JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 90, NO. B1, PAGES 581-591, JANUARY 10, 1985

A BAYESIAN APPROACH TO NONLINEAR INVERSION

David D. Jackson

Institute of Geophysics and Planetary Physics, University of California, Los Angeles

Mitsuhiro Matsu'ura

Geophysical Institute, Faculty of Science, Faculty of Science
University of Tokyo, Japan

Abstract. Iterated linear inversion theory can often solve nonlinear inverse problems. Also, linear inversion theory provides convenient error estimates and other interpretive measures. But are these interpretive measures valid for nonlinear problems? We address this question in terms of the joint probability density function (pdf) of the estimated parameters. Briefly, linear inversion theory will be valid if the observations are linear functions of the parameters within a reasonable (say, 95%) confidence region about the optimal estimate, if the optimal estimate is unique. We use Bayes' rule to show how prior information can improve the uniqueness of the optimal estimate, while stabilizing the iterative search for this estimate. We also develop quantitative criteria for the relative importance of prior and observational data and for the effects of nonlinearity. Our method can handle any form of pdf for observational data and prior information. The calculations are much easier (about the same as required for the Marquardt method) when both observational and prior data are Gaussian. We present calculations for some simple one- and two-parameter nonlinear inverse problems. These examples show that the asymptotic statistics (those based on the linear theory) may in some cases be grossly erroneous. In other cases, accurate observations, prior information, or a combination of the two may effectively linearize an otherwise nonlinear problem.

1. Introduction

We now have powerful methods for solving linear parametric inverse problems and expressing the degree of uncertainty and nonuniqueness in their solutions. Unfortunately, many inverse problems that arise in geophysics are nonlinear. Fortunately, most of these can still be treated using linear perturbation theory and linear inversion. But there has been no convenient method to assess the importance of nonlinearity in these quasi-linear problems. In this paper we present such a method.

Backus [1970, 1971] pointed out that Bayesian methods could be applied to geophysical inverse problems. Box and Tiao [1973] give a rather thorough review of the philosophy and the theory of Bayesian estimation. Tarantola and Valette [1982a] discuss the application of probability theory to nonlinear inverse problems. They provide much more detail than we do, and we recommend this paper to the reader. Tarantola and Valette [1982b] give a general formalism for nonlinear inverse problems involving Gaussian data. All of the references above take a more theoretical approach than we do; we emphasize practical aspects, useful approximations, and validity tests for these approximations. The last three of the above references deal almost exclusively with "noninformative" prior probability distributions, designed to minimize the bias introduced by the prior estimates. We take the approach that the prior distribution may provide valid information and that bias is not necessarily to be avoided. In a comparison paper [Matsu'ura and Jackson, 1984] we present a simple algorithm for evaluating the asymptotic covariance matrix for estimation errors and we apply the methods of the present paper to a geophysical problem with actual observations.

Paper number 4B1198.
0148-0227/85/004B-1198$05.00

2. Linear Inversion

Let us summarize some important results for linear inversion theory in introducing notation and ideas to be used below.

2.1. Linear Parametric Inverse Problems

Consider a linear parametric inverse problem of the form

$$\underline{y} = \underline{A}\underline{\chi} + \underline{\varepsilon} \qquad (1)$$

where $\underline{y}$ is an n-vector of observed data, $\underline{A}$ is an n × m matrix (called the design matrix) of known coefficients, $\underline{\chi}$ is an m-vector of unknown parameters, and $\underline{\varepsilon}$ is an n-vector of random errors with known probability density function (pdf) equal to $p(\underline{\varepsilon})$.

We know values of neither $\underline{\chi}$ nor $\underline{\varepsilon}$, but we are free to propose some hypothetical parameter vector $\underline{x}$, and ask if it could be the true solution. If it were, then the residuals

$$\underline{e} = \underline{y} - \underline{A}\underline{x} \qquad (2)$$

would be equal to $\underline{\varepsilon}$. Thus

$$p(\underline{y}|\underline{x}) = p(\underline{e}) = p(\underline{\varepsilon}) \qquad (3)$$

that is, if $\underline{x} = \underline{\chi}$, then the event $\underline{y} = \underline{A}\underline{x} + \underline{e}$ is equivalent to the event that $\underline{e} = \underline{\varepsilon}$. Equation (3) provides a consistency test for the hypothesis that $\underline{x} = \underline{\chi}$: if $\underline{e}$ is an unlikely outcome of $\underline{\varepsilon}$, then $\underline{x}$ is an unlikely value of $\underline{\chi}$.

Several methods exist for finding $\underline{x}$, such that the corresponding residuals $\underline{e}$ are relatively consistent with $p(\underline{e})$. A good one is the maximum likelihood method in which we maximize $p(\underline{y}|\underline{x})$ in (3) with respect to $\underline{x}$.

Suppose the errors $\underline{e}$ are Gaussian, with zero mean and covariance $\underline{E}$. We express this assumption as

$$\underline{e} \sim N(\underline{0},\underline{E}) \tag{4}$$

Then

$$p(\underline{y}|\underline{x}) = a \exp(-S^2/2) \tag{5}$$

where

$$\underline{S}^2 = \underline{e}^T\underline{E}^{-1}\underline{e} \tag{6}$$

Here, a is a normalization constant independent of $\underline{x}$, and $\underline{e}$ is defined in (2). Maximizing (5) is equivalent to minimizing (6), which leads to the weighted least squares procedure.

The maximum likelihood procedure does not require that errors be Gaussian, but the calculations are much simpler when they are. In the simple form above, the maximum likelihood procedure only maximizes consistency; it cannot find the most probable solution by itself. The result may be subject to nonuniqueness. There may be infinitely many solutions $\underline{x}$ which give the same residuals $\underline{e}$ and the same maximum value for $p(\underline{e})$.

2.2. Nonuniqueness

For any linear estimator $\underline{H}$, as is pointed out by Matsu'ura and Hirata [1982], a formal solution to (1) is given by

$$\hat{\underline{x}} = \underline{x}_0 + \underline{H}\,(\underline{y} - \underline{Ax}_0) \tag{7}$$

Here $\underline{x}$ is the final estimate of $\underline{\chi}$, $\underline{x}_0$ is a starting guess or "prior estimate" (corresponding to the null hypothesis). The estimator $\underline{H}$ is an m × n matrix, and its particular form is determined so as to satisfy a certain criterion for best estimate. Reorganizing slightly, we have

$$\hat{\underline{x}} = \underline{Hy} + \underline{Kx}_0 \tag{8}$$

where

$$\underline{K} = \underline{I} - \underline{HA} \tag{9}$$

The matrix product $\underline{HA}$ is commonly called the "resolution matrix" [Jackson, 1972; Wiggins, 1972]. If $\underline{HA} = \underline{I}$, then $\underline{K} = \underline{0}$ and the final estimate depends only on the observations $\underline{y}$. In this case, the observations are said to "resolve" the parameters uniquely and the estimator $\underline{H}$ is said to be "unbiased." Otherwise, the observations do not uniquely resolve the solution vector, the final estimate depends on the starting guess, and the estimator $\underline{H}$ is biased. This second situation occurs whenever nontrivial solutions exist to the homogeneous equations $\underline{Av} = \underline{0}$; then there will be an infinite number of hypothetical solutions $\underline{x}$, each producing the same residuals $\underline{e}$, and thus having the same likelihood $p(\underline{e})$. Nonuniqueness occurs whenever there are not enough linearly independent observations to determine the needed parameters.

Three basic approaches deal with nonuniqueness. The first approach, exemplified by the Monte Carlo method, is to examine a large number of candidate solutions for common features. The second approach, based on the work of Backus and Gilbert [1968, 1970] is to estimate linear combinations of the parameters that are physically meaningful and uniquely defined by the observations. The third approach, elaborated here, is to incorporate prior information directly in the estimation procedure.

We begin by formulating the prior estimate $\underline{x}_0$ as data, subject to unknown errors:

$$\underline{x}_0 = \underline{\chi} + \underline{\delta} \tag{10}$$

where $\underline{\delta}$ is a vector of unknown errors in the prior estimate. As in (1), we have a known quantity on the left equal to the sum of two unknown quantities on the right. Combining (1), (8), (9), and (10), we have

$$\hat{\underline{x}}-\underline{\chi} = \underline{H}\,\underline{\varepsilon} + \underline{K}\,\underline{\delta} \tag{11}$$

which says that for any linear estimation procedure, the estimation error is a linear combination of the observation errors
$\underline{\varepsilon}$ and the prior estimation errors $\underline{\delta}$.
If we have probabilistic information about $\underline{\delta}$, we can use it in designing an estimator $\underline{H}$ to minimize the total errors in (11).

For a hypothetical solution $\underline{x}$, we define residuals to the prior estimate,

$$\underline{d} = \underline{x}_0 - \underline{x} \tag{12}$$

If $\underline{x} = \underline{\chi}$, then $\underline{d} = \underline{\delta}$ by (10).

2.3. Prior Information

Bayes' rule [Hoel, 1971, p. 20] provides a framework for incorporating prior information. It reads

$$p(\underline{x}|\underline{y}) = p(\underline{y}|\underline{x})p(\underline{x}) \;/\; p(\underline{y}) \tag{13}$$

Here $p(\underline{x}|\underline{y})$ is the "posterior" joint pdf for the parameters, given data $\underline{y}$; $p(\underline{y}|\underline{x}) = p(\underline{e})$ as above; $p(\underline{x})$ is the prior joint pdf of the parameters; and $p(\underline{y})$ is a function independent of $\underline{x}$ that serves as a normalizing factor so that $p(\underline{x}|\underline{y})$ integrates to unity over the parameter space.

Suppose we have prior estimates of the parameters as in (10), and that these estimates are subject to Gaussian errors with zero mean and covariance matrix $\underline{D}$; that is,

$$\underline{d} \sim (\underline{0},\underline{D}) \tag{14}$$

If we also have Gaussian observations as assumed in (4), (5), and (6), and if the observation errors $\underline{e}$ are statistically independent of $\underline{d}$, then the posterior pdf is

$$p(\underline{x}|\underline{y}) = a \exp(-T^2/2) \tag{15}$$

where

$$T^2 = \underline{e}^T\underline{E}^{-1}\underline{e} + \underline{d}^T\underline{D}^{-1}\underline{d} \tag{16}$$

Note that $\underline{e}$ and $\underline{d}$ are functions of $\underline{x}$ by virtue of (2) and (12).

The maximum likelihood solution will minimize (16). The solution is of the form (8), with

$$\underline{H} = \underline{M}^{-1}\underline{A}^{T}\underline{E}^{-1} \tag{17}$$

$$\underline{K} = \underline{M}^{-1}\underline{D}^{-1} \tag{18}$$

and

$$\underline{M} = \underline{A}^{T}\underline{E}^{-1}\underline{A} + \underline{D}^{-1} \tag{19}$$

Then, from (11), the posterior covariance matrix is

$$\underline{C} \equiv H\ E\ H^{T} + K\ D\ K^{T} = M^{-1} \tag{20}$$

We assume that $\underline{E}$ and $\underline{D}$ are both positive definite, so that $\underline{M}$ is also positive definite and nonsingular. Thus M^{-1} exists, and the data are sufficient to define the solution uniquely, subject to random errors described by the covariance matrix, in (20).

Equations (17) through (20) could also be obtained by simply appending the prior data $\underline{x}$ to the observations $\underline{y}$, appending the design matrix $\underline{I}$ for the prior data to the design matrix $\underline{A}$ for the observations, and solving by maximum likelihood. In this way, the prior data are used like any other except that the corresponding design matrix is very simple, and the prior data can determine any linear combination of the parameters to within a finite uncertainty. We can join $\underline{H}$ and $\underline{K}$ to form a partitioned estimator operating on both the observations and prior data. We can write the corresponding resolution matrix as the sum of two parts [Jackson, 1979],

$$\underline{I} = \underline{HA} + \underline{KI} \tag{21}$$

The equality follows from (9). With prior data included, all of the parameters are uniquely resolved, and the two parts of the resolution matrix show the relative contributions of the observational and prior data, respectively.

Observational data reduce the variance of any linear combination of parameters from its prior variance. Suppose we are interested in some scalar feature $z = \underline{b}^{T}\underline{x}$, where $\underline{b}$ is a vector of known coefficients. Then the prior variance of z is $\underline{b}^{T}\underline{Db}$, and the posterior variance is $\underline{b}^{T}\underline{Cb}$. From (20) and (19), it follows that

$$\underline{b}^{T}\underline{Cb} \leq \underline{b}^{T}\underline{Db} \tag{22}$$

for any $\underline{b}$. Thus the posterior variance is never larger than the prior variance. Furthermore, prior data reduce the variance of any linear combination of parameters, compared to its variance without the prior data. That is,

$$\underline{b}^{T}\underline{Cb} \leq \underline{b}^{T}(\underline{A}^{T}\underline{E}^{-1}\underline{A})^{-1}\underline{b} \tag{23}$$

for any $\underline{b}$. The left-hand side is always finite, but the right side may be infinite if the observed data are insufficient to resolve the parameters uniquely.

2.4. Confidence Limits

We may assign confidence limits to the maximum likelihood solution of (15) as follows. Let

$$\hat{\underline{e}} = \underline{y} - \underline{A}\hat{\underline{x}} \tag{24}$$

and

$$\hat{\underline{d}} = \underline{x}_0 - \hat{\underline{x}} \tag{25}$$

be the residuals to the observational and prior data, respectively, obtained for the maximum likelihood solution. Then let

$$\hat{T}^2 = \hat{\underline{e}}^{T}\underline{E}^{-1}\hat{\underline{e}} + \hat{\underline{d}}^{T}\underline{D}^{-1}\hat{\underline{d}} \tag{26}$$

be the minimum obtainable error criterion. For any other hypothetical solution $\underline{x}$, we have

$$T^2 = \hat{T}^2 + (\underline{x}-\hat{\underline{x}})^{T}\underline{M}(\underline{x}-\hat{\underline{x}}) \tag{27}$$

and the pdf is given by (15). Thus the form of the confidence limits for the Bayesian estimate is the same as that for the maximum likelihood solution without prior data. The primary difference is that the normal matrix $\underline{M}$ will have an extra term in the Bayesian case, and this extra term will narrow the confidence limits.

2.5. Conditional and Marginal Statistics

Sometimes we wish to isolate a few parameters, or linear combinations of parameters, for close examination. We may do this to test a hypothesis, to examine nonlinearity as discussed below, or simply to plot. We begin by making a reversible transformation

$$\underline{z} = \underline{B}^{T}(\underline{x}-\hat{\underline{x}}) \tag{28}$$

where $\underline{B}^{T}$ is a known, nonsingular m × m matrix. The pdf for $\underline{z}$ is

$$p(\underline{z}) = a\ p\ (\underline{x}) \tag{29}$$

where $p(\underline{x})$ is the posterior pdf for the parameters, and the constant a is the inverse of the determinant of $\underline{B}$. We now partition the vector $\underline{z}$ into the interesting part

$$\underline{z}_1 = \underline{B}_1^{T}\ (\underline{x}-\hat{\underline{x}}) \tag{30}$$

and a complementary part

$$\underline{z}_2 = \underline{B}_2^{T}\ (\underline{x}-\hat{\underline{x}}) \tag{31}$$

where $\underline{B}_1^{T}$ is the upper k × m submatrix of $\underline{B}^{T}$, and $\underline{B}_2^{T}$ is the lower (m-k) × m submatrix. We do not mean to imply that the complementary part is uninteresting, and in fact we may take several different linear combinations of $\underline{x}-\hat{\underline{x}}_0$ for special study. For example, we may wish to plot contours of constant probability density for each parameter pair in turn. To do so, we take k=2, let $\underline{B}_1^{T}$ contain two selected rows of an identity matrix, then repeat for all desired pairs. Sometimes, more complicated linear combinations may be selected for their special physical significance. The rows of the complementary part $\underline{B}_2^{T}$ are arbitrary except that they should be orthogonal to the interesting vectors,

$$\underline{B}_2{}^T\underline{B}_1 = \underline{0} \tag{32}$$

and they should complement $\underline{B}_1{}^T$ so that $\underline{B}^T$ has an inverse.

The marginal pdf of $\underline{z}_1$ is defined to be

$$p(\underline{z}_1) = \int p(\underline{z})\, d\underline{z}_2 \tag{33}$$

where the integral is over all the dummy variables $\underline{z}_2$. The integral may be nasty in general, but the covariance matrix of $\underline{z}_1$ follows easily from (30); the marginal covariance of $\underline{z}_1$ is

$$\underline{Z}_1 = \underline{B}_1{}^T\underline{C}\underline{B}_1 \tag{34}$$

where $\underline{C}$ is given by (20). If the parameter estimates are Gaussian, then the linear combinations $\underline{z}_1$ will also be Gaussian, so that

$$\underline{z}_1 \sim N(\underline{0},\underline{Z}_1) \tag{35}$$

defines the marginal pdf of $\underline{z}_1$.

For comparison with the conditional covariance derived below, it helps to relate $\underline{Z}_1$ to the normal matrix $\underline{M}$. Let

$$\underline{V}_1 = \underline{B}_1(\underline{B}_1{}^T\underline{B}_1)^{-1}, \quad \underline{V}_2 = \underline{B}_2(\underline{B}_2{}^T\underline{B}_2)^{-1} \tag{36}$$

and

$$\underline{M}_{11} = \underline{V}_1{}^T\underline{M}\underline{V}_1\ , \quad \underline{M}_{12} = \underline{V}_1{}^T\underline{M}\underline{V}_2$$

$$\underline{M}_{22} = \underline{V}_2{}^T\underline{M}\underline{V}_2 \tag{37}$$

then

$$\underline{Z}_1 = (\underline{M}_{11} - \underline{M}_{12}\underline{M}_{22}^{-1}\,\underline{M}_{12}{}^T)^{-1} \tag{38}$$

We generally need the marginal pdf for testing hypotheses or assigning confidence limits. The marginal pdf is easy to compute when the posterior estimates $\hat{\underline{x}}$ are Gaussian, but otherwise the integration over (m-k) parameters in (33) may have to be performed numerically. To test linear constraints on the model, or to assess nonlinearity effects as discussed below, we may use the conditional pdf. The conditonal pdf is given by

$$p(\underline{z}_1|\underline{z}_2) = p(\underline{z}_1,\underline{z}_2)\ /\ p(\underline{z}_2) \tag{39}$$

where $\underline{z}_1$ is a vector of k variable parameters, $\underline{z}_2$ is held fixed, and $p(\underline{z}_2)$ is effectively a normalizing constant. We can evaluate (39) more easily than (33), because there is no need to integrate over the complementary variables. If we set $\underline{z}_2 = \underline{0}$, then

$$(\underline{x}-\hat{\underline{x}}) = \underline{V}_1\underline{z}_1 \tag{40}$$

From (27) and (37), we can see that the conditional covariance matrix is

$$\underline{Z}_1{}' = \underline{M}_{11}{}^{-1} \tag{41}$$

The difference between the marginal and conditional covariance matrices shows up clearly when the parameters of interest are simply the first k elements of $\underline{x}$-$\hat{\underline{x}}$. Then $\underline{B}$ is an identity matrix, and $\underline{B}_1{}^T$ contains the top k rows of the identity matrix. The marginal covariance matrix is just the upper left k by k submatrix of $\underline{C}$. The conditional covariance matrix is the inverse of the upper left k by k submatrix of the normal matrix.

2.6. Relative Importance of the Prior and Observational Data

Straightforward methods will quantify the relative importance of any data subset, including the prior data.

1. The easiest method is to compare the posterior covariance matrix $\underline{C}$ with the prior covariance $\underline{D}$. If the posterior covariance is nearly the same as the prior covariance, then the prior data provide nearly all the information, and the observational data nearly none. One may focus on specific linear features of the parameters (called "moments") of the form $z=\underline{b}^T\underline{x}$. If both sides of (22) are nearly equal, then the observations have contributed little new information. On the other hand, if both sides of (23) are nearly equal, then the prior data made no important contribution to resolving $\underline{z}$.

2. The sensitivity of any parameter estimate $\hat{x}_k$ to any datum y_i is given by (8) to be

$$\partial\hat{x}_k/\partial y_i = H_{ki} \tag{42}$$

where H_{ki} is the corresponding element of $\underline{H}$. Similarly, the sensitivity to the prior data is given by

$$\partial\hat{x}_k/\partial x_{0j} = K_{kj} \tag{43}$$

where K_{kj} is the corresponding element of $\underline{K}$, defined in (18). Thus one can compare elements in the k'th rows of $\underline{H}$ and $\underline{K}$, respectively, to see their relative contributions in determining $\hat{\underline{x}}_k$. Remember that $\underline{H}$ is influenced by the prior data, and $\underline{K}$ is influenced by the observations, by (17)-(19). Note also that the elements of $\underline{H}$ and $\underline{K}$ may have different units; we should relate $\underline{H}$ and $\underline{K}$ to the covariance matrices $\underline{E}$ and $\underline{D}$ respectively, for meaningful comparison. A systematic method for standardizing the data and parameters is given below.

3. The matrix products $\underline{HA}$ and $\underline{KI}$ from (21) may be interpreted as partial resolution matrices, corresponding to the observational and prior data, respectively. They sum to $\underline{I}$, because the observational and prior data together resolve the parameter estimates completely to within finite errors having covariance $\underline{C}$. The relative sizes of these dimensionless elements give the relative weights of the two data types in resolving the parameters. Even better, the diagonal elements of the identity sum to m, the number of parameters resolved. Thus the sums of the diagonal elements of $\underline{HA}$ and $\underline{KI}$ give the equivalent number of parameters resolved by each data type.

4. We may also write the covariance matrix $\underline{C}$ as the sum of two parts corresponding to the observed and prior data:

$$\underline{C} = \underline{HEH}^T + \underline{KDK}^T \qquad (44)$$

Again, the relative element sizes indicate the contributions from the two data types.

Which of the above methods provides the best diagnostic? We prefer the third method, comparing the partial resolution matrices, because we can sum the diagonal elements to obtain a meaningful scalar quantity.

2.7. Standardized Variables

The four diagnostic tests above mean more if the data and parameters have been standardized so that their prior covariance matrices are identity matrices. Jackson [1972] discussed the basic idea; we summarize the method only briefly here. Let $\underline{F}$ and $\underline{G}$ be two matrices such that

$$\underline{F}^T\underline{F} = \underline{E}^{-1} \quad \text{and} \quad \underline{G}^T\underline{G} = \underline{D}^{-1} \qquad (45)$$

Such matrices can always be found if $\underline{E}$ and $\underline{D}$ are positive definite, as we assume here. Then we may write

$$\underline{y}' = \underline{A}'\underline{x}' + \underline{e}' \qquad (46)$$

and

$$\underline{x}_0' = \underline{I}\underline{x}' + \underline{d}' \qquad (47)$$

where

$$\underline{y}' = \underline{F}\,\underline{y} \qquad (48)$$

$$\underline{e}' = \underline{F}\,\underline{e} \qquad (49)$$

$$\underline{x}' = \underline{G}\,\underline{x} \qquad (50)$$

$$\underline{x}_0' = \underline{G}\,\underline{x}_0 \qquad (51)$$

$$\underline{d}' = \underline{G}\,\underline{d} \qquad (52)$$

and

$$\underline{A}' = \underline{F}\,\underline{A}\,\underline{G}^{-1} \qquad (53)$$

Equations (46) and (47) are simplei than (1) and (10), because the primed quantities are dimensionless and the covariance matrices for $\underline{e}'$ and $\underline{d}'$ are standardized

$$\underline{E}' = \underline{F}\,\underline{E}\,\underline{F}^T = \underline{I} \qquad (54)$$

and

$$\underline{D}' = \underline{G}\,\underline{D}\,\underline{G}^T = \underline{I} \qquad (55)$$

The various matrices described above have analogies in the standardized coordinate system. The transformations are

$$\underline{M}' = (\underline{G}^{-1})^T\,\underline{M}\,\underline{G}^{-1} = (\underline{A}')^T\,\underline{A}' + \underline{I} \qquad (56)$$

$$\underline{C}' = \underline{G}\,\underline{C}\,\underline{G}^T = (\underline{M}')^{-1} \qquad (57)$$

$$\underline{H}' = \underline{G}\,\underline{H}\,\underline{F}^{-1} = \underline{C}'\,(\underline{A}')^T \qquad (58)$$

$$\underline{K}' = \underline{G}\,\underline{K}\,\underline{G}^{-1} = \underline{C}' \qquad (59)$$

and

$$\underline{I} = \underline{G}\,\underline{H}\,\underline{A}\,\underline{G}^{-1} + \underline{G}\,\underline{K}\,\underline{G}^{-1} = \underline{H}'\,\underline{A}' + \underline{K}' \qquad (60)$$

Equations (22) through (44) become much simpler in the standardized coordinate system because of (54) and (55). Furthermore, the diagnostic relationships become simpler for the relative importance of observed and prior data:

1. Because the prior covariance matrix $\underline{D}'$ is an identity, the diagonal elements of $\underline{C}'$ can be compared directly to one.

2. The elements of $\underline{H}'$ and $\underline{K}'$ are dimensionless and directly comparable. The element H_{ki}' is the partial derivative of the x_k' with respect to y_i', where now x_k' and y_i' are each measured in units of the prior uncertainty.

3. The partial resolution matrices, $\underline{H}'\underline{A}'$ and $\underline{K}'$, are symmetric.

3. Nonlinear Inversion

Suppose the data $\underline{y}$ are related to the unknown parameters by a nonlinear transformation:

$$\underline{y} = \underline{f}(\underline{x}) + \underline{\varepsilon} \qquad (61)$$

where $\underline{f}$ is an n-vector of known functions depending on the arguments $\underline{x}$. Suppose again that we have prior estimates given by (10) and that we have probabilistic information about $\underline{\delta}$. We can use the same logic as in the linear problem; we let

$$\underline{e} = \underline{y} - \underline{f}(\underline{x}) \qquad (62)$$

be the residual for a hypothetical solution $\underline{x}$, and maximize the likelihood function (15).

3.1. Maximum Likelihood Solution

If both the observational and prior data are Gaussian with properties described by (4) and (14), then the maximum likelihood solution minimizes (16) with $\underline{e}$ given by (62). In the linear case, T^2 is a quadratic function of $\underline{x}$, so that minimizing (16) leads to linear equations in $\underline{x}$. Furthermore, when the observations and prior data are Gaussian, so are the final estimates. In the nonlinear case, (16) is no longer quadratic in $\underline{x}$, and the final estimates are no longer Gaussian, even if the observations and prior data are Gaussian. The example below will demonstrate clearly this non-Gaussian behavior.

3.2. An Algorithm

There are several numerical methods to minimize (6), or (16), in the Bayesian approach. Dahlquist and Bjork [1974] give a comprehensive summary. The best method to use in practice will depend on the existence and ease of computing the first- and higher-order partial derivatives of $\underline{f}(\underline{x})$. The most popular method in geophysical applications is the Gauss-Newton method.

The basic idea is to expand the data equations (61) in a Taylor series about the starting guess $\underline{x}_0$, then estimate a correction to $\underline{x}_0$ using linear inversion theory. The partial derivatives $\partial f_i/\partial x_j$ make up the design matrix $\underline{A}$, and if the functions $\underline{f}$ were linear, then the method would converge to the maximum likelihood solution in one iteration. If the functions $\underline{f}(\underline{x})$ are continuous and differentiable, then the maximum likelihood solution, which minimizes (16), satisfies

$$\underline{A}^T\underline{E}^{-1}\underline{e} + \underline{D}^{-1}\underline{d} = \underline{0} \tag{63}$$

where $\underline{e}$ and $\underline{d}$ are given in (62) and (12), and $\underline{A}$ is an $\underline{n}$ by m matrix of elements

$$A_{ij} = [\ \partial f_i/\partial x_j\ ]_x \tag{64}$$

It should be noted that $\underline{e}$, $\underline{d}$ and $\underline{A}$ in (63) are functions of $\underline{x}$. To solve the implicit equations (63) we expand the observation equations (61) in a Taylor series about a starting model $\underline{x}_k$, then estimate a correction to $\underline{x}_k$ using linear theory, and repeat the same process until it converges. If the functions $\underline{f}$ were linear, then the procedure would converge to the maximum likelihood solution in one iteration. A simple algorithm that converges rapidly to the maximum likelihood solution in some strongly underdetermined examples [see Matsu'ura and Jackson, 1984] is as follows:

$$\underline{x}_{k+1} = \underline{x}_k + b\,\underline{M}_k^{-1}\underline{r}_k \tag{65}$$

where

$$\underline{M}_k = \underline{A}_k^T\underline{E}^{-1}\underline{A}_k + \underline{D}^{-1} \tag{66}$$

$$\underline{r}_k = \underline{A}_k^T\underline{E}^{-1}\underline{e}_k + \underline{D}^{-1}\underline{d}_k \tag{67}$$

and $\underline{A}_k$ is given by (64) with $\underline{x} = \underline{x}_k$. Similarly, $\underline{e}$ and $\underline{d}$ are given by (62) and (12) with $\underline{x} = \underline{x}_k$.

The factor b is an adjustable parameter to be set between zero and one at each step. For nearly linear problems it can be set safely to one, but for strongly nonlinear problems it should be smaller than one to keep the perturbations from wantonly overstepping the linear range of $\underline{f}(\underline{x})$. The iterations begin at $\underline{x} = \underline{x}_0$ with k = 0, and proceed until $\underline{r}_k$ is acceptably small. This convergence criterion is quite reasonable, since it is clear that $\underline{x}_k$ which gives $\underline{r}_k = \underline{0}$ satisfies the maximum likelihood condition (63). Let $\hat{\underline{x}}$ be limiting values of $\underline{x}_k$, satisfying (63). For the final estimate $\hat{\underline{x}}$, we can evaluate the covariance of estimation errors directly from (63) under the assumption of linearity for $\underline{f}(\underline{x})$ at $\underline{x} = \hat{\underline{x}}$:

$$\underline{C} = \underline{M}^{-1} \tag{68}$$

with

$$\underline{M} = \underline{A}^T\underline{E}^{-1}\underline{A} + \underline{D}^{-1} \tag{69}$$

where $\underline{A}$ is given by (64) with $\underline{x} = \hat{\underline{x}}$. The matrix $\underline{C}$ is often called the asymptotic covariance matrix for the estimation errors.

In principle, the Bayesian estimation scheme is no different from standard nonlinear least squares. By simply renaming the prior data as observational data, we could write (12) in the form (62), and achieve the same result as (65) by using almost any nonlinear least squares algorithm. We prefer to spotlight the prior data with a separate name. Because the prior covariance matrix $\underline{D}$ is positive definite, the normal matrix (66) must be nonsingular, and the prior data will thus stabilize the search procedure. The degree of stabilization increases as the prior data become less uncertain, so that $\underline{D}^{-1}$ increases.

The algorithm described by (65) through (67) resembles the Marquardt [1963] method, in which a diagonal term is added to the normal matrix to stabilize the iteration procedures. The Marquardt method could also be described by (65)-(67), but with

$$\underline{D}^{-1} = \lambda_k\,\underline{I} \tag{70}$$

$$\underline{d}_k = \underline{0} \tag{71}$$

and b = 1. The quantity λ_k is decreased as the iterations proceed and $\underline{r}_k$ is reduced. The Marquardt method differs significantly from ours in principle and practice. In the Marquardt method, the primary intention is to stabilize the iterative search for the least squares estimate in the early stage, until the estimate becomes linearly close to the optimal solution. Furthermore, the optimality criterion includes only the observational data in the Marquardt method: the vector $\underline{d}_k$ is merely the step between successive iterations, and the stabilizing term $\underline{D}^{-1}$ in (66) is made to decrease as the iterations progress. In our method, the prior data are treated as legitimate, to be fit by the prior model. Thus $\underline{d}_k$ measures the distance between $\underline{x}_k$ and the prior estimate $\underline{x}_0$, and $\underline{D}^{-1}$ does not depend on $\underline{x}_k$. The final Marquardt estimate may depend strongly on the starting model, if the normal matrix is nearly singular. Our final estimate depends on the prior estimate but not upon the starting or intermediate models. Also, our normal matrix cannot be singular because $\underline{D}^{-1}$ is nonsingular. If the functions $\underline{f}(\underline{x})$ are linear, our algorithm converges to the final solution in just one iteration; not so for the Marquardt algorithm.

We could use the Marquardt method to stabilize the iterative search procedure by adding to (66) a term such as $\lambda_k\,\underline{I}$, instead of introducing the adjustable parameter b, where λ_k decreases to zero with iteration number k. The effect of the additional term would be like that of setting b<1 in (65); it simply reduces the step size.

Inclusion of adequate prior data will guarantee a unique solution for linear problems, but not for nonlinear problems. In the optimization criterion (16), the term in $\underline{d}$ has a unique minimum, but the other term may have many isolated minima. The Bayesian approach is no panacea, but prior data can help reduce nonuniqueness in many nonlinear problems.

3.3. Marginal and Conditional Statistics

Let $\underline{x}$, $\underline{M}$, and $\underline{C}$ be limiting values of $\underline{x}_k$, $\underline{M}_k$, and $\underline{C}_k$, satisfying the maximum likelihood condition (63). The matrix $\underline{C}$ is often called the asymptotic covariance matrix for the estimation errors. When the relationship between the data and the parameters is only mildly nonlinear, then the asymptotic covariance matrix $\underline{C}$ in (68) may be close to the exact covariance matrix for

TABLE 1. One-Dimensional Inverse Problem

Case	σ_y	σ_x	x_0	$\hat{x}$	σ
(a)	0.2	0.2	0.424	0.8635	0.100
(b)	0.2	0.5	0.212	0.9685	0.101
(c)	0.5	0.2	0.212	0.2993	0.195
(d)	0.5	0.5	0.0	0.7071	0.289

estimation errors. We could evaluate the exact covariance matrix using

$$C_{ij} = \int (x_i - \hat{x}_i)(x_j - \hat{x}_j)\, p(\underline{x}|\underline{y})\, d\underline{x} \qquad (72)$$

using (13) for $p(\underline{x}|\underline{y})$ in the general case, or (15), (16), (12), and (62) if both the observational and prior data are Gaussian. Enormous calculations would be required for most problems. We can test the linearity more easily if we assume the asymptotic covariance matrix is valid, then compare the asymptotic pdf to the exact pdf. For Gaussian data, the asymptotic pdf is given by (15) and (27) with $\underline{M}$ being the asymptotic normal matrix in (69). The exact pdf is given by (15) with (16), (12), and (62). For plotting purposes, we may want to select parameters one or two at a time, and plot the marginal or conditional pdf. Generally, the marginal pdf is more useful, but computing it by (33) for the exact case involves an integration over m-k variables. So it may be necessary to compromise and use the conditional pdf instead. If the exact and asymptotic versions of the conditional pdf disagree strongly, then probably no asymptotic statistic is valid. If the exact and asymptotic conditional pdf agree closely, then the marginal pdf's might agree also, especially if the interesting variables are uncorrelated with the complementary variables. If the interesting and complementary variables are strongly correlated, then the marginal distribution may differ greatly from the conditional. Thus two conditions support the validity of the asymptotic statistics: (1) the exact and asymptotic conditional pdf's agree closely, and (2) the interesting and complementary variables are uncorrelated according to the asymptotic covariance matrix.

4. Examples

4.1. One Parameter Nonlinear Problem

Suppose we estimate the velocity of a particle from its measured kinetic energy. We designate the velocity by x and assume that the particle moves in one dimension only. The observed kinetic energy is

$$y = ax^2 + e \qquad (73)$$

an equation of the form (61). We take units such that a = 1, let the measured kinetic energy be 1, and assume the error e has mean 0 and variance σ_y^2.

In the absence of prior information we could use (5) with

$$S^2 = (1 - x^2)^2 / \sigma_y^2 \qquad (74)$$

and $a = (2\pi\sigma_y^2)^{-1/2}$. Equation (74) has minima at x = 1 and x = -1.

Now assume that we have prior information about the velocity, say, $x_0 = x + d$, where d is an estimation error with mean 0 and variance σ_x^2. We use (15) with

$$T^2 = (1 - x^2)^2 / \sigma_y^2 + (x_0 - x)^2 / \sigma_x^2 \qquad (75)$$

The optimality condition (63) is

$$2x(1-x^2) / \sigma_y^2 + (x_0 - x) / \sigma_x^2 = 0 \qquad (76)$$

There are three solutions to (76); when all are real, they correspond to two minima with an intervening maximum. Otherwise there is one minimum and two complex solutions. For any solution $\hat{x}$ which corresponds to a minimum, (64), (69), and (68) give

$$A = 2\hat{x} \qquad (77)$$

$$M = 4\hat{x}^2 / \sigma_y^2 + 1/\sigma_x^2 \qquad (78)$$

and

$$C = \sigma^2 = (4\hat{x}^2/\sigma_y^2 + 1/\sigma_x^2)^{-1} \qquad (79)$$

We chose several different values of σ_y, σ_x, and x_0 to illustrate the properties of the solu-

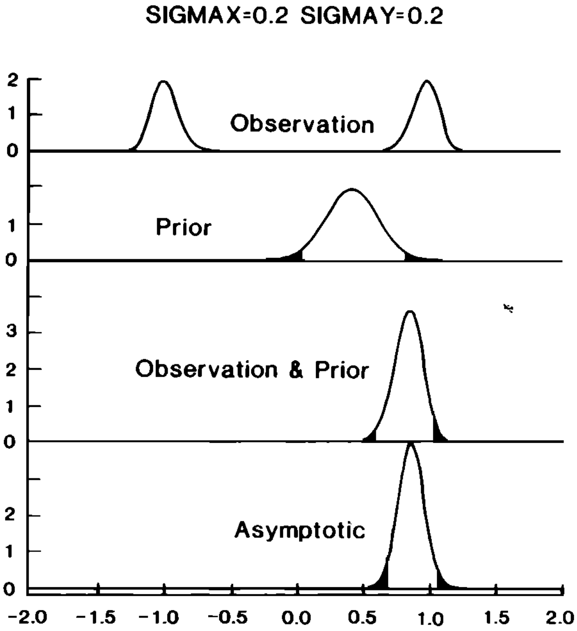

Fig. 1. Probability density functions for example 1, case (a). From top to bottom, curves represent p(y|x), p(x), p(x|y), and the asymptotic Gaussian approximation to p(x|y). Shaded areas indicate values outside the equal-tailed 95% confidence interval. For case (a), p(x|y) and the asymptotic Gaussian are nearly identical, indicating that the asymptotic error estimate is adequate.

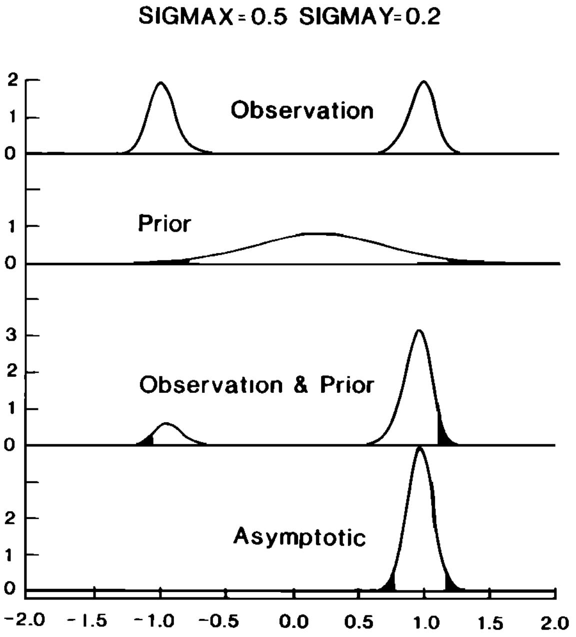

Fig. 2. Results for case (b), in same format as Figure 1. The asymptotic error estimate is misleading, because it ignores the possible solutions near x = -1.

tions and the various probability density functions. We took x = (1 - σ_x - σ_y) / $\sqrt{2}$, so that the prior datum is weakly inconsistent with the observational datum. The values we used, the value $\hat{x}$ that minimizes (75), and the asymptotic standard deviation σ, are shown in Table 1.

Results for all cases are shown in Figures 1-4. In each case the upper curve is p(y|x), given by (5) and (74). Although this curve is not strictly a pdf of x, we have normalized it with respect to x for plotting convenience. The second curve is the prior pdf, a Gaussian with mean x_0 and variance σ_x^2 as described above.

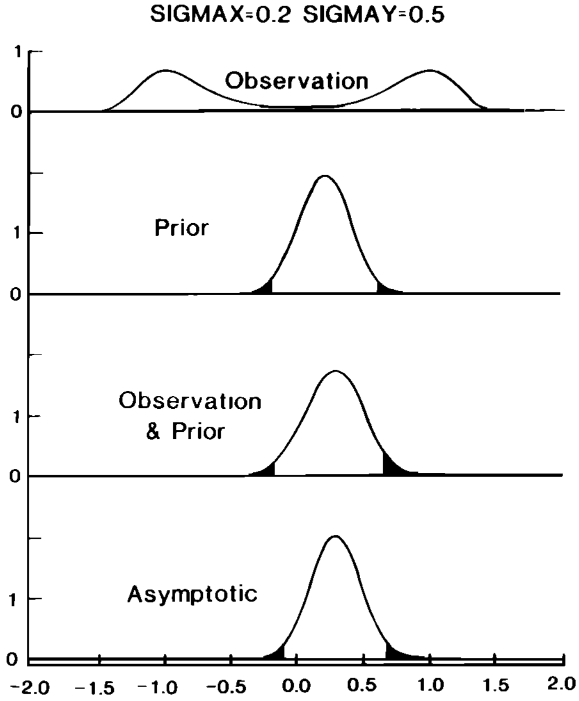

Fig. 3. Results for case (c) in same format as Figure 1. The asymptotic error estimate is adequate, because the posterior pdf is dominated by the Gaussian prior pdf.

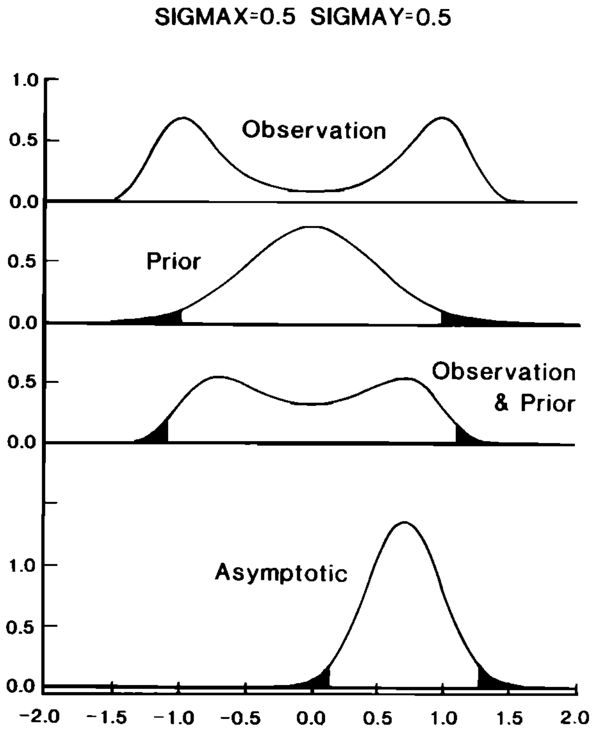

Fig. 4. Results for case (d) in same format as Figure 1. The asymptotic Gaussian poorly represents p(x|y), indicating that linear theory gives misleading error estimates.

The third curve is the posterior pdf, given by (13) as the normalized product of the upper two curves. The bottom curve is the asymptotic Gaussian pdf with mean value $\hat{x}$ and variance σ^2 determined from (79). Shaded regions on each of the lower three curves show values of x outside the equal tailed 95% confidence interval.

Case (a) reveals the advantages of having accurate observations combined with accurate prior data. Even though the prior datum is inconsistent with the observational datum, it is sufficient to exclude the solution at x = -1 allowed by the observation of x^2. The asymptotic Gaussian pdf of x is hardly distinguishable from the exact posterior pdf. The 95% confidence limits agree closely. Thus the asymptotic error estimate is very accurate for this case.

In case (b) the prior uncertainty is larger, and the prior estimate is closer to zero than in case (a). The posterior pdf (labeled "observation & prior") clearly favors solutions with positive velocity but does not exclude negative velocities. The posterior pdf is bimodal: it resembles a Gaussian curve near its positive peak, but the 95% confidence limits differ strongly from those of the asymptotic pdf because of the negative solutions. Case (b) differs from case (a) only in the prior data, so clearly the prior data contribute significantly to case (a). In case (b) the asymptotic error estimate is slightly misleading.

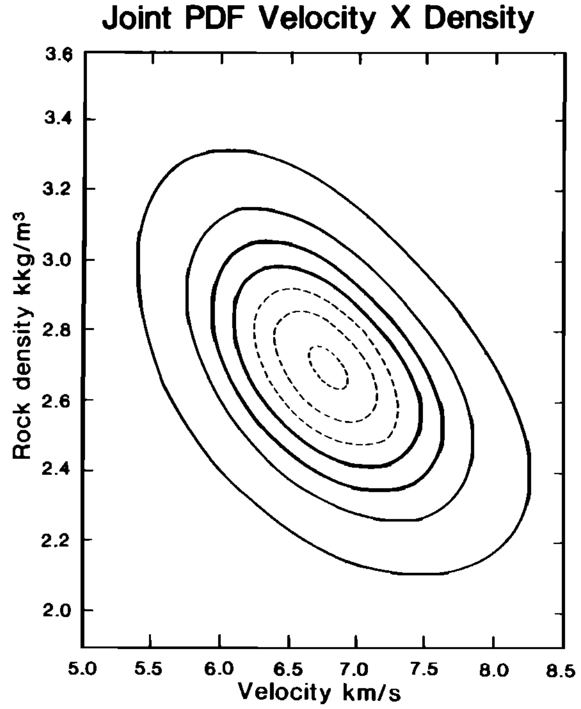

Fig. 5. Contours of constant pdf versus density and velocity for example 2. The asymptotic Gaussian approximation is not plotted because it is indistinguishable within the 95% confidence region.

In case (c) we have accurate prior data but less accurate observational data. The posterior pdf strongly resembles the prior, so the observational datum does not contribute much to the final answer. The asymptotic pdf is a fair representation of the exact posterior pdf, so the asymptotic procedure gives reasonable error bounds. However, this is true only because the prior datum dominates.

In case (d) neither the observational nor the prior datum is very accurate. The prior datum has a mean value of 0, so it does not discriminate at all between the positive and negative solutions. To compute the asymptotic pdf, we arbitrarily chose the positive solution to the optimization problem, although a nonlinear least squares procedure might find the negative solution just as well. The asymptotic pdf differs drastically from the exact posterior pdf, so the asymptotic procedure gives quite misleading error estimates for this case.

Taken together, cases (a)-(d) show that the validity of the asymptotic estimates depends on both the observational and prior uncertainty. For some problems, we might assess the adequacy of linearized inversion by examining the nonlinearity of f(x) within a few posterior standard deviations of the maximum likelihood solution. In this problem, a very accurate observation might assure that f(x) is linear where it counts, but the observation could never resolve the ambiguity between the positive and negative solutions.

The asymptotic standard deviations for cases (a) and (b) are nearly identical, yet the posterior pdfs differ. For this reason we recommend plotting an exact posterior pdf, even if it must be the conditional pdf of a single variable at a time, to assess nonlinearity.

We tested the importance of prior data using the methods of section 2.7. Results appear in Table 2.

Because we have only one parameter and one observation, the matrices $\underline{H}'$, $\underline{K}'$, and $\underline{H}'\underline{A}'$ become scalars. Remember that all items in Table 2 are based on linear theory, which is reasonably valid for case (a) and (c), marginally valid for case (b), and hardly applicable to case (d). For all cases, the resolutions H'A' and K' add to 1.0, so we can view H'A' as the fraction of information provided by the observation, and K' as the fraction provided by the prior datum. As expected from the discussion above, the contribution of the prior data is dominant in case (c), significant in cases (a) and (d), and barely noticeable in case (b).

TABLE 2. Importance of Prior and Observational Data: One-Dimensional Inverse Problem

Case	H'	K'	H'A'
(a)	0.434	0.251	0.749
(b)	0.198	0.041	0.959
(c)	0.226	0.946	0.054
(d)	0.471	0.333	0.667

4.2. Two-Parameter Nonlinear Problem

Suppose we observe the acoustic impedance of a rock layer and wish to determine the density and seismic velocity of the layer. Letting y be the acoustic impedance, x_1 the density, and x_2 the seismic velocity, we have

$$y = a\, x_1\, x_2 + e \tag{80}$$

where e is a random error. The problem is clearly nonlinear and without prior information has infinitely many solutions. With reasonable prior information, the density and velocity become uniquely resolved, and linear inverse theory is adequate. We take

$$x_{10} = x_1 + d_1, \quad x_{20} = x_2 + d_2 \tag{81}$$

where x_{10} = 2.8 10^3kg/m^3, x_{20} = 7.0 km/s, y = 17.6, and a = 10^{-6} m^{2}s/kg. We assume covariances $E_{11} = (2.0)^2$; D_{11} = (0.3 10^3kg/m^3)^2, D_{22} = (0.7 km/s)^2, and D_{12} = 0. Thus the relative uncertainties are about 10% for both observational and prior data, not unreasonable for most regions of the earth's crust. Note also that the observation is different from, but not inconsistent with, the value of 19.6 predicted from the prior estimates.

We used the algorithm of section 3.2 to find the maximum likelihood estimate of density and velocity, starting at the prior solution $\underline{x}_0$. At each step we have

$$\underline{A} = (x_2,\, x_1) \tag{82}$$

and

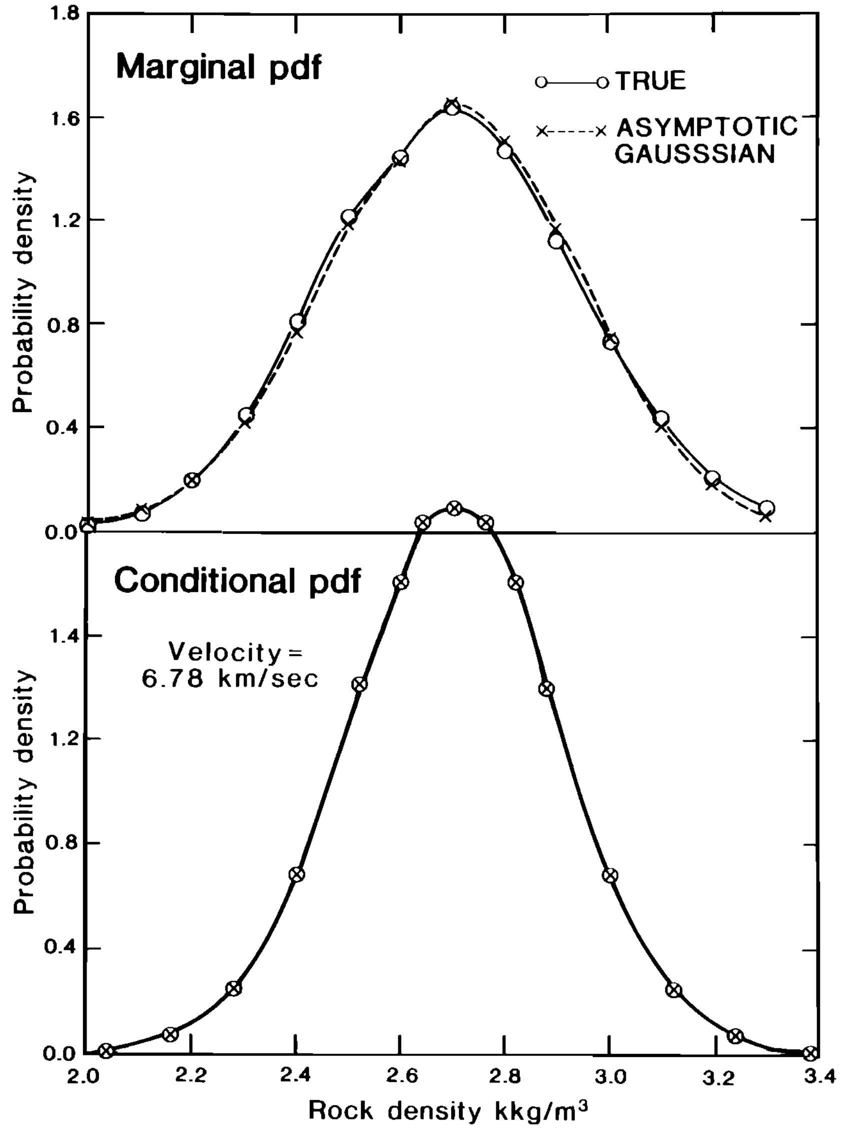

Fig. 6. Marginal (top) and conditional (bottom) pdf of density for example 2. The exact pdf is shown solid, while the asymptotic Gaussian is dashed.

$$\underline{M} = \begin{vmatrix} x_2{}^2/E_{11}+ 1/D_{11} & x_1x_2/E_{11} \\ & \\ x_1x_2/E_{11} & x_1{}^2/E_{11}+ 1/D_{22} \end{vmatrix} \qquad (83)$$

The maximum likelihood solution is $x_1 = 2.70\ 10^3\text{kg/m}^3$, $x_2 = 6.78$ km/s, $C_{11} = (0.241\ 10^3\text{kg/m})^2$, $C_{22} = (0.584\ \text{km/s})^2$, with a correlation coefficient of -0.49.

The posterior pdf is contoured in Figure 5 as a function of density and velocity. We used (15), (16), (12), and (62) in the calculations We also used (27), based on the asymptotic normal matrix $\underline{M}$ in (69), to calculate the pdf under the linear assumption. The exact and asymptotic pdf are indistinguishable within their 95% confidence intervals, so we have not plotted the asymptotic version.

Figure 6 shows both exact and asymptotic versions of the marginal and conditional pdf of density. We followed the prescription of section (2.5), with $\underline{B}_1{}^T = (1,0)$, $\underline{B}_2{}^T = (0,1)$, so that $z_1 = x_1$ and $z_2 = x_2$. The density has marginal and conditional variances of $(0.241\ 10^3\text{kg/m}^3)^2$ and $(0.210\ 10^3\text{kg/m}^3)^2$, respectively. The similarity between the exact and asymptotic pdf's shows that the asymptotic error limits suffice here.

Clearly the prior data must be important in this problem, because without them the problem would be nonlinear and nonunique. We used the tools of section 2.7 to quantify the importance of the prior information. We found $\Sigma\ (\underline{H}'\underline{A}')_{ii}{}'$ $= 0.66$, $\Sigma K_{ii}{}' = 1.34$, which we interpret to mean that the observation provided 33% of the information, while the rest came from the prior data. Table 3 shows a more detailed breakdown. All parameters in Table 3 are standardized, so that observations are measured in units of their standard deviations, and the parameters are measured in units of their prior standard deviations. The matrices $\underline{H}'$ and $\underline{K}'$ reveal the estimates' sensitivity to the observational and

TABLE 3. Importance of Prior and Observational Data: Two-Dimensional Inverse Problem

	H'	K'		H'A'	
x_1 (density)	0.348	0.647	-0.328	0.353	0.328
x_2 (velocity)	0.323	-0.328	0.695	0.328	0.305

prior data, respectively. We see that the estimated density is most sensitive to the prior estimate of density, and that a 1.0 unit increase in the prior velocity estimate will cause a decrease of 0.328 units in the estimated density. Likewise the velocity depends most strongly on its own prior estimate, and decreases with the increasing prior density estimate. Ironically, the final estimate of the velocity x_2 is slightly more sensitive to the prior density estimate (-0.328) than it is to the observation. But the relationship is symbiotic; without one, the other would not help much. Remembering that $\underline{C}' = \underline{K}'$, and $\underline{D}' = \underline{I}$, we see that the posterior variances of density and velocity have been reduced from the prior variances by 35% and 30%, respectively. We interpret these figures as the percentages of information provided by the observational datum.

5. Conclusions

To solve nonunique inverse problems, we must supply some form of prior information. Bayes' rule provides a quantitative way to do this and to make meaningful error estimates for the solution. In general the application of Bayes' rule may require laborious computations. However, if the observations and the prior data have Gaussian errors, then Bayes' rule leads to a simple solution having the same form as the standard weighted least squares solution. For linear inverse problems, we can assess the relative importance of any datum from the standardized estimation matrix $\underline{H}'$ for observational data or $\underline{K}'$ for prior data. The resolution matrix $\underline{I}$ is the sum of the partial resolution matrices $\underline{H}'\underline{A}'$ and $\underline{K}'$ corresponding to the observational and prior data, respectively.

For nonlinear problems the posterior pdf may be strongly non-Gaussian even if both observations and prior data are Gaussian. Our one-parameter example illustrated this well. The asymptotic error estimates derived from linear theory are based on Gaussian parameter errors, so those error estimates may be invalid for some highly nonlinear problems. However, prior data and accurate observations can make intrinsically nonlinear problems into effectively linear ones, as our two-parameter example showed. Comparison of the exact pdf with the asymptotic pdf is the best way to assess the importance of nonlinearity. We can make this comparison without computing the full multidimensional pdf. Instead, we use a simple algorithm for iterative search to find the maximum likelihood parameter estimates, and compute the exact and asymptotic marginal pdf's for one or two parameters at a time. If these agree within reasonable confidence limits, then the problem is effectively linear.

Acknowledgments. We gratefully acknowledge support from NASA (contract NAS5-26895), USGS (contract 14-08-0001-21243), and UCLA (intramural computing account IBI8).

References

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

Backus, G. E., Inference from inadequate and inaccurate data, *Lectures in App. Math.*, 14, 1-105, 1971.

Backus, G. E., and F. J. Gilbert, The resolving power of gross earth data, *Geophys. J. R. Astron. Soc.*, 16, 169-205, 1968.

Backus, G. E., and F. J. Gilbert, Uniqueness in the inversion of gross earth data, *Philos. Trans. R. Soc. London*, 266, 123-192, 1970.

Box, G. E. P., and G. C. Tiao, *Bayesian Inference*, Addison-Wesley, Reading, Mass., 1973.

Dahlquist, G., and A. Bjork, *Numerical Methods*, 573 pp., Prentice-Hall, Englewood Cliffs, N. J., 1974.

Hoel, P. Q., *Introduction to Mathematical Statistics*, 409 pp., John Wiley, New York, 1971.

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

Jackson, D. D., The use of a priori data to resolve nonuniqueness in linear inversion, *Geophys. J. R. Astron. Soc.*, 57, 137-157, 1979.

Marquardt, D. L., An Algorithm for least-squares estimation of nonlinear parameters, *J. Soc. Ind. Appl. Math*, 2, 431-441, 1963.

Matsu'ura, M., and N. Hirata, Generalized least-squares solutions to quasi-linear inverse problem with a priori information, *J. Phys. Earth*, 30, 451-468, 1982.

Matsu'ura, M., and D. Jackson, Inverse dislocation model for Hollister trilateration data, submitted to *J. Geophys. Res.*, 1984.

Tarantola, A., and B. Valette, Inverse problems: Quest for information, *J. Geophys.*, 50, 159-170, 1982a.

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

Wiggins, R. A., The general linear inverse problems: Implication of surface waves and free oscillations for earth structure, *Rev. Geophys. Space Phys.*, 10, 251-285, 1972.

D. D. Jackson, Institute of Geophysics and Planetary Physics, Department of Earth and Space Sciences, University of California, Los Angeles, CA 90024.

M. Matsu'ura, Geophysical Institute, Faculty of Science, University of Tokyo, Tokyo, Japan.

(Received December 1, 1983
revised July 3, 1984
accepted September 6, 1984.)