NASA-CR-184839) CONFIDENCE SET INFERENCE WITH A PRIOR QUADRATIC BOUND (Scripps Institution of Oceanography) 102 p CSCL 08G
N89-21435
Unclas
G3/46 0197309

Confidence Set Inference with a Prior Quadratic Bound

George E. Backus

Institute of Geophysics and Planetary Physics, Scripps Institution of Oceanography, University of California, USA

Accepted _____. Received _____; in original form _____

SUMMARY

In the uniqueness part of a geophysical inverse problem, the observer wants to predict all likely values of P unknown numerical properties $\mathbf{z}=(z_1,...,z_P)$ of the earth from measurement of D other numerical properties $\mathbf{y}^{(0)}=(y_1^{(0)},...,y_D^{(0)})$, using full or partial knowledge of the statistical distribution of the random errors in $\mathbf{y}^{(0)}$. The data space Y containing $\mathbf{y}^{(0)}$ is D-dimensional, so when the model space X is infinite-dimensional the linear uniqueness problem usually is insoluble without prior information about the correct earth model $\mathbf{x}$. If that information is a quadratic bound on $\mathbf{x}$ (e.g., energy or dissipation rate), Bayesian inference (BI) and stochastic inversion (SI) inject spurious structure into $\mathbf{x}$, implied by neither the data nor the quadratic bound. Confidence set inference (CSI) provides an alternative inversion technique free of this objection. The first step in CSI is to estimate unmodelled systematic errors in $\mathbf{y}^{(0)}$ and $\mathbf{z}$. The second step is to choose any finite-dimensional subspace X_N of X, and to use the prior quadratic bound to estimate the truncation error when the full data function $F:X\to Y$ in the forward problem is approximated by restricting it to X_N to give a finite-dimensional function $F_N:X_N\to Y$. Step three calculates the eigenstructure (singular value decomposition) of F_N. Step 4 uses

this eigenstructure to find for each positive $\rho \leq 1$ a Neyman subset $K^Z(\rho)$ of the P-dimensional prediction space Z such that either the correct value of the prediction vector $\mathbf{z}$ is a member of the confidence set $K^Z(\rho)$ or an event has occurred whose probability was no more than ρ. In contrast to SI and BI, CSI offers no incentive for considering any value of P except 1. CSI is illustrated in the problem of estimating the geomagnetic field B at the core-mantle boundary (CMB) from components of B measured on or above the earth's surface. Neither the heat flow nor the energy bound is strong enough to permit estimation of B_r at single points on the CMB, but the heat flow bound permits estimation of uniform averages of B_r over discs on the CMB, and both bounds permit weighted disc-averages with continous weighting kernels. Both bounds also permit estimation of low-degree Gauss coefficients at the CMB. The heat flow bound resolves them up to degree 8 if the crustal field at satellite altitudes must be treated as a systematic error, but can resolve to degree 11 under the most favorable statistical treatment of the crust. These two limits produce circles of confusion on the CMB with diameters of 25° and 19° respectively.

Key words: confidence sets, inverse problems, core-mantle boundary field

1 INTRODUCTION

In geophysical inverse problems, the data are finitely many real numbers $y_1^{(0)}, ..., y_D^{(0)}$ measured with unknown observational errors $\delta y_1^{(0)}, ..., \delta y_D^{(0)}$. The goal is to use the data to estimate properties of the real earth. A common intermediate step is the existence problem, to find one physically reasonable earth model $\mathbf{x}^{(0)}$ which fits the data acceptably (i.e., within the random errors of observation and the systematic errors of modelling). Usually the model space X of all possible earth models $\mathbf{x}$ is infinite dimensional, so many physically reasonable models besides $\mathbf{x}^{(0)}$ will fit the data acceptably. Thus besides the existence problem there is a uniqueness problem, to estimate the "error" in $\mathbf{x}^{(0)}$, i.e., to discover how much the real earth might deviate from $\mathbf{x}^{(0)}$.

The uniqueness problem must be carefully formulated if it is to be amenable to quantitative solution. Collect the data into an observed data vector $\mathbf{y}^{(0)} = (y_1^{(0)}, ..., y_D^{(0)})$, and let Y be the data space, the vector space (linear space) of all possible D-dimensional data vectors. The theory of the forward problem provides a function F which assigns to each earth model $\mathbf{x}$ in X the value $F(\mathbf{x})$ which $\mathbf{y}^{(0)}$ would take if $\mathbf{x}$ were the correct earth model and there were no errors. The errors in $\mathbf{y}^{(0)}$ are of two types: the random error of observation, $\delta y = (\delta y_1, ..., \delta y_D)$; and the systematic error $\eta = (\eta_1, ..., \eta_D)$ due to the inadequacy of the model space X. If $\mathbf{x}$ is the correct model, then

$$\mathbf{y}^{(0)} = F(\mathbf{x}) + \delta \mathbf{y} + \eta \ . \tag{1.1a}$$

The hope is to use the data $\mathbf{y}^{(0)}$ to estimate properties of the real earth. The complete information content of an arbitrary infinite-dimensional model vector $\mathbf{x}$ can be neither registered in a finite computer nor comprehended by a finite observer, so in any real inverse problem the data $\mathbf{y}^{(0)}$ will be used to estimate a *finite* number of properties of the earth. In the present paper it will be assumed that these properties can be described by a finite list of real numbers which together constitute a "prediction vector," $\mathbf{z} = (z_1, ..., z_P)$, in the P-dimensional "prediction space" Z. The theory of the forward problem provides a function G which assigns to each model $\mathbf{x}$ in X the value $G(\mathbf{x})$ that $\mathbf{z}$ would take if $\mathbf{x}$ were the correct model and there were no errors. Thus

$$\mathbf{z} = G(\mathbf{x}) + \zeta \tag{1.1b}$$

where ζ is the systematic error produced by inadequacies in the model space. Since (1.1b) is a prediction rather than an observation, it contains no random error δz of measurement. That random error would enter only in a later attempt to verify the prediction by direct measurement, and it need not encumber the inverse problem.

The uniqueness part of the inverse problem consists in using $\mathbf{y}^{(0)}$ in (1.1a) to put limits on $\mathbf{x}$, and transferring those limits to $\mathbf{z}$ via (1.1b). Of course this program is doomed unless estimates are available for the errors $\delta y, \eta$ and ζ. By definition, a systematic error like η or ζ is one about which only inequalities are known, while the random error δy can be thought of as drawn at random from a population of error vectors in Y with a probability distribution p_E on Y.

Even if p_E for δy and inequalities for η and ζ are known, the uniqueness problem is usually insoluble for another reason, basically because $\dim Y < \dim X$, so (1.1) has more unknowns than equations. The difficulty is obvious in the linear case. Let F_i and G_j be the real-valued functions on X defined by

$$F(\mathbf{x}) = (F_1(\mathbf{x}), ..., F_P(\mathbf{x})) \tag{1.2a}$$

and

$$G(\mathbf{x}) = (G_1(\mathbf{x}), ..., G_P(\mathbf{x})) . \tag{1.2b}$$

If F and G are linear, so are the data functionals F_i and the prediction functionals G_j. If each G_j is a linear combination of $F_1, ..., F_D$, then estimates of $\mathbf{z}$ are possible (Backus & Gilbert, 1968, 1970), but otherwise the set of $\mathbf{z}$'s permitted by the data is unbounded (Backus, 1970a).

When the prediction functionals are not linear combinations of the data functionals, the observer who wants to estimate (i.e., bound) $\mathbf{z}$ from (1.1b) must invoke more information about the earth than is contained in (1.1a). This "prior information" comes from previous knowledge about the earth in particular and about physics and chemistry in general. For example, in

attempting to invert seismic data to find the density ρ and seismic velocities V_P and V_S, the seismologist is quite sure that ρ, V_P and V_S are positive everywhere, and is rather confident of upper bounds for them based on laboratory measurements, the theory of condensed matter, and solar and astronomical observations which contribute to the discussion of the likely range of chemical compositions of the earth. Examples of prior information in modelling the geomagnetic field are that its total energy cannot have a rest mass greater than that of the earth (the energy bound), and that the minimum ohmic heating rate required to sustain it is probably less than the total rate of geothermal heat flow observed at the earth's surface (the heat flow bound).

The seismic prior information just mentioned can be stated in terms of "linear bounds." There are M linear functions f_i, $i = 1,...,M$, which assign real numbers $f_i(\mathbf{x})$ to all earth models $\mathbf{x}$, and there are $2N$ real numbers $a_i < b_i$ such that the correct $\mathbf{x}$ is known to satisfy

$$a_i \le f_i(\mathbf{x}) \le b_i \tag{1.3}$$

for $i = 1,...,M$. In the seismic example, the model $\mathbf{x}$ is the triple of functions ρ, V_P, V_S, and $f_i(\mathbf{x})$ is the value of one of ρ, V_P or V_S at a particular location $\mathbf{r}_i$ in the earth. Evidently $M = \infty$ in the seismic case. This is essential, because for $M < \infty$ the $f_1,...,f_M$ simply augment the number of data functionals from D to $D + M$, and if $\dim X = \infty$ it is still true that $D + M < \dim X$. When $M = \infty$, linear constraints are extremely useful in resolving nonuniqueness. Seismic examples are given by Wiggins *et al.* (1973), Garmany (1979), Orcutt (1980), Stark *et al.* (1986), Stark (1987, 1988), and Stark and Parker (1987). Examples from gravity inversion appear in Parker (1974, 1975), and examples from geomagnetic prospecting are given by Oldenburg (1983) and Huestis and Parker (1977). McNutt and Royden (1988) give an application to the geotherm. In all these cases, the numerical inversion technique is linear programming (Gass, 1985) but many of the problems are nonlinear and involve very adroit use of the constraints.

The energy and heat flow bounds in the geomagnetic modelling problem have a mathematical character different from (1.3). They are both of the form

$$Q(\mathbf{x},\mathbf{x}) \le q \tag{1.4a}$$

where q is a known real number and Q is a known positive-definite quadratic form in the model $\mathbf{x}$. That is, $Q(\mathbf{x},\mathbf{x})$ is a homogeneous quadratic polynomial in $\mathbf{x}$, and

$$Q(\mathbf{x},\mathbf{x}) > 0 \tag{1.4b}$$

for every nonzero model $\mathbf{x}$. The bound (1.4) is a prior quadratic bound on the correct earth model $\mathbf{x}$. A single prior quadratic bound on $\mathbf{x}$ always confines the prediction vector $\mathbf{z}$ to a bounded subset K^Z of the prediction space Z, whereas infinitely many linear bounds are needed to do so if $\dim X = \infty$. Whether K^Z is small enough to be useful must usually be settled by numerical calculation. One of the conclusions reached in this paper is that very lax prior quadratic bounds on $\mathbf{x}$ can produce surprisingly tight bounds on $\mathbf{z}$ if the data are adequate. The use of a prior quadratic bound was suggested by Backus (1970a), Jackson (1979) and Gubbins and Bloxham (1985).

Jackson (1979) calls (1.3) and (1.4) "hard bounds" on the correct earth model $\mathbf{x}$, as opposed to "soft bounds." A soft bound is a probability distribution p_X on the model space X, which describes where in X the correct model $\mathbf{x}$ is likely to be. A prior soft bound p_X can be subjective or objective. If subjective, it is the observer's personal probability distribution for $\mathbf{x}$, incorporating all his or her knowledge about $\mathbf{x}$ except the data $\mathbf{y}^{(0)}$. Measuring $\mathbf{y}^{(0)}$ increases the observer's knowledge, and changes p_X from a prior to a posterior personal probability distribution. An objective p_X arises when there really are many different earth models, for example many different ocean-bottom magnetic anomaly patterns, all sampled by surface magnetometers, and some sampled by deep-towed magnetometers. The deep-towed samples constitute an objective data base from which p_X can be estimated by standard statistical techniques. This p_X can then be used to interpret a surface record from a new area.

The two currently popular methods for incorporating prior information into the inverse problem (1.1) are stochastic inversion (SI) and Bayesian inference (BI). Both methods use soft prior information, a probability distribution p_X for models $\mathbf{x}$ in X. Stochastic inversion is essentially minimum variance estimation, while Bayesian inference uses Bayes' theorem to combine p_X with $\mathbf{y}^{(0)}$ and the error distribution p_E to produce a new probability distribution $\tilde{p}_X$ for $\mathbf{x}$ in X. Both SI and BI can be applied to either a subjective or an objective prior p_X. When p_X and p_E are Gaussian and (1.1) is linear, BI and SI produce the same results. A review of the two methods and a bibliography appears in Backus (1988a). Tarantola's (1987) book about BI appeared after that bibliography was assembled.

In the geophysical literature it has become common practice to apply BI and SI to hard prior information like (1.3) and (1.4). This requires that first the hard information be "softened" to a prior probability distribution p_X. Backus (1970c, 1988a), Jackson (1979) and Gubbins and Bloxham (1985) discuss the softening process. Backus (1988b) recently showed that when $\dim X = \infty$, softening the hard quadratic bound (1.4) to a probability distribution inevitably adds spurious information about the correct model $\mathbf{x}$. This new information is subtle, and can easily escape the observer's attention. It is definitely not implied by the original inequality (1.4), and is of a character likely to be unacceptable to most observers. For example, every p_X which adds no other information to (1.4) will assign prior probability zero to the set of models $\mathbf{x}$ for which $Q(\mathbf{x}, \mathbf{x})$ is finite. When $\dim X = \infty$ and the prior information is a quadratic bound (1.4), neither SI nor BI is a suitable technique for resolving nonuniqueness in the inverse problem (1.1).

The present paper develops Neyman's (1937) theory of confidence sets as a replacement for BI and SI when the prior information is a hard quadratic bound (1.4). The general idea of confidence set inference (CSI) is described by Backus (1987) and Stark (1988), and in fact is a special case of a well-established scheme for statistical inference, as will be shown in sections 3 and 4 below. With CSI, the inevitable and often very large uncertainty in the value of q in (1.4a)

cannot be dealt with by "softening" (1.4a) to a probability distribution on X. Therefore, conclusions drawn with CSI must be tested for sensitivity to q over a range of values. In the geomagnetic problem, it will be seen that altering q by several factors of 10 does not appreciably affect the conclusions.

The present paper advocates replacing BI and SI by CSI only in certain circumstances. If the prior information really is a probability distribution p_X for $\mathbf{x}$ on X, and the observer has confidence in it, then BI and SI are preferable to CSI. To use CSI on p_X, the observer must convert p_X to a quadratic inequality (hardening the soft information, in the terminology of Jackson, 1979). When $\dim X = \infty$, this hardening process discards large amounts of prior information (Backus, 1988b).

In the existence problem as well, SI and BI are unobjectionable techniques. When $\dim X \gg \dim Y$, computational schemes for constructing an $\mathbf{x}^{(0)}$ which acceptably fits the data are prone to numerical instability because so many models are acceptable. Tikhonov (1963) and Tikhonov and Arsenin (1972) suggested a general remedy now called regularization: add another condition on $\mathbf{x}$ besides (1.1) which will make $\mathbf{x}^{(0)}$ unique, so the computer does not go into hunting oscillation or wander off to infinity in the high wave-number part of the model space. The oldest regularization scheme is simple truncation of the model space to an N-dimensional subspace X_N, from which an $\mathbf{x}^{(0)}$ is selected by a least-squares fit to the data. If the fit is unsatisfactory, N is increased. Tikhonov considered other regularization schemes, most of which seek the $\mathbf{x}$ that acceptably fits the data and minimizes some norm on X. Geophysical examples occur in Backus and Gilbert (1967) and Shure *et al.* (1982). Both BI and SI offer such techniques, and when so used they solve the existence part of the inverse problem but not the uniqueness part. This distinction is not always made clear in the literature.

One of the most convincing examples of prior quadratic information in the form (1.4) arises in trying to estimate the geomagnetic field $\mathbf{B}$ at the core-mantle boundary (CMB) from

measurements of **B** made on and above the surface of the earth. Therefore, the geomagnetic problem will be used to illustrate the general theory of CSI. The next section describes the geomagnetic problem in detail, and uses it to clarify some general issues in inversion. The remainder of the paper is a description of CSI, followed by its detailed application to the geomagnetic problem.

2 THE GEOMAGNETIC INVERSE PROBLEM

In the geomagnetic inverse problem used here to illustrate CSI, the model space X is the linear space of all possible geomagnetic fields **B** produced outside the core-mantle boundary (CMB) by electric currents inside the core. Every such **B** is defined only in the region from the CMB to infinity, and can be written there as

$$\mathbf{B} = -\nabla \psi \tag{2.1a}$$

where

$$\psi(\mathbf{r}) = a \sum_{l=1}^{\infty} (a/r)^{l+1} \sum_{m=-l}^{l} \beta_l^m(a)\, Y_l^m(\hat{\mathbf{r}}) . \tag{2.1b}$$

Here a is the radius of the CMB, $\mathbf{r}$ is the position vector measured from the center of the earth, r is $|\mathbf{r}|$, $\hat{\mathbf{r}}$ is $\mathbf{r}/r$, and $\{Y_l^{-l}, ..., Y_l^l\}$ is any orthogonal basis for the spherical harmonics of degree l, normalized so that $|Y_l^m|^2$ averages to $(2l+1)^{-1}$ on the surface of the unit sphere. The $\beta_l^m(a)$ are the internal Gauss coefficients of **B** at the CMB. They can be used to parameterize the model space X. Alternatively, X can be parameterized by any parameterization of the set of scalar-valued functions on the CMB, because **B** outside the core is completely determined by the radial component B_r on the CMB (Backus, 1986, gives several parametrizations). In this formulation of the inverse problem, the true values of B_r on and above the CMB will include unmodelled contributions from sources outside the core. These are treated as errors in the `data`.

In this geomagnetic inverse problem, the data are

$$y_i = \hat{\nu}_i \cdot \mathrm{B}(\mathrm{r}_i) \tag{2.2}$$

where, for $i = 1, ..., D$, $\hat{\nu}_i$ is a unit vector and r_i is a position on or above the surface of the earth. Equations (2.1) and (2.2) describe how to calculate the data produced by any model B if the errors vanish. Therefore, (2.1) and (2.2) give the forward data function F in (1.1a). This F is linear, but it would be nonlinear if some of the data were intensities or angles rather than Cartesian components of B.

For any i among $1, ..., D$, if $\hat{\nu}_i$ is the unit vector appearing in (2.2) then the $y_i^{(0)}$ of (1.1a) is the measured $\hat{\nu}_i$ component of the magnetic field at r_i. Thus $y_i^{(0)}$ includes the instrument errors, position errors, and the field produced by all sources outside the core (since (2.1) models only internal sources). These are magnetization in the crust and the satellite, and electric currents in the mantle, crust, ocean, ionosphere, satellite, and magnetosphere. The observer has wide latitude to apportion these contributions to $\mathbf{y}^{(0)}$ among the three terms in (1.1a): $F(\mathbf{x})$, $\delta\mathbf{y}$ and $\boldsymbol{\eta}$. For example, the instrument and navigation errors are likely to be treated as part of the random error $\delta\mathbf{y}$. The contribution to $\mathbf{y}^{(0)}$ from crustal magnetization might be unknown, but an upper bound on its magnitude might be available; then it would belong in the systematic error $\boldsymbol{\eta}$. Alternatively, perhaps the crustal magnetization is well-described by a two-dimensional random process on the surface of the earth. Then its contribution to $\mathbf{y}^{(0)}$ could be included in $\delta\mathbf{y}$. This statistical hypothesis about p_E, the probability distribution for $\delta\mathbf{y}$, could be tested by the methods described in section 3. Finally, if a representation for B more general than (2.1) were adopted (Backus, 1986), any one of the sources of B could be included as part of the earth model $\mathbf{x}$, and its contribution to B computed as part of $F(\mathbf{x})$.

The predictions z_i in (1.1b) might be the 224 Gauss coefficients of degree less than 15, or the values of B_r at 300 points on or above the CMB, or the flux of B through certain null-flux curves on the CMB (Backus, 1968; Gubbins and Bloxham, 1985). In the first two cases, G in

(1.1b) is linear, and in the first case $\zeta = \mathbf{0}$ in (1.1b) because the prediction can be made exactly once the model is known. In the last two cases, $\zeta \neq 0$ because the prediction includes a contribution from magnetic fields with sources outside the core.

In the geomagnetic example, prior inequalities (1.4) can be obtained for both energy and dissipation rate. Einstein's relativity requires that the rest mass of the energy in B be part of the total mass of the earth, as measured by surface gravity and the Cavendish experiment for Newton's gravitational constant. In the notation of (2.1), it follows that (Backus, 1988a)

$$\sum_{l=1}^{\infty} (2l+1)(l+1)^{-1} \sum_{m=-l}^{l} |\beta_l^m(a)|^2 \leq 2 \times 10^{33}\ \mathrm{nT}^2 \tag{2.3}$$

if the $\beta_l^m(a)$ are measured in nanoTesla. Alternatively, the electrical conductivity of the core probably is known to within a factor of 10, and the total geothermal heat flux is known with somewhat better accuracy (Stacey, 1977). If ohmic dissipation in the core does not exceed the geothermal heat flux then (Backus, 1988a, based on Parker, 1972, and Gubbins, 1975)

$$\sum_{l=1}^{\infty} (l+1)(2l+1)(2l+3)l^{-1} \sum_{m=-l}^{l} |\beta_l^m(a)|^2 \leq 3 \times 10^{17}\ \mathrm{nT}^2\ . \tag{2.4}$$

Since the gauss coefficients are probably of the order of 10^5 nT at the CMB, bounds like (2.3) and (2.4) appear at first sight to be unhelpful. It is one of the counterintuitive properties of BI on model spaces of high dimension that geographically well-distributed measurements of B can use a probabilistic (softened) version of (2.4) to find $\beta_1^0(a)$ from satellite data correct within about one part in 10^5, and to find $q_{10}^m(a)$ within five percent (R.A. Langel and R.H. Estes, private communication). Whether this accuracy survives the replacement of BI by CSI will be answered by future computations, to be reported elsewhere.

Other quadratic forms besides (2.3) and (2.4) have been considered in geomagnetic modelling (see, e.g., Shure *et al.*, 1982), but all have been isotropic, i.e., invariant under rotations about the center of the earth, and so have had expressions of the form

$$\sum_{l=1}^{\infty} C(l) \sum_{m=-l}^{l} |\beta_l^m(a)|^2 \le q \tag{2.5}$$

for various choices of $C(l)$.

3 CONFIDENCE SET INFERENCE IN GENERAL

Confidence set inference (CSI) is an extension of the parametric method of statistical inference originally proposed by Neyman (1937). As a formal procedure, it applies to all inverse problems, linear or not. Useful results are extracted in particular problems by exploiting their special characteristics, such as linearity or the availability of prior information. The following discussion of CSI draws heavily on Chapter 34 of Cramér (1946). The necessary measure- and set-theoretic notations are listed in Appendix A.

In CSI, the observer has measured a data vector $\mathbf{y}^{(0)} = (y_1^{(0)}, ..., y_D^{(0)})$ in the D-dimensional data space $Y = R^D$. (R is the real line and R^D is the space of the real $1 \times D$ matrices.) The observer knows that

$$\mathbf{y}^{(0)} = \tilde{\mathbf{y}}^{(0)} + \boldsymbol{\eta} \tag{3.1a}$$

where $\tilde{\mathbf{y}}^{(0)}$ and $\boldsymbol{\eta}$ are two vectors about which only the following partial information is available: there is a given (usually small) subset S^Y of Y such that

$$\boldsymbol{\eta} \in S^Y ; \tag{3.1b}$$

and $\tilde{\mathbf{y}}^{(0)}$ was drawn at random from a population whose probability distribution on Y is $\tilde{p}_Y$. This $\tilde{p}_Y$ is not known, but it is known to belong to a given $\tilde{m}$-parameter family of probability distributions on Y. The parameters, $\theta_1, ..., \theta_{\tilde{m}}$, will be collected into a parameter vector

$$\tilde{\boldsymbol{\theta}} := (\theta_1, ..., \theta_{\tilde{m}}) \tag{3.1c}$$

in the $\tilde{m}$-dimensional space $\tilde{\Theta} = R^{\tilde{m}}$ of parameter vectors. (":=" means "is defined as.") The pro-

bability distribution whose parameter vector is $\tilde{\theta}$ will be written $\tilde{p}(\tilde{\theta}, \cdot)$. All the $\tilde{p}(\tilde{\theta}, \cdot)$, for all $\tilde{\theta} \in \tilde{\Theta}$, are assumed to have the same domain Σ_Y, a particular σ-ring of subsets of Y. If $Q \in \Sigma_Y$, then the probability assigned to Q by $\tilde{p}(\tilde{\theta}, \cdot)$ will be written $\tilde{p}(\tilde{\theta}, Q)$. There is no restriction on how the $\tilde{p}(\tilde{\theta}, \cdot)$ depend on $\tilde{\theta}$. They could all be Gaussian, with means and variance matrices in Y specified as functions of $\tilde{\theta}$; or they could be spline fits to an empirical distribution obtained from the data (Constable and Parker, 1988). Neyman (1937) requires that the function $\tilde{\theta} \mapsto \tilde{p}(\tilde{\theta}, \cdot)$ be injective, but no such demand need be made here; $\tilde{p}(\tilde{\theta}_1, \cdot) = \tilde{p}(\tilde{\theta}_2, \cdot)$ will not imply $\tilde{\theta}_1 = \tilde{\theta}_2$. In the formal development of CSI, $\tilde{m}$ in (3.1c) is unrestricted and can even be infinite. In practical applications, usually

$$\tilde{m} \ll \dim Y \,, \tag{3.1d}$$

a restriction which permits use of the observed data vector $\mathbf{y}^{(0)}$ to test the hypothesis that $\tilde{p}_Y$ really does belong to the family $\tilde{p}(\tilde{\theta}, \cdot)$ (Kendall and Stuart, 1979, Ch. 30).

The observer's goal is to use $\mathbf{y}^{(0)}$ to predict P numbers $z_1, ..., z_P$. Let $\mathbf{z} := (z_1, ..., z_P)$ be the "prediction vector," a member of the prediction space $Z = R^P$. Nothing is known about $\mathbf{z}$ except that

$$\mathbf{z} = \tilde{G}(\tilde{\theta}_0) + \zeta \tag{3.1e}$$

where $\tilde{G} : \tilde{\Theta} \to Z$ is a known function, $\tilde{\theta}_0$ is unknown but is known to be one of the (possibly many) parameter vectors $\tilde{\theta}$ such that

$$\tilde{p}_Y = \tilde{p}(\tilde{\theta}, \cdot) \,, \tag{3.1f}$$

and ζ is an unknown vector about which no information is available except that

$$\zeta \in S^Z \tag{3.1g}$$

where S^Z is a given subset of Z. The vectors η in (3.1a) and ζ in (3.1e) can be thought of as unknown systematic errors.

The CSI solution to the problem (3.1) is to look for a subset $K^Z \subseteq Z$, which is small in some useful sense (to be discussed later) and which has a high probability of containing the true prediction vector **z**. This idea of probability requires careful examination because it is not quite the usual one axiomatized by Kolmogorov (1950). Unlike Bayesian inference, CSI assigns probabilities to statements rather than to subsets of Z, and the number assigned to a statement is not unique. Typically, CSI produces a statement Σ, a number $\rho \in (0, 1]$, and a proof that either Σ is true or some event E has occurred whose probability was ρ. If $\rho \ll 1$, the observer usually will feel safe in accepting Σ as true. The value of ρ is the "failure rate" for Σ, and $1-\rho$ is the "confidence level" for Σ. The statement Σ is said to hold with failure rate ρ or at confidence level $1-\rho$. An observer who always accepts statements when they have failure rate ρ will be wrong, in the long run, in a fraction of cases approximately no more than ρ. The nonuniqueness arises because if Σ is true with failure rate ρ, then obviously Σ is true with every larger failure rate. Ideally, one would like to calculate the greatest lower bound of the failure rates for Σ, but this is often a very difficult calculation.

The calculus of failure rates is somewhat different from the ordinary calculus of probability. To illustrate this, suppose that Σ_1 and Σ_2 are statements with failure rates ρ_1 and ρ_2. Let E_i be the event with probability ρ_i which must have occurred if Σ_i is false. Let $\rho_1 \wedge \rho_2$ denote the smaller of ρ_1 and ρ_2. If Σ_1 implies Σ_2, then failure of Σ_2 means that Σ_1 is false, so E_1 has occurred. Therefore Σ_2 has failure rate ρ_1. Consequently, if Σ_1 implies Σ_2, then Σ_2 has failure rate $\rho_1 \wedge \rho_2$ as well as rate ρ_2. The statement "Σ_1 or Σ_2" is false whenever both Σ_1 and Σ_2 are false. (Here "or" has its technical, logical meaning.) Then both E_1 and E_2 must have occurred. The probability of this compound event could be as high as $\rho_1 \wedge \rho_2$, but not higher, so "Σ_1 or Σ_2" holds with failure rate $\leq \rho_1 \wedge \rho_2$. The statement "$\Sigma_1$ and Σ_2" is false whenever one of Σ_1 or Σ_2 fails, i.e., whenever one of E_1 or E_2 occurs. The probability of this compound event is no greater than $\rho_1 + \rho_2$, so "Σ_1 and Σ_2" holds with failure rate $\leq \rho_1 + \rho_2$. If Σ_1 is certainly true, then $\rho_1 = 0$, and as

special cases of the above calculations, "Σ_1 or Σ_2" holds with failure rate 0 (in fact it is certain), "Σ_1 and Σ_2" holds with failure rate ρ_2, and if Σ_1 implies Σ_2 then Σ_2 holds with failure rate 0. In fact, if Σ_1 implies Σ_2 and Σ_1 is certain, then obviously so is Σ_2.

The goal of CSI is to construct for each $\rho \in (0, 1]$ a set $K^Z(\rho) \subseteq Z$, and to prove that if $\mathbf{z}$ is the true prediction vector then

$$\mathbf{z} \in K^Z(\rho) \tag{3.2}$$

with failure rate $\leq \rho$. It is desirable that the set $K^Z(\rho)$, called a confidence set for $\mathbf{z}$, be as small as possible, in a sense to be discussed later. The proof that (3.2) holds with failure rate $\leq \rho$ will make heavy use of the rules of inference described in the preceding paragraph.

Given $\rho \in (0, 1]$, the observer begins the construction of $K^Z(\rho)$ by choosing for each $\tilde{\theta} \in \tilde{\Theta}$ a set $\tilde{K}^Y(\rho, \tilde{\theta}) \in \Sigma_Y$ such that

$$\tilde{p}(\tilde{\theta}, \tilde{K}^Y(\rho, \tilde{\theta})) \geq 1 - \rho \ . \tag{3.3a}$$

For each $\tilde{\theta} \in \tilde{\Theta}$, the set $\tilde{K}^Y(\rho, \tilde{\theta})$ can be chosen in any way whatever, subject only to (3.3a).

Now the $\tilde{\theta}_0$ which appears in (3.1e) solves (3.1f), so $\tilde{p}(\tilde{\theta}_0, \cdot)$ is the true probability distribution p_Y. Therefore the probability that $\tilde{\mathbf{y}}^{(0)} \in \tilde{K}^Y(\rho, \tilde{\theta}_0)$ is at least $1 - \rho$, so the statement

$$\tilde{\mathbf{y}}^{(0)} \in \tilde{K}^Y(\rho, \tilde{\theta}_0) \tag{3.3b}$$

holds with failure rate $\leq \rho$. By hypothesis,

$$\eta \in S^Y \ . \tag{3.3c}$$

Since (3.3b) has failure rate $\leq \rho$ and (3.3c) has failure rate 0, the conjunctive statement "(3.3b) and (3.3c) are both true" has failure rate $\leq \rho + 0 = \rho$. But the truth of this conjunctive statement, together with (3.1a), implies

$$y^{(0)} \in \tilde{K}^Y(\rho, \tilde{\theta}_0) + S^Y \ . \tag{3.3d}$$

Therefore (3.3d) has failure rate $\leq \rho$. The offending event E with probability $\leq \rho$ which must

occur if (3.3d) fails is the event that (3.3b) is false; i.e.,

$$\mathbf{y}^{(0)} \in Y \setminus \tilde{K}^Y(\rho, \tilde{\theta}_0) . \tag{3.3e}$$

Next, for each $\mathbf{y} \in Y$ the observer defines a set, namely

$$\tilde{K}^{\tilde{\Theta}}(\rho, \mathbf{y}) := \{\tilde{\theta} : \tilde{\theta} \in \tilde{\Theta} \text{ and } \mathbf{y} \in K^Y(\rho, \tilde{\theta}) + S^Y\} . \tag{3.3f}$$

Clearly $\tilde{K}^{\tilde{\Theta}}(\rho, \mathbf{y}) \subseteq \tilde{\Theta}$. Furthermore, $\tilde{\theta} \in \tilde{K}^{\tilde{\Theta}}(\rho, \mathbf{y})$ iff (if and only if) $\mathbf{y} \in K^Y(\rho, \tilde{\theta}) + S^Y$. Therefore (3.3d) is true iff

$$\tilde{\theta}_0 \in \tilde{K}^{\tilde{\Theta}}(\rho, \mathbf{y}^{(0)}) , \tag{3.3g}$$

so (3.3g) has failure rate $\leq \rho$. Finally, the observer defines the set

$$K^Z(\rho) := \tilde{G}(\tilde{K}^{\tilde{\Theta}}(\rho, \mathbf{y}^{(0)})) + S^Z . \tag{3.3h}$$

Statements (3.1e) and (3.1g) are certain and (3.3g) has failure rate $\leq \rho$. If $K^Z(\rho)$ is defined by (3.3h), then (3.2) is a consequence of the simultaneous validity of (3.1e), (3.1g) and (3.3g). Therefore, by the failure rate rules of inference, when $K^Z(\rho)$ is defined by (3.3h) then (3.2) holds with failure rate $\leq \rho$. The offending event with probability $\leq \rho$ which must occur if (3.2) fails is (3.3e).

The observer has great freedom in choosing $\tilde{K}^Y(\rho, \tilde{\theta})$, the only restriction being (3.3a). This freedom can be exploited to make the set $K^Z(\rho)$ in (3.3h) as small as possible. But what does "small" mean for a subset of Z if $\dim Z > 1$? No length or volume has been defined on Z, and "small" may mean different things for different components z_i, since those components may be measured in different physical units. If $\dim Z = 1$, the "size" K^Z of a subset $K^Z \subseteq Z$ is easy to define because Z is simply the real line R. A convenient definition is

$$|K^Z| := \tfrac{1}{2}(\sup K^Z - \inf K^Z) \tag{3.4a}$$

where $\sup K^Z$ is the supremum or least upper bound of K^Z and $\inf K^Z$ is the infimum or greatest lower bound of K^Z. The quantity $2|K^Z|$ is called the "diameter" of K^Z because

$$2|K^Z| = \sup\{|u-w| : u, w \in K^Z\}. \tag{3.4b}$$

Therefore $|K^Z|$ might be called the "radius" of K^Z. Definition (3.4b) could be extended to the case $\dim Z > 1$ if $|u|$ were defined in that case, but it is not. A useful consequence of (3.4a) is that if K_1^Z and K_2^Z are any subsets of R then

$$|K_1^Z + K_2^Z| = |K_1^Z| + |K_2^Z|, \tag{3.4c}$$

with equality, not merely inequality.

Bayesian inference and stochastic inversion lead naturally to prediction spaces Z with $\dim Z > 1$, but in confidence set inference it is both possible and desirable to consider only $\dim Z = 1$. The goal of any inversion of data is numerical prediction. The point of considering the predictions $z_1, ..., z_P$ together as a single prediction vector $\mathbf{z} = (z_1, ..., z_P)$ is that for any function $h : Z \to R$ a confidence set for the numerical prediction $h(\mathbf{z})$ is obtained easily from the confidence set for $\mathbf{z}$. If $\mathbf{z} \in K^Z(\rho)$ with failure rate ρ, then clearly $h(\mathbf{z}) \in h(K^Z(\rho))$ with failure rate ρ. But trying to treat all functions $h : Z \to R$ simultaneously has a cost. It forces a compromise in choosing the $\tilde{K}^Y(\rho, \tilde{\theta})$ for (3.3a): none of the confidence set radii $|h(K^Z(\rho))|$ should be immoderately large, whatever the function $h : Z \to R$. Clearly, for any particular h, the smallest value of $|h(K^Z(\rho))|$ is achieved by ignoring other functions h' when choosing $\tilde{K}^Y(\rho, \tilde{\theta})$ to satisfy (3.3a). For any particular prediction $h(\mathbf{z})$, $\tilde{K}^Y(\rho, \tilde{\theta})$ should be tailored to the particular function $h : Z \to R$. But then the observer is simply replacing the function $\tilde{G} : \tilde{\Theta} \to Z$ in (3.1e) by the function $\tilde{g} : \tilde{\Theta} \to R$, where $\tilde{g} = h \circ \tilde{G}$, the composite of h with $\tilde{G}$. Thus in CSI the observer is really dealing with the case $\dim Z = 1$ in (3.1). This point of view has been made practical only by the advent of inexpensive and powerful computers. Before them, economy required that the data be "reduced" to a multidimensional model and its error estimates. All individual numerical predictions were then computed from the model.

There may remain cases where an observer really needs a multidimensional prediction for itself rather than as an intermediate step toward various one-dimensional predictions. The present paper does not discuss these cases. Henceforth, it will be assumed that $\dim Z = 1$, so $Z = R$, and the function $\tilde{G} : \tilde{\Theta} \to R$ in (3.1c) will be written $\tilde{g} : \tilde{\Theta} \to R$, while $\mathbf{z}$ will be written z and $\boldsymbol{\zeta}$ will be written ζ. Since the codomain of $\tilde{g}$ is R, $\tilde{g}$ is a functional, the prediction functional.

4 CONFIDENCE SET INFERENCE IN GEOPHYSICAL INVERSE PROBLEMS

Most geophysical inverse problems, even nonlinear ones, permit considerable simplification of the general formalism for CSI given in section 3. To see this, it is necessary to state the typical geophysical inverse problem at an almost pedantic level of precision, as follows.

The observer has measured the components of the data vector $\mathbf{y}^{(0)}=(y_1^{(0)},...,y_D^{(0)})$ in the data space $Y=R^D$, and wants to use $\mathbf{y}^{(0)}$ to predict the value z of a certain numerical property of the earth. As discussed at the end of section 3, the prediction is only one real number rather than several. The tools available to the observer are these: (1) an arbitrary set X called the model space, whose members $\mathbf{x}$ will be called earth models; (2) a function $F:X\to Y$ called the data function; (3) a functional $g:X\to R$ called the prediction functional; (4) a known subset S^Y of Y; (5) a known subset of S^Z of R; (6) an m-parameter family of probability distributions on Y. The case $m=0$ is permitted, and then the "family" will consist of a single known probability distribution p on Y. If $m>0$, the m parameters will be collected into a single parameter vector $\boldsymbol{\theta}:=(\theta_1,...,\theta_m)$ in the parameter space $\Theta=R^m$, and the probability distribution with parameter vector $\boldsymbol{\theta}$ will be written $p(\boldsymbol{\theta},\cdot)$. The probability which $p(\boldsymbol{\theta},\cdot)$ assigns to the subset $Q\subseteq Y$ will be written $p(\boldsymbol{\theta},Q)$. All the $p(\boldsymbol{\theta},\cdot)$ must have the same domain Σ_Y, a σ-ring of subsets of Y. Moreover, Σ_Y must be closed under various operations of Euclidean geometry. Rather than list these in detail, the present paper assumes simply that Σ_Y consists of all subsets of Y to which a Euclidean volume can be assigned, i.e., the Lebesgue-measurable sets (Halmos, 1950), so Σ_Y will include all open and all closed subsets of Y.

The information available to the observer is the following: a vector $\delta\mathbf{y}$ was drawn at random from a population in Y whose probability distribution is p_E. This p_E may be unknown but is known to belong to the given m-parameter family. That is, there is at least one $\boldsymbol{\theta}\in\Theta$ such that

$$p_E=p(\boldsymbol{\theta},\cdot)\,. \tag{4.1a}$$

The observer also knows that there are vectors $\mathbf{x}\in X$, $\boldsymbol{\eta}\in Y$ and a real number ζ such that

$$\boldsymbol{\eta}\in S^Y\,, \tag{4.1b}$$

$$\zeta \in S^Z \,, \tag{4.1c}$$

$$y^{(0)} = F(\mathbf{x}) + \eta + \delta y \,, \tag{4.1d}$$

and

$$z = g(\mathbf{x}) + \zeta \,. \tag{4.1e}$$

In (4.1), $\mathbf{x}$ will be called a correct earth model, η will be called the systematic error in modelling the data with $\mathbf{x}$, ζ will be called the systematic error in modelling the prediction with $\mathbf{x}$, and δy will be called the random error of observation.

The geophysical inverse problem (4.1) is a special case of the statistical inference problem (3.1). To see this, introduce the following definitions:

$$\tilde{\Theta} := \Theta \times X \tag{4.2a}$$

$$\tilde{\boldsymbol{\theta}} := (\boldsymbol{\theta}, \mathbf{x}) \tag{4.2b}$$

$$\tilde{\mathbf{y}}^{(0)} := F(\mathbf{x}) + \delta \mathbf{y} \tag{4.2c}$$

$$\tilde{g}(\tilde{\boldsymbol{\theta}}) := g(\mathbf{x}) \tag{4.2d}$$

and, for every $Q \in \Sigma_Y$ and every $(\boldsymbol{\theta}, \mathbf{x}) \in \tilde{\Theta}$,

$$\tilde{p}(\tilde{\boldsymbol{\theta}}, Q) := p(\boldsymbol{\theta}, Q - F(\mathbf{x})) \,. \tag{4.2e}$$

To use this correspondence between (4.1) and (3.1), the "parameter vector" $\tilde{\boldsymbol{\theta}}$ must be permitted to be a member of an arbitrary set $\tilde{\Theta}$, not necessarily an $\tilde{m}$-tuple like (3.1c). The assumption (3.1c) was never used in section 3, so abandoning it does no harm, and now (4.2b) makes sense. In practice, X is usually a linear space or a manifold of some dimension N, and then $\tilde{\boldsymbol{\theta}}$ in (4.2b) is exactly of the form (3.1c) with $\tilde{m} = \underline{m} + N$. In this special case, the desirable but not essential condition (3.1d) becomes

$$m + \dim X \ll \dim Y \,. \tag{4.3}$$

The formal application of CSI to (4.1) requires neither (3.1c) nor (4.3), but (4.3) does permit tests of the statistical hypothesis (4.1a), that δy was drawn at random from a population described by some member of the family $p(\boldsymbol{\theta}, \cdot)$.

In principle, the family $p(\theta,\cdot)$ can also depend on $\mathbf{x}$, but the simpler problem (4.1) covers most geophysical inversions. In that simpler problem, some of the flexibility available in choosing a $\tilde{K}^Y(\rho,\tilde{\theta})$ is not very useful, and can be sacrificed to the goal of economy. To construct the special kind of $\tilde{K}^Y(\rho,\tilde{\theta})$ appropriate to (4.1), for each $\theta \in \Theta$ choose a set $K^Y(\rho,\theta) \in \Sigma_Y$ such that

$$p(\theta, K^Y(\rho,\theta)) \geq 1-\rho \tag{4.4a}$$

and define (in the notation of A.3d)

$$\tilde{K}^Y(\rho,\tilde{\theta}) := F(\mathbf{x}) + K_Y(\rho,\theta) \tag{4.4b}$$

where $\tilde{\theta} = (\theta,\mathbf{x})$ as in (4.2b). Then (3.3f) becomes

$$\tilde{K}^{\tilde{\Theta}}(\rho,\mathbf{y}) := \{(\theta,\mathbf{x}) : \theta \in \Theta \text{ and } \mathbf{x} \in X \text{ and } \mathbf{y} \in F(\mathbf{x}) + K^Y(\rho,\theta) + S^Y\}. \tag{4.4c}$$

Now, since $\dim Z = 1$, (3.3h) is written

$$K^Z(\rho) = \tilde{g}(\tilde{K}^{\tilde{\Theta}}(\rho,\mathbf{y}^{(0)})) + S^Z \tag{4.4d}$$

and, at confidence level $\geq 1-\rho$ or with failure rate $\leq \rho$

$$z \in K^Z(\rho). \tag{4.4e}$$

The description of $K^Z(\rho)$ can be simplified further. For any subset $\tilde{W} \subseteq \tilde{\Theta}$ define $\Pi^X(\tilde{W})$ as

$$\Pi^X(\tilde{W}) := \{\mathbf{x} : \mathbf{x} \in X \text{ and there is at least one } \theta \in \Theta \text{ such that } (\theta,\mathbf{x}) \in \tilde{W}\}. \tag{4.5a}$$

Heuristically, $\Pi^X(\tilde{W})$ can be thought of as the shadow of $\tilde{W}$ cast on X by light rays parallel to Θ. According to (4.2d), for every subset $\tilde{W} \subseteq \tilde{\Theta}$,

$$\tilde{g}(\tilde{W}) = g(\Pi^X(\tilde{W})). \tag{4.5b}$$

Define $K^X(\rho,\mathbf{y}) \subseteq X$ as

$$K^X(\rho,\mathbf{y}) := \Pi^X(\tilde{K}^{\tilde{\Theta}}(\rho,\mathbf{y})). \tag{4.5c}$$

Then the various definitions imply that

$$K^X(\rho, \mathbf{y}) = \{\mathbf{x} : \mathbf{x} \in X \text{ and there is at least one } \boldsymbol{\theta} \in \Theta \text{ such that } \mathbf{y} \in F(\mathbf{x}) + K^Y(\rho, \boldsymbol{\theta}) + S^Y\} . \tag{4.5d}$$

Furthermore, from (4.5b) and (4.5c),

$$\tilde{g}(\tilde{K}^{\tilde{\Theta}}(\rho, \mathbf{y}^{(0)})) = g(K^X(\rho, \mathbf{y}^{(0)})) \tag{4.5e}$$

so finally

$$K^Z(\rho) = g(K^X(\rho, \mathbf{y}^{(0)})) + S^Z \tag{4.5f}$$

and, with failure rate $\leq \rho$,

$$z \in K^Z(\rho) . \tag{4.5g}$$

The observer still has great freedom in constructing the sets $K^Y(\rho, \boldsymbol{\theta})$ to satisfy (4.4a), and this freedom can be used to minimize $|g(K^X(\rho, \mathbf{y}^{(0)}))|$. By (4.5f) and (3.4c), the result will be to minimize $|K^Z(\rho)|$. To achieve the minimum possible value of $|g(K^X(\rho, \mathbf{y}^{(0)}))|$ can require intricate statistical theory; the present paper will try to make $|g(K^X(\rho, \mathbf{y}^{(0)}))|$ small, but not always as small as possible.

5 LINEAR INVERSE PROBLEMS WITH KNOWN ERROR STATISTICS AND AN INJECTIVE DATA FUNCTION

The solution (4.5) to the inverse problem (4.1) is that $z \in K^Z(\rho)$ at confidence level $\geq 1-\rho$, or with failure rate $\leq \rho$, where $K^Z(\rho)$ is given by (4.5f). This "solution" is incomplete, because no particular choice has been made for the sets $K^Y(\rho, \boldsymbol{\theta})$ in (4.4a). The present section shows one way to choose the sets $K^Y(\rho, \boldsymbol{\theta})$ and thus complete (4.5f) to an explicit solution, in what is perhaps the simplest inverse problem of all, the linear problem with known random error statistics and an injective data function.

In this simplest version of problem (4.1), X is a linear space, $F : X \to Y$ is a linear injection, $g : X \to R$ is a linear functional, and the observer knows p_E, the probability distribution in Y for the random error vector $\delta\mathbf{y}$ which appears in (4.1d).

As noted in appendix B, p_E induces a unique dot product on Y, and this makes Y a finite-dimensional Hilbert space. Therefore p_E can be discussed in the language of tensors on Hilbert spaces, as set out in appendix C. The expected value of δy, $E[\delta y]$, is known. The observer can redefine $y^{(0)}$ as $y^{(0)} - E[\delta y]$, thus achieving the result

$$E[\delta y] = 0 . \tag{5.1a}$$

Then the variance tensor of p_E is simply $E[(\delta y)(\delta y)]$, and in terms of the dot product on Y determined by p_E,

$$E[(\delta y)(\delta y)] = I_Y \tag{5.1b}$$

where I_Y is the identity tensor on Y.

Let $P_{F(X)}$ and $Q_{F(X)}$ be the orthogonal projectors of Y onto $F(X)$ and its orthogonal complement, $F(X)^{\perp}$. Then

$$P_{F(X)}(F(x)) = F(x) \tag{5.2a}$$

and

$$Q_{F(X)}(F(x)) = 0 \tag{5.2b}$$

for every $x \in X$. Applying $P_{F(X)}$ and $Q_{F(X)}$ to (4.1d) gives

$$P_{F(X)}(y^{(0)}) = F(x) + P_{F(X)}(\eta) + P_{F(X)}(\delta y) \tag{5.2c}$$
$$Q_{F(X)}(y^{(0)}) = Q_{F(X)}(\eta) + Q_{F(X)}(\delta y) . \tag{5.2d}$$

Now $F : X \to Y$ is an injection, and thus so is $(F|X) : X \to F(X)$. But $F|X$ is also a surjection, and hence is a bijection. Therefore, $(F|X)^{-1} : F(X) \to X$ exists. For brevity, this inverse function will be written simply F^{-1}, so

$$F^{-1} : F(X) \to X . \tag{5.3a}$$

Note that $F^{-1}(y)$ is defined only for $y \in F(X)$, not for all $y \in Y$. Therefore, F^{-1} could not be applied to (4.1d). But (4.1d) has been split into the two pieces (5.2c) and (5.2d), and F^{-1} can be applied to (5.2c) to give

$$\mathbf{x} = \mathbf{x}^{(0)} - F^{-1} \circ P_{F(X)}(\eta) - F^{-1} \circ P_{F(X)}(\delta \mathbf{y}) \tag{5.3b}$$

where

$$\mathbf{x}^{(0)} = F^{-1} \circ P_{F(X)}(\mathbf{y}^{(0)}) \tag{5.3c}$$

and $F^{-1} \circ P_{F(X)}$ is the composite (A.1a) of F^{-1} with $P_{F(X)}$. Substituting (5.3b) in (4.1e) gives

$$z = z^{(0)} + \zeta - g \circ F^{-1} \circ P_{F(X)}(\eta) - g \circ F^{-1} \circ P_{F(X)}(\delta \mathbf{y}) \tag{5.3d}$$

where

$$z^{(0)} = g(\mathbf{x}^{(0)}) = g \circ F^{-1} \circ P_{F(X)}(\mathbf{y}^{(0)}) . \tag{5.3e}$$

In this simplest inverse problem, (5.3d) contains the germ of the solution (4.5e,f). Since p_E is known, there is no parameter vector θ in (4.1a), and m, the number of parameters, is zero. Then in (4.4a), $K^Y(\rho, \theta)$ does not depend on any θ, so (4.4a) becomes the requirement that a set $K^Y(\rho) \subseteq Y$ be chosen so that

$$p_E(K^Y(\rho)) \geq 1 - \rho . \tag{5.4}$$

To motivate the particular choice of $K^Y(\rho)$ to be suggested here, consider the linear functional $g \circ F^{-1} \circ P_{F(X)} : Y \to R$. Since Y is a finite-dimensional Hilbert space, so is its subspace $F(X)$. Therefore there is a unique vector $\boldsymbol{\gamma} \in F(X)$, such that for all $\mathbf{y} \in F(X)$, $g \circ F^{-1}(\mathbf{y}) = \boldsymbol{\gamma} \,\square\, \mathbf{y}$. But then, for all $\mathbf{y} \in Y$,

$$g \circ F^{-1} \circ P_{F(X)}(\mathbf{y}) = \boldsymbol{\gamma} \,\square\, \mathbf{y} \tag{5.5a}$$

because $g \circ F^{-1} \circ P_{F(X)}(\mathbf{y}) = \boldsymbol{\gamma} \,\square\, P_{F(X)}(\mathbf{y}) = P_{F(X)}(\boldsymbol{\gamma}) \,\square\, \mathbf{y} = \boldsymbol{\gamma} \,\square\, \mathbf{y}$. It is also true that

$$\| g \circ F^{-1} \| = \| \boldsymbol{\gamma} \| \tag{5.5b}$$

where $\| \boldsymbol{\gamma} \| := (\boldsymbol{\gamma} \,\square\, \boldsymbol{\gamma})^{1/2}$ and $\| g \circ F^{-1} \|$ is defined as in appendix C. Hence, if $\hat{\boldsymbol{\gamma}} := \boldsymbol{\gamma} / \| \boldsymbol{\gamma} \|$, then

$$\boldsymbol{\gamma} = \| g \circ F^{-1} \| \hat{\boldsymbol{\gamma}} \tag{5.5c}$$

and

$$\| \hat{\boldsymbol{\gamma}} \| = 1 . \tag{5.5d}$$

Then

$$g \circ F^{-1} \circ P_{F(X)}(\delta y) = \|g \circ F^{-1}\|(\hat{\gamma} \square \delta y) . \tag{5.5e}$$

Now consider the real random variable $\hat{\gamma} \square \delta y$. Its probability distribution, $p_{\hat{\gamma}}$, is the marginal distribution of p_E on the one-dimensional subspace of Y spanned by $\hat{\gamma}$. If V is any Lebesgue-measurable subset of the real line then

$$p_{\hat{\gamma}}(V) = p_E(V\hat{\gamma} + \{\gamma\}^{\perp}) . \tag{5.6a}$$

The mean and variance of $\hat{\gamma} \square \delta y$ are

$$E[\hat{\gamma} \square \delta y] = 0 \tag{5.6b}$$

and

$$E[(\hat{\gamma} \square \delta y)^2] = 1 , \tag{5.6c}$$

because $E[\hat{\gamma} \square \delta y] = \hat{\gamma} \square E[\delta y] = \hat{\gamma} \square 0 = 0$, and $E[(\hat{\gamma} \square \delta y)^2] = E[(\hat{\gamma} \square \delta y)(\delta y \square \hat{\gamma})] = E[\hat{\gamma} \square (\delta y)(\delta y) \square \hat{\gamma}] = \hat{\gamma} \square E[(\delta y)(\delta y)] \square \hat{\gamma} = \hat{\gamma} \square I_Y \square \hat{\gamma} = \hat{\gamma} \square \hat{\gamma} = 1$.

To construct a $K^Y(\rho)$ in (4.4), use the notation of appendix A for open and half-open intervals. For any $\nu \in [0, \infty)$, define

$$\rho(\hat{\gamma}, \nu) = p_{\hat{\gamma}}([\nu, \infty)) + p_{\hat{\gamma}}((-\infty, -\nu]) . \tag{5.7a}$$

For simplicity, suppose that $p_{\hat{\gamma}}(V) = 0$, whenever V contains only one point, and that $p_{\hat{\gamma}}(V) > 0$ whenever V is an open interval. Then

$$\rho(\hat{\gamma}, 0) = 1 , \tag{5.7b}$$

and as ν increases from 0 to ∞, $\rho(\hat{\gamma}, \nu)$ decreases monotonically from 1 to 0. Therefore the function $\rho(\hat{\gamma}, \cdot) : [0, \infty) \to (0, 1]$ has an inverse function, $\rho(\hat{\gamma}, \cdot)^{-1} : (0, 1] \to [0, \infty)$, which will be written $\nu(\hat{\gamma}, \cdot) : (0, 1] \to [0, \infty)$. The probability is ρ that $|\hat{\gamma} \square \delta y| \geq \nu(\hat{\gamma}, \rho)$. This motivates the definition

$$K^Y(\rho) := \{y : y \in Y \text{ and } |\hat{\gamma} \square y| < \nu(\hat{\gamma}, \rho)\} . \tag{5.7c}$$

Then $\delta y \notin K^Y(\rho)$ with probability ρ. Therefore

$$|\hat{\gamma} \square \delta y| < \nu(\hat{\gamma},\rho) \tag{5.7d}$$

at confidence level $1-\rho$, or with failure rate ρ.

Now for any positive $a \in R$, let

$$B^Z(a) := \{z : z \in R \quad \text{and} \quad |z| \le a\}. \tag{5.8a}$$

Clearly

$$|B^Z(a)| = a \tag{5.8b}$$

in the sense of (3.4a). Also, from (5.5d) and (5.7d),

$$g \circ F^{-1} \circ P_{F(X)}(\delta y) \in B^Z(\|g \circ F^{-1}\| \nu(\hat{\gamma},\rho)) \tag{5.8c}$$

with failure rate ρ. Therefore, from (5.3d),

$$z \in K^Z(\rho) \tag{5.8d}$$

with failure rate ρ, where

$$K^Z(\rho) = z^{(0)} + S^Z + g \circ F^{-1} \circ P_{F(X)}(S^Y) + B^Z(\|g \circ F^{-1}\| \nu(\hat{\gamma},\rho)). \tag{5.8e}$$

Often, S^Z and S^Y, and hence $g \circ F^{-1} \circ P_{F(X)}(S^Y)$, will be symmetric about the origin; that is, $S^Z = -S^Z$ and $S^Y = -S^Y$. In this case, the notation (3.4a) provides a simple way to restate (5.8d,e), namely

$$|z - z^{(0)}| \le |S^Z| + |g \circ F^{-1} \circ P_{F(X)}(S^Y)| + \|g \circ F^{-1}\| \nu(\hat{\gamma},\rho) \tag{5.8f}$$

with failure rate ρ.

Of the three terms on the right in (5.8f), $|S^Z|$ will be easy to evaluate because a symmetric S^Z is usually defined simply by giving $|S^Z|$. For the second term on the right of (5.8f) it is always true that

$$|g \circ F^{-1} \circ P_{F(X)}(S^Y)| \le \|g \circ F^{-1}\| \; |S^Y|. \tag{5.9a}$$

where

$$|S^Y| := \sup\{\|y\| : y \in S^Y\}.$$

The inequality in (5.9a) is equality if S^Y is a ball, i.e., if there is a real β such that $S^Y = B^Y(\beta)$ where

$$B^Y(\beta) := \{\eta : \eta \in Y \quad \text{and} \quad \|\eta\| \leq \beta\} . \tag{5.9b}$$

However, (5.9a) will not be the best estimate if, as often happens, S^Y is a parallelipiped. In this case there is a $\beta = (\beta_1, ..., \beta_D) \in Y$ such that $S^Y = C^Y(\beta)$ where

$$C^Y(\beta) := \{\eta : \eta \in Y \quad \text{and} \quad |\eta_i| \leq |\beta_i| \quad \text{for} \quad i = 1, ..., D\} . \tag{5.9c}$$

If $S^Y = C^Y(\beta)$ but the δy_i are correlated, then the edges of S^Y are not mutually orthogonal in Y, and calculating $|g \circ F^{-1} \circ P_{F(X)}(S_Y)|$ becomes a problem in linear programming on a graph with 2^D vertices in N dimensions, where $N = \dim X$.

It remains to discuss the last term on the right in (5.8f). The easiest way to evaluate $\|g \circ F^{-1}\|$ is to use the identity (5.5b). To evaluate $\nu(\hat{\gamma},\rho)$, one must do the integrals with respect to dp_E in Y required to evaluate $p_{\hat{\gamma}}$ in (5.6a), and then one must carry out the integrals over the real intervals in (5.7a), to evaluate $\rho(\hat{\gamma},\nu)$. The inverse function of $\rho(\hat{\gamma},\cdot)$ is $\nu(\hat{\gamma},\cdot)$. In one special case, all these calculations can be done almost in closed form. If p_E is Gaussian, then $p_{\hat{\gamma}}$ is a Gaussian distribution on R with mean 0 and variance 1. There is only one such one-dimensional Gaussian, so $p_{\hat{\gamma}}$ does not depend on $\hat{\gamma}$, and (5.7a) becomes

$$\rho(\nu) = (2/\pi)^{1/2} \int_{\nu}^{\infty} d\xi\, e^{-1/2\xi^2} . \tag{5.10}$$

The inverse function, $\nu(\rho)$, is tabulated; a partial list of its values appears in Table 1. Rational approximations are given by Abramowitz and Stegun (1964, p. 298).

++++++++++Table 1 near here++++++++++

The solution (5.8d,e) or (5.8f) has a simple interpretation in terms of least-squares fitting. If V is any subspace of Y, and $\mathrm{y} \in Y$, then the projection $P_V(\mathrm{y})$ is that $\mathrm{v} \in V$ which minimizes $\|\mathrm{y} - \mathrm{v}\|$ (Halmos, 1958). This is the $\mathrm{v} \in V$ which provides the best fit to y in the sense of weighted least squares, the weight matrix being the inverse of the variance matrix of the random

errors (see appendix B). Thus the $\mathbf{x}^{(0)}$ of (5.3c) is that member of X such that $F(\mathbf{x}^{(0)})$ gives the best fit to $\mathbf{y}^{(0)}$ in the sense of weighted least squares. The $z^{(0)}$ of (5.3e) is the value the prediction would have if $\mathbf{x}^{(0)}$ were actually the correct earth model and there were no errors.

Of the two equations (5.2c,d) into which (4.1d) was partitioned, only (5.2c) has been used so far. If $F(X)=Y$, then $Q_{F(X)}=0$, and (5.2d) has no content. On the other hand, if

$$\dim X \ll \dim Y \tag{5.11}$$

then, since $m=0$, (4.3) holds, and it is possible to use the data vector $\mathbf{y}^{(0)}$ to test the hypothesis that $\delta\mathbf{y}$ was drawn at random from a population whose probability distribution is p_E. In the simplest case, $|S_Y| \ll 1$ so $\|Q_{F(X)}(\eta)\| \ll 1$ in (5.2d). Since any component of $\delta\mathbf{y}$ has variance 1, it follows that $Q_{F(X)}(\eta)$ can be neglected in (5.2d). Then that equation asserts that $Q_{F(X)}(\mathbf{y}^{(0)})$, the residual part of $\mathbf{y}^{(0)}$ which remains after removing the best least squares fit, is itself a random sample from the population $Q_{F(X)}(\delta\mathbf{y})$. The probability distribution of this population is the marginal distribution of p_E on $F(X)^{\perp}$. Many tests of this hypothesis are available in the extensive literature on tests of fit (see, e.g., Kendall and Stuart, 1979, Ch. 30). As an example, suppose that p_E is Gaussian. Let $N=\dim X=\dim F(X)$, and let $\{\hat{\mathbf{y}}_{N+1},...,\hat{\mathbf{y}}_D\}$ be an orthonormal basis for $F(X)^{\perp}$. Then the $D-N$ numbers $\hat{\mathbf{y}}_i \,\square\, \mathbf{y}^{(0)}$ with $i=D+1,...,N$ should be independent random samples from a one-dimensional Gaussian population with mean 0 and variance 1. There are a variety of tests of this hypothesis (Kendall and Stuart, loc cit), one being the Kolmogorov-Smirnoff test. Another test uses the fact that

$$s^2 := (D-N)^{-1} \sum_{i=N+1}^{D} (\hat{\mathbf{y}}_i \,\square\, \mathbf{y}^{(0)})^2 \tag{5.12a}$$

is the sum of $D-N$ independent identically distributed random variables, so the probability distribution for s^2 is nearly Gaussian (the exact distribution for $(D-N)s^2$ is chi-square with $D-N$ degrees of freedom). The mean of s^2 is 1, and its variance is $2(D-N)^{-1}$, so the probability is nearly $1-\rho$ that

$$|s^2-1| \le \nu(\rho)[2/(D-N)]^{1/2}. \tag{5.12b}$$

That is, if p_E is Gaussian then (5.12b) holds with failure rate approximately ρ. In (5.12b), $\nu(\cdot)$ is the inverse of the function $\rho(\cdot)$ defined by (5.10).

6 USING A PRIOR QUADRATIC BOUND TO MAKE THE DATA FUNCTION INJECTIVE

The preceding section is based on three crucial assumptions: that the observer knows the probability distribution p_E for the random error $\delta\mathbf{y}$ in the data vector $\mathbf{y}^{(0)}$; that $F:X\to Y$ and $g:X\to R$ are linear; and that the data function $F:X\to Y$ is an injection. The present section shows how to relax the third assumption, so as to treat realistic linear inverse problems in which $\dim X=\infty$. Section 5 cannot treat such problems directly because if F is injective then $\dim X=\dim F(X)$. Since $F(X)\subseteq Y$, it follows that $\dim X\le\dim Y=D<\infty$ when F is injective.

When $\dim X=\infty$ (or more generally, when F is not injective even though $\dim X<\infty$) the idea is to approximate the inverse problem (4.1) by another problem with an injective data function. To generate this subproblem, it is necessary to invoke prior information beyond that contained in the original problem (4.1). In the present section, the new prior information will be a quadratic bound (1.4). The observer knows a constant q and a positive definite quadratic form Q on X; and the observer knows that there are vectors $\mathbf{x}\in X$, $\boldsymbol{\eta}\in S^Y$ and a scalar $\zeta\in S^Z$ such that $\mathbf{x},\boldsymbol{\eta},\zeta$ satisfy (4.1) and $\mathbf{x}$ also satisfies (1.4).

Appendix B shows that $q^{-1}Q$ defines a dot product on X. The dot product of $\mathbf{x}$ and $\tilde{\mathbf{x}}$ will be written $\mathbf{x}\,\square\,\tilde{\mathbf{x}}$, and $\|\mathbf{x}\|$, the norm (or length) of the vector $\mathbf{x}$, will be defined as $\|\mathbf{x}\|=(\mathbf{x}\,\square\,\mathbf{x})^{1/2}$. Then (1.4) can be written as

$$\|\mathbf{x}\|\le 1\,. \tag{6.1}$$

Even knowing (6.1), the observer cannot reduce (4.1) to a problem with an injective data function unless certain further information is available about the quadratic form Q which appears in the prior bound, (1.4) or (6.1). Specifically, the bound (1.4) must be a strong enough constraint on $\mathbf{x}$ that it makes the given data function $F:X\to Y$ and the given prediction functional

$g: X \to R$ continuous in the norm $\|\mathbf{x}\|$ which $q^{-1}Q$ defines. That is, $\lim_{n\to\infty} \|\mathbf{x}_n - \mathbf{x}\| = 0$ must always imply $\lim_{n\to\infty} \|F(\mathbf{x}_n) - F(\mathbf{x})\| = 0$ and $\lim_{n\to\infty} |g(\mathbf{x}_n) - g(\mathbf{x})| = 0$. Since F and g are linear, continuity at $\mathbf{x} = \mathbf{0}$ is equivalent to continuity at every $\mathbf{x} \in X$ (Halmos, 1951). And since F and g are linear, continuity at $\mathbf{x} = \mathbf{0}$ is equivalent to boundedness (see appendix C for definitions and Halmos, 1951, for proofs).

When $\dim X < \infty$, F and g will always be continuous because they are linear, but when $\dim X = \infty$ linearity does not imply continuity (Halmos, 1951). Continuity of F and g with respect to the norm $\|\mathbf{x}\|$ is an extra assumption about that norm. When $\dim X = \infty$ and F is not continuous, CSI requires the observer to reexamine her or his prior knowledge of the earth and try to replace (1.4) with another such inequality strong enough to make F continuous. Backus (1970b) describes some methods ("quellings") for doing so. When $\dim X = \infty$ and g is not continuous, the observer has two choices: finding a stronger Q for (1.4), or accepting a slightly different prediction z_A for which (4.1e) can be replaced by $z_A = g_A(\mathbf{x}) + \zeta_A$, with a new prediction functional g_A which is continuous. This is the tactic adopted by Backus and Gilbert (1968, 1970).

Besides assuring that F and g are continuous, which may require reexamining the prior geophysical information, the observer must also make sure that X is a Hilbert space with the dot product defined by the quadratic form $q^{-1}Q$. This is merely a mathematical technicality, and requires no new geophysics. To say that the dot product space X is a Hilbert space is merely to say that it is complete in the norm $\|\mathbf{x}\|$; that is, all its Cauchy sequences have limits. This is a trivial requirement because if X is not complete it can always be completed (Halmos, 1951). Therefore it can be assumed without loss of generality that X in (4.1), (6.1) is a Hilbert space with the dot product defined by $q^{-1}Q$. In the present paper, it will be assumed that X is separable (Halmos, 1951). This assumption is not essential, but does simplify the notation, since it amounts to assuming that all orthonormal bases for X can be indexed by the integers. In every Hilbert space inverse problem known to the author, X is separable.

Since X is a Hilbert space and g is a bounded linear functional on X, there is a unique vector $\mathbf{g} \in X$ such that

$$g(\mathbf{x}) = \mathbf{g} \square \mathbf{x} \tag{6.2a}$$

for every $\mathbf{x} \in X$. Moreover, if $\|\mathbf{g}\| := (\mathbf{g} \square \mathbf{g})^{1/2}$, then

$$\|g\| = \|\mathbf{g}\| \tag{6.2b}$$

where $\|g\|$ is defined as in appendix C:

$$\|g\| := \sup \{ |g(\mathbf{x})| : \mathbf{x} \in B^X(1) \}, \tag{6.2c}$$

the ball $B^X(\alpha)$ being defined for any real α as

$$B^X(\alpha) := \{ \mathbf{x} : \mathbf{x} \in X \text{ and } \|\mathbf{x}\| \le \alpha \}. \tag{6.2d}$$

Also, since X and Y are both Hilbert spaces and $F : X \to Y$ is bounded and linear, there is a unique tensor $\mathbf{F} \in Y \otimes X$ (see appendix C) such that

$$F(\mathbf{x}) = \mathbf{F} \square \mathbf{x} \tag{6.2e}$$

for every $\mathbf{x} \in X$. Moreover,

$$\|F\| = \|\mathbf{F}\| \tag{6.2f}$$

where, as in appendix C

$$\|F\| := \sup \{ \|F(\mathbf{x})\| : \mathbf{x} \in B^X(1) \} \tag{6.2g}$$

and

$$\|\mathbf{F}\| := \sup \{ |\mathbf{y} \square \mathbf{F} \square \mathbf{x}| : \mathbf{y} \in B^Y(1) \text{ and } \mathbf{x} \in B^X(1) \}. \tag{6.2h}$$

Now the inverse problem (4.1), (6.1) can be written as follows:

$$p_E = p(\theta, \cdot) \tag{6.3a}$$
$$\eta \in S^Y \tag{6.3b}$$
$$\zeta \in S^Z \tag{6.3c}$$
$$\mathbf{y}^{(0)} = \mathbf{F} \square \mathbf{x} + \eta + \delta\mathbf{y} \tag{6.3d}$$
$$z = \mathbf{g} \square \mathbf{x} + \zeta \tag{6.3e}$$

$$\|\mathbf{x}\| \leq 1 \,. \tag{6.3f}$$

In the present section, the number m of parameters in the parameter vector $\boldsymbol{\theta}$ is 0, so (6.3a) simply means that p_E is known. The precise statement of the inverse problem (6.3) is this: the observer wants to estimate z from what is known. What is known is p_E, S^Y, S^Z, $\mathbf{y}^{(0)}$, $\mathbf{F}$, $\mathbf{g}$, the fact that $\delta\mathbf{y}$ was drawn at random from a population with probability density p_E, and the fact that there is a triple $(\mathbf{x},\boldsymbol{\eta},\zeta)$ which satisfies (6.3).

The observer can approximate (6.3) in many different ways by a problem with an injective data function, and so is free to choose an approximation which minimizes the error in estimating z. Section 7 will describe one technique for minimizing this error. The present section discusses how to find all the approximations to (6.3) which have injective data functions.

To produce any such approximation, let N be a positive integer, and let X_N be any N-dimensional subspace of X. It will not be necessary to assume that $\dim X = \infty$, but that is the most interesting case, so the notation will be tailored to it. Define

$$X_{(0,N)} := X_N \tag{6.4a}$$

and

$$X_{(N,\infty)} := X_N^{\perp} \,. \tag{6.4b}$$

Since X is a Hilbert space,

$$X = X_{(0,N)} \oplus X_{(N,\infty)} \tag{6.4c}$$

This conventional notation for orthogonal direct sums is described in appendix D, and simply abbreviates the following remarks. For every $\mathbf{x} \in X$ there are unique vectors $\mathbf{x}_{(0,N)} \in X_{(0,N)}$ and $\mathbf{x}_{(N,\infty)} \in X_{(N,\infty)}$ such that

$$\mathbf{x} = \mathbf{x}_{(0,N)} + \mathbf{x}_{(N,\infty)} \,. \tag{6.4d}$$

These vectors have the additional property that

$$\mathbf{x}_{(0,N)} \square \mathbf{x}_{(N,\infty)} = 0 \tag{6.4e}$$

so that

$$\|\mathbf{x}\|^2 = \|\mathbf{x}_{(0,N)}\|^2 + \|\mathbf{x}_{(N,\infty)}\|^2 . \tag{6.4f}$$

To approximate F and g, define

$$F_{(0,N)} := F \mid X_{(0,N)} \tag{6.5a}$$
$$F_{(N,\infty)} := F \mid X_{(N,\infty)} \tag{6.5b}$$
$$g_{(0,N)} := g \mid X_{(0,N)} \tag{6.5c}$$
$$g_{(N,\infty)} := g \mid X_{(N,\infty)} , \tag{6.5d}$$

and let the corresponding tensors and vectors be $\mathbf{F}_{(0,N)} \in Y \otimes X_{(0,N)}$, $\mathbf{F}_{(N,\infty)} \in Y \otimes X_{(N,\infty)}$, $\mathbf{g}_{(0,N)} \in X_{(0,N)}$ and $\mathbf{g}_{(N,\infty)} \in X_{(N,\infty)}$. Now an apparent ambiguity of notation must be resolved. There are two ways to construct $\mathbf{g}_{(0,N)}$ and $\mathbf{g}_{(N,\infty)}$ from g: the construction $g \mapsto \mathbf{g}$ in (6.2a) followed by $\mathbf{g} \mapsto (\mathbf{g}_{(0,N)}, \mathbf{g}_{(N,\infty)})$ as in (6.4d); or the construction $g \mapsto (g_{(0,N)}, g_{(N,\infty)})$ of (6.5c,d) followed by $g_{(0,N)} \mapsto \mathbf{g}_{(0,N)}$ and $g_{(N,\infty)} \mapsto \mathbf{g}_{(N,\infty)}$ as in (6.2a). Obviously both constructions give the same result, so in fact the notation is unambiguous.

Since F and g are linear, $F(\mathbf{x}) = F(\mathbf{x}_{(0,N)}) + F(\mathbf{x}_{(N,\infty)}) = F_{(0,N)}(\mathbf{x}_{(0,N)}) + F_{(N,\infty)}(\mathbf{x}_{(N,\infty)})$, and similarly for g. Therefore

$$\mathbf{F} \square \mathbf{x} = \mathbf{F}_{(0,N)} \square \mathbf{x}_{(0,N)} + \mathbf{F}_{(N,\infty)} \square \mathbf{x}_{(N,\infty)} \tag{6.6a}$$

and

$$\mathbf{g} \square \mathbf{x} = \mathbf{g}_{(0,N)} \square \mathbf{x}_{(0,N)} + \mathbf{g}_{(N,\infty)} \square \mathbf{x}_{(N,\infty)} . \tag{6.6b}$$

Substituting the expressions in (6.3d,e) suggests the following definitions:

$$\eta_{(N,\infty)} := \mathbf{F}_{(N,\infty)} \square \mathbf{x}_{(N,\infty)} + \eta \tag{6.7a}$$
$$\zeta_{(N,\infty)} := \mathbf{g}_{(N,\infty)} \square \mathbf{x}_{(N,\infty)} + \zeta \tag{6.7b}$$
$$S^Y_{(N,\infty)} := F_{(N,\infty)}(B^{X_{(N,\infty)}}(1)) + S^Y \tag{6.7c}$$
$$S^Z_{(N,\infty)} := g_{(N,\infty)}(B^{X_{(N,\infty)}}(1)) + S^Z \tag{6.7d}$$

where $B^{X_{(N,\infty)}}$ is defined by replacing X with $X_{(N,\infty)}$ in (6.2d). Note that

$$g_{(N,\infty)}(B^{X_{(N,\infty)}}(1)) = B^Z(\|\mathbf{g}_{(N,\infty)}\|) \tag{6.7e}$$

where B^Z is defined by (5.8a). The preceding definitions permit (6.3) to be recast in the following form:

$$p_E = p(\theta, \cdot) \tag{6.8a}$$
$$\eta_{(N,\infty)} \in S^Y_{(N,\infty)} \tag{6.8b}$$
$$\zeta_{(N,\infty)} \in S^Z_{(N,\infty)} \tag{6.8c}$$
$$\mathbf{y}^{(0)} = \mathbf{F}_{(0,N)} \square \mathbf{x}_{(0,N)} + \eta_{(N,\infty)} + \delta\mathbf{y} \tag{6.8d}$$
$$z = \mathbf{g}_{(0,N)} \square \mathbf{x}_{(0,N)} + \zeta_{(N,\infty)} \tag{6.8e}$$
$$\|\mathbf{x}_{(0,N)}\| \le 1 . \tag{6.8f}$$

To obtain (6.8) from (6.3), note that (6.8a) is (6.3a) and that (6.8d,e) are obvious consequences of (6.3d,e), (6.6), and (6.7a,b). The bound (6.8f) follows from (6.3f) and (6.4f). Assertions (6.8b,c) follow from (6.3b,c) once it is shown that $\mathbf{F}_{(N,\infty)} \square \mathbf{x}_{(N,\infty)} \in F_{(N,\infty)}(B^X(1))$ and $\mathbf{g}_{(N,\infty)} \square \mathbf{x}_{(N,\infty)} \in g_{(N,\infty)}(B^X(1))$. These claims are established by noting that (6.3f) and (6.4f) imply $\|\mathbf{x}_{(N,\infty)}\| \le 1$.

The problem (6.8) is formally identical with (6.3), but now the data space is $X_{(0,N)}$ with $\dim X_{(0,N)} = N < \infty$, and the data function and prediction functional are $F_{(0,N)} : X_{(0,N)} \to Y$ and $g_{(0,N)} : X_{(0,N)} \to R$. These are approximations to the original F and g, and the errors of approximation, the truncation errors, are included in (6.8b,c).

Nothing in the foregoing construction of $F_{(0,N)}$ assures that it will be injective. However, it can always be restricted to be so. Let $\mathbf{N}_{(0,N)}$ be the null space of $F_{(0,N)}$. That is,

$$\mathbf{N}_{(0,N)} := \{ \mathbf{x} : \mathbf{x} \in X_{(0,N)} \text{ and } F_{(0,N)}(\mathbf{x}) = 0 \} . \tag{6.9a}$$

Then $(F_{(0,N)} \mid \mathbf{N}^{\perp}_{(0,N)}) : \mathbf{N}^{\perp}_{(0,N)} \to Y$ is necessarily injective. To prove this, it suffices to prove that $F_{(0,N)} \mid \mathbf{N}^{\perp}_{(0,N)}$ never assigns 0 to a nonzero vector in its domain (Halmos, 1958). But if $\mathbf{x} \in \mathbf{N}^{\perp}_{(0,N)}$ and $F_{(0,N)}(\mathbf{x}) = 0$, then also $\mathbf{x} \in \mathbf{N}_{(0,N)}$, so $\mathbf{x}$ is orthogonal to itself, and hence $\mathbf{x} = 0$.

Now let

$$M = \dim \mathbf{N}^{\perp}_{(0,N)} \tag{6.9b}$$

and in the argument leading from (6.3) to (6.8) use

$$X_M = \mathrm{N}_{(0,N)}^{\perp} \tag{6.9c}$$

instead of X_N. Then $X_{(0,M)} = X_M = \mathrm{N}_{(0,N)}^{\perp}$ and the new $F_{(0,M)}$ will be an injection of $X_{(0,M)}$ into Y. Thus section 5 becomes applicable. The numerical calculation of $X_{(0,M)}$ and $F_{(0,M)}$ from $X_{(0,N)}$ and $F_{(0,N)}$ can be done quite explicitly, since $F_{(0,M)} = F_{(0,N)} \mid X_{(0,M)} = F \mid \mathrm{N}_{(0,N)}^{\perp}$. For any vector $\mathbf{x} \in X_{(0,N)}$, the vector $F_{(0,N)}(\mathbf{x})$ in Y can be written $(F^1_{(0,N)}(\mathbf{x}), \ldots, F^D_{(0,N)}(\mathbf{x}))$. For any $i \in \{1, \ldots, D\}$ the real number $F^i_{(0,N)}(\mathbf{x})$ depends linearly and continuously on $\mathbf{x} \in X_{(0,N)}$. Therefore $F^i_{(0,N)} : X_{(0,N)} \to R$ is a continuous linear functional on $X_{(0,N)}$. Hence there is a unique vector $\mathbf{F}^i_{(0,N)} \in X_{(0,N)}$ such that

$$F^i_{(0,N)}(\mathbf{x}) = \mathbf{F}^i_{(0,N)} \square \mathbf{x} \tag{6.9d}$$

for every $\mathbf{x} \in X_{(0,N)}$. The functional $F^i_{(0,N)}$ will be called the ith restricted data functional and $\mathbf{F}^i_{(0,N)}$ will be called the ith restricted data kernel. The space $\mathrm{N}_{(0,N)}^{\perp}$ is spanned by the D restricted data kernels $\mathbf{F}^1_{(0,N)}, \ldots, \mathbf{F}^D_{(0,N)}$.

One particular choice of X_N will assure from the outset that $F_{(0,N)} : X_{(0,N)} \to Y$ is injective, so that the extra step described in the preceding two paragraphs is unnecessary. Let N_F be the whole null space of F, that is

$$\mathrm{N}_F := \{ \mathbf{x} : \mathbf{x} \in X \text{ and } F(\mathbf{x}) = 0 \} . \tag{6.10a}$$

Writing $F(\mathbf{x}) = (F^1(\mathbf{x}), \ldots, F^D(\mathbf{x}))$ defines the D data functionals $F^i : X \to R$, and (6.2a) provides the corresponding data kernels $\mathbf{F}^i \in X$ (Backus & Gilbert, 1967). The space $\mathrm{N}_F^{\perp}$ is spanned by $\mathbf{F}^1, \ldots, \mathbf{F}^D$, so $\dim \mathrm{N}_F^{\perp} \le D$. Take $N = \dim \mathrm{N}_F^{\perp}$ and

$$X_N = \mathrm{N}_F^{\perp} . \tag{6.10b}$$

When this X_N is used to obtain (6.8) from (6.3), clearly $F_{(0,N)} : X_{(0,N)} \to Y$ will be an injection.

Choosing $X_N = \mathrm{N}_F^{\perp}$ avoids the extra step (6.9) in the approximation process for (6.3), and has the further advantage that with this choice

$$\mathbf{F}_{(N,\infty)} = \mathbf{0} \tag{6.10c}$$

and (6.7c) becomes

$$S^Y_{(N,\infty)} := S^Y \ . \tag{6.10d}$$

In other words, (6.10b) produces the smallest possible truncation error (namely **0**) among all choices of subspaces of X. However, (6.10b) has one serious disadvantage: it requires at least some numerical computation with $D \times D$ full matrices, and D is often very large. For satellite magnetic data, D is typically of the order of 10^4. Another problem is almost certain to appear in (6.10) but can also arise in (6.9). The data functionals $F^1, ..., F^N$ in (6.10) or $F^1, ..., F^M$ in (6.9) may be linearly dependent or nearly dependent. Then a very small subset of them may carry all the information in $\mathbf{y}^{(0)}$ which exceeds the noise, $\eta + \delta\mathrm{y}$. Computations with the whole linearly independent set of data functionals will be numerically ill-conditioned and wasteful of computer resources. All these questions are addressed in the next section, which shows a way better than (6.9c) to choose X_M, whether X_N is arbitrary or is given by (6.10).

7 RESOLUTION IN CONFIDENCE SET INFERENCE

When $\dim X = \infty$ and the observer accepts a prior quadratic bound on the correct earth model $\mathbf{x}$, the choices (6.9c) and (6.10b) provide two approximations to the linear inverse problem (6.3). Each approximation provides a finite-dimensional model space X, an injective data function $F : X \to Y$, and known confining sets S^Y and S^Z for the systematic errors η and ζ. Therefore both approximations are amenable to treatment by section 5. However, both approximations are potentially subject to a serious numerical defect. Even though $F^{-1} : F(X) \to X$ exists in both cases, $\|F^{-1}\|$ may be so large that $\|g \circ F^{-1}\|$ is too large to make (5.8d,e) or (5.8f) a useful estimate of z. Since F is linear, all that is required to make F injective is that $F(\hat{\xi}) \neq 0$ for all $\hat{\xi} \in \partial B^X(1)$, where $\partial B^X(1) := \{\hat{\mathbf{x}} : \hat{\mathbf{x}} \in X \text{ and } \|\hat{\mathbf{x}}\| = 1\}$. Nothing prevents $F(\hat{\xi})$ from being very small for some such $\hat{\xi}$, so that $\|F^{-1}\|$ is large. Physically speaking, if $\hat{\xi} \in \partial B^X(1)$ and $F(\hat{\xi})$ is smaller than the error $\eta + \delta\mathbf{y}$, then whatever component $\hat{\xi} \,\square\, \mathbf{x}$ of $\hat{\xi}$ is present in the correct earth model $\mathbf{x}$, that component will contribute to $\mathbf{y}^{(0)}$ a "signal" which is below the noise, because (6.3f) requires $|\hat{\xi} \,\square\, \mathbf{x}| \leq 1$. Then, even though $F(\hat{\xi}) \neq 0$, it is fruitless to try to resolve $\hat{\xi} \,\square\, \mathbf{x}$ and use it in estimating z.

In Bayesian inference and stochastic inversion, the computations required to pick out the resolvable parts of the earth model $\mathbf{x}$ in problem (6.3) are well-known (Jackson, 1979; Gubbins and Bloxham, 1985; Backus, 1988a). The present section describes the corresponding calculations of resolution for confidence set inference (CSI).

The first step is to use the prior quadratic bound to approximate (6.3) by a finite-dimensional problem (6.8). There is no restriction on the X_N used here, and either (6.9c) or (6.10b) is acceptable. It will be assumed, however, that

$$N = \dim X_N \leq \dim Y = D \ . \tag{7.1a}$$

A more complicated notation would make (7.1a) avoidable, but the effort is of doubtful value, since (7.1a) holds in most practical inversions, and certainly in (6.10b).

Resolution in CSI is based on the eigenstructure or singular value decomposition (SVD) of the finite-dimensional approximate data function $F_{(0,N)}: X_{(0,N)} \to Y$. The eigenstructure is available because both $X_{(0,N)}$ and Y are Hilbert spaces; and the eigenstructure is physically useful because the dot products on X and Y arise from real physical information about the correct earth model X and the random error δy. Appendix D describes SVD in the form most convenient for the present discussion.

Let the eigenfactors or singular values of $F_{(0,N)}$ in descending order be

$$\phi_1 \geq \phi_2 \geq \cdots \geq \phi_N \geq 0 \tag{7.1b}$$

where multiple eigenfactors are repeated according to their multiplicities. Let $\{\hat{\mathbf{x}}_1 ..., \hat{\mathbf{x}}_N\}$ and $\{\hat{y}_1, ..., \hat{y}_N\}$ be orthonormal subsets of $X_{(0,N)}$ and Y such that

$$F_{(0,N)}(\hat{\mathbf{x}}_i) = \phi_i \, \hat{y}_i \tag{7.1c}$$

for $i = 1, ..., N$. Since $N = \dim X_{(0,N)}$, $\{\hat{\mathbf{x}}_1, ..., \hat{\mathbf{x}}_N\}$ is an orthonormal basis for $X_{(0,N)}$ and can be extended to an orthonormal basis for all of X, written $\{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \cdots\}$. For each integer i define

$$g_i := \mathbf{g} \,\square\, \hat{\mathbf{x}}_i \tag{7.1d}$$

and

$$x_i := \mathbf{x} \,\square\, \hat{\mathbf{x}}_i \tag{7.1e}$$

so that

$$\mathbf{x} = \sum_{i=1}^{\infty} x_i \, \hat{\mathbf{x}}_i \tag{7.1f}$$

and

$$\mathbf{g} = \sum_{i=1}^{\infty} g_i \, \hat{\mathbf{x}}_i \, . \tag{7.1g}$$

Here $\mathbf{x}$ and $\mathbf{g}$ are the earth model and prediction functional appearing in (6.3).

Calculation of resolution is facilitated by the following notation. For any integers p and q such that $0 \leq p < q$, define $X_{(p,q)}$ as the subspace of X spanned by $\{\hat{\mathbf{x}}_{p+1}, ..., \hat{\mathbf{x}}_q\}$. Define $X_{(p,p)} = \{0\}$, and let $X_{(p,\infty)}$ be the subspace of X spanned by $\{\hat{\mathbf{x}}_{p+1}, \hat{\mathbf{x}}_{p+2}, \cdots\}$. For the special

cases $X_{(0,N)}$ and $X_{(N,\infty)}$ these definitions agree with those already introduced in (6.4). If $0 \le p \le q \le \infty$, define

$$F_{(p,q)} := F \mid X_{(p,q)}; \tag{7.2a}$$

$$g_{(p,q)} := g \mid X_{(p,q)}; \tag{7.2b}$$

$$Y_{(p,q)} := F(X_{(p,q)}); \tag{7.2c}$$

$$P^X_{(p,q)} := P_{X_{(p,q)}}, \tag{7.2d}$$

the orthogonal projector of X onto $X_{(p,q)}$;

$$P^Y_{(p,q)} := P_{Y_{(p,q)}}, \tag{7.2e}$$

the orthogonal projector of Y onto $Y_{(p,q)}$;

$$\mathbf{x}_{(p,q)} := P^X_{(p,q)}(\mathbf{x}); \tag{7.2f}$$

and

$$\mathbf{y}^{(0)}_{(p,q)} := P^Y_{(p,q)}(\mathbf{y}^{(0)}). \tag{7.2g}$$

This notation leads to a number of useful identities, some of which are expressed most easily in the tensor language of appendices C and D. If

$$0 \le p \le q \le \infty \tag{7.3a}$$

then

$$\mathbf{x}_{(p,q)} = \sum_{i=p+1}^{q} x_i \, \mathbf{x}_i \tag{7.3b}$$

$$\mathbf{g}_{(p,q)} = \sum_{i=p+1}^{q} g_i \, \hat{\mathbf{x}}_i \tag{7.3c}$$

and

$$\mathbf{P}^X_{(p,q)} = \sum_{i=p+1}^{q} \hat{\mathbf{x}}_i \, \hat{\mathbf{x}}_i \tag{7.3d}$$

where a sum is understood to vanish if its upper limit is less than its lower limit. If

$$0 \le p \le q \le N \tag{7.4a}$$

then

$$\mathbf{P}^{Y}_{(p,q)} = \sum_{i=p+1}^{q} \mathfrak{y}_i \, \mathfrak{y}_i \tag{7.4b}$$

and

$$\mathbf{F}_{(p,q)} = \sum_{i=p+1}^{q} \phi_i \, \mathfrak{y}_i \, \mathfrak{x}_i \, . \tag{7.4c}$$

If $0 \le p \le q$ and q is such that

$$\phi_q > 0 \tag{7.5a}$$

then (7.1b) implies that $F_{(p,q)} : X_{(p,q)} \to Y_{(p,q)}$ is a bijection whose inverse $F^{-1}_{(p,q)}$ is represented by the tensor

$$\mathbf{F}^{-1}_{(p,q)} = \sum_{i=p+1}^{q} \phi_i^{-1} \, \mathfrak{x}_i \, \mathfrak{y}_i \, . \tag{7.5b}$$

Here $F^{-1}_{(p,q)}$ has as its domain $Y_{(p,q)}$, but $\mathbf{F}^{-1}_{(p,q)}$ can be interpreted as a member of either $X_{(p,q)} \otimes Y_{(p,q)}$ or $X \otimes Y$. Equation (7.5b) is simply (D.11e) in appendix D. It is also important to observe that (7.5b), (7.4b) and the orthonormality of $\{\mathfrak{y}_1, ..., \mathfrak{y}_N\}$ imply

$$\mathbf{F}^{-1}_{(p,q)} \square \mathbf{P}^{Y}_{(p,q)} = \mathbf{F}^{-1}_{(p,q)} \, . \tag{7.5c}$$

If n is any integer satisfying

$$0 \le n < N \, , \tag{7.6a}$$

then in the standard notation described in appendix D

$$X_{(0,N)} = X_{(0,n)} \oplus X_{(n,N)} \tag{7.6b}$$

$$X_{(n,\infty)} = X_{(n,N)} \oplus X_{(N,\infty)} \tag{7.6c}$$

$$X = X_{(0,n)} \oplus X_{(n,N)} \oplus X_{(N,\infty)} \, . \tag{7.6d}$$

Now the idea is to approximate (6.3) by each of the $X_{(0,n)}$ for $n = 1, ..., N$ and to try to select n to minimize $|K^{Z}_{(0,n)}(\rho)|$, the radius of the confidence interval with failure rate ρ for the prediction z. For each n in (7.6a), definitions (6.7a,b) become

$$\eta_{(n,\infty)} := \mathbf{F}_{(n,\infty)} \square \mathbf{x}_{(n,\infty)} + \eta \tag{7.6e}$$

$$\zeta_{(n,\infty)} := \mathbf{g}_{(n,\infty)} \square \mathbf{x}_{(n,\infty)} + \zeta \, . \tag{7.6f}$$

If n is small enough so that

$$\phi_n > 0 \tag{7.7a}$$

then $F_{(0,n)}: X_{(0,n)} \to Y$ is an injection, and section 5 can be applied verbatim to the linear, finite-dimensional injective inverse problem whose terms in (4.1) are $X := X_{(0,n)}$, $F := F_{(0,n)}$, $g := g_{(0,n)}$, $\mathbf{x} := \mathbf{x}_{(0,n)}$, $\eta := \eta_{(n,\infty)}$, and $\zeta = \zeta_{(n,\infty)}$. The solution, (5.3d) or (5.8e), can be slightly simplified because now $X_{(0,n)}$ as well as Y has a dot product, and tensors are available. Using the tensor notation, (5.3c,e) become

$$\mathbf{x}^{(0)}_{(0,n)} := \mathbf{F}^{-1}_{(0,n)} \square \mathbf{y}^{(0)}_{(0,n)} \tag{7.7b}$$

and

$$z^{(0)}_{(0,n)} := \mathbf{g}_{(0,n)} \square \mathbf{x}^{(0)}_{(0,n)} . \tag{7.7c}$$

Using (7.5c) makes it possible to write the solution (5.3d) in the form

$$z = z^{(0)}_{(0,n)} + \zeta_{(n,\infty)} - \mathbf{g}_{(0,n)} \square \mathbf{F}^{-1}_{(0,n)} \square (\eta_{(n,\infty)} + \delta\mathbf{y}) . \tag{7.7d}$$

This last equation can be slightly simplified. Clearly $F_{(n,\infty)} = F_{(n,N)} + F_{(N,\infty)}$ and, from (7.4c) and (7.5b),

$$\mathbf{F}^{-1}_{(0,n)} \square \mathbf{F}_{(n,N)} = \mathbf{0} , \tag{7.8a}$$

so

$$\mathbf{F}^{-1}_{(0,n)} \square \mathbf{F}_{(n,\infty)} = \mathbf{F}^{-1}_{(0,n)} \square \mathbf{F}_{(N,\infty)} . \tag{7.8b}$$

Therefore, from definition (7.6e) applied to n and N

$$\mathbf{F}^{-1}_{(0,n)} \square \boldsymbol{\eta}_{(n,\infty)} = \mathbf{F}^{-1}_{(0,n)} \square \boldsymbol{\eta}_{(N,\infty)} \tag{7.8c}$$

and (7.7d) can be replaced by

$$z = z^{(0)}_{(0,n)} + \zeta_{(n,\infty)} - \mathbf{g}_{(0,n)} \square \mathbf{F}^{-1}_{(0,n)} \square (\eta_{(N,\infty)} + \delta\mathbf{y}) . \tag{7.8d}$$

For the particular choice $X_N = \mathrm{N}_F^{\perp}$, i.e., (6.10b), clearly $\mathbf{F}_{(N,\infty)} = \mathbf{0}$, so in (7.8d) $\eta_{(N,\infty)}$ can be replaced by η, the original systematic error in (4.1b) and (4.1d).

For any $\rho \in (0, 1]$ a confidence interval $K^Z_{(0,n)}(\rho)$ can be written down from (7.8d) such that

$$z \in K^Z_{(0,n)}(\rho) \tag{7.9a}$$

with failure rate ρ. Here z is the correct value of the prediction. To construct $K^Z_{(0,n)}(\rho)$, first define

$$\gamma_{(0,n)} := g_{(0,n)} \square \Gamma^{-1}_{(0,n)} \tag{7.9b}$$

and

$$\hat{\gamma}_{(0,n)} := \gamma_{(0,n)} / \|\gamma_{(0,n)}\| \,. \tag{7.9c}$$

Let

$$\nu_n(\rho) = \nu(\hat{\gamma}_{(0,n)}, \rho) \tag{7.9d}$$

as calculated from p_E via (5.7). Then (5.8c) and (7.5c) permit the assertion that, with failure rate ρ,

$$|\, g_{(0,n)} \square \Gamma^{-1}_{(0,n)} \square \delta y \,| \le \|\gamma_{(0,n)}\| \, \nu_n(\rho) \,. \tag{7.9e}$$

For any set $Q \subseteq Y$ and any $\gamma \in Y$, define

$$\gamma \square Q := \{\gamma \square y : y \in Q\} \,.$$

Then (7.8d) implies that with failure rate $\le \rho$

$$z \in z^{(0)}_{(0,n)} + S^Z_{(n,\infty)} - \gamma_{(0,n)} \square S^Y_{(N,\infty)} + B^Z(\|\gamma_{(0,n)}\| \nu_n(\rho)) \tag{7.9f}$$

where $\nu_n(\rho)$ is given by (7.9d), $\gamma_{(0,n)}$ by (7.9b), and B^Z by (5.8a). Using (6.7d,e) to evaluate $S^Z_{(n,\infty)}$ shows that the set on the right of (7.9f) is

$$K^Z_{(0,n)}(\rho) := z^{(0)}_{(0,n)} + S^Z + B^Z(\|g_{(n,\infty)}\| + \nu_n(\rho)\|\gamma_{(0,n)}\|) - \gamma_{(0,n)} \square S^Y_{(n,\infty)} \,. \tag{7.9g}$$

Thus, when $K^Z_{(0,n)}(\rho)$ is defined by (7.9g), (7.9a) holds with failure rate $\le \rho$. Usually, S^Z and $S^Y_{(N,\infty)}$ are symmetrical about the origin, and then (7.9a) and (7.9g) together are equivalent to the assertion that, with failure rate $\le \rho$,

$$|\, z - z^{(0)}_{(0,n)} \,| \le |\, S^Z \,| + \|g_{(n,\infty)}\| + \nu(n,\rho)\|\gamma_{(0,n)}\| + |\, \gamma_{(0,n)} \square S^Y_{(N,\infty)} \,| \,. \tag{7.9h}$$

The foregoing calculation can be repeated for any n satisfying (7.7a), and the idea is to choose n to minimize the right side of (7.9h). Therefore it is necessary to understand how the terms in (7.9h) vary with n. In the interest of simplicity it will be assumed that

$$S^Y_{(N,\infty)} = B^Y(\beta)\,, \tag{7.10a}$$

the ball defined by (5.9b). Then, by (7.8c),

$$|\gamma_{(0,n)} \square S^Y_{(n,\infty)}| = \|\gamma_{(0,n)}\|\beta \tag{7.10b}$$

so (7.9h) becomes

$$|z - z^{(0)}_{(0,n)}| \le |S^Z| + T_\rho(n) \tag{7.11a}$$

where

$$T_\rho(n) := \|g_{(n,\infty)}\| + \|\gamma_{(0,n)}\|\kappa_n(\rho) \tag{7.11b}$$

and

$$\kappa_n(\rho) := \beta + \nu(n,\rho)\,. \tag{7.11c}$$

Let M be the largest n such that $\phi_n > 0$. The value of $T_\rho(n)$ can be calculated for $n = 0, 1, , \ldots, M$, and the minimum can be chosen by inspection.

Why might the observer expect to find values of $n \in \{0, \ldots, M\}$ for which $T_\rho(n)$ is substantially less then $T_\rho(0)$ or $T_\rho(M)$? A rigorous general proof seems unattainable, but a suggestive discussion can be given. From (7.11b), $T_\rho(0) = \|g_{(0,\infty)}\| = \|g\|$. In this case, the estimate of z in (7.11a) comes entirely from the prior information (6.3f). The data vector $\mathbf{y}^{(0)}$ contributes nothing. If the data are at all relevant to the prediction, then $T_\rho(n)$ should decrease initially as n increases from 0 and the data begin to contribute to the estimate of z. But why should $T_\rho(n)$ start to increase again before n reaches M?

The reason lies in a characteristic property of most non-trivial linear inverse problems. In an ideal problem, all possible data could be collected without error, and would determine a unique earth model $\mathbf{x}$. In this ideal case, both the model space X and the data space Y are infinite dimensional Hilbert spaces, and the data function $F : X \to Y$ is a linear injection. In the non-

trivial ideal problems, however, F is compact (Kato, 1976), so that $\phi_n \to 0$ as $n \to \infty$. In a real inverse problem with $\dim Y < \infty$, but with Y large enough to give some hope of approximating the ideal problem and obtaining useful limits on the prediction, the observer should expect to find

$$\phi_n \ll 1 \tag{7.12a}$$

as n increases toward M, and indeed this is commonly observed (Gubbins and Bloxham, 1985; Langel and Estes, private communication).

To see more clearly how $T_\rho(n)$ will vary with n, introduce the abbreviations $R(n) := \|g_{(n,\infty)}\|^2$ and $S(n) := \|\gamma_{(0,n)}\|^2$. Then from (7.3c)

$$R(n) = \sum_{i=n+1}^{\infty} g_n^2 \tag{7.12b}$$

and from (7.4c), (7.5b) and (7.10b)

$$S(n) = \sum_{i=1}^{n} \phi_i^{-2} g_i^2 . \tag{7.12c}$$

Furthermore,

$$T_\rho(n) = R(n)^{1/2} + S(n)^{1/2} \kappa_n(\rho) . \tag{7.12d}$$

Define

$$D_\rho(n) := g_n^{-2} [T_\rho(n-1) - T_\rho(n)] . \tag{7.13a}$$

Then, if the data are relevant, one can expect $D_\rho(n) > 0$ for small n. If $T_\rho(n)$ does have a minimum, then eventually $D_\rho(n)$ will become negative. But a simple calculation gives

$$D_\rho(n) = [R(n-1)^{1/2} + R(n)^{1/2}] - \phi_n^{-2} \kappa_n(\rho) [S(n-1)^{1/2} + S(n)^{1/2}] . \tag{7.13b}$$

Thus $D_\rho(n) > 0$ exactly as long as

$$\phi_n > \frac{\kappa_n(\rho)[R(n-1)^{1/2} + R(n)^{1/2}]}{\phi_n [S(n-1)^{1/2} + S(n)^{1/2}]} . \tag{7.13c}$$

Clearly, if $0 \le n \le M$,

$$R(n-1)^{1/2} + R(n)^{1/2} \geq 2\,\|\mathbf{g}_{(M,\infty)}\|$$

while (7.1b) assures that

$$\phi_n\,[S(n-1)^{1/2} + S(n)^{1/2}] \leq 2\,\|\mathbf{g}_{(0,M)}\|\ .$$

Except for distributions (5.6a) with pathologically long tails (e.g., a superposition of two Gaussians with quite different variances) it will happen that $\nu(\hat{\gamma}_{(0,n)},\rho) > 1$ for the interesting values of ρ (see Table 1). Then $\kappa_n(\rho) > 1$, and the right side of (7.13c) will always be at least $\|\mathbf{g}_{(M,\infty)}\| / \|\mathbf{g}_{(0,M)}\|$. When n becomes so large that

$$\phi_n < \|\mathbf{g}_{(M,\infty)}\| / \|\mathbf{g}_{(0,M)}\| \tag{7.13d}$$

then $D_\rho(n) < 0$, and $T_\rho(n)$ will increase with n. The minimum value of $T_\rho(n)$ will certainly have been passed.

The foregoing calculation of a confidence interval for z makes no explicit attempt to enforce the constraint (6.1), and the n which minimizes $T_\rho(n)$ may not satisfy

$$\|\mathbf{x}_{(0,n)}\| \leq 1\ . \tag{7.14a}$$

As a result, the confidence interval $K^Z(\rho)$ is conservatively long, and could be shortened by enforcing (7.14a) as a further constraint. It is not clear whether the required calculations are worth doing. In practise, q in (1.4a) will not be known with any precision, and therefore the CSI estimate of $|K^Z(\rho)|$ will not inspire much confidence unless

$$\|\mathbf{x}_{(0,n)}\| \ll 1 \tag{7.14b}$$

at the value of n which minimizes $T_\rho(n)$. The cautious observer will compute $\|\mathbf{x}_{(0,n)}\|$ and check (7.14b) as an essential part of CSI. For values of n near M, even (7.14a) is likely to fail, but such values of n are of no interest in estimating z.

8 ESTIMATING THE PROBABILITY DISTRIBUTION OF THE RANDOM ERRORS

In sections 5, 6, and 7 it was assumed that the probability distribution p_E for the random error vector δy is known. In fact, under certain conditions p_E can be estimated from the data vector $\mathbf{y}^{(0)}$. The conditions are that p_E be known to belong to a given m-parameter family of distributions $p(\boldsymbol{\theta}, \cdot)$; that there be a model space X such that $F: X \to Y$ is an injection; that (4.3) holds; and that $|S^Y| \ll 1$. These conditions must often be verified a posteriori, because $|S^Y|$ depends on p_E, that is, on the unknown value of $\boldsymbol{\theta}$. The optimal choice $X = X_{(0,n)}$ in section 7 will also depend on this unknown $\boldsymbol{\theta}$.

The recursive character of the foregoing statement of conditions under which p_E can be estimated from the data is one result of the nonlinearity of the problem of estimating $\boldsymbol{\theta}$. It is difficult to make general statements about nonlinear inverse problems, but one class of problems admits a general solution by the method of maximum likelihood. This is the class in which all the probability distributions $p(\boldsymbol{\theta}, \cdot)$ in the given family have density functions $f(\boldsymbol{\theta}, \cdot): Y \to R$. That is, for any Lebesgue measurable subset Q of Y,

$$p(\boldsymbol{\theta}, Q) = \int_Q f(\boldsymbol{\theta}, \mathbf{y})\, dy_1 \cdots dy_D \,, \tag{8.1a}$$

a result usually abbreviated as

$$dp(\boldsymbol{\theta}, \mathbf{y}) = f(\boldsymbol{\theta}, \mathbf{y})\, dy_1 \cdots dy_D \,. \tag{8.1b}$$

Then the probability distributions (4.2e) all have densities given by

$$d\bar{p}(\tilde{\boldsymbol{\theta}}, \mathbf{y}) = f(\boldsymbol{\theta}, \mathbf{y} - F(\mathbf{x}))\, dy_1 \cdots dy_D \,, \tag{8.1c}$$

where $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}, \mathbf{x})$. The method of maximum likelihood estimates $\boldsymbol{\theta}$ and $\mathbf{x}$ from $\mathbf{y}^{(0)}$ by choosing $\boldsymbol{\theta}^{(0)}$ and $\mathbf{x}^{(0)}$ to maximize $f(\boldsymbol{\theta}, \mathbf{y}^{(0)} - F(\mathbf{x}))$. The estimate is useful when

$$\lambda = \dim Y - \dim X - m \tag{8.1d}$$

satisfies $\lambda \gg 1$. The λ in (8.1d) is called the number of degrees of freedom in the inverse problem. If $f(\boldsymbol{\theta}, \mathbf{y}^{(0)} - F(\mathbf{x}))$ does have a unique maximum $(\boldsymbol{\theta}^{(0)}, \mathbf{x}^{(0)})$, then $\boldsymbol{\theta}^{(0)}$ and $\mathbf{x}^{(0)}$ are random

variables. Under mild conditions on f, the following facts are known (Cramér, 1946, ch. 33). Suppose that $\lambda \gg 1$ and that for each $\theta \in \Theta$, $p(\theta,\cdot)$ makes $\delta y_1, ..., \delta y_D$ independent; or alternatively, suppose that all the $p(\theta,\cdot)$ are Gaussian. Non-diagonal correlation matrices $E[(\delta y_i)(\delta y_j)]$ are permitted in the Gaussian case. The random variables $\theta^{(0)}$ and $\mathbf{x}^{(0)}$ are approximately Gaussian, with variance matrices of order λ^{-1} and means correct to order $\lambda^{-½}$. From these facts, confidence sets $\tilde{K}^Y(\rho,\tilde{\theta})$ can be constructed as in (3.3a), using (4.2b). The author has not yet carried out this program for any $m > 1$, but believes that the outcome will be the same as in the case $m = 1$ to be described below. That is, for failure rates ρ which are not too small in a sense determined by the λ of (8.1d) (see (8.5a)), there will be a positive $\rho_1 \ll \rho$ and a confidence set $K^\Theta(\rho, \mathbf{y}^{(0)})$ with two crucial properties: (1) the true θ will be a member of $K^\Theta(\rho, \mathbf{y}^{(0)})$ with failure rate ρ_1; (2) all the $p(\theta,\cdot)$ with $\theta \in K^\Theta(\rho, \mathbf{y}^{(0)})$ will be nearly the same. Then p_E can be taken to be any one of the $p(\theta,\cdot)$ with $\theta \in K^\Theta(\rho, \mathbf{y}^{(0)})$, and this known p_E can be used in sections 5, 6, and 7 with failure rate $\rho - \rho_1$ to produce a confidence set for z with failure rate ρ.

This phenomenon is visible in one common formulation of the geomagnetic modelling problem, where $m = 1$. Here the data function $F : X \to Y$ and the prediction functional $g : X \to R$ are linear, and the probability distribution p_E is believed to belong to the one-parameter family $p(\theta,\cdot)$, all of whose members are Gaussian on Y with mean 0 and with variance matrices which are θ^2 times the known variance matrix of $p(1,\cdot)$. This latter need not be diagonal. The usual statement of this problem is that the unknown parameter θ is to be estimated along with the earth model $\mathbf{x}^{(0)}$. In CSI, estimating θ and z are separate problems, and neither requires the other, although the scheme presented here begins by producing an estimate for θ whose failure rate is much less than the failure rate ρ desired for z.

To begin the estimates, note that if Q is any Lebesgue-measurable subset of Y then

$$p(\theta, Q) = p(1, \theta^{-1}Q). \tag{8.2a}$$

If $h : Y \to R$, then $E_\theta[h(\delta \mathbf{y})]$ will denote the expected value of $h(\delta \mathbf{y})$ under $p(\theta,\cdot)$. That is,

$$E_\theta[h(\delta \mathbf{y})] := \int_Y dp(\theta, \mathbf{y}) h(\mathbf{y}). \tag{8.2b}$$

Then

$$E_\theta[\delta y_i] = 0 \tag{8.2c}$$

and, because of (8.2a),

$$E_\theta[(\delta y_i)(\delta y_j)] = \theta^2 E_1[(\delta y_i)(\delta y_j)] . \tag{8.2d}$$

Let $(\mathbf{y} \square \bar{\mathbf{y}})_\theta$ denote the dot product defined on Y by $p(\theta, \cdot)$ as in appendix B. Then (8.2d) makes clear that

$$(\mathbf{y} \square \bar{\mathbf{y}})_\theta = \theta^{-2}(\mathbf{y} \square \bar{\mathbf{y}})_1 \tag{8.2e}$$

and, in an obvious notation,

$$\|\mathbf{y}\|_\theta = \theta^{-1} \|\mathbf{y}\|_1 \tag{8.2f}$$

In particular, if $p(1, \cdot)$ makes $\hat{\mathbf{y}}_1$ a unit vector in Y, then $p(\theta, \cdot)$ will make $\theta \hat{\mathbf{y}}_1$ a unit vector, which prompts the abbreviation

$$\hat{\mathbf{y}}_\theta = \theta \hat{\mathbf{y}}_1 \tag{8.2g}$$

The advantage of (8.2e) is that the notion of orthogonality in Y does not depend on θ. Therefore the orthogonal projectors of Y onto $F(X)$ and $F(X)^\perp$ are independent of θ. All of sections 5, 6, and 7 can be carried out for $\theta = 1$, and the corresponding results for any θ can be obtained by using (8.2e) to introduce appropriate powers of θ in the various terms in sections 5, 6, and 7. For example, if $\phi_i(\theta)$ is the ith eigenfactor of $F : X \to Y$ in the sequence (7.1b) when $p_E = p(\theta, \cdot)$, then

$$\phi_i(\theta) = \theta^{-1} \phi_i(1) . \tag{8.2h}$$

For any positive θ, let $s(\theta)^2$ denote the observed value of the random variable (5.12a) calculated under the assumption that $p_E = p(\theta, \cdot)$. Then (8.2e) implies

$$s(\theta)^2 = \theta^{-2} s(1)^2 \tag{8.3a}$$

and (5.12b) becomes

$$|\theta^{-2} s(1)^2 - 1| \le \nu(\rho_1) 2^{1/2} (D - N)^{-1/2} , \tag{8.3b}$$

If p_E really is $p(\theta, \cdot)$, then (8.3b) is true with failure rate ρ_1. Then so are all its consequences, including the assertion that

$$\theta \le s(1)[1-v(\rho_1)2^{1/2}(D-N)^{-1/2}]^{-1/2}. \tag{8.3c}$$

If $p_E = p(\theta, \cdot)$, then (5.7d) becomes

$$|(\hat{\gamma}_\theta \square \delta y)_\theta| < v(\rho_2) \tag{8.4a}$$

where $v(\cdot)$ is the inverse of the function $\rho(\cdot)$ defined by (5.11), and $\hat{\gamma}_\theta$ is as in (8.2g). Then (8.2e) and (8.2g) show that (8.4a) is equivalent to

$$|(\hat{\gamma}_1 \square \delta y)_1| < \theta v(\rho_2). \tag{8.4b}$$

Since (8.4a) holds at failure rate ρ_2, so does (8.4b).

Confidence sets can be combined by the rules given in the paragraph following (3.3f). According to those rules, at a failure rate no larger than

$$\rho = \rho_1 + \rho_2 \tag{8.4c}$$

both (8.4b) and (8.3c) are true. Therefore, at a failure rate no greater than ρ,

$$|(\hat{\gamma}_1 \square \delta y)_1| < v(\rho_2)s(1)[1-v(\rho_1)2^{1/2}(D-N)^{-1/2}]^{-1/2}. \tag{8.4d}$$

If ρ is fixed, (8.4d) will be satisfied with failure rate $\le \rho$ for any choice of ρ_1 and ρ_2 satisfying (8.4c), so ρ_1 and ρ_2 can be chosen to minimize the right side of (8.4d), and then (8.4d) can be used as in section 5 to obtain $K^Z(\rho)$, a confidence set for the prediction z at failure rate ρ.

In fact, unless $(D-N)^{1/2} \gg 1$, the $v(\rho_1)$ in (8.4d) really should be calculated from (5.10), the Gaussian density being replaced by the density appropriate to the chi-square distribution with $D-N$ degrees of freedom. Here it will be assumed that $(D-N)^{1/2} \gg 1$. Then, if the observer is willing to accept a failure rate ρ large enough to permit

$$v(\rho) \ll (D-N)^{1/2}, \tag{8.5a}$$

it will be possible to choose ρ_1 so that

$$\rho_1 \ll \rho \tag{8.5b}$$

and yet

$$\nu(\rho_1) \ll (D-N)^{1/2} . \tag{8.5c}$$

With this choice of ρ_1,

$$\rho \approx \rho_2 \tag{8.5d}$$

and with an error of order $\nu(\rho_1)(D-N)^{-1/2}$, (8.4d) becomes

$$|(\hat{\gamma}_1 \square \delta y)_1| < \nu(\rho) s(1) . \tag{8.5e}$$

For any θ satisfying (8.3b), (8.5e) is equivalent to

$$|(\hat{\gamma}_1 \square \delta y)_1| < \nu(\rho)\theta \tag{8.5f}$$

with an error of order $\nu(\rho_1)(D-N)^{-1/2}$. By (8.2e) and (8.2g), (8.5f) is the same as

$$|(\hat{\gamma}_\theta \square \delta y)_\theta| < \nu(\rho) . \tag{8.5g}$$

Therefore, except for a fractional inaccuracy of order $\nu(\rho_1)(D-N)^{-1/2}$, the statement (8.5g) is correct with failure rate ρ if θ is any parameter value which satisfies (8.3b).

In other words, if the observer will accept a failure rate ρ which satisfies (8.5a), then he or she can estimate θ by choosing any value which satisfies (8.3b), and can use $p_E = p(\theta, \cdot)$ to define a dot product on Y. All of sections 5, 6, and 7 can then be invoked with the p_E, and the confidence sets thus calculated will be in error by factors of the order of $1 + \nu(\rho_1)(D-N)^{-1/2}$.

Certainly there are interesting problems where $\dim \Theta > 1$. For example, the distribution of data errors may be the sum of two Gaussians, one with much smaller amplitude and larger variance then the other, so that there are outliers in the data. Alternatively, the mean $\langle \delta y \rangle$ may not be known to vanish. If nothing is known about $\langle \delta y \rangle$, of course the data are useless. Often, however, the unknown mean can be parametrized by a family of distributions for which $\dim \Theta \ll \dim Y$, and then the mean can be estimated from the data. This is probably true, for example, of the orbit errors in satellite geomagnetism, although the method has not been tried there. When $\dim \Theta > 1$, the estimation of θ is more complicated then the one-dimensional illustration considered in this section. For example, the dot product in Y will depend on θ. Further work is required to discuss

definitively the general case, that dim $\Theta > 1$.

9 THE GEOMAGNETIC EXAMPLE

As an application of confidence set inference, consider the inverse problem described in section 2: to estimate the geomagnetic field **B** at the CMB from measurements of Cartesian components of **B** at and above the earth's surface.

The infinite-dimensional model space X consists of all the magnetic fields (2.1) outside the CMB whose sources are inside the core. Only prior quadratic bounds of the form (2.5) will be considered, so the dot product in X of the two fields **B** and $\tilde{\mathbf{B}}$ with Gauss coefficients $\beta_l^m(a)$ and $\tilde{\beta}_l^m(a)$ is

$$\mathbf{B} \,\square\, \tilde{\mathbf{B}} := q^{-1} \sum_{l=1}^{\infty} C(l) \sum_{m=-l}^{l} \beta_l^m(a) \tilde{\beta}_l^m(a) . \tag{9.1}$$

To study whether the data function $F : X \to Y$ and prediction functional $g : X \to R$ are continuous under (9.1) it will be very helpful to have the particular orthonormal basis for X generated by the individual spherical harmonics Y_l^m. Let $\hat{\mathbf{x}}_l^m$ denote the model field **B** all of whose Gauss coefficients vanish with the one exception that

$$\beta_l^m(a) = q^{1/2} C(l)^{-1/2} . \tag{9.2a}$$

Then (9.1) evidently implies that

$$\hat{\mathbf{x}}_l^m \,\square\, \hat{\mathbf{x}}_L^M = \delta_{lL}\, \delta_{mM} . \tag{9.2b}$$

By (2.1), if $\mathbf{B} \,\square\, \hat{\mathbf{x}}_l^m = 0$ for all l and m then $\mathbf{B} = \mathbf{0}$. Hence the collection of all the $\hat{\mathbf{x}}_l^m$ is an orthonormal basis for X in the sense of Hilbert space. For any model **B** with Gauss coefficients $\beta_l^m(a)$, (9.1) and (9.2a) imply that

$$\hat{\mathbf{x}}_l^m \,\square\, \mathbf{B} = q^{-1/2} C(l)^{1/2} \beta_l^m(a) , \tag{9.2c}$$

so

$$\mathbf{B} = q^{-1/2} \sum_{l=1}^{\infty} C(l)^{1/2} \sum_{m=-l}^{l} \beta_l^m(a) \hat{\mathbf{x}}_l^m , \tag{9.2d}$$

the series being convergent in the norm defined by (9.1) (Halmos, 1951).

To study the continuity of the data function and prediction functional it will also be helpful to introduce the definitions

$$\Xi_L := \{\hat{\mathbf{x}}_l^m : 1 \le l < L+1 \quad \text{and} \quad -l \le m \le l\} \tag{9.3a}$$

and

$$X_{[L]} := sp\ \Xi_L\ . \tag{9.3b}$$

Here $sp\ \Xi_L$ is the "span" of Ξ_L, the set of all *finite* linear combinations of members of Ξ_L. The set $X_{[L]}$ is a subspace of X, and if $L < \infty$ then $\dim X_{[L]} = L(L+2)$. The subspace $X_{[\infty]}$ consists of all the series (9.2d) which have only finitely many nonzero terms, and X is the closure of $X_{[\infty]}$ in the norm generated by (9.1).

Now suppose a linear functional $g : X \to R$ is given. Is it continuous? To find out, define

$$g_l^m := q^{-1/2} C(l)^{1/2} g(\hat{\mathbf{x}}_l^m)\ . \tag{9.4a}$$

If $\mathbf{B} \in X_{[L]}$ then evidently

$$g(\mathbf{B}) = \sum_{l=1}^{L} \sum_{m=-l}^{l} g_l^m \beta_l^m(a) \tag{9.4b}$$

where $\beta_l^m(a)$ are the Gauss coefficients of $\mathbf{B}$. By the Riesz-Fisher representation theorem, (Lorch, 1962, p.63), the linear functional g is continuous iff there is a $\mathbf{g} \in X$ such that for every $\mathbf{B} \in X$

$$g(\mathbf{B}) = \mathbf{g} \,\square\, \mathbf{B}\ . \tag{9.4c}$$

From (9.4a), it then follows that $g(\hat{\mathbf{x}}_l^m) = \mathbf{g} \,\square\, \hat{\mathbf{x}}_l^m$, so

$$\mathbf{g} = q^{1/2} \sum_{l=1}^{\infty} C(l)^{-1/2} \sum_{m=-l}^{l} g_l^m \hat{\mathbf{x}}_l^m\ . \tag{9.4d}$$

Therefore

$$\|\mathbf{g}\|^2 = q \sum_{l=1}^{\infty} C(l)^{-1} \sum_{m=-l}^{l} (g_l^m)^2\ . \tag{9.4e}$$

If $\mathbf{g} \in X$, $\|\mathbf{g}\| < \infty$, so g is continuous iff the series (9.4e) converges. The coefficients g_l^m required for this test can be obtained from either (9.4a) or (9.4b). If g is continuous, then (9.4b) holds for all $\mathbf{B} \in X$; that is

$$g(\mathbf{B}) = \sum_{l=1}^{\infty} \sum_{m=-l}^{l} g_l^m \beta_l^m(a) , \tag{9.4f}$$

the series being absolutely convergent for all $\mathbf{B} \in X$.

It is of great interest to try to use the data to reconstruct B_r, the radial component of $\mathbf{B}$, on the CMB. To do so, consider the various functionals $g : X \to R$ which give weighted averages of B_r over the CMB, with various weighting kernels. It will suffice to consider axisymmetric kernels. To describe these requires more notation. If $c \in R$, define

$$S(c) := \{\mathbf{r} : \mathbf{r} \in R^3 \quad \text{and} \quad |\mathbf{r}| = c\} ; \tag{9.5a}$$

then $S(c)$ is the surface of the three-dimensional sphere of radius c centered on $\mathbf{0}$. If $f : S(1) \to R$, define the average of f over $S(1)$ by

$$\langle f(\hat{\mathbf{r}}) \rangle_{\hat{\mathbf{r}}} := (4\pi)^{-1} \int_{S(1)} d^2\hat{\mathbf{r}} \, f(\hat{\mathbf{r}}) \tag{9.5b}$$

where $d^2\hat{\mathbf{r}}$ is the element of surface area on $S(1)$. Every function $G : [-1, 1] \to R$ produces an axisymmetric weighting kernel $\mathbf{g}(\hat{\mathbf{s}})$ at each point $a\hat{\mathbf{s}} \in S(a)$. The linear functional $g(\hat{\mathbf{s}}) : X \to R$ is defined by requiring for each $\mathbf{B} \in X$ that

$$\mathbf{g}(\hat{\mathbf{s}}) \square \mathbf{B} := \langle G(\hat{\mathbf{r}} \cdot \hat{\mathbf{s}}) B_r(a\hat{\mathbf{r}}) \rangle_{\hat{\mathbf{r}}} . \tag{9.5c}$$

The functional $g(\hat{\mathbf{s}})$ has coefficients $g_l^m(\hat{\mathbf{s}})$ as defined by (9.4f). To find them, write

$$G(\mu) = \sum_{l=1}^{\infty} G_l \, P_l(\mu) \tag{9.5d}$$

where $P_l(\mu)$ is the Legendre polynomial of degree l and

$$G_l = (2l+1) \int_{-1}^{1} d\mu \, G(\mu) P_l(\mu) . \tag{9.5e}$$

The addition theorem for real scalar spherical harmonics (Erdelyi *et al.*, 1953) permits (9.5d) with

$\mu = \mathfrak{f} \cdot \S$ to be written

$$G(\mathfrak{f} \cdot \S) = \sum_{l=0}^{\infty} G_l \sum_{m=-l}^{l} Y_l^m(\mathfrak{f})\, Y_l^m(\S) \tag{9.5f}$$

(recall that Y_l^m are real). Then, from (9.5c),

$$\mathbf{g}(\S) \square \mathbf{B} = \sum_{l=0}^{\infty} G_l \sum_{m=-0}^{l} Y_l^m(\S)\, \langle Y_l^m(\mathfrak{f}) B_r(a\mathfrak{f}) \rangle_{\mathfrak{f}} \,. \tag{9.5g}$$

Because of the Schmidt normalization of Y_l^m, (2.1) gives

$$\langle Y_l^m(\mathfrak{f}) B_r(a\mathfrak{f}) \rangle_{\mathfrak{f}} = (l+1)(2l+1)^{-1} \beta_l^m(a) \,. \tag{9.5h}$$

Comparing (9.5g) with (9.4f) leads to

$$g_l^m(\S) = (l+1)(2l+1)^{-1} G_l\, Y_l^m(\S) \,. \tag{9.5i}$$

Two useful consequences of (9.5i) should be noted. By the addition theorem for scalar spherical harmonics,

$$\sum_{m=-l}^{l} |\, g_l^m(\S)\,|^2 = (l+1)^2 (2l+1)^{-2} G_l^2 \tag{9.5j}$$

and so (9.4e) becomes

$$\|\mathbf{g}(\S)\|^2 = q \sum_{l=1}^{\infty} (l+1)^2 (2l+1)^{-2} C(l)^{-1} G_l^2 \,. \tag{9.5k}$$

As pointed out in section 6, CSI will fail if the prediction functional $g : X \to R$ is discontinuous under (9.1). The weighted averaging functional $g(\S)$ defined by (9.5c) is continuous iff (9.5k) converges. If (9.5k) diverges, then the prior quadratic bound (2.5) is too weak to overcome the non-uniqueness of the prediction z in (4.1e) produced by the fact that $\dim X = \infty$. Then the particular average (9.5c) cannot be estimated from surface and satellite data with only the prior information (2.5). One special case of (9.5c) is obtained by putting $G(\mu) = \delta(\mu)$, the Dirac delta function, so $G_l = 2l+1$. In this case, $g(\S) \square \mathbf{B} = B_r(a\S)$ (Backus, 1986). It follows that the values of B_r at single points $a\S$ on the CMB can be estimated from surface and satellite data iff (if and only if) the prior quadratic bound (2.5) satisfies

$$\sum_{l=1}^{\infty} (l+1)^2 C(l)^{-1} < \infty . \tag{9.6}$$

For the energy bound (2.3), $C(l)=(l+1)(2l+1)^{-1}$, and for the heat-flow bound (2.4), $C(l)=(l+1)(2l+1)(2l+3)l^{-1}$. For both these bounds, (9.6) diverges, so neither bound is strong enough to permit estimation of B_r at the CMB from surface and satellite measurements of B. The best one can expect with those bounds is to obtain localized averages of B_r.

One such localized average is the unweighted average of B_r over a disc of radius α radians, or $a\alpha$ km, centered on the point $a\hat{\mathbf{s}} \in S(a)$. For this average,

$$G(\mu) = (1-\cos\alpha)^{-1} \tag{9.7}$$

if $\cos\alpha \le \mu \le 1$ and $G(\mu)=0$ otherwise. When $l \gg 1$, $P_l(\cos\theta)$ is asymptotically the Bessel function $J_0[\theta(l+\frac{1}{2})]$, so (9.5e) becomes (Abramowitz & Stegun, 1964, pp.336 and 361)

$$\begin{aligned} G_l &\approx 2\alpha(1-\cos\alpha)^{-1} J_1[\alpha(l+\tfrac{1}{2})] \\ &\approx (1-\cos\alpha)^{-1}(8\pi\alpha/l)^{\frac{1}{2}} \cos[\alpha(l+\tfrac{1}{2})-3\pi/4] \end{aligned}$$

as $l \to \infty$. Then (9.5k) converges for the heat-flow bound (2.4) but not for the energy bound (2.3). The heat-flow bound is strong enough to permit estimating uniform averages of B_r over circular patches on the CMB, but the energy bound is not. If the energy bound is to be used to estimate a weighted average of B_r over circular patches, the coefficients G_l in (9.5d) must approach 0 faster than $l^{-\frac{1}{2}}$ as l approaches infinity, so the kernel function $G(\mu)$ must be smoother than the boxcar (9.7). For example, $G(\mu)$ might be a tent function, so that G_l behaves like $l^{-3/2}$ as $l \to \infty$.

Now consider the data space Y. For simplicity it will be supposed that the data (2.2) consist of all three Cartesian components of B measured at $D/3$ locations $\mathbf{r}_i$ on and above $S(b)$, the surface of the earth. The random errors δy_i of the individual components will be assumed to be Gaussian, independent, and identically distributed with mean 0 and variance θ^2. Then for any data vectors $\mathbf{y}$ and $\tilde{\mathbf{y}}$, their dot product is

$$\mathbf{y} \boxdot \tilde{\mathbf{y}} = \theta^{-2} \sum_{i=1}^{D} y_i \tilde{y}_i \tag{9.8a}$$

or

$$\mathbf{y} \square \tilde{\mathbf{y}} = \theta^{-2} \sum_{i=1}^{D/3} \mathbf{B}(\mathbf{r}_i) \cdot \tilde{\mathbf{B}}(\mathbf{r}_i) \tag{9.8b}$$

where $\mathbf{B}(\mathbf{r}_i)$ and $\tilde{\mathbf{B}}(\mathbf{r}_i)$ are the two measured vector magnetic fields at the location $\mathbf{r}_i$. The dot product on the right in (9.8b) is the ordinary vector dot product in R^3.

The data function $F : X \to Y$ is very simple in the geomagnetic example. If $\mathbf{y} = F(\mathbf{B})$ then the D components of $\mathbf{y}$ are the D Cartesian components of $\mathbf{B}(\mathbf{r})$ at the $D/3$ locations $\mathbf{r}_i$. In particular, (9.8b) implies that for two models $\mathbf{B}$ and $\tilde{\mathbf{B}} \in X$,

$$F(\mathbf{B}) \square F(\tilde{\mathbf{B}}) = \theta^{-2} \sum_{i=1}^{D/3} \mathbf{B}(\mathbf{r}_i) \cdot \tilde{\mathbf{B}}(\mathbf{r}_i) \,. \tag{9.9a}$$

Therefore,

$$\|F(\mathbf{B})\|^2 = \theta^{-2} \sum_{i=1}^{D/3} |\mathbf{B}(\mathbf{r}_i)|^2 \,. \tag{9.9b}$$

To use CSI, the observer must verify that the data function $F : X \to Y$ is bounded (continuous) in the norms on X and Y defined by (9.1) and (9.8a). To prove this, note that for any $r > a$, (2.1) implies

$$\mathbf{B}(r\mathbf{f}) = \sum_{l=1}^{\infty} (a/r)^{l+2} \sum_{m=-l}^{l} \beta_l^m(a)\, \mathbf{b}_l^m(\mathbf{f}) \tag{9.9c}$$

where

$$\mathbf{b}_l^m(\mathbf{f}) := -\nabla \left[r^{-l-1} Y_l^m(\mathbf{f}) \right]_{r=1} .$$

Furthermore, by the addition theorem for vector spherical harmonics (Backus, 1988a, Appendix A), if $\mathbf{f} \in S(1)$ then

$$\sum_{m=-l}^{l} |\mathbf{b}_l^m(\mathbf{f})|^2 = (l+1)(2l+1) \,. \tag{9.9d}$$

Therefore, by Schwarz's inequality,

$$\left| \sum_{m=-l}^{l} \beta_l^m(a)\, \mathbf{b}_l^m(\mathbf{f}) \right|^2 \le (l+1)(2l+1) \sum_{m=-l}^{l} |\beta_l^m(a)|^2$$

and hence, using Schwarz's inequality once more,

$$|\mathbf{B}(r\mathbf{f})|^2 \le q\,\|\mathbf{B}\|^2 \sum_{l=1}^{\infty} (l+1)(2l+1)C(l)^{-1}(a/r)^{2l+4} \tag{9.9e}$$

where $\|\mathbf{B}\|^2 = \mathbf{B}\square\mathbf{B}$ as defined by (9.1). If $r > a$ and $C(l)$ is any rational function of l, then (9.9e) converges, so the linear function $\mathbf{B} \mapsto \mathbf{B}(r\mathbf{f})$ with domain X and codomain R^3 is bounded for any fixed $r\mathbf{f}$ with $r > a$. Then the data function $F : X \to Y$ is also bounded, and in fact (9.9e) implies that

$$\|F\|^2 \le (qD/3)\vartheta^{-2} \sum_{l=1}^{\infty} (l+1)(2l+1)C(l)^{-1}(a/c)^{2l+4} \tag{9.9f}$$

as long as $|\mathbf{r}_i| \ge c > a$ for all observation points $\mathbf{r}_i$. In particular, $\|F\| < \infty$ and F is continuous when the prior quadratic bound is either the energy bound or the heat-flow bound. Both these bounds permit the use of CSI as long as the prediction functional is bounded.

To cast the geomagnetic problem in the form (4.1), it remains to produce sets $S^Y \subseteq Y$ and $S^Z \subseteq R$ such that if η and ζ are the systematic errors in modelling the data and the prediction then the observer is confident that $\eta \in S^Y$ and $\zeta \in S^Z$. For simplicity, it will be assumed here that the prediction z in (4.1e) is a numerical property of the magnetic field produced outside the core by currents in the core. Nothing else is involved in z. In that case, z is exactly calculable from the correct model $\mathbf{x}$, so $\zeta = 0$ and

$$S^Z = \{0\}. \tag{9.10a}$$

The specification of S^Y is more difficult, since η includes contributions to the data from the mantle, crust, ionosphere and magnetosphere. To treat the crustal contributions, one can model them (Langel, Estes, and Meade, 1982), but here they will be controlled by assuming that all the data are collected from satellites on or above the spherical surface $S(c)$, which lies at an altitude $c - b$ above $S(b)$, the surface of the earth. Several workers (Lowes, 1974; Langel and Estes, 1982; Cain, Wang, and Schmitz, 1988) have used the satellite data to estimate the spatial power spectrum of the geomagnetic field, as defined by Mauersberger (1956), Lucke (1957) and Lowes (1974). They find the spectrum's dependence on spherical harmonic degree l to be well fitted by

a sum of two terms representing spatially white sources slightly below $S(a)$ and slightly below $S(b)$. If the second term is interpreted as that produced by the magnetization of the crust, then the satellite data provide a bound on η_c, the contribution to η from the crust. At altitude $c-b=400$ km, the parameters of fit obtained by Cain *et al.* (1988) give the rms total intensity of the crustal field to be about

$$u = 12 \text{ nT} . \tag{9.10b}$$

If **B** is measured at $D/3$ points on or above $S(c)$, then

$$\|\eta_c\| \le (u/\theta)(D/3)^{1/2} . \tag{9.10c}$$

Estimates of the contributions to η from the mantle, ionosphere and magnetosphere are more difficult, and will be neglected here. In practice, ionospheric and magnetospheric effects have been minimized by selectively discarding data particularly sensitive to these effects. For example, in modelling the MAGSAT data, Langel and Estes (1985) used vector data only at geomagnetic latitudes below 50°. In the polar caps at higher latitudes they used only total intensity, which is less seriously affected than are the horizontal components of **B** by the field-aligned currents in the magnetosphere. Using total intensity makes the data function $F: X \to Y$ nonlinear, but it is nearly linear because above 50° geomagnetic latitude the radial component of **B** accounts for at least ninety percent of the total intensity. If only the crust contributes significantly to the systematic error, then (9.10c) implies that one can take

$$S^Y = B^Y(\beta) \tag{9.10d}$$

where B^Y is the ball defined by (5.9b), and

$$\beta = (u/\theta)(D/3)^{1/2} . \tag{9.10e}$$

The foregoing discussion casts the geomagnetic problem in the form (4.1) with the prior bound (6.1). Now section 6 can be used to estimate the truncation errors produced when X is replaced by an N-dimensional subspace X_N, so as to permit numerical calculations. Of course the answer depends on how X_N is chosen. One choice (Backus and Gilbert, 1968) is to follow (6.10b) and take $X_N = \mathrm{N}_F^{\perp}$. Then $X_N = sp\ \{\mathbf{F}_1, ..., \mathbf{F}_D\}$, F_i being the data kernels defined by

(1.2a). This choice leads to no truncation error, so

$$S^Y_{(N,\infty)} = S^Y \,, \tag{9.10f}$$

but it has the disadvantage that N is likely to be D or nearly so. Parker and Shure (1982) chose X_N to be the span of a subset of $\{\mathrm{F}_1, ..., \mathrm{F}_D\}$, chosen so that $N \ll D$ but with the observation points sufficiently well-distributed to make the truncation error small. Parker and Shure used numerical evidence rather than a rigorous theory to estimate the truncation error. Still another X_N is obtained by triangulating $S(a)$, giving B_r at the nodes (vertices), and interpolating B_r into each triangle from the vertices (R. L. Parker, private communication). The oldest choice of X_N in geomagnetism, and the only one to be examined in the present paper, is truncation of (2.1) at some degree L. Thus X_N will be one of the spaces $X_{[L]}$ defined by (9.3b), so that $N = L(L+2)$. The space $X_{(N,\infty)} = X^{\perp}_{[L]}$ will be written $X^{[L]}$, and the functions $F_{[L]}: X_{[L]} \to Y$ and $F^{[L]}: X^{[L]} \to Y$ will be defined by

$$F_{[L]} := F \mid X_{[L]} \,, \tag{9.11a}$$

$$F^{[L]} := F \mid X^{[L]} \,. \tag{9.11b}$$

Thus $X_{[L]}$ and $F_{[L]}$ are the $X_{(0,N)}$ and $F_{(0,N)}$ of (6.5a), while $X^{[L]}$ and $F^{[L]}$ are the $X_{(N,\infty)}$ and $F_{(N,\infty)}$ of (6.5b).

To calculate the value of $\|F_{(N,\infty)}\|$ needed in (6.7c), note that if $\mathrm{B} \in X^{[L]}$ then (9.9e) remains true when the lower limit of the sum over l is $L+1$ instead of 1. Therefore (9.9f) becomes

$$\|F^{[L]}\|^2 \le (qD/3)\theta^{-2} \sum_{l=L+1}^{\infty} (l+1)(2l+1)C(l)^{-1}(a/c)^{2l+4} \,. \tag{9.11c}$$

Backus (1988a, appendix A) sums this infinite series when $C(l)$ comes from the energy bound (2.3), and provides an upper limit when $C(l)$ comes from the heat flow bound (2.4). The latter implies that for the heat flow bound

$$\|F^{[L]}\| \le \theta^{-1}(qD/6)^{1/2}(a/c)^{L+3}[1-(a/c)^2]^{-1/2} \,. \tag{9.11d}$$

(In equation (A.16) and the inequality preceding (A.15) of Backus, 1988a, the factor $l+2$ should be $2l+1$. This misprint does not affect the subsequent calculations.)

Nothing is gained by taking L so large that the truncation error (9.11d) is several orders of magnitude smaller than the crustal systematic error (9.10c). The truncation error is less than ten percent of the crustal error when

$$(c/a)^{L+3} \geq (10/u)(q/2)^{1/2}[1-(a/c)^2]^{-1/2} . \tag{9.11e}$$

With $u = 12$ nT, $q = 3\times 10^{17}$ nT^2 and $c = 6771$ km, (9.11e) leads to

$$L \geq 27 . \tag{9.11f}$$

It remains to carry out the analysis of resolution described in section 7. For this purpose, it is necessary to find the eigenstructure (singular value decomposition) of $F_{[L]}: X_{[L]} \to Y$, as described in appendix D. In general, this will require numerical calculations based on the actual locations of the $D/3$ observation points on or above $S(c)$. To avoid such calculations, and to obtain answers in closed analytic form, two assumptions will be made:

$$2L + 1 \ll D^{1/2} ; \tag{9.12a}$$

and the $D/3$ observation points will be all on $S(c)$ and so uniformly distributed that the sum (9.9a) can be approximated by an integral when $\mathbf{B}$ and $\tilde{\mathbf{B}} \in X_{[L]}$. Then

$$F_{[L]}(\mathbf{B}) \square F_{[L]}(\tilde{\mathbf{B}}) = \theta^{-2}(D/3)\langle \mathbf{B}(c\hat{\mathbf{r}}) \cdot \tilde{\mathbf{B}}(c\hat{\mathbf{r}}) \rangle_{\hat{\mathbf{r}}} . \tag{9.12b}$$

This result is obtained from (9.9a) by assigning to each $\mathbf{r}_i \in S(c)$ a small patch of private territory of area $12\pi c^2/D$, the patches being chosen to cover $S(c)$ without overlap. If the $\mathbf{r}_i$ are not nearly uniformly distributed, the data $\mathbf{B}(\mathbf{r}_i)$ can be weighted by areas of private patches which differ from one $\mathbf{r}_i$ to another, so as to achieve (9.12b) when (9.12a) holds. This complication will not be explored here.

In particular, if $l \leq L$ and $\tilde{l} \leq L$ then (9.12b) applies to $\mathbf{B} = \mathfrak{X}_l^m$ and $\tilde{\mathbf{B}} = \mathfrak{X}_{\tilde{l}}^{\tilde{m}}$, whose Gauss coefficients are given by (9.2a). Lowes (1966) and Backus (1986) evaluate the integral on the right of (9.12b) in this case, with the result that

$$F_{[L]}(\mathfrak{X}_l^m) \square F_{[L]}(\mathfrak{X}_{\tilde{l}}^{\tilde{m}}) = \delta_{l\tilde{l}}\,\delta_{m\tilde{m}}\,(qD/3)\theta^{-2}(a/c)^{2l+4}(l+1)C(l)^{-1} \tag{9.12c}$$

Equations (9.12c) and (9.2b) show that when L satisfies (9.12a) then Ξ_L is an orthonormal

eigenbasis for $F_{[L]}$ in $X_{[L]}$. The corresponding eigenfactor, $\phi_l^m = \|F(\hat{\mathbf{x}}_l^m)\| / \|\hat{\mathbf{x}}_l^m\|$, does not depend on m; from (9.12c) it is

$$\phi_l = \theta^{-1}(qD/3)^{1/2}(a/c)^{l+2}(l+1)^{1/2}C(l)^{-1/2}. \tag{9.12d}$$

The cobasis for $F_{[L]}$ in Y corresponding to the eigenbasis Ξ_L consists of the vectors

$$\hat{\mathbf{y}}_l^m = \phi_l^{-1} F(\hat{\mathbf{x}}_l^m). \tag{9.12e}$$

Therefore, by (D.11e),

$$\mathbf{F}_{[L]}^{-1} = \sum_{l=1}^{L} \phi_l^{-1} \sum_{m=-l}^{l} \hat{\mathbf{x}}_l^m \hat{\mathbf{y}}_l^m. \tag{9.12f}$$

It remains to consider the spaces $X_{(0,n)} \subseteq X_{[L]}$ which appear in section 7, and to choose n to minimize the length of the confidence interval $K_{(0,n)}^{Z}(\rho)$ for z defined by (7.9a) and (7.11a). Because ϕ_l^m is independent of m, all the $\hat{\mathbf{x}}_l^m$ with the same l will be treated together, so only spaces $X_{(n,0)} = X_{[l]}$ defined by (9.3b) will be considered. Here $1 \le l \le L$ and $n = l(l+2)$. The models $\mathbf{g}_{(0,n)}$ and $\mathbf{g}_{(n,\infty)}$ will be written $\mathbf{g}_{[l]}$ and $\mathbf{g}^{[l]}$ so, by (9.4d),

$$\mathbf{g}_{[l]} = q^{1/2} \sum_{j=1}^{l} C(j)^{-1/2} \sum_{m=-j}^{j} g_j^m \hat{\mathbf{x}}_j^m, \tag{9.13a}$$

and

$$\mathbf{g}^{[l]} = q^{1/2} \sum_{j=l+1}^{\infty} C(j)^{-1/2} \sum_{m=-j}^{j} g_j^m \hat{\mathbf{x}}_j^m, \tag{9.13b}$$

Therefore,

$$\|\mathbf{g}^{[l]}\|^2 = q \sum_{j=l+1}^{\infty} C(j)^{-1} \sum_{m=-j}^{j} |g_j^m|^2. \tag{9.13c}$$

In (9.13a,b,c), g_j^m is defined by (9.4b) rather than as a Gauss coefficient of $\mathbf{g}$. Similarly, $\boldsymbol{\gamma}_{(0,n)} := \mathbf{g}_{(0,n)} \square \mathbf{F}_{(0,n)}^{-1}$ will be written as

$$\boldsymbol{\gamma}_{[l]} = \mathbf{g}_{[l]} \square \mathbf{F}_{[l]}^{-1} \tag{9.13d}$$

so that, by (9.12f), (9.13a) and (9.12d),

$$\gamma_{[l]} = \theta(3/D)^{1/2} \sum_{j=1}^{l} (c/a)^{j+2}(j+1)^{-1/2} \sum_{m=-j}^{j} g_j^m \hat{y}_j^m \tag{9.13e}$$

and

$$\|\gamma_{[l]}\|^2 = (3\theta^2/D) \sum_{j=1}^{l} (c/a)^{2j+4}(j+1)^{-1} \sum_{m=-j}^{j} |g_j^m|^2 . \tag{9.13f}$$

In (7.11), with $n = l(l+2)$, $K^Z_{(0,n)}(\rho)$, $T_\rho(n)$ and $\kappa_n(\rho)$ will be written $K^Z_{[l]}(\rho)$, $T_\rho[l]$ and $\kappa_{[l]}(\rho)$. Since $S^Z = \{0\}$ in (7.11a), it follows that

$$|K^Z_{[l]}(\rho)| = T_\rho[l] = \|\mathbf{g}^{[l]}\| + \|\gamma_{[l]}\| \kappa_{[l]}(\rho) . \tag{9.13g}$$

If L is chosen to make the truncation error (9.11d) one tenth of the crustal error (9.10c), the β in (7.11c) is given by

$$\beta = (1.1)(u/\theta)(D/3)^{1/2} . \tag{9.13h}$$

Now the values of $T_\rho[l]$ for $1 \le l \le L$ must be examined, to see which is smallest.

Two extreme examples of predictions z will illustrate the foregoing calculations. First, suppose the prediction functional $g : X \to R$ corresponds to the kernel g(ŝ) defined by (9.5c). In this case (9.5j) converts (9.13c,f) to

$$\|\mathbf{g}^{[l]}\|^2 = q \sum_{j=l+1}^{\infty} (j+1)^2(2j+1)^{-2} G_j^2 C(j)^{-1} \tag{9.14a}$$

and

$$\|\gamma_{[l]}\|^2 = (3\theta^2/D) \sum_{j=1}^{l} (c/a)^{2j+4}(j+1)(2j+1)^{-2} G_j^2 . \tag{9.14b}$$

The sum (9.14a) can be computed once for all when $l = L$, and then only $L = 27$ expressions (9.13g), for $1 \le l \le L$, need be computed to minimize the confidence interval for z at failure rate ρ.

In the other extreme example, z is a single Gauss coefficient of the true core field: $z = \beta_l^m(a)$. Then in (9.4b) $g_l^{\bar{m}} = \delta_{l\bar{l}}\,\delta_{m\bar{m}}$. One must take $L \ge l$, and then only the single subspace $X_{[l]}$ need be considered. Clearly

$$\|\mathbf{g}^{[l]}\| = 0 \tag{9.15a}$$

$$\|\gamma_{[l]}\| = \theta(3/D)^{1/2}(l+1)^{-1/2}(c/a)^{l+2}\,. \tag{9.15b}$$

Then (9.13g) becomes

$$|K^{Z}_{[l]}(\rho)| = (l+1)^{-1/2}(c/a)^{l+2}[(1.1)u + (3/D)^{1/2}\theta v(\hat{\gamma}_{[l]},\rho)]\,. \tag{9.15c}$$

If the crust can be treated as a two-dimensional random process on $S(a)$, then the systematic error in (9.15c) is only the truncation error, $(0.1)u$ instead of $(1.1)u$. The crustal error can be included in δy with other random errors, and θ in (9.15c) must be replaced by $(\theta^2+\alpha^2u^2)^{1/2}$ where $\alpha \geq 1$, and $\alpha = 1$ only if the crustal statistics introduce no correlations between the δy_i at different measuring points (Langel, Estes, and Sabaka, 1988). In the most favorable case, $\alpha = 1$, and (9.15c) is replaced by

$$|\tilde{K}^{Z}_{[l]}(\rho)| = (l+1)^{-1/2}(c/a)^{l+2}[(0.1)u + (3/D)^{1/2}(\theta^2+u^2)^{1/2}v(\hat{\gamma}_{[l]},\rho)]\,. \tag{9.15d}$$

++++++++++++++++++Table 2 near here++++++++++++++

Table 2 gives $|K^{Z}_{[l]}(\rho)|$ and $|\tilde{K}^{Z}_{[l]}(\rho)|$ in microTeslas. The values quoted in the table are for a failure rate $\rho = 10^{-4}$. Using the failure rate $\rho = 10^{-2}$ does not change $|K^{Z}_{[l]}(\rho)|$ by an amount sufficient to affect the table, while that larger failure rate lowers $|\tilde{K}^{Z}_{[l]}(\rho)|$ by about ten percent. For comparison, Table 2 also gives the rms value $\langle \beta_l^m(a)^2 \rangle_m^{1/2}$ of the Gauss coefficients of degree l at the CMB, where

$$\langle \beta_l^m(a)^2 \rangle_m := (2l+1)^{-1} \sum_{m=-l}^{l} g_l^m(a)^2\,. \tag{9.15d}$$

The values of $\langle \beta_l^m(a)^2 \rangle_m^{1/2}$ are from Langel and Estes (1982). In Table 2, $u = 12$ nT (Cain *et al.*, 1988), $\theta = 6$ nT (Langel, *et al.*, 1982), and $D = 26{,}500$ (Langel and Estes, 1982). The random errors are assumed to be Gaussian, so $v(\hat{\gamma}_{[l]},\rho)$ is independent of $\hat{\gamma}_{[l]}$ and is given by (3.9) or Table 1.

The implications of the computation leading to Table 2 are the following: the heat-flow bound permits truncation at $L = 27$ with a truncation error of 1.2 nT in each measured component of **B**. Suppose a failure rate $\rho = 10^{-4}$ is acceptable. If the satellite data on a sphere of radius c at

altitude $c-b=400$ km are fitted by least squares with a model (2.1) truncated at $l=L$, and if (9.10c) is accepted as a bound on the crustal field at satellite altitude (or as its standard deviation if it can be viewed as random) then each Gauss coefficient $\beta_l^m(a)$ on the CMB has a confidence interval $K_{[l]}^Z(\rho)$ or $\tilde{K}_{[l]}^Z(\rho)$ whose half-length is given by Table 2. These half-lengths do not depend on m. If the crustal error must be treated as systematic, column 3 of Table 2 is appropriate, and shows that Gauss coefficients with $l \geq 9$ are not resolvable at the CMB. If the crustal error can be treated as random, with uncorrelated contributions at different measuring points, then column 4 of Table 2 is appropriate, and Gauss coefficients at the CMB become unresolvable when $l \geq 12$. If only Gauss coefficients of degree $\leq l$ are known, the circle of confusion on the CMB has a diameter of about $4(l+1)^{-1}$ radians (Booker, 1969). For $l=8$ this is about 25° and for $l=11$ it is about 19°.

Notice that the confidence interval specified by (9.15c) or (9.15d) does not explicitly involve the value $q=3\times10^{17}\ \text{nT}^2$ from (2.4). That value enters only in choosing the truncation level (9.11f). The relationship between q and the minimum acceptable truncation level which produces a truncation error no more than ten percent of the crustal error is (9.11e). If q is changed from q_1 to q_2, L must change from L_1 to L_2 where

$$L_2 - L_1 = \log(q_2/q_1)\,[2\log(c/a)]^{-1}\,. \tag{9.15f}$$

If q is increased by a factor of 3.8, L must be increased by 1. If q is increased by a factor of 14.2, L must be increased by 2. Clearly the conclusions of CSI in this case are quite insensitive to the numerical value of q in (2.5). What is important is that the quadratic form $Q(\mathbf{x},\mathbf{x})$ is known, and that the bound q is known to within a few orders of magnitude.

10 CONCLUSIONS

The uniqueness question in geophysical inversion is the question of inference: to predict from D observed numerical properties of the earth, $y_1^{(0)}, ..., y_D^{(0)}$, not only values but error bounds for P other numerical properties of the earth, $z_1, ..., z_P$. Except in very special cases, infinite dimensionality of the model space X makes the uniqueness problem insoluble without prior information to supplement the data. Such information is available from laboratory experiments, physical theory, and other studies of the earth. Even if the prior information is not certain, accepting it as a working hypothesis is a useful way to approach the data.

One particular kind of prior information is a quadratic bound (1.4a) on the correct earth model, usually a bound on energy or dissipation rate. In earlier studies, such prior information has been "softened" to a probability distribution p_X on the model space X, so that stochastic inversion (SI) or Bayesian inference (BI) could be used to incorporate the prior information into the inversion. Recent work (Backus, 1988a) has shown that this "softening" process always adds spurious new prior "information" not implied by the original inequality (1.4a). Neyman's (1937) theory of confidence sets provides a technique, confidence set inference (CSI), for directly incorporating unsoftened hard prior bounds into the data inversion without generating spurious information. The constant q in (1.4a) is usually known only to within a few orders of magnitude, so an essential part of CSI is the analysis of the sensitivity of its conclusions to variations in q. With CSI the uncertainty in q cannot be dealt with by "softening" (1.4a) to a probability distribution, as is done in BI and SI.

CSI requires that the formal description of the linear inverse problem include the following ingredients, all explicitly known to the observer: a linear model space X, a positive-definite quadratic form $Q(\mathbf{x}, \mathbf{x})$ on X, a data space $Y = R^D$, a prediction space $Z = R^P$, a linear data function $F : X \to Y$, a linear prediction function $G : X \to Z$, a subset $S^Y \subseteq Y$, a subset $S^Z \subseteq Z$, and an m-parameter family of probability distributions $p(\theta, \cdot)$ on Y, where $\theta = (\theta_1, ..., \theta_m)$ is the parameter vector. All the $p(\theta, \cdot)$ must have as their domain the σ-ring Σ_Y of Lebesgue measurable subsets

of Y. The linear functions F and G must be continuous (and therefore bounded) in the norm defined on X by $\|\mathbf{x}\| = Q(\mathbf{x},\mathbf{x})^{1/2}$.

To use CSI, the observer needs six more objects: p_E, δy, θ_0, $\mathbf{x}$, η and ζ. These may be unknown, but the observer must know that they exist and that they have the following properties: p_E is a probability distribution on Y with domain Σ_Y; $\delta\mathbf{y} = (\delta y_1, ..., \delta y_D)$ is a vector drawn at random from Y according to the probability law p_E; and

$$\theta_0 \in R^m \,, \tag{10.1a}$$
$$\mathbf{x} \in X \,, \tag{10.1b}$$
$$\eta \in S^Y \,, \tag{10.1c}$$
$$\zeta \in S^Z \,, \tag{10.d}$$
$$Q(\mathbf{x},\mathbf{x}) \le 1 \,, \tag{10.1e}$$
$$p_E = p(\theta_0, \cdot) \,, \tag{10.1f}$$
$$\mathbf{y}^{(0)} = F(\mathbf{x}) + \eta + \delta\mathbf{y} \,, \tag{10.1g}$$
$$\mathbf{z} = G(\mathbf{x}) + \zeta \,. \tag{10.1h}$$

In (10.1g), $\mathbf{y}^{(0)}$ is the observed data vector $(y_1^{(0)}, ..., y_D^{(0)}) \in Y$; and in (10.1h), $\mathbf{z}$ is the desired prediction vector $(z_1, ..., z_P) \in Z$. The vectors η and ζ are the systematic errors in modelling the data and predictions, and $\delta\mathbf{y}$ is the random error in the data. The vector $\mathbf{x}$ is the "correct" earth model, and θ_0 is the "correct" parameter vector describing the probability distribution of the random error vector $\delta\mathbf{y}$ in Y. Neither $\mathbf{x}$ nor θ_0 need be unique.

The final result of CSI is to produce for each $\rho \in (0, 1]$ a "confidence set" $K^Z(\rho) \subseteq Z$ such that either

$$\mathbf{z} \in K^Z(\rho) \tag{10.2}$$

or an event E has occurred whose probability is no more than ρ. The statement (10.2) is said to hold at confidence level $1-\rho$, or with failure rate ρ. Section 3 describes a rudimentary calculus which keeps track of the failure rates of the statements in any chain of argument whose hypotheses have known upper bounds on their failure rates. This calculus of failure rates greatly facilitates construction of the confidence sets $K^Z(\rho)$. Unlike BI or SI, CSI gains nothing by permitting $\dim Z > 1$, so in the present paper Z is always the real line R, $G : X \to Z$ is always a

functional $g : X \to R$, and $K^Z(\rho)$ is a subset of R. Therefore the size $|K^Z(\rho)|$ is easy to define, and here is taken to be half the diameter of $K^Z(\rho)$. Usually $K^Z(\rho)$ is an interval, and $|K^Z(\rho)|$ is half its length. Many subsets $K^Z(\rho) \subseteq R$ will make (10.2) correct with failure rate ρ, and one goal of CSI is to find a $K^Z(\rho)$ for which $|K^Z(\rho)|$ is small enough to make the estimate of the prediction z useful.

Section 5 shows how to construct $K^Z(\rho)$ in the simplest special case of (10.1), in which p_E is known, $\dim X \le \dim Y$, and $F : X \to Y$ is injective. In this special case, (10.1e) is not needed, and the continuity of F and g is an automatic consequence of the fact they are linear and that $\dim X < \infty$. The construction of $K^Z(\rho)$ leans heavily on the fact that p_E provides a natural dot product on Y, the only dot product under which the variance tensor of p_E is the identity tensor on Y. When S^Y and S^Z are balls or parallelipipeds centered on the origins in Y and Z, $K^Z(\rho)$ is an interval whose center, $z^{(0)}$, is $g(\mathbf{x}^{(0)})$ where $\mathbf{x}^{(0)}$ is the model in X which best fits the observed data vector $\mathbf{y}^{(0)}$ in the sense of least squares, weighted by the inverse of the variance matrix $E[(\delta y_i)(\delta y_j)]$.

When $\dim X = \infty$, (10.1e) definitely is needed, and the dot product $\mathbf{x}_1 \square \mathbf{x}_2 = Q(\mathbf{x}_1, \mathbf{x}_2)$ for $\mathbf{x}_1, \mathbf{x}_2 \in X$ plays an essential role in the argument. Continuity of F and g is not automatic, and any Q which makes F or g discontinuous is too weak to resolve the nonuniqueness inherent in trying to make infinite-dimensional inferences from finitely many data. When F is discontinuous, the observer has no option but to seek a stronger Q. When g is discontinuous, a second option is available: to replace g by a continuous g_A which resembles g closely enough that the new prediction, z_A, is an acceptable substitute for z.

When $\dim X = \infty$ and p_E is known, CSI proceeds in two steps. The first (section 6) produces a finite-dimensional approximation to (10.1), and the second (section 7) produces an injective approximation to the first, thus subsuming the approximate problem under section 5. In the first step, the observer chooses any finite N and any N-dimensional subspace X_N of X. Then (10.1e) provides an N-dimensional approximation to the original infinite-dimensional problem. The new model space is $X_{(0,N)} = X_N$, the new data function is $F_{(0,N)} = F \mid X_{(0,N)}$, and the new

prediction functional is $g_{(0,N)} = g \mid X_{(0,N)}$. The approximation process adds a "truncation error" to each of the systematic errors η and ζ in (10.1), and so changes S^Y and S^Z to larger sets $S^Y_{(N,\infty)}$ and $S^Z_{(N,\infty)}$, but otherwise leaves (10.1) unchanged. One possible choice of X_N is $\mathbf{N}_F^\perp$, the orthogonal complement of the null space of F. This choice has two advantages: its data function $F_{(0,N)}$ is automatically injective, so section 7 can be bypassed; and it may or may not add a truncation error to ζ in (10.1h), but never adds a truncation error to η in (10.1g). Thus it assures $S^Y_{(N,\infty)} = S^Y$. The disadvantages of the choice $X_F = \mathbf{N}_F^\perp$ are that it requires manipulation of $D \times D$ full matrices (D = 26,500 for many MAGSAT studies) and that the $|K^Z(\rho)|$ it produces may be unnecessarily large by several orders of magnitude. If the data are relevant to the predictions, usually there will be an X_N with $N \ll D$ which produces much tighter error bounds on z than does $\mathbf{N}_F^\perp$.

Section 7 describes the second step in approximating (10.1) by a finite-dimensional problem. This step depends on computing the eigenstructure (singular value decomposition) of $F_{(0,N)}: X_{(0,N)} \to Y$, so dot products on both $X_{(0,N)}$ and Y are essential. If $\phi_1 \geq \phi_2 \geq \cdots \geq \phi_M > 0$ are the nonzero eigenfactors (singular values) of $F_{(0,N)}$, and if $\{\hat{\mathbf{x}}_1, ..., \hat{\mathbf{x}}_M\}$ is a corresponding orthonormal eigenbasis for $Y \square \mathbf{F}_{(0,N)}$, with cobasis $\{\hat{\mathbf{y}}_1, ..., \hat{\mathbf{y}}_M\}$, then for $i = 1, ..., M$,

$$F(\hat{\mathbf{x}}_i) = \phi_i \hat{\mathbf{y}}_i \ , \tag{10.3a}$$

while if $\mathbf{x} \in X_{(0,N)} \cap \{\mathbf{x}_1, ..., \hat{\mathbf{x}}_M\}^\perp$ then $F(\mathbf{x}) = 0$. The tensor $\mathbf{F}_{(0,N)} \in Y \otimes X_{(0,N)}$ which corresponds to $F_{(0,N)}: X_{(0,N)} \to Y$ is

$$\mathbf{F}_{(0,N)} = \sum_{i=1}^{M} \phi_i \hat{\mathbf{y}}_i \hat{\mathbf{x}}_i \ . \tag{10.3b}$$

Let $X_{(0,n)}$ be the subspace of $Y \square \mathbf{F}_{(0,N)}$ spanned by $\{\hat{\mathbf{x}}_1, ..., \hat{\mathbf{x}}_n\}$, and let $F_{(0,n)} := F \mid X_{(0,n)}$. If $1 \leq n \leq M$, then $F_{(0,n)}: X_{(0,n)} \to F(X_{(0,n)})$ is an injection, so section 5 produces a confidence set $K^Z_{(0,n)}(\rho)$. For each n,

$$z \in K^Z_{(0,n)}(\rho) \tag{10.3c}$$

with failure rate ρ. Now the observer can choose the n in $1 \leq n \leq M$ which minimizes

$|K^Z_{(0,n)}(\rho)|$. If the data are relevant to the predictions and X_N is large enough to give a good approximation to the original problem (10.1), then this optimal n will be neither 1 nor M, and must be found by numerical computation. In the approximate inverse problem whose data function is $F_{(0,n)}: X_{(0,n)} \to Y$, the model components $\hat{\mathbf{x}}_j \square \mathbf{x}$ with $j \le n$ are estimated from the data vector $\mathbf{y}^{(0)}$, and provide an estimate $z^{(0)}_{(0,n)}$ for the prediction z. The components with $j > n$ are not used in estimating $z^{(0)}_{(0,n)}$. All the model components with $1 \le j \le N$ contribute to $|K^Z_{(0,n)}(\rho)|$, the error estimate for z. The contributions of the components with $j \le n$ arise from the random error $\delta\mathbf{y}$ and the systematic error $\boldsymbol{\eta}$ in the data vector, while the contributions of the components with $j > n$ come from the prior quadratic bound, (10.1e). If n is too small, useful data are being discarded, and if n is too large, some data are being used whose errors are so large that better error control in z is provided by the prior quadratic bound on $\mathbf{x}$.

If p_E is not known, it can be estimated from the data as long as $m + n \ll D$. This problem is not examined here for $m \ge 2$, but the case $m = 1$ is studied in one common situation: the variance matrix $E[(\delta y_i)(\delta y_j)]$ is supposed to be an unknown multiple $\theta^2 \mathbf{V}_1$ of a known $D \times D$ matrix $\mathbf{V}_1$. The result, which seems likely to generalize to any $m \ll D - n$, is that if ρ is not too small relative to $(D-n)^{-1/2}$ in the sense of (5.8a), then a confidence set $K^\Theta(\rho_1)$ for θ can be found such that $\rho_1 \ll \rho$ and yet $K^\Theta(\rho_1)$ is so small that every $\theta \in K^\Theta(\rho_1)$ produces nearly the same $p(\theta, \cdot)$. Then for any $\theta \in K^\Theta(\rho_1)$, p_E can be assumed equal to $p(\theta, \cdot)$, and sections 5, 6, and 7 can be invoked for this known p_E.

The problem of geomagnetic modelling at the CMB illustrates CSI. The data are measurements of Cartesian components of the geomagnetic field $\mathbf{B}$ at points $\mathbf{r}_i$ on or above the earth's surface. The model space X is the set of all $\mathbf{B}$ outside the CMB which can be produced by electric currents inside it. Both the energy bound (2.3) and the heat-flow bound (2.4) make the data function $F: X \to Y$ continuous. If the prediction functional $g: X \to R$ is defined by choosing a fixed $\mathbf{r}$ on the CMB and requiring $g(\mathbf{B}) = B_r(\mathbf{r})$ for every $\mathbf{B} \in X$, then both the energy bound and the heat-flow bound make g discontinuous (unbounded). Therefore neither bound enables the observer to estimate $B_r(\mathbf{r})$ on the CMB from the data. If $g(\mathbf{B})$ is the unweighted average of B_r

over a disc on the CMB, the heat-flow bound makes g continuous and $g(\mathrm{B})$ observable (i.e., estimatable from the surface data), while the energy bound makes g discontinuous and $g(\mathrm{B})$ unobservable. If $g(\mathrm{B})$ is the disc average weighted by a tent function, both prior bounds make g continuous and $g(\mathrm{B})$ observable.

Among the many choices for X_N which have appeared in the literature of geomagnetic modelling, section 9 examines only the oldest, $X_N = X_{[L]}$, the space of models (2.1) whose Gauss coefficients $\beta_l^m(a)$ vanish for $l > L$. Then $N = \dim X_{[L]} = L(L+2)$ and $F_{[L]} := F_{(0,N)} = F|X_{[L]}$. To avoid numerical computation in finding the eigenstructure of $F_{[L]}$, some simplifying assumptions are accepted. The data consist of all three Cartesian components of B measured by satellite at $D/3$ positions $\mathbf{r}_i$ on a single spherical surface $S(c)$ of radius $c = 6771$ km. The positions $\mathbf{r}_i$ are nearly uniformly distributed on $S(c)$, and only truncation levels L are considered for which $2L+1 \ll (D/3)^{1/2}$, so the necessary sums over the $\mathbf{r}_i$ can be approximated by surface integrals over $S(c)$. The random errors in the D individual component measurements of B are assumed to be independently and identically distributed as Gaussians with mean 0 and variance θ^2. Then an orthonormal eigenbasis for $F_{[L]}$ in $X_{[L]}$ consists of the magnetic fields with a single non-vanishing Gauss coefficient, and the data vectors generated by these fields are the cobasis. The eigenfactors for $F_{[L]}$ are calculated analytically to yield explicit formulas for $|K^Z(\rho)|$. Table 2 shows the results when the satellite data are used to predict the Gauss coefficients $\beta_l^m(a)$ at the CMB. In that table, MAGSAT parameters are used: D = 26,500; $\theta = 6$ nT; and a systematic rms error $u = 12$ nT is produced in the data by the crustal field. The possibility is also considered that the crust might be amenable to treatment as a two-dimensional random process, so that u becomes a standard deviation rather than an error bound. The truncation level $L = 27$ is chosen to make the truncation error no more than 1.2 nT. The actual value of q in (2.4) enters only through this truncation level. If q were 14 times its value in (2.4), the truncation level would have to be $L = 29$. Thus the conclusions are quite insensitive to q. Table 2 is calculated for a failure rate $\rho = 10^{-4}$, but relaxing this to $\rho = 10^{-2}$ would not affect the table entries for the systematic crust and would decrease those for the random crust by ten percent. An improvement of accuracy by a

factor of about 8 is produced if the crustal error can be treated as random rather than systematic. The smallest circle of confusion in observing B_r on the CMB has diameter about 26° for a systematic crustal error, and about 19° for a random crustal error, corresponding to highest resolvable degrees of $l = 11$ and $l = 8$ respectively.

11 ACKNOWLEDGMENTS

This work was supported in part by US National Science Foundation grants EAR 85-21453 and EAR 86-07614 and by the National Aeronautics and Space Agency grant NAG 5-818. The author thanks J. Bloxham, R. Estes, D. Gubbins, and R. Langel for helpful discussions. Special thanks are due to J. Park, P. Stark, and A. Tarantola for detailed and substantive comments on the manuscript.

APPENDIX A. NOTATION FOR SET THEORY

The notation for set theory will be that advocated by MacLane & Birkhoff (1967) and Birkhoff & Bartee (1970). For any objects, u, v, w, $\{u, v, w\}$ denotes the set whose members are u, v and w, while (u, v, w) is the ordered triple. Thus $\{u, v, w\} = \{v, w, u\}$ but $(u, v, w) \neq (v, w, u)$. The null set, the set without members, is written $\varnothing$. If X is a set, $U \subseteq X$ means that U is a subset of X, i.e., every member of U is a member of X; and $x \in X$ means that x is a member of X. Thus $u \in \{u, v, w\}$ and $\{w, u\} \subseteq \{u, v, w\}$, but $(u, v) \notin \{u, v, w\}$ and $(u, v) \not\subseteq \{u, v, w\}$. The symbol "$x \in X$" also abbreviates the noun clause "x which is a member of X." If $P[x]$ is a statement about x, such as "x is an odd integer," then $\{x : P[x]\}$ is the set of all x for which $P[x]$ is true. The symbol ":=" means "is defined as." If X and Y are sets, then $X \cap Y := \{u : u \in X \text{ and } u \in Y\}$, $X \cup Y := \{u : u \in X \text{ or } u \in Y \text{ or both}\}$, $X \setminus Y := \{u : u \in X \text{ and } u \notin Y\}$, and $X \times Y := \{u : u = (x, y) \text{ and } x \in X \text{ and } y \in Y\}$. This last definition can be abbreviated $X \times Y := \{(x, y) : x \in X \text{ and } y \in Y\}$. The sets $X \cap Y$, $X \cup Y$, $X \setminus Y$ and $X \times Y$ are called the intersection, union, difference and Cartesian product of X and Y.

Suppose X and Y are sets and F is a function which assigns to each $x \in X$ a value $F(x) \in Y$. Then one writes "$F : X \to Y$," usually read as "F maps X into Y." The symbol "$F : X \to Y$" also stands for the noun clause "the function F which maps X into Y." The sets X and Y are the domain and codomain of F. Two functions F and G are equal iff (if and only if) they have the same domain X, the same codomain Y, and $F(x) = G(x)$ for every $x \in X$. When F can be given by an explicit formula, another notation is available. For example, if R is the set of real numbers and $F : R \to R$ is defined by requiring $F(u) = u^2 + 3u$ for every $u \in R$, then "$F : X \to Y$" can be replaced by "$F : u \mapsto u^2 + 3u$," or simply by "$u \mapsto u^2 + 3u$." If $P[x]$ is a statement about x, and $F : X \to Y$, then the set $\{y : y = F(x)$ for at least one $x \in X$ which makes $P[x]$ true$\}$ is abbreviated as $\{F(x) : x \in X \text{ and } P[x]\}$, or simply as $\{F(x) : P[x]\}$.

If $F : X \to Y$ and $U \subseteq X$, then $F(U)$ stands for the set $\{F(u) : u \in U\}$. The set $F(X)$ is the "image" of F. If $U \subseteq X$, then $F \mid U$ denotes the function $\tilde{F} : U \to F(U)$ such that $\tilde{F}(u) = F(u)$ for every $u \in U$. The function $F \mid U$ is called the "restriction of F to U." Its image coincides

with its codomain.

If $F:X \to Y$ and $G:Y \to Z$, then the "composite" of G with F is the function $(G \circ F):X \to Z$ defined by requiring that for each $x \in X$

$$(G \circ F)(x) := G(F(x)). \tag{A.1a}$$

If also $H:Z \to W$, then (A.1a) implies

$$H \circ (G \circ F) = (H \circ G) \circ F. \tag{A.1b}$$

The identity mapping on X is the function $I_X:X \to X$ such that $I_X(x)=x$ for every $x \in X$. If $F:X \to Y$, then clearly $I_Y \circ F = F = F \circ I_X$.

If $F:X \to Y$ and $F(X)=Y$, then F is a "surjection." If $F(x)=F(\tilde{x})$ always implies $x=\tilde{x}$, then F is an "injection." If F has both properties, it is a "bijection of X to Y." If $F:X \to Y$ is a bijection, then it has an inverse function, $F^{-1}:Y \to X$, defined by requiring that for each $y \in Y$, $F^{-1}(y)$ is the unique $x \in X$ such that $F(x)=y$.

A function $G:Y \to X$ is a "left inverse" of $F:X \to Y$ if $G \circ F = I_X$, and a "right inverse" of F if $F \circ G = I_Y$. If $F:X \to Y$ has a left inverse, then F is an injection; and if F has a right inverse, then F is a surjection. If $F:X \to Y$ has both a left inverse $G:Y \to X$ and a right inverse $H:Y \to X$, then F is a bijection and $G=H=F^{-1}$. For example, $G = G \circ I_Y = G \circ (F \circ F^{-1}) = (G \circ F) \circ F^{-1} = I_X \circ F^{-1} = F^{-1}$.

R will always denote the set of real numbers. If a and $b \in R$ then $a \wedge b$ is the smaller of a and b, while $a \vee b$ is the larger. The four kinds of intervals are $[a,b] := \{u : u \in R \text{ and } a \le u \le b\}$, $(a,b] := \{u : a < u \le b\}$, $[a,b) := \{u : a \le u < b\}$, and $(a,b) := \{u : a < u < b\}$. Suppose Y is a real vector space, N is a positive integer, and for each $i \in \{1,...,N\}$, $a_i \in R$ and $F_i:X \to Y$. Then the linear combination, the function $(\Sigma_{i=1}^{N} a_i F_i):X \to Y$, is defined by requiring that for each $x \in X$

$$\left[\sum_{i=1}^{N} a_i F_i\right](x) := \sum_{i=1}^{N} a_i F_i(x). \tag{A.2a}$$

If both X and Y are vector spaces, a function $F:X \to Y$ is called "linear" if for every positive

integer N and any $a_1, ..., a_N \in R$ and any $\mathbf{x}_1, ..., \mathbf{x}_N \in X$

$$F\left[\sum_{i=1}^{N} a_i \mathbf{x}_i\right] = \sum_{i=1}^{N} a_i F(\mathbf{x}_i). \tag{A.2b}$$

Any linear combination of linear functions $F_i : X \to Y$ is itself linear. Since R is a one-dimensional real vector space, the definitions and remarks of the foregoing paragraph apply to functionals (functions with codomain R).

If X, Y and Z are real vector spaces and $\mathbf{T} : X \times Y \to Z$, then $\mathbf{T}$ is "bilinear" if $\mathbf{T}(\mathbf{x}, \mathbf{y})$ depends linearly on each of $\mathbf{x}$ and $\mathbf{y}$ when the other is fixed. Linear combinations of bilinear functions are defined by (A.2a) and are themselves bilinear.

Suppose U and V are subsets of real vector space X, and A and B are subsets of R. Then

$$AU + BV := \{a\mathbf{u} + b\mathbf{v} : a \in A, b \in B, \mathbf{u} \in U \text{ and } \mathbf{v} \in V\}. \tag{A3a}$$

If $c \in R$ and $\mathbf{x} \in U$ then

$$cU := \{c\}U, \tag{A3b}$$
$$-U := (-1)U, \tag{A3c}$$
$$\mathbf{x} + U := \{\mathbf{x}\} + U, \tag{A3d}$$
$$A\mathbf{x} := A\{\mathbf{x}\}. \tag{A3e}$$

In particular, $0\,U = \{0\}$ and $1\,U = U$. Notice the distinction between $U - V$ and $U \setminus V$.

A σ-ring (Halmos, 1950) is a non-empty set Σ whose members are subsets of another set X. To qualify as a σ-ring, Σ must have two properties: (1) P and $Q \in \Sigma$ implies $P \setminus Q \in \Sigma$; (2) if all sets of the countably infinite sequence $Q_1, Q_2, Q_3, \cdots$ are members of Σ, then so is their union, $Q_1 \cup Q_2 \cup Q_3 \cup \cdots$ (abbreviated as $\cup_{i=1}^{\infty} Q_i$).

A measure μ on a set X is a function whose domain is a σ-ring Σ of subsets of X, whose codomain is R^+, the set of non-negative real numbers together with $+\infty$, and which satisfies

$$\mu(\cup_{i=1}^{\infty} Q_i) = \sum_{i=1}^{\infty} \mu(Q_i) \tag{A.4a}$$

whenever $Q_i \in \Sigma$ and $Q_i \cap Q_j = \varnothing$ for all i and all $j \neq i$. Property (A.4a) is called "countable

additivity," and the members of Σ are then called the μ-measurable subsets of X (Halmos, 1950). If the whole set X is μ-measurable and if

$$\mu(X) = 1 \tag{A.4b}$$

then μ is called a "probability measure" or a "probability distribution" on X. The μ-measurable subsets P and Q of X are then called "events." P is called "the event P," "the event that $x \in P$," or "the event that P happens." Thus $\mu_X(P \cup Q)$ is "the probability that at least one of P and Q happens," while $\mu_X(P \cap Q)$ is "the probability that both P and Q happen," and $\mu_X(P \setminus Q)$ is "the probability that P happens and Q fails." Clearly

$$\mu_X(P \cup Q) \le \mu_X(P) + \mu_X(Q) \tag{A.4c}$$

and

$$\mu_X(P \cap Q) \le \mu_X(P) \wedge \mu_X(Q) . \tag{A.4d}$$

If μ_X and μ_Y are probability measures on sets X and Y, their σ-rings being Σ_X and Σ_Y, then the product measure $\mu_{X\times Y}$ on $X \times Y$ is defined as follows. Its σ-ring is the intersection of all the σ-rings of subsets of $X \times Y$ which include as members all the sets $P \times Q$ with $P \in \Sigma_X$ and $Q \in \Sigma_Y$. For such sets,

$$\mu_{X\times Y}(P \times Q) := \mu_X(P)\mu_Y(Q) \tag{A.5}$$

and (A.4a) permits the calculation of $\mu_{X\times Y}(W)$ for any other set $W \in \Sigma_{X\times Y}$ (Halmos, 1950). Setting $P = X$ and $Q = Y$ in (A.5) shows that $\mu_{X\times Y}$ is a probability measure on $X \times Y$. In this context, if $P \in \Sigma_X$ and $Q \in \Sigma_Y$, then the event $P \times Y$ is called the event P, and the event $X \times Q$ is called the event Q, so (A.4c,d) remain true. The probability that at least one of events P and Q happens is no more than the sum of their probabilities, and the probability that both P and Q happen is no more than the smaller of their probabilities.

APPENDIX B. REAL HILBERT SPACES FOR MODELS AND DATA

In linear inversion, the model space X is a real linear space (vector space; Halmos, 1958), and its dimension, $\dim X$, is infinite. A quadratic inequality like (2.3) or (2.4) induces on X a natural dot product, which makes X a pre-Hilbert space. Then (Halmos, 1951) X can be completed to a Hilbert space. On the data space Y, the probability distribution for the random error vector δy induces a dot product. Since $D = \dim Y < \infty$, Y is automatically complete, and so a finite-dimensional Hilbert space. These natural dot products on X and Y will be constructed in this appendix.

On X, a quadratic inequality like (2.3) or (2.4) produces a symmetric, positive-definite bilinear functional Q. This Q assigns to any pair $(\mathbf{x}, \tilde{\mathbf{x}})$ of models in X a real number $Q(\mathbf{x}, \tilde{\mathbf{x}})$ such that

$$Q(\mathbf{x}, \tilde{\mathbf{x}}) = Q(\tilde{\mathbf{x}}, \mathbf{x}); \tag{B.1a}$$

$$Q(\mathbf{x}, \mathbf{x}) > 0 \tag{B.1b}$$

if $\mathbf{x} \neq \mathbf{0}$; and for any positive integer N, real numbers $a_1, \ldots, a_N$, and models $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_N$,

$$Q(\mathbf{x}_0, \sum_{i=1}^{N} a_i \mathbf{x}_i) = \sum_{i=1}^{N} a_i Q(\mathbf{x}_0, \mathbf{x}_i). \tag{B.1c}$$

Q is "bilinear" because (B.1a) and (B.1c) imply that $Q(\mathbf{x}, \tilde{\mathbf{x}})$ is also linear in $\mathbf{x}$ when $\tilde{\mathbf{x}}$ is fixed. The observer's prior information is the knowledge of a real number q such that the correct earth model $\mathbf{x}$ must satisfy $Q(\mathbf{x}, \mathbf{x}) \leq q$, or

$$q^{-1} Q(\mathbf{x}, \mathbf{x}) \leq 1. \tag{B.2}$$

For example, in (2.3) if $\beta_l^m(a)$ and $\tilde{\beta}_l^m(a)$ are the Gauss coefficients of the magnetic fields $\mathbf{B}$ and $\tilde{\mathbf{B}}$ at the CMB, then

$$q^{-1} Q(\mathbf{B}, \tilde{\mathbf{B}}) = (2 \times 10^{33})^{-1} \sum_{l=1}^{\infty} \sum_{m=-l}^{l} \left[\frac{2l+1}{l+1} \right] \beta_l^m(a) \tilde{\beta}_l^m(a).$$

On X, define the inner product or dot product $\mathbf{x} \square \tilde{\mathbf{x}}$ by

$$\mathbf{x} \square \tilde{\mathbf{x}} := q^{-1} Q(\mathbf{x}, \tilde{\mathbf{x}}) \tag{B.3a}$$

and the length $\|\mathbf{x}\|$ by

$$\|\mathbf{x}\| := (\mathbf{x} \square \mathbf{x})^{1/2} \tag{B.3b}$$

The notation $\mathbf{v} \cdot \tilde{\mathbf{v}}$ will be reserved for use with the ordinary real three-dimensional vectors. On being completed in the norm $\|\mathbf{x}\|$, X becomes a Hilbert space. The investigator's prior quadratic information (B.2) can now be written

$$\|\mathbf{x}\| \le 1 . \tag{B.4}$$

In geomagnetism, the simplicity of the recursion relations for complex spherical harmonics sometimes makes it useful to consider a complex model space X^*, one whose scalars are complex. A real Hilbert space X can always be extended to a complex Hilbert space X^*. The members of X^* are the formal symbols $\mathbf{x} + i\tilde{\mathbf{x}}$ where $\mathbf{x}$ and $\tilde{\mathbf{x}} \in X$ and $i = (-1)^{1/2}$. Linear combinations with complex scalars are defined in the obvious way using $i^2 = -1$, and the complex conjugate of $\mathbf{x} + i\tilde{\mathbf{x}}$ is $(\mathbf{x} + i\tilde{\mathbf{x}})^C = \mathbf{x} - i\tilde{\mathbf{x}}$. If $\mathbf{x}_1$, $\mathbf{x}_2$, $\tilde{\mathbf{x}}_1$, $\tilde{\mathbf{x}}_2 \in X$, the dot product of $\mathbf{w}_1 = \mathbf{x}_1 + i\tilde{\mathbf{x}}_1$ and $\mathbf{w}_2 = \mathbf{x}_2 + i\tilde{\mathbf{x}}_2$ is defined as $\mathbf{w}_1 \square \mathbf{w}_2 := \mathbf{x}_1 \square \mathbf{x}_2 - \tilde{\mathbf{x}}_1 \square \tilde{\mathbf{x}}_2 + i[\tilde{\mathbf{x}}_1 \square \mathbf{x}_2 + \mathbf{x}_1 \square \tilde{\mathbf{x}}_2]$ and the inner product of $\mathbf{w}_1$ and $\mathbf{w}_2$ is $(\mathbf{w}_1 \mid \mathbf{w}_2) := \mathbf{w}_1^C \square \mathbf{w}_2$. Complexifying X in this way introduces a distinction between inner and dot products which engenders complications in the notation, so only vector spaces with real scalars will be admitted here as model or data spaces.

On the data space Y, the inner product comes via integration from p_E, the probability distribution of the random error vector $\delta\mathbf{y}$. As usual, integrals of real-valued functions of $\delta\mathbf{y}$ with respect to p_E will be written as expected values. Thus

$$E[\delta y_i] := \int_Y dp_E(\mathbf{y}) y_i \tag{B.5a}$$

$$E[(\delta y_i)(\delta y_j)] := \int_Y dp_E(\mathbf{y}) y_i y_j . \tag{B.5b}$$

If p_E is known and $E[\delta y_i] \neq 0$, then its known value can be subtracted from both sides of (1.1). This redefines $\mathbf{y}^{(0)}$ and $\delta\mathbf{y}$ in such a way that

$$E[\delta y_i] = 0 . \tag{B.6a}$$

If p_E is being estimated from the data, $E[\delta y_i]$ will be estimated. Alternatively, p_E can be parametrized so that $E[\delta y_i] = 0$, and the necessary correction can be added to η_i or $F_i(\mathbf{x})$ and subtracted from δy_i in (1.1a). This slightly increases the bound on $\boldsymbol{\eta}$ or the amount of information carried by the models in X, and again achieves (B.6a). Henceforth, (B.6a) will be accepted. Then the $D \times D$ variance matrix V of $\delta \mathbf{y}$ has entries

$$V_{ij} = E[(\delta y_i)(\delta y_j)] . \tag{B.6b}$$

This variance matrix is positive semi-definite. If it is not positive-definite, then, with probability 1, $\delta \mathbf{y}$ lies in a subspace U of Y (Cramér, 1946). Then there are only dim U linearly independent linear combinations of $\delta y_1, ..., \delta y_D$. The corresponding linear combinations of $y_1, ..., y_D$ can be taken as the new data, and U is the new data space. The remaining $D -$ dim U linear combinations of $y_1, ..., y_D$ have no random error. In the real world, this must mean that their values are obtained from theory rather than measurement. They are not really data, and can be discarded. (Usually, they will all vanish.) Henceforth, the matrix V of (B.6b) will be supposed positive-definite.

Now let $W := V^{-1}$. The $D \times D$ matrix W is the "weight matrix" generated by the random errors. It is symmetric and positive-definite. On the data space Y define the dot product of $\mathbf{y} = (y_1, ..., y_D)$ and $\tilde{\mathbf{y}} = (\tilde{y}_1, ..., \tilde{y}_D)$ as

$$\mathbf{y} \,\square\, \tilde{\mathbf{y}} := \sum_{i,j=1}^{D} y_i \, W_{ij} \, \tilde{y}_j . \tag{B.7}$$

This dot product measures the data in units of their random errors, so it is a dimensionless pure number. It will now be shown to have the property that for any fixed $\mathbf{y}$ and $\tilde{\mathbf{y}}$ in Y

$$E[(\mathbf{y} \,\square\, \delta \mathbf{y})\,(\delta \mathbf{y} \,\square\, \tilde{\mathbf{y}})] = \mathbf{y} \,\square\, \tilde{\mathbf{y}} . \tag{B.8}$$

Conversely, no other dot product on Y has the property (B.8). That (B.7) does imply (B.8) follows immediately from (B.6b) and the definition of W. For the converse, suppose that $\mathbf{y} \circ \tilde{\mathbf{y}}$ is any dot product on Y. There will be a symmetric, positive-definite $D \times D$ matrix U such that

$$\mathbf{y} \circ \tilde{\mathbf{y}} = \sum_{i,j-1}^{D} y_i \, U_{ij} \, \tilde{y}_j \tag{B.9}$$

(Halmos, 1958). If the dot product (B.9) has property (B.8), then

$$\sum_{i,j,k,l=1}^{D} y_i \, U_{ij} \, V_{jk} \, U_{kl} \, \tilde{y}_l = \sum_{i,l=1}^{D} y_i \, U_{il} \, \tilde{y}_l \, .$$

Since $\mathbf{y}$ and $\tilde{\mathbf{y}}$ are arbitrary, it follows that $UVU = U$. Since U is positive-definite, U^{-1} exists, so $UV = I$, where I is the $D \times D$ identity matrix. Hence $U = V^{-1} = W$, so (B.9) and (B.7) agree, and $\mathbf{y} \circ \tilde{\mathbf{y}} = \mathbf{y} \square \tilde{\mathbf{y}}$.

APPENDIX C. TENSORS AND LINEAR MAPPINGS ON HILBERT SPACE

This appendix is a self-contained list of the facts and sometimes slightly unconventional Hilbert space notation used in the present paper. Proofs of all assertions can be found in Halmos (1951), Dunford and Schwartz (1958), or Lorch (1962). Some of the shorter proofs are given here to aid the exposition.

A linear mapping $F : X \to Y$ from Hilbert space X to Hilbert space Y is "bounded" if there is a real number K such that for every $\mathbf{x} \in X$,

$$\|F(\mathbf{x})\| \le K \, \|\mathbf{x}\| \, . \tag{C.1a}$$

Halmos (1951) shows that a linear mapping of one Hilbert space into another is bounded iff (if and only if) it is norm-continuous (i.e., $\lim_{n\to\infty} \|\mathbf{x}_n - \mathbf{x}\| = 0$ implies $\lim_{n\to\infty} \|F(\mathbf{x}_n) - F(\mathbf{x})\| = 0$). The smallest K which works in (C.1a) for all $\mathbf{x} \in X$ is called the norm of F, written $\|F\|$. Then the best possible version of (C.1a) is

$$\|F(\mathbf{x})\| \le \|F\| \, \|\mathbf{x}\| \, . \tag{C.1b}$$

A bilinear functional $\mathbf{T} : X \times Y \to R$ is "bounded" if there is a real number K such that for each $(\mathbf{x}, \mathbf{y}) \in X \times Y$

$$|\mathbf{T}(\mathbf{x}, \mathbf{y})| \le K \, \|\mathbf{x}\| \, \|\mathbf{y}\| \, . \tag{C.1c}$$

A bilinear $\mathbf{T}$ is norm-continuous iff it is bounded. The smallest K which works in (C.1c) for all $(\mathbf{x}, \mathbf{y}) \in X \times Y$ is called the norm of $\mathbf{T}$ and written $\|\mathbf{T}\|$. If X and Y are Hilbert spaces whose dot products are written with "$\square$", and if $\mathbf{T} : X \times Y \to R$ is a bounded bilinear functional, $\mathbf{T}(\mathbf{x}, \mathbf{y})$ will often be written as $\mathbf{x} \square \mathbf{T} \square \mathbf{y}$. Thus

$$\mathbf{x} \square \mathbf{T} \square \mathbf{y} := \mathbf{T}(\mathbf{x}, \mathbf{y}) \, , \tag{C.1d}$$

and the best possible version of (C.1c) is

$$|\mathbf{x} \square \mathbf{T} \square \mathbf{y}| \le \|\mathbf{x}\| \, \|\mathbf{T}\| \, \|\mathbf{y}\| \, . \tag{C.1e}$$

The real number $\mathbf{x} \square \mathbf{T} \square \mathbf{y}$ depends linearly on each of $\mathbf{x}$, $\mathbf{T}$ and $\mathbf{y}$ if the other two are fixed.

If X and Y are Hilbert spaces, the set of all bounded linear mappings $F : X \to Y$ will be denoted by $\mathrm{BL}(X \to Y)$, and the set of all bounded bilinear functionals $\mathrm{T} : X \times Y \to R$ will be denoted by $X \otimes Y$. The members of $X \otimes Y$ are "tensors over $X \times Y$." If linear combinations are defined by (A.2a), both $\mathrm{BL}(X \to Y)$ are $X \otimes Y$ are real vector spaces. In fact, if N is any positive integer, and $F_i \in \mathrm{BL}(X \to Y)$, $\mathrm{T}_i \in X \otimes Y$, and $a_i \in R$ for each $i \in \{1, ..., N\}$, then

$$\| \sum_{i=1}^{N} a_i F_i \| \leq \sum_{i=1}^{N} |a_i| \, \|F_i\| \tag{C.2a}$$

$$\| \sum_{i=1}^{N} a_i \mathrm{T}_i \| \leq \sum_{i=1}^{N} |a_i| \, \|\mathrm{T}_i\| . \tag{C.2b}$$

Moreover, if Z is also a real Hilbert space, and $G \in \mathrm{BL}(Y \to Z)$, then

$$\|G \circ F\| \leq \|G\| \, \|F\| . \tag{C.2c}$$

Inequalities (C.2b,c) follow immediately from the definitions of the norms, while (C.2a) is a consequence of two facts about any Hilbert space Y: if $a \in R$ and $\mathbf{y}, \tilde{\mathbf{y}} \in Y$ then

$$\|a\,\mathbf{y}\| = |a| \, \|\mathbf{y}\| \tag{C.3a}$$

and

$$\|\mathbf{y} + \tilde{\mathbf{y}}\| \leq \|\mathbf{y}\| + \|\tilde{\mathbf{y}}\| . \tag{C.3b}$$

The triangle inequality, (C.3b), is itself an immediate consequence of the Schwarz inequality,

$$|\mathbf{y} \square \tilde{\mathbf{y}}| \leq \|\mathbf{y}\| \, \|\tilde{\mathbf{y}}\| , \tag{C.3c}$$

which in turn follows from the fact that $\mathbf{y}' \square \mathbf{y}' \geq 0$ for all $\mathbf{y}' \in Y$ (Halmos, 1951).

When X and Y are Hilbert spaces, every function in $\mathrm{BL}(X \to Y)$ can be thought of as a tensor in $Y \otimes X$ and vice-versa. This correspondence is useful both to simplify notation and because it permits any operation applicable to either a tensor or a linear mapping to be transferred immediately to the other. To establish the correspondence, let $F \in \mathrm{BL}(X \to Y)$, and define $\mathrm{T}_F : Y \times X \to R$ by requiring that for each $(\mathbf{y}, \mathbf{x}) \in Y \times X$

$$\mathrm{T}_F(\mathbf{y}, \mathbf{x}) = \mathbf{y} \square F(\mathbf{x}) \tag{C.4a}$$

Clearly T_F is a bilinear functional on $Y \times X$. Moreover, by (C.1b) and (C.3c), $|T_F(y,x)| \le \|y\| \, \|F\| \, \|x\|$, so $T_F \in Y \otimes X$ and $\|T_F\| \le \|F\|$. In fact,

$$\|T_F\| = \|F\| . \tag{C.4b}$$

To see this, let $x_1, x_2, \cdots$ be an infinite sequence of nonzero vectors in X such that $\lim_{n\to\infty} \|F(x_n)\| / \|x_n\| = \|F\|$. Let $y_n := F(x_n)$. Then $T_F(y_n, x_n) = \|y_n\| \, \|F(x_n)\|$ so

$$\lim_{n\to\infty} |T_F(y_n, x_n)| / \|x_n\| \, \|y_n\| = \|F\| .$$

Therefore $\|T_F\| \ge \|F\|$. Since also $\|T_F\| \le \|F\|$, (C.4b) is established.

Equation (C.4a) is a rule which assigns to each bounded linear function $F \in BL(X \to Y)$ a tensor $T_F \in Y \otimes X$. That is, (C.4a) defines a function $\mathcal{J}: BL(X \to Y) \to Y \otimes X$ such that for each $F \in BL(X \to Y)$,

$$\mathcal{J}(F) = T_F . \tag{C.4c}$$

It follows immediately from (A.2a) that $\mathcal{J}$ is linear, and (C.4b) shows that $\mathcal{J}$ preserves norms.

To show that $\mathcal{J}$ is a bijection from $BL(X \times Y)$ to $Y \otimes X$, an inverse for it will be constructed. First, suppose that $f \in BL(Y \to R)$. Then (Halmos, 1951) there is a unique vector $y_f \in Y$ such that for every $y \in Y$

$$f(y) = y \,\square\, y_f . \tag{C.5a}$$

Now suppose $T \in Y \otimes X$. For any fixed $x \in X$, consider the functional $f_x : Y \to R$ defined by requiring that for each $y \in Y$

$$f_x(y) = T(y,x) . \tag{C.5b}$$

Denote this functional f_x by $T(\cdot, x)$. Clearly $f_x \in BL(Y \to R)$, and in fact $\|f_x\| \le \|T\| \, \|x\|$. Therefore (C.5a) applies to $f = f_x$. In Y there is a unique vector y_{f_x}, or $y_{T(\cdot,x)}$, such that for every $y \in Y$, $f_x(y) = y \,\square\, y_T(\cdot, x)$. That is

$$T(y,x) = y \,\square\, y_T(\cdot, x) .$$

Now define $F_{\mathbf{T}}: X \to Y$ by requiring that for each $\mathbf{x} \in X$

$$F_{\mathbf{T}}(\mathbf{x}) = \mathbf{y}_{\mathbf{T}}(\cdot, \mathbf{x}) . \tag{C.5c}$$

Then $F_{\mathbf{T}}$ is uniquely determined by the fact that for each $\mathbf{x} \in X$ and $\mathbf{y} \in Y$,

$$\mathbf{T}(\mathbf{y}, \mathbf{x}) = \mathbf{y} \,\square\, F_{\mathbf{T}}(\mathbf{x}) . \tag{C.5d}$$

Finally, define $\mathcal{F}: Y \otimes X \to \mathrm{BL}(X \to Y)$ by requiring that for each $\mathbf{T} \in Y \otimes X$,

$$\mathcal{F}(\mathbf{T}) = F_{\mathbf{T}} . \tag{C.5e}$$

Comparing (C.4a) with (C.5d) shows that $F = F_{\mathbf{T}}$ iff (if and only if) $\mathbf{T} = \mathbf{T}_F$. That is, $\mathcal{F}: Y \otimes X \to \mathrm{BL}(X \to Y)$ is both a left and a right inverse for $\mathcal{J}: \mathrm{BL}(X \to Y) \to Y \otimes X$. Therefore, $\mathcal{F}$ and $\mathcal{J}$ are both bijections, and each is the inverse of the other. Since $\mathcal{J}$ is linear, its inverse must be linear. The linearity of $\mathcal{F}$ can also be inferred directly from the uniqueness of the $F_{\mathbf{T}}$ defined by (C.5d). Of course, (C.4b) now implies

$$\|F_{\mathbf{T}}\| = \|\mathbf{T}\| .$$

Henceforth it will be convenient to confuse the bounded linear mapping $F: X \to Y$ with the tensor $\mathbf{T}_F \in Y \otimes X$, and to write $\mathbf{T}_F$ as $\mathbf{F}$. Then, in the notation of (C.1d),

$$\mathbf{y} \,\square\, \mathbf{F} \,\square\, \mathbf{x} = \mathbf{y} \,\square\, F(\mathbf{x}) \tag{C.6a}$$

for every $\mathbf{x} \in X$ and $\mathbf{y} \in Y$. If either F or $\mathbf{F}$ is given (C.6a) uniquely determines the other. Furthermore, (C.4b) can be written

$$\|\mathbf{F}\| = \|F\| . \tag{C.6b}$$

Equation (C.6a) suggests writing the vector $F(\mathbf{x})$ as $\mathbf{F} \,\square\, \mathbf{x}$, so

$$\mathbf{F} \,\square\, \mathbf{x} := F(\mathbf{x}) . \tag{C.6c}$$

Then (C.6a) becomes

$$\mathbf{y} \,\square\, \mathbf{F} \,\square\, \mathbf{x} = \mathbf{y} \,\square\, (\mathbf{F} \,\square\, \mathbf{x}) . \tag{C.6d}$$

The vector $\mathbf{F} \,\square\, \mathbf{x}$ depends linearly on each of $\mathbf{F}$ and $\mathbf{x}$ when the other is fixed, and moreover (C.1b)

takes the Schwarz-like form

$$\|\mathbf{F} \square \mathbf{x}\| \leq \|\mathbf{F}\| \, \|\mathbf{x}\| \, . \tag{C.6e}$$

A special case of (C.6) is that in which $Y = R$. Then F is a linear functional $f : X \to R$, $\mathbf{F}$ is the corresponding vector $\mathbf{x}_f \in X$, and (C.6a) reduces to (C.5a).

A first application of the correspondence (C.6a) between tensors and bounded linear mappings will be to define the transpose of each. For $\mathbf{F} \in Y \otimes X$, the definition of the transpose $\mathbf{F}^T$ is simple: for every $(\mathbf{y}, \mathbf{x}) \in Y \times X$,

$$\mathbf{F}^T(\mathbf{x}, \mathbf{y}) = \mathbf{F}(\mathbf{y}, \mathbf{x}) \, . \tag{C.7a}$$

Evidently $\mathbf{F}^T \in X \otimes Y$,

$$\|\mathbf{F}\| = \|\mathbf{F}^T\| \tag{C.7b}$$

and

$$(\mathbf{F}^T)^T = \mathbf{F} \, . \tag{C.7c}$$

Moreover, if $\mathbf{F}$, $\mathbf{G} \in Y \otimes X$ and $a, b \in R$, then clearly

$$(a\mathbf{F} + b\mathbf{G})^T = a\mathbf{F}^T + b\mathbf{G}^T \, . \tag{C.7d}$$

Now suppose $F \in \mathrm{BL}(X \to Y)$. Then F^T, the transpose of F, can be defined as the mapping in $\mathrm{BL}(Y \to X)$ corresponding to the tensor $\mathbf{F}^T \in X \otimes Y$. Thus $F^T : Y \to X$ is the unique function with domain Y and codomain X such that for every $(\mathbf{x}, \mathbf{y}) \in X \times Y$,

$$\mathbf{y} \square F(\mathbf{x}) = \mathbf{x} \square F^T(\mathbf{y}) \, . \tag{C.7e}$$

Without a discussion equivalent to the introduction of tensors, it is not clear that a function $F^T : Y \to X$ with property (C.7e) exists at all. The tensor argument shows that a unique F^T exists, is linear and bounded, and satisfies (C.7b,c,d).

It will be useful to write $F^T(\mathbf{y})$ as $\mathbf{y} \square \mathbf{F}$. Thus, if $\mathbf{y} \in Y$ and $F \in \mathrm{BL}(X \to Y)$, then invoking (3.9c) gives

$$\mathbf{y} \square \mathbf{F} := F^T(\mathbf{y}) = \mathbf{F}^T \square \mathbf{y} \tag{C.8a}$$

The "dot product" $\mathbf{y} \square \mathbf{F}$ is bilinear in $\mathbf{y}$ and $\mathbf{F}$, and (C.6e) and (C.7b) imply

$$\|\mathbf{y} \square \mathbf{F}\| \le \|\mathbf{y}\| \, \|\mathbf{F}\| \, . \tag{C.8b}$$

Equation (C.7e) can now be rewritten as $\mathbf{y} \blacksquare (\mathbf{F} \blacksquare \mathbf{x}) = \mathbf{x} \square (\mathbf{y} \blacksquare \mathbf{F})$, which is $(\mathbf{y} \blacksquare \mathbf{F}) \blacksquare \mathbf{x}$. Thus, (C.6d) and (C.7e) now yield

$$\mathbf{y} \blacksquare (\mathbf{F} \square \mathbf{x}) = (\mathbf{y} \square \mathbf{F}) \square \mathbf{x} = \mathbf{y} \blacksquare \mathbf{F} \square \mathbf{x} = \mathbf{x} \square \mathbf{F}^T \square \mathbf{y} \, .$$

The correspondence (C.6a) also provides a simple way to define symmetric and positive definite linear mappings. A mapping $F \in \mathrm{BL}(X \to X)$ is symmetric if $F^T = F$, and positive definite if $\mathbf{x} \square \mathbf{F} \square \mathbf{x} > 0$ whenever $\mathbf{x} \in X$ and $\mathbf{x} \neq \mathbf{0}$.

Composition is an operation more easily defined for functions and then transferred to tensors. Suppose that X, Y, and Z are Hilbert spaces and $\mathbf{G} \in Z \otimes Y$, and $\mathbf{F} \in Y \otimes X$. Then $\mathbf{G} \square \mathbf{F}$ is defined as the member of $Z \otimes X$ such that for each $\mathbf{x} \in X$

$$(\mathbf{G} \square \mathbf{F}) \square \mathbf{x} := \mathbf{G} \blacksquare (\mathbf{F} \square \mathbf{x}) \, . \tag{C.9a}$$

If W is another Hilbert space and $\mathbf{H} \in W \otimes Z$, then (A.1b) implies

$$\mathbf{H} \blacksquare (\mathbf{G} \square \mathbf{F}) = (\mathbf{H} \square \mathbf{G}) \square \mathbf{F} \, . \tag{C.9b}$$

The identity

$$(\mathbf{G} \square \mathbf{F})^T = \mathbf{F}^T \blacksquare \mathbf{G}^T \tag{C.9c}$$

and the corresponding result for G and F come immediately from the following tedious but simple application of the definitions and the various associative laws: for every $\mathbf{x} \in X$ and $\mathbf{z} \in Z$, $\mathbf{x} \square (\mathbf{G} \square \mathbf{F})^T \square \mathbf{z} = \mathbf{z} \square (\mathbf{G} \square \mathbf{F}) \square \mathbf{x} = \mathbf{z} \square [(\mathbf{G} \square \mathbf{F}) \square \mathbf{x}] = \mathbf{z} \square [\mathbf{G} \square (\mathbf{F} \square \mathbf{x})] = (\mathbf{z} \square \mathbf{G}) \square (\mathbf{F} \square \mathbf{x}) = (\mathbf{F} \square \mathbf{x}) \square (\mathbf{z} \square \mathbf{G}) = (\mathbf{x} \square \mathbf{F}^T) \square (\mathbf{G}^T \square \mathbf{z}) = \mathbf{x} \square [\mathbf{F}^T \square (\mathbf{G}^T \square \mathbf{z})] = \mathbf{x} \square [(\mathbf{F}^T \square \mathbf{G}^T) \square \mathbf{z}] = \mathbf{x} \blacksquare (\mathbf{F}^T \blacksquare \mathbf{G}^T) \square \mathbf{z}$. Then (C.9c) and (C.8a) imply that for every $\mathbf{z} \in Z$

$$(\mathbf{z} \square \mathbf{G}) \square \mathbf{F} = \mathbf{z} \square (\mathbf{G} \square \mathbf{F}) \, . \tag{C.9d}$$

The identity tensor on X is defined as the tensor $\mathbf{I}_X \in X \otimes X$ which corresponds via (C.6a) to the identity mapping $I_X \in \mathrm{BL}(X \to X)$. To any $(\mathbf{x}, \bar{\mathbf{x}}) \in X \times X$, $\mathbf{I}_X$ assigns the value

$$\mathbf{I}_X(\mathbf{x},\tilde{\mathbf{x}}) = \mathbf{x} \,\square\, \mathbf{I}_X \,\square\, \tilde{\mathbf{x}} = \mathbf{x} \,\square\, \tilde{\mathbf{x}} \,. \tag{C.10a}$$

Then evidently

$$I_X^T = I_X \,. \tag{C.10b}$$

If $F \in \mathrm{BL}(X \to Y)$ has an inverse, $F^{-1}: Y \to X$, then so does $F^T \in \mathrm{BL}(Y \to X)$, and

$$(F^T)^{-1} = (F^{-1})^T \,. \tag{C.10c}$$

The proof is this. Since X and Y are complete, the Banach inverse theorem (Luenberger, 1969, p. 149) says that $F^{-1} \in \mathrm{BL}(Y \to X)$. The tensor in $X \otimes Y$ corresponding to F^{-1} is written, of course, as $\mathbf{F}^{-1}$. Now $(F^{-1})^T \in \mathrm{BL}(X \to Y)$ does exist, and $(F^{-1})^T \circ F^T = (F \circ F^{-1})^T = I_Y^T = I_Y$, while $F^T \circ (F^{-1})^T = (F^{-1} \circ F)^T = I_X^T = I_X$. Thus $(F^{-1})^T$ is a right and a left inverse for F^T. Hence $(F^T)^{-1}$ exists, and satisfies (C.10c)

Another application of the correspondence (C.6a) is in the construction of dyads, which are very easy to define as tensors. Let X and Y be real Hilbert spaces and let $\tilde{\mathbf{x}} \in X$ and $\tilde{\mathbf{y}} \in Y$. Define the "dyad" $\tilde{\mathbf{x}}\tilde{\mathbf{y}} \in X \otimes Y$ by requiring that for every $\mathbf{x} \in X$ and $\mathbf{y} \in Y$

$$\mathbf{x} \,\square\, (\tilde{\mathbf{x}}\tilde{\mathbf{y}}) \,\square\, \mathbf{y} = (\mathbf{x} \,\square\, \tilde{\mathbf{x}})(\tilde{\mathbf{y}} \,\square\, \mathbf{y}) \,. \tag{C.11a}$$

The dyad $\tilde{\mathbf{x}}\tilde{\mathbf{y}}$ is often written $\tilde{\mathbf{x}} \otimes \tilde{\mathbf{y}}$. Evidently

$$(\tilde{\mathbf{x}}\tilde{\mathbf{y}})^T = \tilde{\mathbf{y}}\tilde{\mathbf{x}} \,. \tag{C.11b}$$

For any $a \in R$ and $\mathbf{x} \in X$, $a\mathbf{x}$ will also be written $\mathbf{x}a$. With this convention, the mapping in $\mathrm{BL}(Y \to X)$ which corresponds to the tensor $\tilde{\mathbf{x}}\tilde{\mathbf{y}}$ is determined by the fact that for each $y \in Y$

$$(\tilde{\mathbf{x}}\tilde{\mathbf{y}}) \,\square\, \mathbf{y} = \tilde{\mathbf{x}}(\tilde{\mathbf{y}} \,\square\, \mathbf{y}) \,, \tag{C.11c}$$

and its transpose is determined by the fact that for each $\mathbf{x} \in X$

$$\mathbf{x} \,\square\, (\tilde{\mathbf{x}}\tilde{\mathbf{y}}) = (\mathbf{x} \,\square\, \tilde{\mathbf{x}})\tilde{\mathbf{y}} \,. \tag{C.11d}$$

There is still another pair of associative laws for dyads. Suppose X, Y, Z are real Hilbert spaces, and $\mathbf{F} \in Z \otimes X$, and $\mathbf{G} \in Y \otimes Z$, and $\tilde{\mathbf{x}} \in X$, and $\tilde{\mathbf{y}} \in Y$. Then for any $\mathbf{y} \in Y$, $[\mathbf{F} \,\square\, (\tilde{\mathbf{x}}\tilde{\mathbf{y}})] \,\square\, \mathbf{y} = \mathbf{F} \,\square\, [(\tilde{\mathbf{x}}\tilde{\mathbf{y}}) \,\square\, \mathbf{y}] = \mathbf{F} \,\square\, [\tilde{\mathbf{x}}(\tilde{\mathbf{y}} \,\square\, \mathbf{y})] = (\mathbf{F} \,\square\, \tilde{\mathbf{x}})(\tilde{\mathbf{y}} \,\square\, \mathbf{y}) = [(\mathbf{F} \,\square\, \tilde{\mathbf{x}})\tilde{\mathbf{y}}] \,\square\, \mathbf{y}$. Therefore

$$\mathbf{F} \square (\tilde{\mathbf{x}}\tilde{\mathbf{y}}) = (\mathbf{F} \square \tilde{\mathbf{x}})\tilde{\mathbf{y}} . \tag{C.11e}$$

Similarly

$$(\tilde{\mathbf{x}}\tilde{\mathbf{y}}) \square \mathbf{G} = \tilde{\mathbf{x}}(\tilde{\mathbf{y}} \square \mathbf{G}) . \tag{C.11f}$$

It will be necessary to define the integral of a function whose codomain is a set of tensors, because variance tensors of probability distributions are such objects. Suppose that Ω is an arbitrary set and p is a probability distribution (measure) on Ω (see appendix A and Halmos, 1950). For a large class of functionals $f : \Omega \to R$, called the p-integrable functionals, the integral $\int_{\Omega} dp(\omega) f(\omega)$ is defined, and is usually written as an expected value,

$$E_p[f(\omega)] := \int_{\Omega} dp(\omega) f(\omega) . \tag{C.12a}$$

Now suppose that X and Y are Hilbert spaces and $\mathbf{T} : \Omega \to X \otimes Y$. For each fixed $(\mathbf{x}, \mathbf{y}) \in X \otimes Y$, $\omega \mapsto \mathbf{x} \square \mathbf{T}(\omega) \square \mathbf{y}$ defines a functional on Ω. Suppose that for each $(\mathbf{x}, \mathbf{y}) \in X \times Y$ this functional is p-integrable. Then it is possible to define the integral of $\mathbf{T}(\omega)$ with respect to p, written either

$$\int_{\Omega} dp(\omega)\mathbf{T}(\omega) \quad \text{or} \quad E_p[\mathbf{T}(\omega)] .$$

By definition, $E_p[\mathbf{T}(\omega)] : X \times Y \to R$, and this functional is defined by requiring that for each $(\mathbf{x}, \mathbf{y}) \in X \times Y$

$$E_p[\mathbf{T}(\omega)](\mathbf{x}, \mathbf{y}) := E_p[\mathbf{x} \square \mathbf{T}(\omega) \square \mathbf{y}] . \tag{C.12b}$$

Clearly $E_p[\mathbf{T}(\omega)]$ is bilinear in $\mathbf{x}$ and $\mathbf{y}$. If it is also bounded, then $E_p[\mathbf{T}(\omega)] \in X \otimes Y$, and the function $\mathbf{T} : \Omega \to X \otimes Y$ is called p-integrable. For example, if the functional $\omega \mapsto \|\mathbf{T}(\omega)\|$ is p-integrable, so is $\mathbf{T}$. If $\mathbf{T}$ is p-integrable, then the definition (C.12b) of its integral can be rewritten

$$\mathbf{x} \square E_p[\mathbf{T}(\omega)] \square \mathbf{y} := E_p[\mathbf{x} \square \mathbf{T}(\omega) \square \mathbf{y}] . \tag{C.13a}$$

One consequence of (C.13a) is that if W and Z are also Hilbert spaces, and $\mathbf{P} \in W \otimes X$ and $\mathbf{Q} \in Y \otimes Z$, then $\mathbf{P} \square \mathbf{T}$ and $\mathbf{T} \square \mathbf{Q}$ are integrable, and

$$\mathbf{P} \square E_p[\mathbf{T}(\omega)] = E_p[\mathbf{P} \square \mathbf{T}(\omega)] \tag{C.13b}$$

and

$$E_p[\mathrm{T}(\omega)] \square \mathrm{Q} = E_p[\mathrm{T}(\omega) \square \mathrm{Q}] . \tag{C.13c}$$

In the special case of (C.12) which arises in CSI, Ω is the data space Y, p is the probability distribution p_E of the random error vector $\delta\mathbf{y}$, and X is either R or Y. Following the convention introduced in (B.5), $E_{p_E}[\mathrm{T}(\mathbf{y})]$ will be written $E[\mathrm{T}(\delta\mathbf{y})]$. Then $E[\delta\mathbf{y}]$ is defined as a member of $R \otimes Y$, i.e., of Y, and clearly (B.6a) implies

$$E[\delta\mathbf{y}] = \mathbf{0} . \tag{C.14a}$$

The variance tensor of $\delta\mathbf{y}$ is defined as

$$\mathbf{V} := E[(\delta\mathbf{y})(\delta\mathbf{y})] , \tag{C.14b}$$

because of (C.14a). From (C.13a), if $\mathbf{y}$ and $\tilde{\mathbf{y}} \in Y$ then

$$\mathbf{y} \square \mathbf{V} \square \tilde{\mathbf{y}} = E[(\mathbf{y} \square \delta\mathbf{y})(\delta\mathbf{y} \square \tilde{\mathbf{y}})] .$$

Therefore, from (B.8),

$$\mathbf{y} \square \mathbf{V} \square \tilde{\mathbf{y}} = \mathbf{y} \square \tilde{\mathbf{y}} .$$

Then, from (C.10a),

$$\mathbf{V} = \mathrm{I}_Y . \tag{C.14c}$$

Thus the dot product (B.7) is the only dot product on Y which makes $\mathbf{V}$, the variance tensor of $\delta\mathbf{y}$, equal to the identity tensor on Y. If p_E is Gaussian, then adopting on Y the dot product (B.7) gives p_E the density function

$$dp_E(\mathbf{y}) = (2\pi)^{-D/2} \exp(-\tfrac{1}{2} \|\mathbf{y}\|^2)\, d^D_{\square}\mathbf{y} \tag{C.15a}$$

where $d^D_{\square}\mathbf{y}$ is the volume element in Y defined by the dot product (B.7). If some other dot product, $\mathbf{y} \circ \tilde{\mathbf{y}}$, is adopted for Y, then $\mathbf{V}$ is defined by (C.14b) and $\mathbf{V} \neq \mathrm{I}_Y$. A Gaussian p_E will have a density function given by

$$dp_E(\mathbf{y}) = (2\pi)^{-D/2} (\det \mathbf{V})^{-1/2} \exp(-\tfrac{1}{2}\mathbf{y} \circ \mathbf{V}^{-1} \circ \mathbf{y})\, d^D_{\circ}\mathbf{y} \tag{C.15b}$$

where $d^D_{\circ}y$ is the volume element in Y defined by the dot product $\circ$.

APPENDIX D. THE EIGENSTRUCTURE OF A BOUNDED LINEAR FUNCTION WITH FINITE DIMENSIONAL CODOMAIN

When the model space X and the data space Y are Hilbert spaces, and the data function $F:X \to Y$ is linear, section 7 describes the calculations needed to optimize resolution in confidence set inference. Those calculations depend on the eigenstructure of F (its singular value decomposition). That eigenstructure is determined by the facts that

$$\dim Y < \infty \tag{D.1a}$$

and

$$\|F\| < \infty . \tag{D.1b}$$

The algebraic discussion in Golub and Van Loan (1983, p. 16) suffices for computations, but the geometrical discussion given here may be a useful supplement for readers who, like the author, find it easier to think geometrically.

The geometrical discussion is based on orthogonality. If X is a real Hilbert space and $\mathbf{x}, \tilde{\mathbf{x}} \in X$, and if $\mathbf{x} \square \tilde{\mathbf{x}} = 0$, then one writes "$\mathbf{x} \perp \tilde{\mathbf{x}}$", read "$\mathbf{x}$ is orthogonal to $\tilde{\mathbf{x}}$." A set $\{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \cdots\} \subseteq X$ is "orthonormal" if

$$\hat{\mathbf{x}}_i \square \hat{\mathbf{x}}_j = \delta_{ij} \tag{D.2}$$

($\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$). Every orthonormal set is linearly independent.

If $\mathbf{x} \in X$, $U \subseteq X$ and $V \subseteq X$, then $\mathbf{x} \perp U$ means that $\mathbf{x} \perp \mathbf{u}$ for every $\mathbf{u} \in U$, and $U \perp V$ means that $\mathbf{u} \perp \mathbf{v}$ for every $\mathbf{u} \in U$ and $\mathbf{v} \in V$. Define

$$\mathbf{x} \square U := \{\mathbf{x} \square \mathbf{u} : \mathbf{u} \in U\}$$

and

$$U \square V := \{\mathbf{u} \square \mathbf{v} : \mathbf{u} \in U \text{ and } \mathbf{v} \in V\} .$$

Then $\mathbf{x} \perp U$ iff $\mathbf{x} \square U = \{0\}$, and $U \perp V$ iff $U \square V = \{0\}$.

If $\mathbf{x} \in X$ and also $\mathbf{x} \perp X$ then $\mathbf{x} \square \mathbf{x} = 0$, so $\mathbf{x} = \mathbf{0}$. The most useful corollary of this observation is that if $\mathbf{x}_1$ and $\mathbf{x}_2 \in X$ and $\mathbf{x}_1 \square \mathbf{x} = \mathbf{x}_2 \square \mathbf{x}$ for all $\mathbf{x} \in X$, then $(\mathbf{x}_1 - \mathbf{x}_2) \perp X$ so $\mathbf{x}_1 = \mathbf{x}_2$.

If $U \subseteq X$ and $V \subseteq X$ and $U \perp V$, then $U + V$ is written $U \oplus V$. Every vector in $U \oplus V$ can be written in exactly one way as $\mathbf{u} + \mathbf{v}$ with $\mathbf{u} \in U$ and $\mathbf{v} \in V$. To see this, suppose $\mathbf{u}, \mathbf{u}' \in U$ and $\mathbf{v}, \mathbf{v}' \in V$. Then $(\mathbf{u} - \mathbf{u}') \square (\mathbf{v}' - \mathbf{v}) = 0$. If also $\mathbf{u} + \mathbf{v} = \mathbf{u}' + \mathbf{v}'$, then $\mathbf{u} - \mathbf{u}' = \mathbf{v}' - \mathbf{v} = \mathbf{0}$.

If $U \subseteq X$, then

$$U^{\perp} := \{\mathbf{x} : \mathbf{x} \in X \quad \text{and} \quad \mathbf{x} \perp U\} .$$

The set $U^{\perp}$ is the "orthogonal complement" of U. It is always a subspace of X, even if U is not, and it is closed in the sense that if $\mathbf{x}_1, \mathbf{x}_2, \cdots \in U^{\perp}$ and $\lim_{n \to \infty} \|\mathbf{x}_n - \mathbf{x}\| = 0$ then $\mathbf{x} \in U^{\perp}$ (Halmos, 1951). Clearly $U \perp U^{\perp}$ so

$$U \subseteq (U^{\perp})^{\perp} .$$

If U is itself a subspace of X, then (Halmos, 1951)

$$U^c = (U^{\perp})^{\perp} \tag{D.3a}$$

where U^c is the closure of U. If U is a closed subspace of X, then $U^c = U$ so

$$U = (U^{\perp})^{\perp} ; \tag{D.3b}$$

in this case

$$X = U \oplus U^{\perp} . \tag{D.3c}$$

In particular, (D.3c) holds if $\dim U < \infty$, because every finite-dimensional subspace of X is closed. Equation (D.3c) means that for each $\mathbf{x} \in X$ there are unique vectors $\mathbf{x}_U$ and $\mathbf{x}_U^{\perp}$ such that $\mathbf{x}_U \in U$, $\mathbf{x}_U^{\perp} \in U^{\perp}$, and

$$\mathbf{x} = \mathbf{x}_U + \mathbf{x}_U^{\perp} . \tag{D.3d}$$

Of course it is also true that

$$\mathbf{x}_U \square \mathbf{x}_U^{\perp} = 0 . \tag{D.3e}$$

If U is a closed subspace of X, (D.3d) permits the introduction of functions $P_U : X \to X$ and $Q_U : X \to X$ defined by requiring that for each $\mathbf{x} \in X$

$$P_U(\mathbf{x}) = \mathbf{x}_U \tag{D.4a}$$

$$Q_U(\mathbf{x}) = \mathbf{x}_U^{\perp} . \tag{D.4b}$$

The function P_U is called the orthogonal projector of X onto U. It is a surjection onto U, and $P_U \mid U = I_U$. Clearly Q_U is $P_U^{\perp}$. From the uniqueness of $\mathbf{x}_U$ in (D.3c) it follows that P_U is linear. From (D.3d) it follows that

$$\|\mathbf{x}\|^2 = \|\mathbf{x}_U\|^2 + \|\mathbf{x}_U^{\perp}\|^2 , \tag{D.5}$$

so $\|\mathbf{x}_U\| \le \|\mathbf{x}\|$, and hence $\|P_U\| \le 1$. Evidently

$$\|P_U\| = 1 \quad \text{if} \quad U \ne \{0\} , \tag{D.6a}$$

and

$$\|Q_U\| = 1 \quad \text{if} \quad U \ne X . \tag{D.6b}$$

Also, from the definitions,

$$P_U + Q_U = I_X \tag{D6c}$$

$$P_U \circ P_U = P_U \tag{D.6d}$$

$$Q_U \circ Q_U = Q_U \tag{D.6e}$$

$$P_U \circ Q_U = Q_U \circ P_U = 0 . \tag{D.6f}$$

Moreover, if $\mathbf{x}$ and $\tilde{\mathbf{x}} \in X$, then $\mathbf{x} \square P_U(\tilde{\mathbf{x}}) = \mathbf{x} \square \tilde{\mathbf{x}}_U = \mathbf{x}_U \square \tilde{\mathbf{x}}_U = \mathbf{x}_U \square \tilde{\mathbf{x}} = P_U(\mathbf{x}) \square \tilde{\mathbf{x}} = \tilde{\mathbf{x}} \square P_U(\mathbf{x})$, so

$$P_U^T = P_U . \tag{D.6g}$$

Suppose U is a finite dimensional subspace of X. Let $\dim U = N < \infty$, and let $\{\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N\}$ be any orthonormal basis for U (many can always be found; Halmos, 1958). Then

$$\mathbf{P}_U = \sum_{i=1}^{N} \hat{\mathbf{x}}_i \, \hat{\mathbf{x}}_i . \tag{D.6h}$$

To prove this, let $\mathbf{x} \in X$. Then $\mathbf{x}_U \in U$, so

$$\mathbf{x}_U = \sum_{i=1}^{N} a_i \, \hat{\mathbf{x}}_i$$

where

$a_i = \hat{\mathbf{x}}_i \,\square\, \mathbf{x}_U$.

But $\hat{\mathbf{x}}_i \in U$ so $\hat{\mathbf{x}}_i \,\square\, \mathbf{x}_U^{\perp} = 0$. Thus

$a_i = \hat{\mathbf{x}}_i \,\square\, \mathbf{x}$.

Therefore

$$\mathbf{P}_U \,\square\, \mathbf{x} = \mathbf{x}_U = \sum_{i=1}^{N} \hat{\mathbf{x}}_i \, a_i = \sum_{i=1}^{N} \hat{\mathbf{x}}_i (\hat{\mathbf{x}}_i \,\square\, \mathbf{x}) = \sum_{i=1}^{N} [(\hat{\mathbf{x}}_i \, \hat{\mathbf{x}}_i) \,\square\, \mathbf{x}] = (\sum_{i=1}^{N} \hat{\mathbf{x}}_i \, \hat{\mathbf{x}}_i) \,\square\, \mathbf{x} .$$

This proves (D.5h).

The foregoing properties of orthogonality can now be used to obtain the eigenstructure of $F : X \to Y$ when X and Y are Hilbert spaces, F is linear and bounded, and (D.1) holds. First consider the null space of F, defined as

$$\mathbf{N}_F := \{\mathbf{x} : \mathbf{x} \in X \quad \text{and} \quad F(\mathbf{x}) = \mathbf{0}\} . \tag{D.7a}$$

Since F is linear, $\mathbf{N}_F$ is a subspace of X. Since F is continuous, $\mathbf{N}_F$ is closed. Therefore

$$X = \mathbf{N}_F \oplus \mathbf{N}_F^{\perp} . \tag{D.7b}$$

In the spirit of the conventions (C.6c) and (C.8a), define

$$\mathbf{F} \,\square\, X := \{\mathbf{F} \,\square\, \mathbf{x} : \mathbf{x} \in X\} \tag{D.8a}$$

and

$$Y \,\square\, \mathbf{F} := \{\mathbf{y} \,\square\, \mathbf{F} : \mathbf{y} \in Y\} . \tag{D.8b}$$

Then, clearly,

$$\mathbf{F} \,\square\, X = F(X) \subseteq Y \tag{D.8c}$$

and

$$Y \,\square\, \mathbf{F} = F^T(Y) \subseteq X . \tag{D.8d}$$

Since $Y \,\square\, \mathbf{F}$ is the image of the linear function $F^T : Y \to X$ whose domain is finite-dimensional,

$\dim (Y \square \mathbf{F}) < \infty$ (Halmos, 1958). Define

$$M := \dim (Y \square \mathbf{F}) . \tag{D.8e}$$

The identity

$$(Y \square \mathbf{F}) \square \mathbf{x} = Y \square (\mathbf{F} \square \mathbf{x})$$

holds for every $\mathbf{x} \in X$. Therefore $\mathbf{x} \perp (Y \square \mathbf{F})$ iff $\mathbf{F} \square \mathbf{x} = 0$, i.e., iff $\mathbf{x} \in \mathbf{N}_F$. It follows that

$$\mathbf{N}_F = (Y \square \mathbf{F})^{\perp} , \tag{D.8f}$$

and, by (D.2a)

$$\mathbf{N}_F^{\perp} = (Y \square \mathbf{F})^c .$$

But, being finite-dimensional, $Y \square \mathbf{F}$ is closed, so $(Y \square \mathbf{F})^c = Y \square \mathbf{F}$, and hence

$$Y \square \mathbf{F} = \mathbf{N}_F^{\perp} . \tag{D.8g}$$

Then (D.7b) can be written

$$X = \mathbf{N}_F \oplus (Y \square \mathbf{F}) . \tag{D.8h}$$

Equation (D.8h) resolves X into the two pieces $\mathbf{N}_F$ and $Y \square \mathbf{F}$. The function $F \mid \mathbf{N}_F$ is trivial, simply $\mathbf{0}$, so to study F it suffices to study the function

$$\tilde{F} := F \mid Y \square \mathbf{F} . \tag{D.9a}$$

The function $\tilde{F}$ has two important properties. Firstly,

$$\tilde{F} (Y \square \mathbf{F}) = F (X) . \tag{D.9b}$$

To see this, take $\mathbf{x} \in X$ and, using (D.7h), write

$$\mathbf{x} = \mathbf{x}_{Y \square \mathbf{F}}^{\perp} + \mathbf{x}_{Y \square \mathbf{F}} .$$

Since $\mathbf{x}_{Y \square \mathbf{F}}^{\perp} \in \mathbf{N}_F$, therefore $F (\mathbf{x}_{Y \square \mathbf{F}}^{\perp}) = 0$, so $F (\mathbf{x}) = F (\mathbf{x}_{Y \square \mathbf{F}}) = \tilde{F} (\mathbf{x}_{Y \square \mathbf{F}})$. This proves (D.9b). The second important property of

$$\tilde{F} : Y \square \mathbf{F} \to F (X) \tag{D.9c}$$

is that it is an injection. For suppose $\mathbf{x} \in Y \square \mathbf{F}$ and $\tilde{F}(\mathbf{x}) = 0$. Then $F(\mathbf{x}) = 0$, so $\mathbf{x} \in \mathbf{N}_F$. Then by (D.8f), $\mathbf{x} \in (Y \square \mathbf{F})^{\perp}$ as well as $Y \square \mathbf{F}$, so $\mathbf{x} = \mathbf{0}$. Since $\tilde{F}$ is linear, it is injective. But this fact and (D.9b) show that $\tilde{F}$ is a bijection, and thus has an inverse,

$$\tilde{F}^{-1} : F(X) \to Y \square \mathbf{F} . \tag{D.9d}$$

Now consider the function $\tilde{F}^T \circ \tilde{F} : Y \square \mathbf{F} \to Y \square \mathbf{F}$. This function is symmetric, for $(\tilde{F}^T \circ \tilde{F})^T = \tilde{F}^T \circ (\tilde{F}^T)^T = \tilde{F}^T \circ \tilde{F}$. It is also positive definite, for if $\mathbf{x} \in Y \square \mathbf{F}$ and $\mathbf{x} \neq \mathbf{0}$ then $\tilde{F}(\mathbf{x}) \neq \mathbf{0}$, so $0 < \|\tilde{\mathbf{F}} \square \mathbf{x}\|^2 = (\tilde{\mathbf{F}} \square \mathbf{x}) \square (\tilde{\mathbf{F}} \square \mathbf{x}) = (\mathbf{x} \square \tilde{\mathbf{F}}^T) \square (\tilde{\mathbf{F}} \square \mathbf{x}) = \mathbf{x} \square (\tilde{\mathbf{F}}^T \square \tilde{\mathbf{F}}) \square \mathbf{x}$. The domain $Y \square \mathbf{F}$ of the symmetric, positive definite function $\tilde{F}^T \circ \tilde{F}$ has finite dimension M. Therefore, counting repetitions of multiple eigenvalues, $\tilde{F}^T \circ \tilde{F}$ has M positive real eigenvalues (Halmos, 1958). These can be ordered as

$$\phi_1^2 \geq \phi_2^2 \geq \cdots \geq \phi_M^2 > 0 . \tag{D.10a}$$

Furthermore, $Y \square \mathbf{F}$ has an orthonormal basis $\{\hat{\mathbf{x}}_1, ..., \hat{\mathbf{x}}_M\}$ consisting of eigenvectors of $\tilde{F}^T \circ \tilde{F}$ with the eigenvalues (D.10a). That is

$$\tilde{F}^T \circ \tilde{F}(\hat{\mathbf{x}}_i) = \phi_i^2 \hat{\mathbf{x}}_i \tag{D.10b}$$

for $i = 1, ..., M$. This last equation can also be written

$$\tilde{\mathbf{F}}^T \square \tilde{\mathbf{F}} \square \hat{\mathbf{x}}_i = \phi_i^2 \hat{\mathbf{x}}_i .$$

Then, by (D.2),

$$\hat{\mathbf{x}}_j \square \tilde{\mathbf{F}}^T \square \tilde{\mathbf{F}} \square \hat{\mathbf{x}}_i = \phi_i^2 \delta_{ij} .$$

Therefore, by (C.8a),

$$(\tilde{\mathbf{F}} \square \hat{\mathbf{x}}_j) \square (\tilde{\mathbf{F}} \square \hat{\mathbf{x}}_i) = \phi_i^2 \delta_{ij} . \tag{D.10c}$$

Next, define

$$\hat{\mathbf{y}}_i := \phi_i^{-1} \tilde{F}(\hat{\mathbf{x}}_i) . \tag{D.11a}$$

By (D.10c), $\{\hat{\mathbf{y}}_1, ..., \hat{\mathbf{y}}_M\}$ is an orthonormal subset of $Y \square \mathbf{F}$. Since $\dim Y \square \mathbf{F} = M$, $\{\hat{\mathbf{y}}_1, ..., \hat{\mathbf{y}}_M\}$ is an orthonormal basis for $Y \square \mathbf{F}$. Moreoever, from the definition (D.11a),

$$\tilde{F}(\tilde{\mathbf{x}}_i) = \phi_i \hat{\mathbf{y}}_i \tag{D.11b}$$

for $i = 1, ..., M$. An immediate consequence of (D.11b) is that in $Y \otimes (Y \square \mathbf{F})$

$$\tilde{\mathbf{F}} = \sum_{i=1}^{M} \phi_i \hat{\mathbf{y}}_i \hat{\mathbf{x}}_i \ . \tag{D.11c}$$

To see this, note that

$$\mathbf{I}_{Y \square \mathbf{F}} = \sum_{i=1}^{M} \hat{\mathbf{x}}_i \hat{\mathbf{x}}_i \tag{D.11d}$$

and that

$$\tilde{\mathbf{F}} = \tilde{\mathbf{F}} \square \mathbf{I}_{Y \square \mathbf{F}} = \sum_{i=1}^{M} (\tilde{\mathbf{F}} \square \hat{\mathbf{x}}_i) \hat{\mathbf{x}}_i = \sum_{i=1}^{M} \tilde{F}(\hat{\mathbf{x}}_i) \hat{\mathbf{x}}_i \ .$$

Now (D.11c) follows from (D.11b). A useful consequence of (D.11c) is that

$$\tilde{\mathbf{F}}^{-1} = \sum_{i=1}^{M} \phi_i^{-1} \hat{\mathbf{x}}_i \hat{\mathbf{y}}_i \ . \tag{D.11e}$$

To prove (D.11e), denote its right-hand side by $\tilde{\mathbf{G}}$, and observe from (D.11d) and (D.11c) that $\tilde{\mathbf{G}} \square \tilde{\mathbf{F}} = \tilde{\mathbf{F}} \square \tilde{\mathbf{G}} = \mathbf{I}_{Y \square \mathbf{F}}$.

Finally, from (D.8h) and (D.11c) it will be shown that

$$\mathbf{F} = \sum_{i=1}^{M} \phi_i \hat{\mathbf{y}}_i \hat{\mathbf{x}}_i \ , \tag{D.12a}$$

where now the right side of (D.12a) is interpreted as a tensor in $Y \otimes X$. Equation (D.12a) exhibits the eigenstructure of F, its singular value decomposition. The positive real numbers $\phi_1, ..., \phi_M$ are the eigenfactors or singular values of F, $\{\hat{\mathbf{x}}_1, ..., \hat{\mathbf{x}}_M\}$ is an eigenbasis for F, and $\{\hat{\mathbf{y}}_1, ..., \hat{\mathbf{y}}_M\}$ is the corresponding cobasis.

To prove (D.12a), let G denote its right-hand side. It will suffice to show that $\mathbf{F} \square \mathbf{x} = \mathbf{G} \square \mathbf{x}$ for every $\mathbf{x} \in X$. But if $\mathbf{x} \in X$ then

$$\mathbf{F} \square \mathbf{x} = F(\mathbf{x}) = F(\mathbf{x}_{Y \square \mathbf{F}}) = \tilde{F}(\mathbf{x}_{Y \square \mathbf{F}}) = \tilde{\mathbf{F}} \square \mathbf{x}_{Y \square \mathbf{F}} = (\sum_{i=1}^{M} \phi_i \hat{\mathbf{y}}_i \hat{\mathbf{x}}_i) \square \mathbf{x}_{Y \square \mathbf{F}} = \sum_{i=1}^{M} \phi_i \hat{\mathbf{y}}_i (\hat{\mathbf{x}}_i \square \mathbf{x}_{Y \square \mathbf{F}}) \ .$$

Since $\hat{\mathbf{x}}_i \in Y \,\square\, \mathbf{F}$, $\mathbf{x}_i \,\square\, \mathbf{x}_{\mathcal{Y}^\perp \,\square\, \mathbf{F}} = 0$, so $\hat{\mathbf{x}}_i \,\square\, \mathbf{x}_{Y \,\square\, \mathbf{F}} = \hat{\mathbf{x}}_i \,\square\, \mathbf{x}$. Therefore

$$\mathbf{F} \,\square\, \mathbf{x} = \sum_{i=1}^{M} \phi_i \, \hat{y}_i \, (\hat{\mathbf{x}}_i \,\square\, \mathbf{x}) = \Big(\sum_{i=1}^{M} \phi_i \, \hat{y}_i \, \hat{\mathbf{x}}_i \Big) \,\square\, \mathbf{x} \, .$$

This completes the proof of (D.12a).

If $\mathrm{N}_F = \{0\}$, so $Y \,\square\, \mathbf{F} = X$, then $\tilde{F} = F$, and (D.11e) will hold for F as well as $\tilde{F}$. In general, $\mathrm{N}_F \neq \{0\}$, and then F^{-1} does not exist. When $\mathrm{N}_F \neq \{0\}$, F is said to have the eigenfactor or singular value 0 as well as the positive eigenfactors $\phi_1 \geq \phi_2 \geq \cdots \geq \phi_M > 0$. Even when $\mathrm{N}_F \neq \{0\}$, $F^T : Y \to X$ exists, and it follows immediately from (D.12a), (C.7d) and (C.11b) that

$$\mathbf{F}^T = \sum_{i=1}^{M} \phi_i \, \hat{\mathbf{x}}_i \, \hat{y}_i \, . \tag{D.12b}$$

Thus F and F^T have the same eigenfactors, and the eigenbasis for either is the cobasis for the other.

The foregoing discussion can be carried out verbatim when $D = \dim Y = \infty$ as long as $F : X \to Y$ is compact (Kato, 1976). The number M of nonzero eigenfactors will always obey

$$M \leq (\dim X) \wedge (\dim Y) \, . \tag{D.13}$$

This point is of only academic interest in practical problems, because they never supply infinitely many data.

The eigenstructure theorem (D.12a) can be applied either to the whole data function $F : X \to Y$ on the infinite-dimensional model space X or its restriction $F_{(0,N)} : X_{(0,N)} \to Y$ to any N-dimensional subspace $X_{(0,N)}$ of X. The calculation of resolution in section 7 will be formally the same in either case. However, if a subspace $X_{(0,N)}$ can be found which permits tight bounds on the truncation errors (6.7) and also satisfies $N \ll D$, then section 7 should be applied to $F_{(0,N)}$ rather than F. Truncation will produce only a small loss of accuracy, and will permit the very large computational economy consequent on manipulating $N \times N$ rather than $D \times D$ matrices.

REFERENCES

Abramowitz, M. & Stegun, I. A., 1964. Handbook of Mathematical Functions, National Bureau of Standards, Washington, D.C., 1046pp.

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

Backus, G., 1970a. Inference from inadequate and inaccurate data, I, *Proc. Natl. Acad. Sci.* **65**, 1–8.

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

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

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

Backus, G., 1987. Objections to stochastic inversion and Bayesian inference, *Trans. AGU EOS* **68**, 1482.

Backus, G., 1988a. Bayesian inference in geomagnetism, *Geophys. Journal* **92**, 125–142.

Backus, G., 1988b. Comparing hard and soft prior bounds in geophysical inverse problems, *Geophys. Journal* **94**, 249–261.

Backus, G. & Gilbert, F., 1967. Numerical applications of a formalism for geophysical inverse problems, *Geophys. J. R. astr. Soc.* **13**, 147–276.

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

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

Birkhoff, G. & Bartee, T. C., 1970. *Modern Applied Algebra*, McGraw-Hill, New York, 431pp.

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

Cain, J. C., Wang, Zh, & Schmitz, D. R., 1988. The geomagnetic modal spectrum for 1980 and

core-crustal separation, *Zeitschrift für Geophysik* (submitted).

Constable, C. B. & Parker, R. L., 1988. Statistics of the geomagnetic secular variation for the past 5 Ma, *J. Geophys. Res.*, in press.

Cramér, H., 1946. *Mathematical Methods of Statistics*, Princeton University Press, Princeton, NJ, 573 pp.

Dunford, N. & Schwartz, J. T., 1958. *Linear Operators, Part I*, Interscience, New York, 858pp.

Erdelyi, A., Magnus, W., Oberhettinger, F., & Tricomi, F., 1953. *Higher Transcendental Functions*, vol. II, McGraw-Hill, New York, 396 pp.

Garmany, J., 1979. On the inversion of travel times, *Geophys. Res. Lett.* **6**, 277–279.

Gass, S. I., 1985. *Linear Programming*, 5th ed., McGraw-Hill, New York, 532pp.

Golub, G. H. & VanLoan, C. F., 1983. *Matrix Computations*, Johns Hopkins, Baltimore, 476 pp.

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

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, 1950. *Measure Theory*, Van Nostrand, New York, 304pp.

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

Halmos, 1958. *Finite Dimensional Vector Spaces*, Van Nostrand, New York, 200pp.

Huestis, S. & 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.

Kato, T., 1976. *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, New York, 592 pp.

Kendall, M. & Stuart, A., 1979. *The Advanced Theory of Statistics*, Vol. 2, 4th ed., Macmillan, New York, 472pp.

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

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

Langel, R. A. & Estes, R. H., 1985. The near-Earth magnetic field at 1980 determined from Magsat data, *J. Geophys. Res.* 90, 2495–2503.

Langel, R. A., Estes, R. H. & Mead, G. D., 1982. Some new methods in magnetic field modelling, *J. Geomag. Geoelectr.* 34, 327–349.

Langel, R. A., Estes R. H. & Sabaka, T. J., 1988. Uncertainty estimates in geomagnetic field modelling, *J. Geophys. Res.* (submitted).

Langel, R. A., Ousley, G., Berbert, J., Murphy, J. & Settle, M., 1982. The MAGSAT mission, *Geophys. Res. Lett.* 9, 243–245.

Lorch, E. R., 1962. *Spectral Theory*, Oxford University Press, New York, 158pp.

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

Lowes, F., 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. Pat. Hoshschule Potsdam, *Math.-Nat. Reihe* 3, 39–46.

Luenberger, D. G., 1969. *Optimization by Vector Space Methods*, Wiley, New York, 326pp.

MacLane, S. & Birkhoff, G., 1967. *Algebra*, MacMillan, New York, 598pp.

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

McNutt, M. & Royden, L., 1987. Extremal bounds on geotherms in eroding mountain belts from metamorphic pressure-temperature conditions, *Geophys. J. R. astr. Soc.* 88, 81–95.

Neyman, J., 1937. Outline of a theory of statistical estimation based on the classical theory of probability, *Philos. Trans. R. Soc. London A* 236, 333–380.

Oldenburg, D., 1983. Funnel functions in linear and nonlinear appraisal, *J. Geophys. Res.* 88, 7387–7398.

Orcutt, J. A., 1980. Joint linear, extremal inversion of seismic kinematic data, *J. Geophys. Res.* **85**, 2649–2660.

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

Parker, R. L., 1974. A new method for modeling marine gravity and magnetic anomalies, *J. Geophys. Res.* **79**, 2014–2016.

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

Parker, R. L. & Shure, L., 1982. Efficient modeling of the earth's magnetic field with harmonic splines, *Geophys. Res. Lett.* **9**, 812–815.

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

Stacey, F., 1977. *Physics of the Earth*, 2nd ed., Wiley, New York, 414pp.

Stark, P., 1987. Rigorous velocity bounds from soft p and $X(p)$ data, Geophys. J. R. astr. Soc. **89**, 987–996.

Stark, P., 1988. Strict bounds and applications, Technical Report 147, Dept. of Statistics, Univ. of Calif., Berkeley, CA, llpp.

Stark, P., Parker, R.L., Masters, G., & Orcutt, J.A., 1986. Strict bounds on seismic velocity in the spherical Earth, *J. Geophys. Res.* **91**, 13,892–13,902.

Stark, P. & Parker, R.L., 1987. Velocity bounds from statistical estimates of p and $X(p)$, *J. Geophys. Res.* **92**, 2713–2719.

Tarantola, A., 1987. *Inverse Problem Theory*, Elsevier, New York, 613pp.

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, 258pp).

Wiggins, R. A., McMechan, G. A. & Toksoz, M. N., 1973. Range of earth structure nonuniqueness implied by body wave observations, *Rev. Geophys.* **11**, 87–114.

Table 1

ρ	$v(\rho)$
10^{-2}	2.58
10^{-3}	3.29
10^{-4}	3.89
10^{-5}	4.42
10^{-6}	4.89
10^{-7}	5.33

Half-length $v(\rho)$ of the symmetric confidence interval with failure rate ρ for a one-dimensional Gaussian with mean 0 and variance 1.

Table 2

l	$\langle \beta_l^m(a)^2 \rangle_m^{1/2}$	$\lvert K^Z_{[l]}(\rho) \rvert$	$\lvert \tilde{K}^Z_{[l]}(\rho) \rvert$
1	107.77	.07	.009
2	22.59	.11	.015
3	23.01	.19	.024
4	17.86	.32	.042
5	11.96	.57	.075
6	9.36	1.03	.135
7	7.93	1.87	.25
8	4.73	3.42	.45
9	6.78	6.30	.83
10	4.59	11.67	1.53
11	4.34	21.69	2.85
12	3.62	40.5	5.32

Confidence sets for the Gauss coefficients when the quadratic bound is the heat flow bound. Column 2 gives the value in microTeslas of the rms of the Gauss coefficients $\beta_l^m(a)$ of degree l at the CMB, calculated from Langel and Estes, 1982. Columns 3 and 4 give the half-lengths of the confidence intervals for $\beta_l^m(a)$ if the crustal error is systematic, or random and uncorrelated. The failure rate ρ is 10^{-4}. Using $\rho = 10^{-2}$ would leave column 3 unaffected and would decrease the entries in column 4 by about ten percent.