we see that (5.7c) implies that for any fixed $\mathbf{x}_1$ and $\mathbf{x}_2$ in X,

$$\mathbf{x}_1 \cdot \mathbf{x}_2 = \langle \tilde{\mathbf{x}}_1(\mathbf{x}) \tilde{\mathbf{x}}_2(\mathbf{x}) \rangle \tag{5.8a}$$

and hence, by (5.6c) and (5.7c),

$$\mathbf{x}_1 \cdot \mathbf{x}_2 = \langle (\mathbf{x}_1 \cdot \mathbf{x})(\mathbf{x}_2 \cdot \mathbf{x}) \rangle. \tag{5.8b}$$

For the applications it will be useful to outline a less abstract approach to Q_X than the foregoing. We choose any fixed basis $\mathbf{b}_1, \ldots, \mathbf{b}_N$ for X. Then for each $\mathbf{x}$ in X there are unique real numbers $\xi^1, \ldots, \xi^N$ such that

$$\mathbf{x} = \xi^i \mathbf{b}_i \tag{5.9a}$$

and these numbers depend linearly on $\mathbf{x}$. That is, there are N linear functionals $\tilde{b}^i : X \to R$ such that in (5.9a)

$$\xi^i = \tilde{b}^i(\mathbf{x}). \tag{5.9b}$$

Thus the ξ^i are random variables, with expected values defined from p_X as in (2.4): $\langle \xi^i \rangle = \langle \tilde{b}^i(\mathbf{x}) \rangle$. It is easy to verify that the vector $\langle \mathbf{x} \rangle$ defined by

$$\langle \mathbf{x} \rangle = \langle \xi^i \rangle \mathbf{b}_i \tag{5.10}$$

is the same for all bases $\mathbf{b}_1, \ldots, \mathbf{b}_N$. Therefore we can shift the origin of X to achieve (5.2b).

Next, we define the $N \times N$ matrix V whose ij entry is

$$V^{ij} = \langle \xi^i \xi^j \rangle. \tag{5.11a}$$

We claim that V^{-1} exists, and we denote its ij entry by $(V^{-1})^{ij}$ and also, when convenient, by V_{ij} (not V_{ij}^{-1}!). Thus

$$V_{ij} = (V^{-1})^{ij}. \tag{5.11b}$$

To prove that V^{-1} exists, it suffices to prove that V is positive definite, i.e. $a_i V^{ij} a_j > 0$ for any real N-tuple $(a_1, \ldots, a_N) \neq (0, \ldots, 0)$. But for any real $a_1, \ldots, a_N$,

$$a^i V^{ij} a_j = \langle (a_i \xi^i)^2 \rangle.$$

Therefore $a_i V^{ij} a_j \geq 0$, and if $a_i V^{ij} a_j = 0$ then the probability is 1 that $\mathbf{x}$ lies in the $(N-1)$-dimensional subspace of X given by

$$a_i \xi_i = a_i \tilde{b}^i(\mathbf{x}) = 0.$$

If this happens, we repace X by that subspace.

Now for any vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ in X, we are able to define a real number

$$Q_X(\mathbf{x}_1, \mathbf{x}_2) = \tilde{b}^i(\mathbf{x}_1) V_{ij} \tilde{b}^j(\mathbf{x}_2). \tag{5.12}$$

It is an exercise in matrix algebra to verify that $Q_X(\mathbf{x}_1, \mathbf{x}_2)$ has the same value, whatever basis $\mathbf{b}_1, \ldots, \mathbf{b}_N$ is used to calculate it. Obviously $Q_X(\mathbf{x}_1, \mathbf{x}_2)$ depends linearly on each of $\mathbf{x}_1$ and $\mathbf{x}_2$ when the other is fixed. Finally, since V is symmetric and positive definite, so is V^{-1}. Hence $Q_X(\mathbf{x}_1, \mathbf{x}_2) = Q_X(\mathbf{x}_2, \mathbf{x}_1)$, and if $Q_X(\mathbf{x}, \mathbf{x}) = 0$ then $\tilde{b}^i(\mathbf{x}) = 0$ for all i, so, from (5.2), $\mathbf{x} = \mathbf{0}$. Thus Q_X is a positive-definite quadratic form constructed on X from p_X, and independent of the basis used in the construction. Now we define a dot product on X by (5.7b). Proving (5.8b) from (5.5) is an exercise in matrix algebra which we omit.

Having obtained from p_X a dot product on X with the crucial property (5.8b), we can choose an orthonormal basis for X, a basis $\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N$ such that

$$\hat{\mathbf{x}}_i \cdot \hat{\mathbf{x}}_j = \delta_{ij}. \tag{5.13a}$$

For every vector $\mathbf{x}$ in X, we can write

$$\mathbf{x} = \sum_{i=1}^{N} x_i \hat{\mathbf{x}}_i \tag{5.13b}$$

with

$$x_i = \hat{\mathbf{x}}_i \cdot \mathbf{x}. \tag{5.13c}$$

Equations (5.2) and (5.13c) imply

$$\langle x_i \rangle = 0, \tag{5.14a}$$

while (5.8), (5.13c) and (5.13a) imply

$$\langle x_i x_j \rangle = \delta_{ij}. \tag{5.14b}$$

Therefore, there is a basis in X which represents the vectors $\mathbf{x}$ in X in terms of coordinates $x_1, \ldots, x_N$ which are uncorrelated random variables, each with zero mean and unit variance.

The dot product on X makes X a Euclidean space with a volume element $d^N\mathbf{x}$, an impossibility when $\dim X = \infty$ (Loewner 1939). If p_X has a density function f with respect to this volume element, i.e. if

$$dp_X(\mathbf{x}) = d^N\mathbf{x} f(\mathbf{x}), \tag{5.15a}$$

and if

$$f(\mathbf{x}) = (2\pi)^{-N/2} \exp[-\|\mathbf{x}\|^2/2] \tag{5.15b}$$

the p_X is called a gaussian distribution. For a gaussian, if $i \neq j$, then x_i and x_j are not only uncorrelated but independent, and not only have zero mean and unit variance but are identically distributed.

The dot product (5.7b) is the natural one to use on X when studying p_X. If $\hat{\boldsymbol{\xi}}$ is any fixed unit vector in X, (5.8b) shows that

$$\langle (\hat{\boldsymbol{\xi}} \cdot \mathbf{x})^2 \rangle = 1, \tag{5.16}$$

so (5.7b) measures each component of $\mathbf{x}$ in units of the standard deviation of that component. When X is the data space Y, and p_X is the probability distribution p_R for the random error $\delta_R\mathbf{y}$ in the data vector $\mathbf{y}$, then (5.1b) says that (5.7b) measures each component $\hat{\boldsymbol{\xi}} \cdot \mathbf{y}$ of $\mathbf{y}$ in units of its error of measurement, the standard deviation of $\hat{\boldsymbol{\xi}} \cdot \delta_R\mathbf{y}$.

The foregoing calculations take a familiar form when X is the data space Y, the data vectors being $\mathbf{y} = (y^1, \ldots, y^d)$. The dual space $\tilde{Y}$ consists of the $d \times 1$ column matrices $\tilde{\mathbf{y}}^T = (\tilde{y}_1, \ldots, \tilde{y}_d)^T$, where T means matrix transpose. The value of $\tilde{\mathbf{y}}$ at $\mathbf{y}$ is

$$\tilde{\mathbf{y}}(\mathbf{y}) = \mathbf{y}\tilde{\mathbf{y}}^T, \tag{5.17}$$

matrix multiplication being intended. The natural basis for Y is $\mathbf{b}_1, \ldots, \mathbf{b}_d$, where $\mathbf{b}_i$ is the $1 \times d$ matrix whose ith

column is 1, the others being zero. The coordinate functional $\vec{b}^i$ is $\tilde{\mathbf{b}}_i^T$. Then

$$\mathbf{y} = y^i \mathbf{b}_i \tag{5.18a}$$

so (5.2b) implies

$$\langle y^i \rangle = 0 \tag{5.18b}$$

and (5.11a) gives

$$V^{ij} = \langle y^i y^j \rangle \tag{5.18c}$$

while (5.12) is

$$\mathbf{y}_1 \cdot \mathbf{y}_2 = y_1^i V_{ij} y_2^j \tag{5.18d}$$

with V_{ij} defined by (5.11b) as $(V^{-1})^{ij}$. In terms of the data $y^1, \ldots, y^d$, p_R is Gaussian if

$$dp_R(\mathbf{y}) = (2\pi)^{-d/2} \exp\left[-\tfrac{1}{2} y^i y^j V_{ij}\right] dy^1, \ldots, dy^d. \tag{5.19}$$

Now we return to (5.14) and the hardening of soft bounds. If $x_1, \ldots, x_N$ are not only uncorrelated but independent, and have not only the same mean and variance but the same 1-D probability distribution, with fourth moment $K+1$, then, as we have seen in section 3, the central limit theorem implies that with probability more than 0.997, $\mathbf{x}$ satisfies

$$|N^{-1} \|\mathbf{x}\|^2 - 1| \le 3(K/N)^{1/2}. \tag{5.20}$$

In (5.20) are two inequalities, or hard bounds, namely

$$N^{-1} \|\mathbf{x}\|^2 \le 1 + 3(K/N)^{1/2} \tag{5.21a}$$

$$N^{-1} \|\mathbf{x}\|^2 \ge 1 - 3(K/N)^{1/2}. \tag{5.21b}$$

If we use only (5.21a), we have discarded at least half the information contained in p_X. If we do so, then for large N we can neglect $(K/N)^{1/2}$ in (5.21a) and write (5.21a) as

$$N^{-1} \|\mathbf{x}\|^2 \le 1; \tag{5.22}$$

but in the same approximation we can write (5.20) as

$$N^{-1} \|\mathbf{x}\|^2 = 1. \tag{5.23}$$

In fact, if we replace the soft bound p_X by the hard bound (5.21a), we have discarded much more than half the information in p_X, because for any n such that $1 \ll n \le N$, with probability 0.997 we will have

$$\left| n^{-1} \sum_{i=1}^{n} x_i^2 - 1 \right| < 3(K/n)^{1/2}, \tag{5.24}$$

and these inequalities do not follow from (5.20). In short, the hardening of the soft bound p_X to a single quadratic inequality discards most of the hard information in p_X. Soft bounds contain much more information than the corresponding hard quadratic bounds.

6 CONCLUSIONS

From the algebraic structure of the linear inverse problem, it is clear that without prior information about the correct earth model $\mathbf{x}_E$, the prediction vector $\mathbf{z}$ is not usefully limited by the data vector $\mathbf{y}$ unless the prediction functionals are mere linear combinations of the data functionals. This conclusion does not require a topology on the model space X, much less a norm or an inner product. Therefore, when there are prediction functionals which are not linear combinations of the data functionals, the linear inverse problem is insoluble without prior information about $\mathbf{x}_E$.

The present paper compares two forms of prior information about $\mathbf{x}_E$. One is a prior personal probability distribution p_X for $\mathbf{x}_E$ in X, a 'soft bound' on $\mathbf{x}_E$. The other is a quadratic inequality

$$Q_X(\mathbf{x}_E, \mathbf{x}_E) \le 1, \tag{6.1}$$

a 'hard bound.' Energy constraints are examples of hard bounds.

We show that a hard bound can be 'softened' to many different probability distributions p_X, but all these p_X's carry large amounts of new information about $\mathbf{x}_E$ which is not present in (6.1). For example, if $\dim X = \infty$ then p_X assigns probability zero to the set of all earth models $\mathbf{x}_E$ for which $Q_X(\mathbf{x}_E, \mathbf{x}_E)$ is finite. When $\dim X$ is very large but finite, p_X assigns very small probability to the truth of (6.1), despite the fact that p_X is supposed to represent a 'fuzzy' version of (6.1). In the inverse problem of downward continuation of the geomagnetic field $\mathbf{B}$, softening the core heat flow bound with a p_X which treats appropriate multiples of the gauss coefficients as independent, identically distributed random variables will lead an observer to convert his estimate of a bound on the ohmic heat production rate in the core to a belief that this rate is infinite.

The same situation is encountered in reverse when we try to 'harden' a probability distribution p_X to a quadratic inequality (6.1). Here p_X generates a positive definite quadratic form Q_X for which (6.1) is true with high probability. However, p_X implies that many other quadratic inequalities for $\mathbf{x}_E$ are true with high probability, and all this information is lost when p_X is replaced by (6.1).

If the data vector $\mathbf{y}$ is to be inverted by means of Bayesian inference or stochastic inversion, the prior information about $\mathbf{x}_E$ must be supplied in the form of a probability distribution p_X. If there is objective evidence or a theoretical basis for p_X, or if p_X is a hypothesis to be tested, then all the prior information about $\mathbf{x}_E$ carried in p_X is legitimate, and an effective inversion will use it. However, if p_X is obtained by softening a hard quadratic bound (6.1), and $\dim X \gg 1$, then p_X contains so much more information than (6.1) that stochastic and Bayesian inversions based on p_X would appear to be suspect. If the prior information is a hard quadratic bound (6.1), the preferred technique for incorporating that information into a data inversion would appear to be confidence set inference (CSI), the multidimensional analogue of the method of confidence intervals (Kendall & Stuart 1979). CSI was explored briefly by Backus (1970a). Work in progress will discuss further details, including resolution, incorporating systematic errors, and questions of computational efficiency.

ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation grant EAR 85–21543 and NASA grant NAG 5-818. The author thanks Gary Egbert for permission to use his proof of (4.20).

REFERENCES

Backus, G., 1968. Kinematics of geomagnetic secular variation in a perfectly conducting core, *Phil. Trans. R. Soc. A*, **263,** 239–266.

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

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

Backus, G., 1975. Gross thermodynamics of heat engines in deep interior of earth, *Proc. Nat. Acad. Sci.*, **72,** 1555–1558.

Backus, G., 1986. Poloidal and toroidal fields in geomagnetic field modelling, *Rev. Geophys.*, **24,** 75–109.

Backus, G., 1988. Bayesian inference in geomagnetism, *Geophys. J. R. astr. Soc.*, **92,** 125–142.

Backus, G., 1987. Isotropic probability measures in infinite dimensional spaces, *Proc. Nat. Acad. Sci.*, **84,** 8755–8757.

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. R. Soc. A*, **266,** 123–192.

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

Bieberbach, L., 1945. *Lehrbuch der Funktionentheorie,* vol. 1, Chelsea, New York, 322 pp.

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. & 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, 304 pp.

Halmos, P. R., 1951. *Introduction to Hilbert Space,* Chelsea, New York, 114 pp.

Halmos, P. R., 1958. *Finite-dimensional Vector Spaces,* Van Nostrand, New York, 200 pp.

Heustis, S. P. & Parker, R. L., 1977. Bounding the thickness of the oceanic magnetized layer, *J. geophys. Res.*, **82,** 5293–5303.

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

Kendall, M. K. & Stuart, A., 1977. *The Advanced Theory of Statistics,* Vol. 1, Macmillan, New York, 472 pp.

Kendall, M. K. & Stuart, A., 1979. *The Advanced Theory of Statistics,* Vol. 2, Macmillan, New York, 748 pp.

Kolmogorov, A. N., 1950. *Foundations of the Theory of Probability,* Chelsea, New York, 84 pp.

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

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

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