Chapter 5

Solution of the Linear, Gaussian Inverse Problem, Viewpoint 3: Maximum Likelihood Methods

5.1 THE MEAN OF A GROUP OF MEASUREMENTS

Suppose that an experiment is performed N times and that each time a single datum d_i is collected. Suppose further that these data are all noisy measurements of the same model parameter m_1. In the view of probability theory, N realizations of random variables, all of which have the same probability density function, have been measured. If these random variables are Gaussian, their joint probability density function can be characterized in terms of a variance σ^2 and a mean m_1 (see Section 2.4) as

$$p(\mathbf{d}) = \sigma^{-N}(2\pi)^{-N/2} \exp\left[-\frac{1}{2}\sigma^{-2}\sum_{i=1}^{N}[d_i - m_1]^2\right] \tag{5.1}$$

The data $\mathbf{d}^{\text{obs}}$ can be represented graphically as a point in the N-dimensional space whose coordinate axes are $d_1, d_2, \ldots, d_N$ (Figure 5.1). The probability density function for the data can also be graphed (Figure 5.2). Note that the probability density function is centered about the line $d_1 = d_2 = \cdots = d_N$, since all the ds are supposed to have the same mean, and that it is spherically symmetric, since all the ds have the same variance.

Suppose that we guess a value for the unknown mean and variance, thus fixing the center and diameter of the probability density function. We can then calculate its numerical value at the data $p(\mathbf{d}^{\text{obs}})$. If the guessed values of mean and variance are close to being correct, then $p(\mathbf{d}^{\text{obs}})$ should be a relatively large number. If the guessed values are incorrect, then the probability, or *likelihood*, of the observed data will be small. We can imagine sliding the cloud of probability in Figure 5.2 up along the line and adjusting its diameter until its probability at the point $\mathbf{d}^{\text{obs}}$ is maximized.

Geophysical Data Analysis: Discrete Inverse Theory. DOI: 10.1016/B978-0-12-397160-9.00005-9

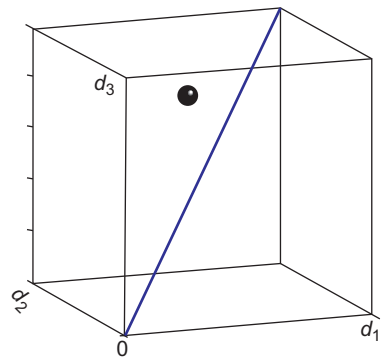

FIGURE 5.1 The data are represented by a single point (black) in a space whose dimensions equal the number of observations (in this case, 3). These data are realizations of random variables with the same mean and variance. Nevertheless, they do not necessarily fall on the line $d_1 = d_2 = d_3$ (blue). *MatLab* script gda05_01.

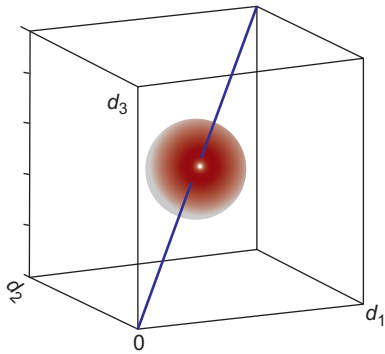

FIGURE 5.2 If the data d_i are assumed to be uncorrelated with equal mean and uniform variance, their probability density function $p(\mathbf{d})$ is a spherical cloud (red), centered on the line $d_1 = d_2 = d_3$ (blue). *MatLab* script gda05_02.

This procedure defines a method of estimating the unknown parameters in the distribution, the *method of maximum likelihood*. It asserts that the optimum values of the parameters maximize the probability that the observed data are in fact observed. In other words, the value of the probability density function at the point $\mathbf{d}^{\text{obs}}$ is made as large as possible. The maximum is located by differentiating $p(\mathbf{d}^{\text{obs}})$ with respect to mean and variance and setting the result to zero as

$$\partial p / \partial m_1 = \partial p / \partial \sigma = 0 \tag{5.2}$$

Maximizing $\log p(\mathbf{d}^{\text{obs}})$ gives the same result as maximizing $p(\mathbf{d}^{\text{obs}})$, since $\log(p)$ is a monotonic function of p. We therefore compute derivatives of the *likelihood function*, $L = \log p(\mathbf{d}^{\text{obs}})$ (Figure 5.3). Ignoring the overall normalization of $(2\pi)^{-N/2}$ we have

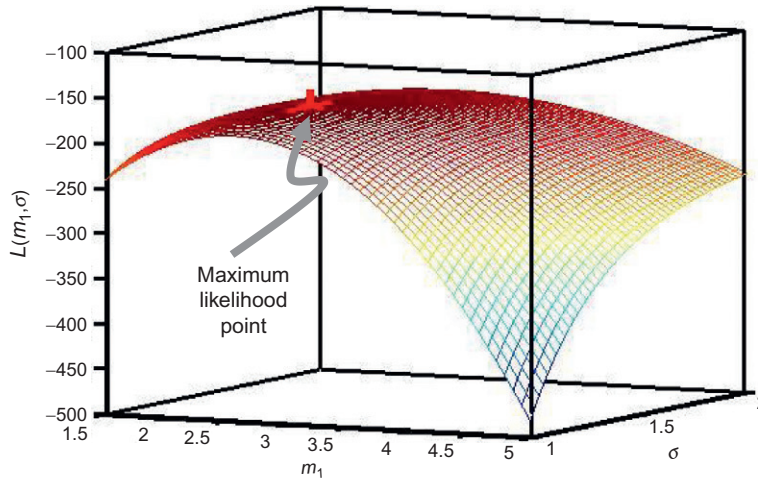

FIGURE 5.3 Likelihood surface for 100 realizations of random variables with equal mean $m_1 = 2.5$ and uniform variance $\sigma^2 = (1.5)^2$. The curvature in the direction of m_1 is greater than the maximum in the direction of the σ, indicating that the former can be determined to greater certainty. *MatLab* script gda05_03.

$$L = \log(p(\mathbf{d}^{\text{obs}})) = -N\log(\sigma) - \frac{1}{2}\sigma^{-2}\sum_{i=1}^{N}(d_i^{\text{obs}} - m_1)^2$$

$$\frac{\partial L}{\partial m_1} = 0 = \frac{1}{2}\sigma^{-2}2m_1\sum_{i=1}^{N}(d_i^{\text{obs}} - m_1) \tag{5.3}$$

$$\frac{\partial L}{\partial \sigma} = 0 = -\frac{N}{\sigma} + \sigma^{-3}\sum_{i=1}^{N}(d_i^{\text{obs}} - m_1)^2$$

These equations can be solved for the estimated mean and variance as

$$m_1^{\text{est}} = \frac{1}{N}\sum_{i=1}^{N}d_i^{\text{obs}} \quad \text{and} \quad \sigma^{\text{est}} = \left[\frac{1}{N}\sum_{i=1}^{N}(d_i^{\text{obs}} - m_1^{\text{est}})^2\right]^{1/2} \tag{5.4}$$

The estimate for m_1 is just the usual formulas for the sample mean. The estimate for σ is the root mean squared error and also is almost the formula for the sample standard deviation, except that it has a leading factor of $1/N$, instead of $1/(N-1)$. We note that these estimates arise as a direct consequence of the assumption that the data possess a Gaussian distribution. If the data distribution were not Gaussian, then the arithmetic mean might not be an appropriate estimate of the mean of the distribution. (As we shall see in Section 8.2, the sample median is the maximum likelihood estimate of the mean of an exponential distribution.)

5.2 MAXIMUM LIKELIHOOD APPLIED TO INVERSE PROBLEM

5.2.1 The Simplest Case

Assume that the data in the linear inverse problem $\mathbf{Gm} = \mathbf{d}$ have a multivariate Gaussian probability density function, as given by

$$p(\mathbf{d}) \propto \exp\left[-\frac{1}{2}(\mathbf{d} - \mathbf{Gm})^{\mathrm{T}}[\mathrm{cov}\ \mathbf{d}]^{-1}(\mathbf{d} - \mathbf{Gm})\right] \tag{5.5}$$

We assume that the model parameters are unknown but (for the sake of simplicity) that the data covariance is known. We can then apply the method of maximum likelihood to estimate the model parameters. The optimum values for the model parameters are the ones that maximize the probability that the observed data are in fact observed. The maximum of $p(\mathbf{d}^{\mathrm{obs}})$ occurs when the argument of the exponential is a maximum or when the quantity given by

$$(\mathbf{d}^{\mathrm{obs}} - \mathbf{Gm})^{\mathrm{T}}[\mathrm{cov}\ \mathbf{d}]^{-1}(\mathbf{d}^{\mathrm{obs}} - \mathbf{Gm}) \tag{5.6}$$

is a minimum. But this expression is just a weighted measure of prediction error. The maximum likelihood estimate of the model parameters is nothing but the weighted least squares solution, where the weighting matrix is the inverse of the covariance matrix of the data (in the notation of Chapter 3, $\mathbf{W}_e = [\mathrm{cov}\ \mathbf{d}]^{-1}$). If the data happen to be uncorrelated and all have equal variance, then $[\mathrm{cov}\ \mathbf{d}] = \sigma^2_d\mathbf{I}$, and the maximum likelihood solution is the simple least squares solution. If the data are uncorrelated but their variances are all different (say, σ^2_{di}), then the prediction error is given by

$$E = \sum_{i=1}^{N} \sigma_{di}^{-2} e_i^2 \tag{5.7}$$

where $e_i = (d_i^{\mathrm{obs}} - d_i^{\mathrm{pre}})$ is the prediction error for each datum. Each individual error is weighted by the reciprocal of its standard deviation; the most certain data are weighted most.

We have justified the use of the L_2 norm through the application of probability theory. The least squares procedure for minimizing the L_2 norm of the prediction error makes sense if the data are uncorrelated, have equal variance, and obey Gaussian statistics. If the data are not Gaussian, then other measures of prediction error may be more appropriate.

5.2.2 *A Priori* Distributions

The least squares solution does not exist when the linear problem is underdetermined. From the standpoint of probability theory, the probability density function of the data $p(\mathbf{d}^{\mathrm{obs}})$ has no well-defined maximum with respect to variations of the model parameters. At best, it has a ridge of maximum probability (Figure 5.4).

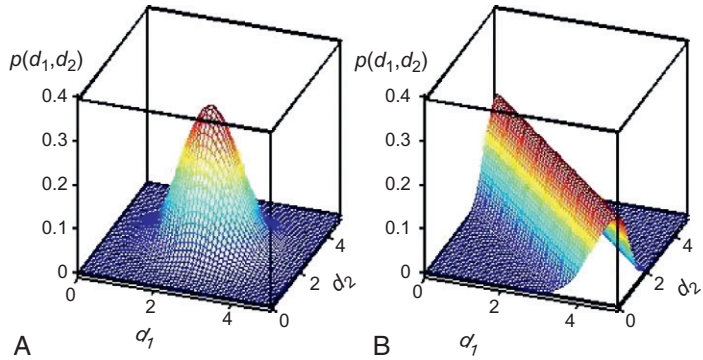

FIGURE 5.4 (A) Probability density function $p(d_1,d_2)$ with a well-defined peak. (B) Probability density function with a ridge. *MatLab* script gda05_04.

We must add *a priori* information that causes the distribution to have a well-defined peak in order to solve an underdetermined problem. One way to accomplish this goal is to write the *a priori* information about the model parameters as a probability density function $p_A(\mathbf{m})$, where the subscript A means "*a priori*." The mean of this probability density function is then the value we expect the model parameter vector to have, and its shape reflects the certainty of this expectation.

A priori distributions for the model parameters can take a variety of forms. For instance, if we expected that the model parameters are close to $\langle \mathbf{m} \rangle$, we might use a Gaussian distribution with mean $\langle \mathbf{m} \rangle$ and variance that reflects the certainty of our knowledge (Figure 5.5). If the *a priori* value of one model parameter were more certain than another, we might use different variances for

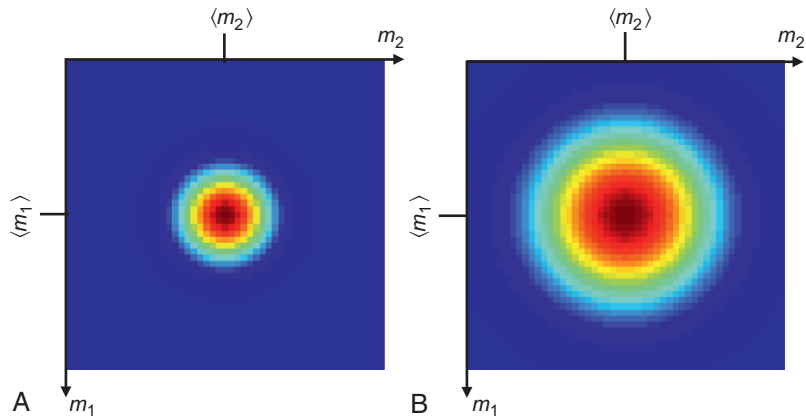

FIGURE 5.5 *A priori* information about model parameters m_1 and m_2, represented with a probability density function $p(m_1,m_2)$. Most probable values are given by means $\langle m_1 \rangle$ and $\langle m_2 \rangle$. Width of the probability density function reflects certainty of knowledge: (A) certain; (B) uncertain. *MatLab* script gda05_05.

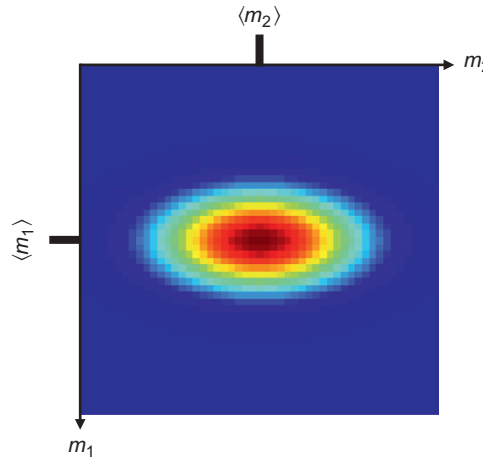

FIGURE 5.6 *A priori* information about model parameters m_1 and m_2 represented with a probability density function $p(m_1,m_2)$. The model parameters are thought to be near $\langle \mathbf{m} \rangle$, with the uncertainty in m_1 less than the uncertainty of m_2. *MatLab* script gda05_06.

the different model parameters (Figure 5.6). The general Gaussian case, with covariance $[\text{cov } \mathbf{m}]_A$, is

$$p_A(\mathbf{m}) \propto \exp\left[-\frac{1}{2}(\mathbf{m}-\langle \mathbf{m} \rangle)^T [\text{cov } \mathbf{m}]_A^{-1} (\mathbf{m}-\langle \mathbf{m} \rangle)\right] \tag{5.8}$$

Equality constraints can be implemented with a distribution that contains a ridge (Figure 5.7). This distribution is non-Gaussian but might be approximated by a Gaussian distribution with nonzero covariance if the expected range of the model parameters were small. Inequality constraints can also be represented by an *a priori* distribution but are inherently non-Gaussian (Figure 5.8).

Similarly, one can summarize the state of knowledge about the measurements with an *a priori* probability density function $p_A(\mathbf{d})$. It simply summarizes the observations, so its mean is $\mathbf{d}^{obs}$ and its covariance is the *a priori* covariance $[\text{cov } \mathbf{d}]$ of the data

$$p_A(\mathbf{d}) \propto \exp\left[-\frac{1}{2}(\mathbf{d}-\mathbf{d}^{obs})^T [\text{cov } \mathbf{d}]^{-1} (\mathbf{d}-\mathbf{d}^{obs})\right] \tag{5.9}$$

One important attribute of $p_A(\mathbf{m})$ and $p_A(\mathbf{d})$ is that they contain information about the model parameters $\mathbf{m}$ and the data $\mathbf{d}$, respectively. The amount of information can be quantified by the *information gain*, a scalar number S defined as

$$\begin{aligned} S[p_A(\mathbf{m})] &= \int p_A(\mathbf{m}) \log\left[\frac{p_A(\mathbf{m})}{p_N(\mathbf{m})}\right] d^M\mathbf{m} \\ S[p_A(\mathbf{d})] &= \int p_A(\mathbf{d}) \log\left[\frac{p_A(\mathbf{d})}{p_N(\mathbf{d})}\right] d^N\mathbf{d} \end{aligned} \tag{5.10}$$

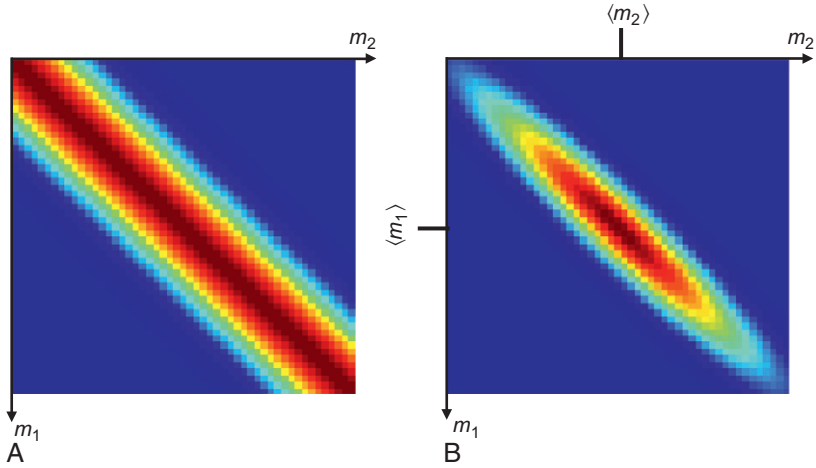

FIGURE 5.7 *A priori* information about model parameters m_1 and m_2, represented with a probability density function $p(m_1,m_2)$. (A) Case when the values of m_1 and m_2 are unknown, but believed to be correlated. (B) Approximation of (A) with a Gaussian probability density function with finite variance. *MatLab* script gda05_07.

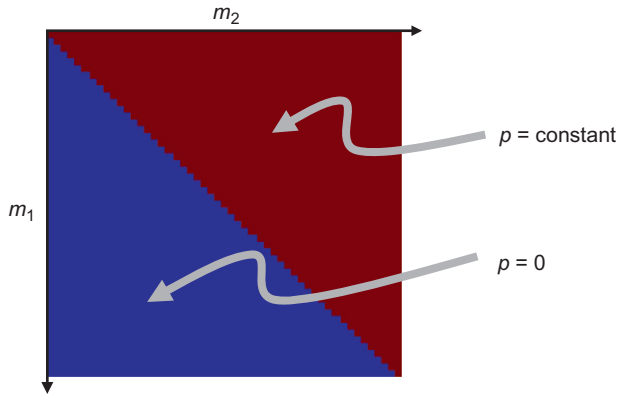

FIGURE 5.8 *A priori* information about model parameters m_1 and m_2 represented with a probability density function $p(m_1,m_2)$. The value of the model parameters are unknown, but the relationship $m_1 \leq m_2$ is believed to hold exactly. This is a non-Gaussian probability density function. *MatLab* script gda05_08.

Here, the *null* probability density functions $p_N(\mathbf{m})$ and $p_N(\mathbf{d})$ express the state of complete ignorance about the model parameters and data, respectively. When the range of **m** and **d** are bounded, the null probability density functions can be taken to be proportional to a constant; that is, $p_N(\mathbf{m}) \propto$ constant and $p_N(\mathbf{d}) \propto$ constant, meaning **m** and **d** "could be anything." However, when **m** and **d** are unbounded, the uniform distribution does not exist, and some other probability density function, such as a very wide Gaussian, must be used, instead.

The quantity $-S$ is sometimes called the *relative entropy* between the two probability density functions. A wide distribution is "more random" than a narrow one; it has more *entropy*.

The information gain is always a nonnegative number and is only zero when $p_A(\mathbf{m}) = p_N(\mathbf{m})$ and $p_A(\mathbf{d}) = p_N(\mathbf{d})$ (Figure 5.9). The information gain S has the following properties (Tarantola and Valette, 1982b): (1) the information gain of the null distribution is zero; (2) all distributions except the null distribution have positive information gain; (3) the more sharply peaked the probability density function becomes, the more its information gain increases; and (4) the information gain is invariant under reparameterizations.

We can summarize the state of knowledge about the inverse problem *before* it is solved by first defining an *a priori* probability density function for the data $p_A(\mathbf{d})$ and then combining it with the *a priori* probability density function for the model $p_A(\mathbf{m})$. The *a priori* data probability density function simply summarizes the observations, so its mean is $\mathbf{d}^{obs}$ and its variance is equal to the *a priori* variance of the data. Since the *a priori* model probability density function is completely independent of the actual values of the data, we can form the joint *a priori* probability density function simply by multiplying the two as

$$p_A(\mathbf{m}, \mathbf{d}) = p_A(\mathbf{m}) p_A(\mathbf{d}) \tag{5.11}$$

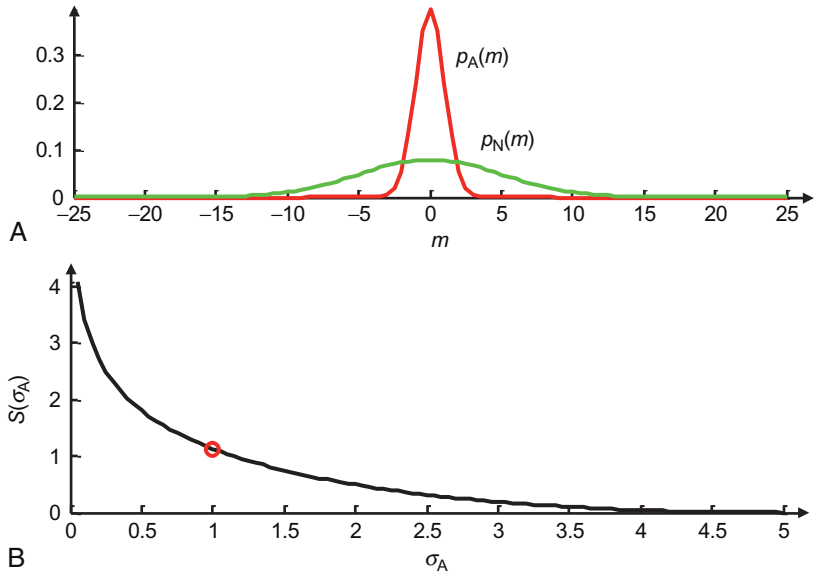

FIGURE 5.9 (A) In this example, a wide Gaussian (green, $\sigma_N = 5$) is used for the null probability density function $p_N(\mathbf{m})$ and a narrow Gaussian (red, $\sigma_N = 1$) is used for the *a priori* probability density function $p_A(\mathbf{m})$. (B) The information gain S decreases as the width of $p_A(\mathbf{m})$ is increased. The case in (A) is depicted with a red circle. *MatLab* script gda05_09.

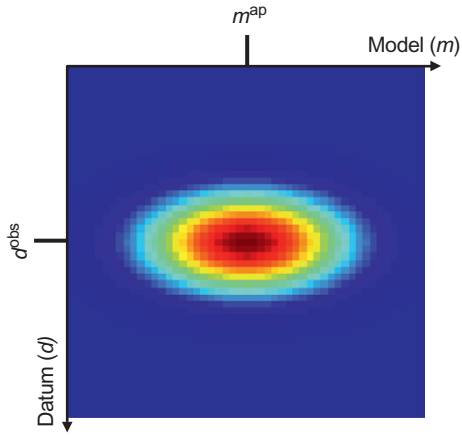

FIGURE 5.10 Joint probability density function $p_A(\mathbf{m}, \mathbf{d})$ for model parameter m and datum d. The distribution is peaked at mean values m^{ap} and d^{obs}. *MatLab* script gda05_10.

This probability density function can be depicted graphically as a "cloud" of probability centered on the observed data and *a priori* model parameters, with a width that reflects the certainty of these quantities (Figure 5.10). If we apply the maximum likelihood method to this distribution, we simply recover the data and *a priori* model. We have not yet applied our knowledge of the model (the relationship between data and model parameters).

5.2.3 Maximum Likelihood for an Exact Theory

Suppose that the model is the rather general equation $\mathbf{g}(\mathbf{m}) = \mathbf{d}$ (which may or may not be linear). This equation defines a surface in the space of model parameters and data along which the solution must lie (Figure 5.11). The maximum likelihood problem then translates into finding the maximum of the joint distribution $p_A(\mathbf{m}, \mathbf{d})$ (or, equivalently, its logarithm) on the surface $\mathbf{d} = \mathbf{g}(\mathbf{m})$ (Tarantola and Valette, B., 1982a):

$$\text{maximize} \quad \log[p(\mathbf{m}, \mathbf{d})] \text{with the constraint } \mathbf{g}(\mathbf{m}) - \mathbf{d} = \mathbf{0} \tag{5.12}$$

Note that if the *a priori* probability density function for the model parameters is much more certain than that of the observed data (that is, if $\sigma_m < \sigma_d$), then the estimate of the model parameters (the maximum likelihood point) tends to be close to the *a priori* model parameters (Figure 5.12). On the other hand, if the data are far more certain than the model parameters (that is, $\sigma_d < \sigma_m$), then the estimates of the model parameters primarily reflect information contained in the data (Figure 5.13).

In the case of Gaussian probability density function, and the linear theory $\mathbf{d} = \mathbf{Gm}$, we need only to substitute $\mathbf{Gm}$ for $\mathbf{d}$ in the expression for $p_A(\mathbf{d})$ to obtain

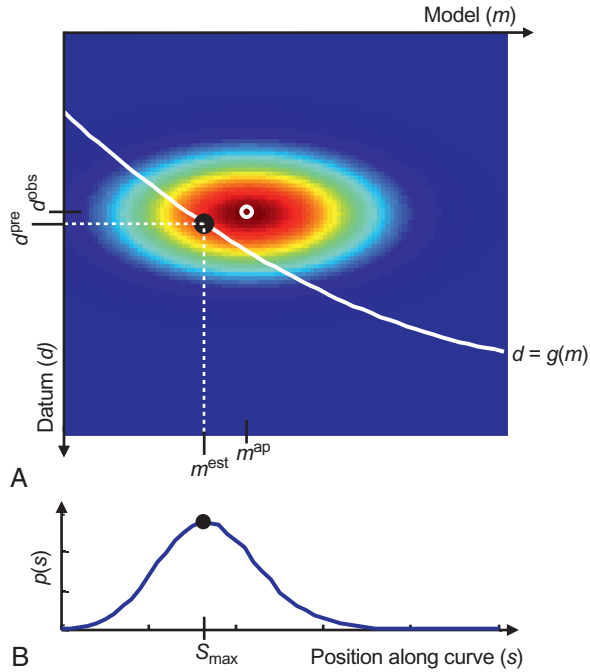

FIGURE 5.11 (A) *A priori* joint probability density function $p(m,d)$ for model parameter m and datum d represents the idea that the model parameter is near its *a priori* value m^{ap} and the datum is near its observed value d^{obs} (white circle). The data and model parameters are believed to be related by an exact theory $d = g(m)$ (white curve). The estimated model parameter m^{est} and predicted datum d^{pre} fall on this curve at the point of maximum probability (black dot). (B) Probability density p evaluated along the curve. The *MatLab* script gda05_11.

$$\begin{aligned} &\text{minimize } \Phi(\mathbf{m}) = L(\mathbf{m}) + E(\mathbf{m}) \text{ with respect to } \mathbf{m} \text{ with} \\ &L(\mathbf{m}) = (\mathbf{m} - \langle \mathbf{m} \rangle)^{\mathrm{T}} [\operatorname{cov} \mathbf{m}]_{\mathrm{A}}^{-1} (\mathbf{m} - \langle \mathbf{m} \rangle) \\ &E(\mathbf{m}) = (\mathbf{G}\mathbf{m} - \mathbf{d}^{\text{obs}})^{\mathrm{T}} [\operatorname{cov} \mathbf{d}]^{-1} (\mathbf{G}\mathbf{m} - \mathbf{d}^{\text{obs}}) \end{aligned} \tag{5.13}$$

Comparison with Section 3.9.3 indicates that this is the weighted damped least squares problem, with

$$\varepsilon^2 \mathbf{W}_m = [\operatorname{cov} \mathbf{m}]_{\mathrm{A}}^{-1} \quad \text{and} \quad \mathbf{W}_e = [\operatorname{cov} \mathbf{d}]^{-1} \tag{5.14}$$

so its solution is the least squares solution of $\mathbf{F}\mathbf{m} = \mathbf{f}$; that is, $\mathbf{F}^{\mathrm{T}}\mathbf{F}\mathbf{m}^{\text{est}} = \mathbf{F}^{\mathrm{T}}\mathbf{f}$ with

$$\mathbf{F} = \begin{bmatrix} [\operatorname{cov} \mathbf{d}]^{-1/2} \mathbf{G} \\ [\operatorname{cov} \mathbf{m}]_{\mathrm{A}}^{-1/2} \mathbf{I} \end{bmatrix} \quad \text{and} \quad \mathbf{f} = \begin{bmatrix} [\operatorname{cov} \mathbf{d}]^{-1/2} \mathbf{d}^{\text{obs}} \\ [\operatorname{cov} \mathbf{m}]_{\mathrm{A}}^{-1/2} \langle \mathbf{m} \rangle \end{bmatrix} \tag{5.15}$$

The matrices $[\operatorname{cov} \mathbf{d}]^{-1/2}$ and $[\operatorname{cov} \mathbf{m}]^{-1/2}$ can be interpreted as the *certainty* of $\mathbf{d}^{\text{obs}}$ and $\langle \mathbf{m} \rangle$, respectively, since they are numerically large when the

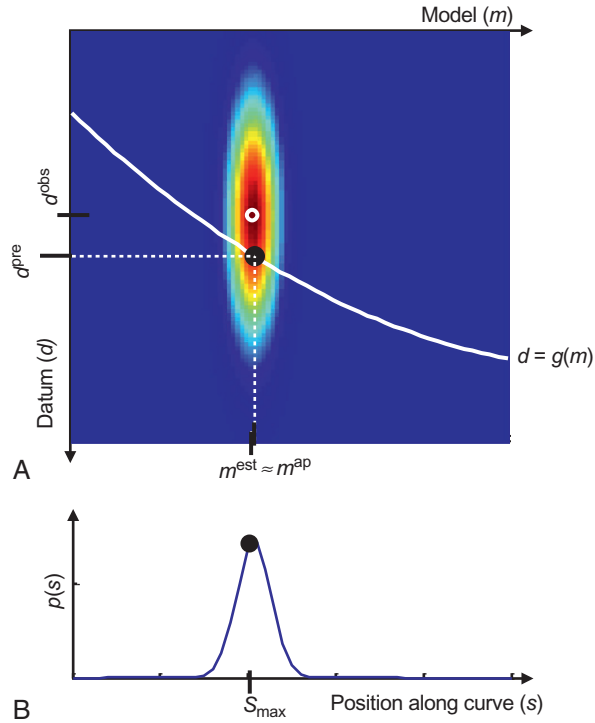

FIGURE 5.12 (A) If the *a priori* model parameter m^{ap} is much more certain than the observed datum d^{obs}, the solution is close to m^{ap} but may be far from d^{obs}. (B) The probability density function p evaluated along the curve. *MatLab* script gda05_12.

uncertainty of these quantities are small. Thus, the top part of the equation $\mathbf{Fm} = \mathbf{f}$ is the data equation $\mathbf{Gm} = \mathbf{d}^{est}$, weighted by its certainty, and the bottom part is the *prior* equation $\mathbf{m} = \langle \mathbf{m} \rangle$, weighted by its certainty. Thus, the matrices $\mathbf{W}_m$ and $\mathbf{W}_e$ of weighted least squares have an important probabilistic interpretation.

The vector $\mathbf{f}$ in Equation (5.15) has unit covariance $[\text{cov}\,\mathbf{f}] = \mathbf{I}$, since its component quantities $[\text{cov}\,\mathbf{d}]^{-1/2}\mathbf{d}^{obs}$ and $[\text{cov}\,\mathbf{m}]_A^{-1/2}\langle \mathbf{m} \rangle$ each have unit covariance (for example, by the usual rules of error propagation, $[\text{cov}\,\mathbf{d}]^{-1/2T}[\text{cov}\,\mathbf{d}][\text{cov}\,\mathbf{d}]^{-1/2} = \mathbf{I}$). Thus, the covariance of the estimated model parameters are

$$[\text{cov}\,\mathbf{m}^{est}] = [\mathbf{F}^T\mathbf{F}]^{-1} \tag{5.16}$$

Somewhat incidentally, we note that, had the *a priori* information involved a linear function $\mathbf{Hm} = \mathbf{h}$ of the model parameters, with covariance $[\text{cov}\,\mathbf{h}]_A$, in contrast to the model parameters themselves, the appropriate form of Equation (5.15) would be

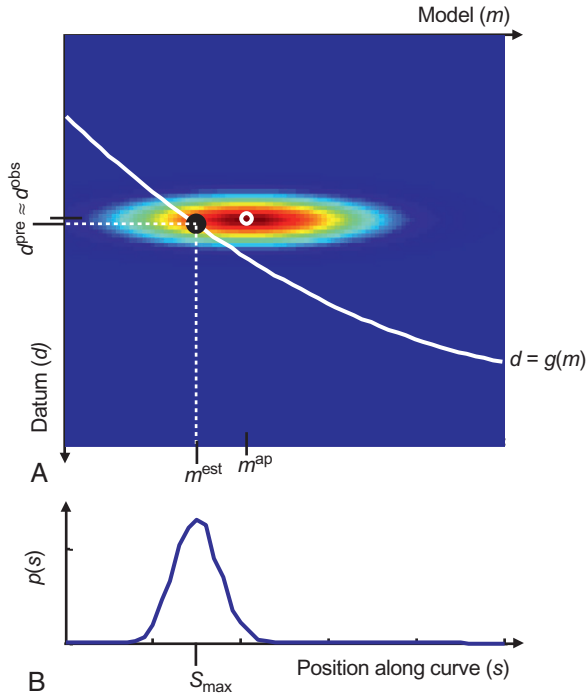

FIGURE 5.13 (A) If the *a priori* model parameter m^{ap} is much less certain than the observed datum d^{obs}, the solution is close to d^{obs} but may be far from m^{ap}. (B) The probability density function p evaluated along the curve. *MatLab* script gda05_13.

$$\mathbf{F} = \begin{bmatrix} [\operatorname{cov} \mathbf{d}]^{-1/2}\mathbf{G} \\ [\operatorname{cov} \mathbf{h}]_A^{-1/2}\mathbf{H} \end{bmatrix} \quad \text{and} \quad \mathbf{f} = \begin{bmatrix} [\operatorname{cov} \mathbf{d}]^{-1/2}\mathbf{d}^{obs} \\ [\operatorname{cov} \mathbf{h}]_A^{-1/2}\mathbf{h} \end{bmatrix} \tag{5.17}$$

This form of weighted damped least squares is especially well suited for computations, especially when $\mathbf{F}^T\mathbf{F}\mathbf{m}^{est} = \mathbf{F}^T\mathbf{f}$ is solved with the biconjugate gradient method.

5.2.4 Inexact Theories

Weighted damped least squares, as it is embodied in Equations (5.15)–(5.17), is extremely useful. Nevertheless, it is somewhat unsatisfying from the standpoint of a probabilistic analysis because the theory has been assumed to be exact. In many realistic problems, there are errors associated with the theory. Some of the assumptions that go into the theory may be unrealistic, or it may be an approximate form of a clumsier but exact theory.

In this case, the model equation $\mathbf{g}(\mathbf{m}) = \mathbf{d}$ can no longer be represented by a simple surface. It has become "fuzzy" because there are now errors associated

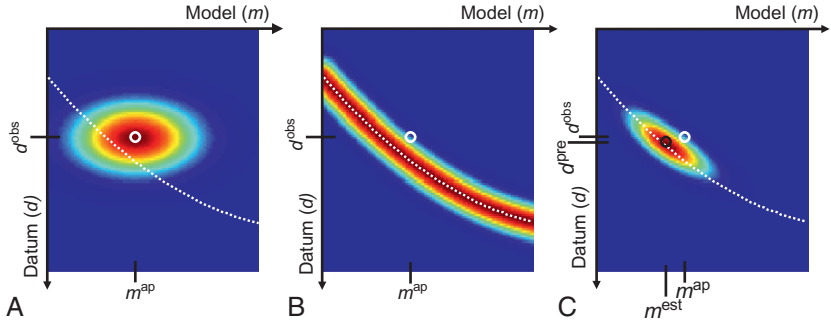

FIGURE 5.14 (A) The *a priori* probability density function $p_A(\mathbf{m}, \mathbf{d})$ represents the state of knowledge before the theory is applied. Its mean (white circle) is the *a priori* model parameter m^{ap} and observed data d^{obs}. (B) An inexact theory is represented by the conditional probability density function $p_g(\mathbf{m}, \mathbf{d})$, which is centered about the exact theory (dotted white curve). (C) The product $p_T(\mathbf{m}, \mathbf{d}) = p_A(\mathbf{m}, \mathbf{d}) p_g(\mathbf{m}, \mathbf{d})$ combines the *a priori* information and theory. Its peak is at the estimated data m^{est} and predicted data d^{pre}. *MatLab* script gda05_14.

with it (Figure 5.14A; Tarantola and Valette, 1982b). Instead of a surface, one might envision a probability density function $p_g(\mathbf{m}, \mathbf{d})$ centered about $\mathbf{g}(\mathbf{m}) = \mathbf{d}$, with width proportional to the uncertainty of the theory. Rather than find the maximum likelihood point of $p_A(\mathbf{m}, \mathbf{d})$ on a surface, we should instead *combine* $p_A(\mathbf{m}, \mathbf{d})$ and $p_g(\mathbf{m}, \mathbf{d})$ into a single distribution and find the maximum likelihood point in the overall volume (Figure 5.13C). To proceed, we need a way of combining two probability density functions, each of which contains information about the data and model parameters.

We have already encountered one special case of a combination in our discussion of Bayesian inference (Section 2.7). After adjusting the variable names to match the current discussion, Bayes' theorem takes the form

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d})} \quad \text{or} \quad p(\mathbf{m}|\mathbf{d}) \propto p(\mathbf{d}|\mathbf{m})\ p(\mathbf{m}) \tag{5.18}$$

Note that the second form omits the denominator, which is not a function of the model parameters and hence acts only as an overall normalization. This second form can be interpreted as updating the information in $p(\mathbf{m})$ (identified now as the *a priori* information) with $p(\mathbf{d}|\mathbf{m})$ (identified now with the data and quantitative model). Thus, in Bayesian inference, probability density functions are combined by multiplication.

In the general case, we denote the process of combining two probability density functions as $p_3 = C(p_1, p_2)$, meaning that functions 1 and 2 are combined into function 3. Then, clearly, the process of combining must have the following properties (adapted from Tarantola and Valette, 1982b):

(a) The order in which two probability density functions are combined should not matter; that is, $C(p_1, p_2)$ should be commutative: $C(p_1, p_2) = C(p_2, p_1)$.

(b) The order in which three probability density functions are combined should not matter; that is, $C(p_1, p_2)$ should be associative: $C(p_1, C(p_2, p_3)) = C(C(p_1, p_2), p_3)$.
(c) Combining a distribution with the null distribution should return the same distribution; that is, $C(p_1, p_N) = p_1$.
(d) The combination $C(p_1, p_2)$ should never be everywhere zero except if p_1 or p_2 is everywhere zero.
(e) The combination $C(p_1, p_2)$ should be invariant under reparameterizations.

These conditions can be shown to be satisfied by the choice (Tarantola and Valette, 1982b):

$$p_3 = C(p_1, p_2) = \frac{p_1 p_2}{p_N} \tag{5.19}$$

(at least up to an overall normalization). Note that if the null distribution is constant (as we shall assume for the rest of this chapter), one combines distributions simply by multiplying them:

$$p_T(\mathbf{m}, \mathbf{d}) = p_A(\mathbf{m}, \mathbf{d})\, p_g(\mathbf{m}, \mathbf{d}) \tag{5.20}$$

Here, the subscript T means the combined or total distribution. Note that as the error associated with the theory increases, the maximum likelihood point moves back toward the *a priori* values of model parameters and observed data (Figure 5.15). The limiting case in which the theory is infinitely accurate is equivalent to the case in which the distribution is replaced by a distinct surface.

The maximum likelihood point of $p_T(\mathbf{m}, \mathbf{d})$ is specified by both a set of model parameters $\mathbf{m}^{est}$ and a set of data $\mathbf{d}^{pre}$. They are determined simultaneously. This approach is different from that of the least squares problem examined in Section 5.2. In that section, we maximized the probability density function with respect to the model parameters only to determine the most probable model parameters $\mathbf{m}^{est}$ and afterward can calculate the predicted data via $\mathbf{d}^{pre} = \mathbf{G}\mathbf{m}^{est}$. The two methods do not necessarily yield the same estimates for the model parameters. To find the likelihood point of $p_T(\mathbf{m}, \mathbf{d})$ with respect to model parameters only, we must sum all the probabilities along lines of equal model parameter. This summation can be thought of as projecting the distribution onto the $\mathbf{d} = 0$ plane (Figure 5.16) and then finding the maximum. The projected distribution $p(\mathbf{m})$ is then

$$p(\mathbf{m}) = \int p_T(\mathbf{m}, \mathbf{d}) \mathrm{d}^N d \tag{5.21}$$

where the integration is performed over the entire range of the ds.

5.2.5 The Simple Gaussian Case with a Linear Theory

To illustrate this method, we rederive the weighted damped least squares solution, with *a priori* probability density function:

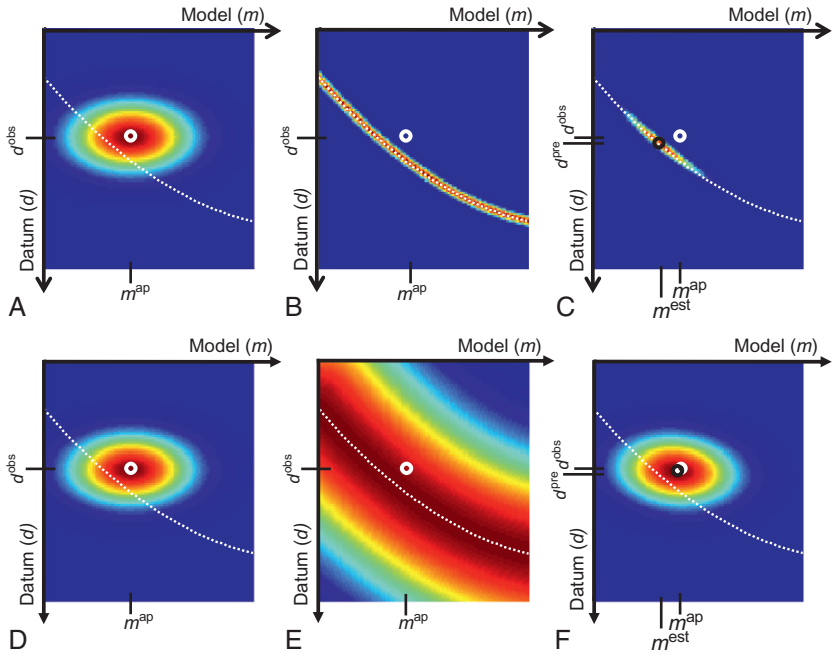

FIGURE 5.15 The rows of the figure have the same format as Figure 5.14. If the theory is made more and more inexact (compare (A–C) with (D–F)), the solution (black circle) moves toward the maximum likelihood point of the *a priori* distribution. *MatLab* script gda05_15.

$$p_A(\mathbf{m},\mathbf{d}) \propto \exp\left[-\frac{1}{2}(\mathbf{m}-\langle\mathbf{m}\rangle)^T[\operatorname{cov}\mathbf{m}]_A^{-1}(\mathbf{m}-\langle\mathbf{m}\rangle)-\frac{1}{2}(\mathbf{d}-\mathbf{d}^{obs})^T[\operatorname{cov}\mathbf{d}]^{-1}(\mathbf{d}-\mathbf{d}^{obs})\right] \quad (5.22)$$

If there are no errors in the theory, then its probability density function is "infinitely narrow" and can be represented by a *Dirac delta function*

$$p_A(\mathbf{m},\mathbf{d}) = \delta(\mathbf{Gm}-\mathbf{d}) \quad (5.23)$$

where we assume that we are dealing with the linear theory $\mathbf{Gm}=\mathbf{d}$. The total distribution is then given by

$$p_T(\mathbf{m},\mathbf{d}) = p_A(\mathbf{m},\mathbf{d})\delta(\mathbf{Gm}-\mathbf{d}) \quad (5.24)$$

Performing the projection "integrates away" the delta function

$$p(\mathbf{m}) = \int p_A(\mathbf{m},\mathbf{d})\delta(\mathbf{Gm}-\mathbf{d})\mathrm{d}^N\mathbf{d} \propto \exp\left[-\frac{1}{2}(\mathbf{m}-\langle\mathbf{m}\rangle)^T[\operatorname{cov}\mathbf{m}]_A^{-1}(\mathbf{m}-\langle\mathbf{m}\rangle) -\frac{1}{2}(\mathbf{Gm}-\mathbf{d}^{obs})^T[\operatorname{cov}\mathbf{d}]^{-1}(\mathbf{Gm}-\mathbf{d}^{obs})\right] \quad (5.25)$$

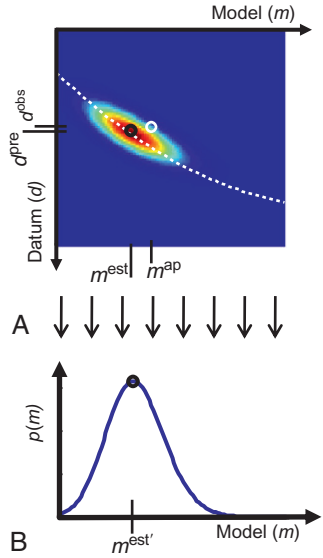

FIGURE 5.16 (A) The joint probability density function $p_T(\mathbf{m}, \mathbf{d})$ can be considered the solution to the inverse problem. Its maximum likelihood point (black circle) gives an estimate of the model parameter m^{est} and a prediction of the data d^{pre}. (B) The function $p_T(\boldsymbol{m}, \mathbf{d})$ is projected onto the m-axis, by integrating over d, to form the probability density function $p(m)$ of the model parameter irrespective of the datum. This function also has a maximum likelihood point $\mathbf{m}^{est'}$ which in general can be different than m^{est}. The distinction points out the difficulty of defining a unique "solution" to an inverse problem. *MatLab* script gda05_16.

This projected distribution is exactly the one we encountered in the weighted damped least squares problem (the logarithm of which is shown in Equation (5.13)).

5.2.6 The General Linear, Gaussian Case

In the general linear, Gaussian case, we assume that all the component probability density functions are Gaussian and that the theory is the linear equation $\mathbf{Gm} = \mathbf{d}$, so that

$$\begin{aligned} p_A(\mathbf{m}) &\propto \exp\left[-\frac{1}{2}(\mathbf{m}-\langle\mathbf{m}\rangle)^T[\operatorname{cov}\mathbf{m}]_A^{-1}(\mathbf{m}-\langle\mathbf{m}\rangle)\right] \\ p_A(\mathbf{d}) &\propto \exp\left[-\frac{1}{2}(\mathbf{d}-\mathbf{d}^{obs})^T[\operatorname{cov}\mathbf{d}]^{-1}(\mathbf{d}-\mathbf{d}^{obs})\right] \\ p_g(\mathbf{m},\mathbf{d}) &\propto \exp\left[-\frac{1}{2}(\mathbf{d}-\mathbf{Gm})^T[\operatorname{cov}\mathbf{g}]^{-1}(\mathbf{d}-\mathbf{Gm})\right] \end{aligned} \tag{5.26}$$

Here, the theory is represented by a Gaussian probability density function with covariance [cov **g**]. The total distribution $p_T(\mathbf{m}, \mathbf{d})$ is the product of these three distributions. As we noted in Section 2.4, products of Gaussian probability density functions are themselves Gaussian, so $p_T(\mathbf{m}, \mathbf{d})$ is Gaussian. We need to determine the mean of this distribution, since it is also the maximum likelihood point that defines $\mathbf{m}^{est}$ and $\mathbf{d}^{pre}$.

To simplify the algebra, we first define a vector $\mathbf{x} = [\mathbf{d}^T, \mathbf{m}^T]^T$ that contains the data and model parameters, a vector $\langle \mathbf{x} \rangle = [\mathbf{d}^{obsT}, \langle \mathbf{m} \rangle^T]^T$ that contains their *a priori* values and a covariance matrix

$$[\text{cov } \mathbf{x}] = \begin{bmatrix} [\text{cov } \mathbf{d}] & 0 \\ 0 & [\text{cov } \mathbf{m}]_A \end{bmatrix} \tag{5.27}$$

The first two products in the total distribution can then be combined into an exponential, with the argument given by

$$-\frac{1}{2}(\mathbf{x} - \langle \mathbf{x} \rangle)^T [\text{cov } \mathbf{x}]^{-1} (\mathbf{x} - \langle \mathbf{x} \rangle) \tag{5.28}$$

To express the third product in terms of $\mathbf{x}$, we define a matrix $\mathbf{F} = [\mathbf{I}, -\mathbf{G}]$ such that $\mathbf{Fx} = \mathbf{d} - \mathbf{Gm} = 0$. The argument of the third product's exponential is then given by

$$-\frac{1}{2}(\mathbf{Fx})^T [\text{cov } \mathbf{g}]^{-1} (\mathbf{Fx}) \tag{5.29}$$

The total distribution is proportional to an exponential with argument

$$\begin{aligned} &-\frac{1}{2}(\mathbf{x} - \langle \mathbf{x} \rangle)^T [\text{cov } \mathbf{x}]^{-1} (\mathbf{x} - \langle \mathbf{x} \rangle) - \frac{1}{2}(\mathbf{Fx})^T [\text{cov } \mathbf{g}]^{-1} (\mathbf{Fx}) = \\ &-\frac{1}{2}\mathbf{x}^T ([\text{cov } \mathbf{x}]^{-1} + \mathbf{F}^T [\text{cov } \mathbf{g}]^{-1} \mathbf{F}) \mathbf{x} - \mathbf{x}^T [\text{cov } \mathbf{x}]^{-1} \langle \mathbf{x} \rangle - \frac{1}{2} \langle \mathbf{x} \rangle^T [\text{cov } \mathbf{x}]^{-1} \langle \mathbf{x} \rangle \end{aligned} \tag{5.30}$$

We would like to manipulate this expression into the standard form of the argument of a Gaussian probability density function; that is, an expression involving just a single vector, say $\mathbf{x}^*$, and a single covariance matrix, say [cov $\mathbf{x}^*$], related by

$$\begin{aligned} &-\frac{1}{2}(\mathbf{x} - \mathbf{x}^*)^T [\text{cov} \mathbf{x}^*]^{-1} (\mathbf{x} - \mathbf{x}^*) = \\ &-\frac{1}{2}\mathbf{x}^T [\text{cov } \mathbf{x}^*]^{-1} \mathbf{x} + \mathbf{x}^T [\text{cov} \mathbf{x}^*]^{-1} \mathbf{x}^* - \frac{1}{2}\mathbf{x}^{*T} [\text{cov} \mathbf{x}^*]^{-1} \mathbf{x}^* \end{aligned} \tag{5.31}$$

We can identify $\mathbf{x}^*$ and [cov $\mathbf{x}^*$] by matching terms of equal powers of $\mathbf{x}$ with Equation (5.30). Matching the quadratic terms yields

$$[\text{cov } \mathbf{x}^*]^{-1} = [\text{cov } \mathbf{x}]^{-1} + \mathbf{F}^T [\text{cov } \mathbf{g}]^{-1} \mathbf{F} \tag{5.32}$$

and matching the linear terms yields

$$[\operatorname{cov}\mathbf{x}^*]^{-1}\mathbf{x}^* = [\operatorname{cov}\mathbf{x}]^{-1}\langle\mathbf{x}\rangle \quad \text{or} \quad \mathbf{x}^* = [\operatorname{cov}\mathbf{x}^*][\operatorname{cov}\mathbf{x}]^{-1}\langle\mathbf{x}\rangle \tag{5.33}$$

These choices do not match the constant term, but such a match is not necessary, because the constant term effects only the overall normalization of the probability density function $p_T(\mathbf{x})$. The vector $\mathbf{x}^*$ corresponds to the maximum likelihood point of $p_T(\mathbf{x})$ and so can be considered the solution to the inverse problem. This solution has covariance [cov $\mathbf{x}^*$].

Equation (5.32) for $[\operatorname{cov}\mathbf{x}^*]^{-1}$ can be explicitly inverted to yield an expression for [cov $\mathbf{x}^*$], using two matrix identities that we now derive (adapted from Tarantola and Valette, 1982a, with permission). Let $\mathbf{C}_1$ and $\mathbf{C}_2$ be two symmetric matrices whose inverses exist, and let $\mathbf{M}$ be a third matrix. The expression $\mathbf{M}^T + \mathbf{M}^T\mathbf{C}_1^{-1}\mathbf{M}\mathbf{C}_2\mathbf{M}^T$ can be written in two ways by grouping terms as $\mathbf{M}^T\mathbf{C}_1^{-1}[\mathbf{C}_1 + \mathbf{M}\mathbf{C}_2\mathbf{M}^T]$ or $[\mathbf{C}_2^{-1} + \mathbf{M}^T\mathbf{C}_1^{-1}\mathbf{M}]\mathbf{C}_2\mathbf{M}^T$. Multiplying by the matrix inverses gives

$$\mathbf{C}_2\mathbf{M}^T[\mathbf{C}_1 + \mathbf{M}\mathbf{C}_2\mathbf{M}^T]^{-1} = [\mathbf{C}_2^{-1} + \mathbf{M}^T\mathbf{C}_1^{-1}\mathbf{M}]^{-1}\mathbf{M}^T\mathbf{C}_1^{-1}. \tag{5.34}$$

Now consider the symmetric matrix expression $\mathbf{C}_2 - \mathbf{C}_2\mathbf{M}^T[\mathbf{C}_1 + \mathbf{M}\mathbf{C}_2\mathbf{M}^T]^{-1}\mathbf{M}\mathbf{C}_2$. By Equation (5.34), this expression equals $\mathbf{C}_2 - [\mathbf{C}_2^{-1} + \mathbf{M}^T\mathbf{C}_1^{-1}\mathbf{M}]^{-1}\mathbf{M}^T\mathbf{C}_1^{-1}\mathbf{M}\mathbf{C}_2$. Factoring out the term in brackets gives $[\mathbf{C}_2^{-1} + \mathbf{M}^T\mathbf{C}_1^{-1}\mathbf{M}]^{-1}\{[\mathbf{C}_2^{-1} + \mathbf{M}^T\mathbf{C}_1^{-1}\mathbf{M}]\mathbf{C}_2 - \mathbf{M}^T\mathbf{C}_1^{-1}\mathbf{M}\mathbf{C}_2\}$. Canceling terms gives $[\mathbf{C}_2^{-1} + \mathbf{M}^T\mathbf{C}_1^{-1}\mathbf{M}]^{-1}$ from which we conclude

$$\begin{aligned}[\mathbf{C}_2^{-1} + \mathbf{M}^T\mathbf{C}_1^{-1}\mathbf{M}]^{-1} &= \mathbf{C}_2 - \mathbf{C}_2\mathbf{M}^T[\mathbf{C}_1 + \mathbf{M}\mathbf{C}_2\mathbf{M}^T]^{-1}\mathbf{M}\mathbf{C}_2 \\ &= \{\mathbf{I} - \mathbf{C}_2\mathbf{M}^T[\mathbf{C}_1 + \mathbf{M}\mathbf{C}_2\mathbf{M}^T]^{-1}\mathbf{M}\}\mathbf{C}_2\end{aligned} \tag{5.35}$$

Equating now $\mathbf{C}_2 = [\operatorname{cov}\mathbf{x}]$, $\mathbf{C}_1 = [\operatorname{cov}\mathbf{g}]$, and $\mathbf{M} = \mathbf{F}$, Equation (5.32) becomes

$$\begin{aligned}[\operatorname{cov}\mathbf{x}^*] = [[\operatorname{cov}\mathbf{x}]^{-1} + \mathbf{F}^T[\operatorname{cov}\mathbf{g}]^{-1}\mathbf{F}]^{-1} = \\ \{\mathbf{I} - [\operatorname{cov}\mathbf{x}]\mathbf{F}^T[[\operatorname{cov}\mathbf{g}] + \mathbf{F}[\operatorname{cov}\mathbf{x}]\mathbf{F}^T]^{-1}\mathbf{F}\}[\operatorname{cov}\mathbf{x}]\end{aligned} \tag{5.36}$$

and Equation (5.33) becomes

$$\mathbf{x}^* = [\operatorname{cov}\mathbf{x}^*][\operatorname{cov}\mathbf{x}]^{-1}\langle\mathbf{x}\rangle = \{\mathbf{I} - [\operatorname{cov}\mathbf{x}]\mathbf{F}^T[[\operatorname{cov}\mathbf{g}] + \mathbf{F}[\operatorname{cov}\mathbf{x}]\mathbf{F}^T]^{-1}\mathbf{F}\}\langle\mathbf{x}\rangle \tag{5.37}$$

Nothing in this derivation requires the special forms of $\mathbf{F}$ and [cov $\mathbf{x}$] assumed above that made $\mathbf{F}\mathbf{x} = 0$ separable into an *explicit* linear inverse problem. Equation (5.37) is in fact the solution to the completely general, *implicit*, linear inverse problem.

When $\mathbf{F}\mathbf{x} = 0$ is an explicit equation, the formula for $\mathbf{x}^*$ in Equation (5.21) can be decomposed into its component vectors $\mathbf{d}^{pre}$ and $\mathbf{m}^{est}$ by substituting the definition of $\mathbf{F}$ and [cov $\mathbf{x}$] (Equation (5.27)) into it and performing the matrix multiplications. An explicit formula for the estimated model parameters is then given by

$$\begin{aligned} \mathbf{m}^{\text{est}} &= \langle \mathbf{m} \rangle + \mathbf{G}^{-g}(\mathbf{d}^{\text{obs}} - \mathbf{G}\langle \mathbf{m} \rangle) = \mathbf{G}^{-g}\mathbf{d}^{\text{obs}} + [\mathbf{I} - \mathbf{R}]\langle \mathbf{m} \rangle \\ \text{with} \quad \mathbf{G}^{-g} &= [\text{cov}\,\mathbf{m}]_{\text{A}}\mathbf{G}^{\text{T}}\{[\text{cov}\,\mathbf{d}] + [\text{cov}\,\mathbf{g}] + \mathbf{G}[\text{cov}\,\mathbf{m}]_{\text{A}}\mathbf{G}^{\text{T}}\}^{-1} \end{aligned} \tag{5.38}$$

where we have used the generalized inverse $\mathbf{G}^{-g}$ and resolution matrix $\mathbf{R} = \mathbf{G}^{-g}\mathbf{G}$ notation for convenience. This generalized inverse is reminiscent of minimum-length inverse $\mathbf{G}^{\text{T}}[\mathbf{G}\mathbf{G}^{\text{T}}]^{-1}$. However, by equating $\mathbf{C}_2 = [\text{cov}\,\mathbf{m}]_{\text{A}}$, $\mathbf{C}_1 = [\text{cov}\,\mathbf{d}] + [\text{cov}\,\mathbf{g}]$, and $\mathbf{M} = \mathbf{G}$ in matrix identity Equation (5.34), we can also write the generalized inverse as

$$\mathbf{G}^{-g} = \{\mathbf{G}^{\text{T}}([\text{cov}\,\mathbf{d}] + [\text{cov}\,\mathbf{g}])^{-1}\mathbf{G} + [\text{cov}\,\mathbf{m}]_{\text{A}}^{-1}\}^{-1}\mathbf{G}^{\text{T}}([\text{cov}\,\mathbf{d}] + [\text{cov}\,\mathbf{g}])^{-1} \tag{5.39}$$

which is reminiscent of least squares generalized inverse $[\mathbf{G}^{\text{T}}\mathbf{G}]^{-1}\mathbf{G}^{\text{T}}$. Both forms of the generalized inverse depend only upon the sum of the covariance of the data $[\text{cov}\,\mathbf{d}]$ and the covariance of the theory $[\text{cov}\,\mathbf{g}]$; that is, they make only a combined contribution. This is the most important insight gained from this problem, for the exact-theory case (Equation (5.15)) can be made identical to the inexact-theory case (Equations (5.38) and (5.39)) with the substitution

$$[\text{cov}\,\mathbf{d}] \rightarrow [\text{cov}\,\mathbf{d}] + [\text{cov}\,\mathbf{g}] \tag{5.40}$$

Hence, from the point of view of computations, one continues to use Equation (5.15) but adjusts the covariance using Equation (5.40).

Since the estimated model parameters are a linear combination of observed data and *a priori* model parameters, we can therefore calculate its covariance as

$$[\text{cov}\,\mathbf{m}^{\text{est}}] = \mathbf{G}^{-g}[\text{cov}\,\mathbf{d}]\mathbf{G}^{-g\text{T}} + [\mathbf{I} - \mathbf{R}][\text{cov}\,\mathbf{m}]_{\text{A}}[\mathbf{I} - \mathbf{R}]^{\text{T}} \tag{5.41}$$

This expression differs from those derived in Chapters 3 and 4 in that it contains a term dependent on the *a priori* model parameter covariance $[\text{cov}\,\mathbf{m}]_{\text{A}}$.

We can examine a few interesting limiting cases of problems which have uncorrelated *a priori* model parameters ($[\text{cov}\,\mathbf{m}]_{\text{A}} = \sigma_m^2\mathbf{I}$), data ($[\text{cov}\,\mathbf{d}]_{\text{A}} = \sigma_d^2\mathbf{I}$), and theory ($[\text{cov}\,\mathbf{g}] = \sigma_g^2\mathbf{I}$).

5.2.7 Exact Data and Theory

Suppose $\sigma_d^2 = \sigma_g^2 = 0$. The solution is then given by

$$\mathbf{m}^{\text{est}} = \mathbf{G}^{\text{T}}[\mathbf{G}\mathbf{G}^{\text{T}}]^{-1}\mathbf{d}^{\text{obs}} = [\mathbf{G}^{\text{T}}\mathbf{G}]^{-1}\mathbf{G}^{\text{T}}\mathbf{d}^{\text{obs}} \tag{5.42}$$

Note that the solution does not depend on the *a priori* model variance, since the data and theory are infinitely more accurate than the *a priori* model parameters. These solutions are just the minimum-length and least squares solutions, which (as we now see) are simply two different aspects of the same solution. The minimum-length form of the solution, however, exists only when the problem

is purely underdetermined; the least squares form exists only when the problem is purely overdetermined.

If the *a priori* model parameters are not equal to zero, then another term appears in the estimated solution:

$$\begin{aligned}\mathbf{m}^{\text{est}} &= \mathbf{G}^{-g}\mathbf{d}^{\text{obs}} + (\mathbf{I}-\mathbf{R})\langle\mathbf{m}\rangle \\ &= \mathbf{G}^{\text{T}}[\mathbf{G}\mathbf{G}^{\text{T}}]^{-1}\mathbf{d}^{\text{obs}} + \{\mathbf{I}-\mathbf{G}^{\text{T}}[\mathbf{G}\mathbf{G}^{\text{T}}]^{-1}\mathbf{G}\}\langle\mathbf{m}\rangle \\ &= [\mathbf{G}^{\text{T}}\mathbf{G}]^{-1}\mathbf{G}^{\text{T}}\mathbf{d}^{\text{obs}}\end{aligned} \tag{5.43}$$

The minimum-length-type solution has been changed by adding a weighted amount of the *a priori* model vector, with the weighting factor being $\{\mathbf{I}-\mathbf{G}^{\text{T}}[\mathbf{G}\mathbf{G}^{\text{T}}]^{-1}\mathbf{G}\}$. This term is not zero, since it can also be written as $\{\mathbf{I}-\mathbf{R}\}$. The resolution matrix of the underdetermined problem never equals the identity matrix. On the other hand, the resolution matrix of the overdetermined least squares problem does equal the identity matrix, so the estimated model parameters of the overdetermined problem are not a function of the *a priori* model parameters. Adding *a priori* information with finite error to an inverse problem that features exact data and theory only affects the underdetermined part of the solution.

5.2.8 Infinitely Inexact Data and Theory

In the case of infinitely inexact data and theory, we take the $\sigma_d^2 \to \infty$ or $\sigma_g^2 \to \infty$ (or both). The solution becomes

$$\mathbf{m}^{\text{est}} = \langle\mathbf{m}\rangle \tag{5.44}$$

Since the data and theory contain no information, we simply recover the *a priori* model parameters.

5.2.9 No *A Priori* Knowledge of the Model Parameters

In this case, the limit is $\sigma_m^2 \to \infty$. The solutions are the same as in Section 5.2.7:

$$\mathbf{m}^{\text{est}} = \mathbf{G}^{\text{T}}[\mathbf{G}\mathbf{G}^{\text{T}}]^{-1}\mathbf{d}^{\text{obs}} + \{\mathbf{I}-\mathbf{G}^{\text{T}}[\mathbf{G}\mathbf{G}^{\text{T}}]^{-1}\mathbf{G}\}\langle\mathbf{m}\rangle = [\mathbf{G}^{\text{T}}\mathbf{G}]^{-1}\mathbf{G}^{\text{T}}\mathbf{d}^{\text{obs}} \tag{5.45}$$

Infinitely weak *a priori* information and finite-error data and theory produce the same results as finite-error *a priori* information and error-free data and theory.

5.3 RELATIVE ENTROPY AS A GUIDING PRINCIPLE

In Section 5.2.2, we introduced the information gain, S (Equation (5.10)), as a way of quantifying the difference in information content between two probability density functions. It can be used as a guiding principle for constructing solutions to inverse problems. The idea is to find the probability density function

$p_T(\mathbf{m})$—the solution to the inverse problem—that minimizes the information gain of $p_T(\mathbf{m})$ relative to the *a priori* probability density function $p_A(\mathbf{m})$. Thus, as little information as possible has been added to the *a priori* information to create the solution. The quantity $-S$ is the relative entropy of the two probability density functions, so this method is called the *Maximum Relative Entropy* method and is often abbreviated MRE (Kapur, 1989). Some authors define the entropy as $+S$, in which case it is called the *Minimum Relative Entropy* method (also abbreviated MRE).

Constraints need to be added to the minimization of S or else the solution would simply be $p_T(\mathbf{m}) = p_A(\mathbf{m})$. One of these constraints must be that the area beneath $p_T(\mathbf{m})$ is unity. The choice of the other constraints depends on the particular type of inverse problem; that is, whether it is under- or overdetermined.

As an example, consider the underdetermined problem, where the equation $\mathbf{d} = \mathbf{Gm}$ can be assumed to hold in the mean. The minimization problem is

$$\begin{aligned} &\text{minimize: } S = \int p_T(\mathbf{m}) \log\left(\frac{p_T(\mathbf{m})}{p_A(\mathbf{m})}\right) \mathrm{d}^M m \text{ with constraints} \\ &\int p_T(\mathbf{m}) \mathrm{d}^M m = 1 \quad \text{and} \quad \int p_T(\mathbf{m})(\mathbf{d} - \mathbf{Gm}) \mathrm{d}^M m = 0 \end{aligned} \tag{5.46}$$

Here, the final distribution $p_T(\mathbf{m})$ is unknown and the *a priori* distribution $p_A(\mathbf{m})$ is prescribed. Note that the second constraint indicates that the mean (expected) value of the error $\mathbf{e} = \mathbf{d} - \mathbf{Gm}$ is zero.

This minimization problem can be solved by using the *Euler-Lagrange method*. It states that the integral $\int F[f(\mathbf{m}), \mathbf{m}] \mathrm{d}^M m$ is minimized subject to the integral constraint $\int G[f(\mathbf{m}), \mathbf{m}] \mathrm{d}^M m$ when $\Phi = F + \lambda G$ is minimized with respect to f. Here, λ is a Lagrange multiplier. In our case, we introduce one Lagrange multiplier λ_0 associated with the first constraint and a vector $\boldsymbol{\lambda}$ of Lagrange multipliers associated with the second

$$\Phi(\mathbf{m}) = p_T \log(p_T) - p_T \log(p_A) + \lambda_0 p_T + \boldsymbol{\lambda}^T(\mathbf{d} - \mathbf{Gm}) p_T \tag{5.47}$$

Differentiating with respect to p_T yields

$$\frac{\partial \Phi}{\partial p_T} = 0 = \log(p_T) + 1 - \log(p_A) + \lambda_0 + \boldsymbol{\lambda}^T(\mathbf{d} - \mathbf{Gm}) \tag{5.48}$$

or

$$p_T(\mathbf{m}) = p_A(\mathbf{m}) \exp\{-(1 + \lambda_0) - \boldsymbol{\lambda}^T(\mathbf{d} - \mathbf{Gm})\}. \tag{5.49}$$

Now suppose that the *a priori* probability density function is Gaussian in form, with *a priori* value $\langle \mathbf{m} \rangle$ and *a priori* covariance $[\operatorname{cov} \mathbf{m}]_A$

$$p_A(\mathbf{m}) \propto \exp\left\{-\frac{1}{2}(\mathbf{m} - \langle \mathbf{m} \rangle)^T [\operatorname{cov} \mathbf{m}]_A^{-1} (\mathbf{m} - \langle \mathbf{m} \rangle)\right\} \tag{5.50}$$

Then

$$p_T(\mathbf{m}) \propto \exp\{-A(\mathbf{m})\} \quad \text{with}$$
$$A(\mathbf{m}) = \frac{1}{2}(\mathbf{m} - \langle \mathbf{m} \rangle)^T [\operatorname{cov} \mathbf{m}]_A^{-1} (\mathbf{m} - \langle \mathbf{m} \rangle) - (1 + \lambda_0) - \boldsymbol{\lambda}^T (\mathbf{d} - \mathbf{G}\mathbf{m}) \tag{5.51}$$

We now assert that the best estimate of the model parameter $\mathbf{m}^{\text{est}}$ is the mean of this distribution, which is also its maximum likelihood point. This point occurs where $A(\mathbf{m})$ is minimum:

$$\frac{\partial A}{\partial m_q} = 0 \quad \text{or} \quad 0 = [\operatorname{cov} \mathbf{m}]_A^{-1} (\mathbf{m}^{\text{est}} - \langle \mathbf{m} \rangle) + \mathbf{G}^T \boldsymbol{\lambda} \tag{5.52}$$

Premultiplying by $\mathbf{G}[\operatorname{cov} \mathbf{m}]_A$, substituting in the constraint equation $\mathbf{d} = \mathbf{G}\mathbf{m}$ (which is assumed to hold in the mean) and rearranging yields

$$\boldsymbol{\lambda} = \{\mathbf{G}[\operatorname{cov} \mathbf{m}]_A \mathbf{G}^T\}^{-1} \{\mathbf{d} - \mathbf{G}\langle \mathbf{m} \rangle\}. \tag{5.53}$$

Substituting this expression for $\boldsymbol{\lambda}$ into Equation (5.52) for $\partial A / \partial \mathbf{m}$ yields the solution

$$\mathbf{m}^{\text{est}} - \langle \mathbf{m} \rangle = [\operatorname{cov} \mathbf{m}]_A \mathbf{G}^T \{\mathbf{G}[\operatorname{cov} \mathbf{m}]_A \mathbf{G}^T\}^{-1} \{\mathbf{d} - \mathbf{G}\langle \mathbf{m} \rangle\}. \tag{5.54}$$

Thus, the principle of maximum relative entropy, when applied to the underdetermined problem, yields the weighted minimum-length solution (compare Equations (5.54) with (3.43) when $\mathbf{W}_m^{-1} = [\operatorname{cov} \mathbf{m}]_A$). Many of the other inverse theory solutions that were developed in this chapter using maximum likelihood techniques can also be derived using the MRE principle (Woodbury, 2011).

5.4 EQUIVALENCE OF THE THREE VIEWPOINTS

We can arrive at the same general solution to the linear inverse problem by three distinct routes.

Viewpoint 1. The solution is obtained by minimizing a weighted sum of L_2 prediction error and L_2 solution simplicity

$$\text{Minimize:} \quad \mathbf{e}^T \mathbf{W}_e \mathbf{e} + \varepsilon^2 [\mathbf{m} - \langle \mathbf{m} \rangle]^T \mathbf{W}_m [\mathbf{m} - \langle \mathbf{m} \rangle] \tag{5.55}$$

where ε^2 is a weighting factor.

Viewpoint 2. The solution is obtained by minimizing a weighted sum of three terms: the Dirichlet spreads of model resolution and data resolution and the size of the model covariance.

$$\text{Minimize:} \quad \alpha_1 \ \text{spread}(\mathbf{R}) + \alpha_2 \ \text{spread}(\mathbf{N}) + \alpha_3 \ \text{size}([\operatorname{cov}_u \mathbf{m}]) \tag{5.56}$$

Viewpoint 3. The solution is obtained by maximizing the likelihood of the joint Gaussian distribution of data, *a priori* model parameters, and theory.

$$\text{Maximize:} \quad L = \log p_{\mathrm{T}}(\mathbf{m}, \mathbf{d}) \tag{5.57}$$

These derivations emphasize the close relationship among the L_2 norm, the Dirichlet spread function, and the Gaussian probability density function.

5.5 THE *F*-TEST OF ERROR IMPROVEMENT SIGNIFICANCE

We sometimes have *two* candidate models for describing an overdetermined inverse problem, one of which is more complicated than the other (in the sense that it possesses a greater number of model parameters). Suppose that Model B is more complicated than Model A and that the total prediction error for Model B is less than the total prediction error for Model A: $E_{\mathrm{B}} < E_{\mathrm{A}}$. Does Model B really fit the data better than Model A?

The answer to this question depends on the variance of the data. Almost any complicated model will fit data better than a less complicated one. The relevant question is whether the fit is *significantly* better, that is, whether the improvement is too large to be accounted for by random fluctuations in the data. For statistical reasons that will be cited, we pretend, in this case, that the two inverse problems are solved with two different realizations of the data.

Suppose that we estimate the variance of the data d_i from the prediction error e_i of each model as

$$(\sigma_d^{\mathrm{est}})^2 = \frac{1}{\nu}\sum_{i=1}^{N} e_i^2 = \frac{E}{\nu} \quad \text{with} \quad \nu = N - M \tag{5.58}$$

This estimate will usually be larger than the true variance of the data, since it also includes a contribution from the (possibly) poor fit of the model. If one model fits the data about as well as the other, then the variance $(\sigma_{d\mathrm{A}}^{\mathrm{est}})^2$ estimated from Model A should be about the same as the variance $(\sigma_{d\mathrm{B}}^{\mathrm{est}})^2$ estimated from Model B. On the other hand, if Model B gives a better fit than Model A, the estimated variances will differ in such a way that the ratio $(\sigma_{d\mathrm{A}}^{\mathrm{est}})^2/(\sigma_{d\mathrm{B}}^{\mathrm{est}})^2$ will be greater than unity. If the ratio is only slightly greater than unity, the difference in fit may be entirely a result of random fluctuations in the data and therefore may not be significant. Nevertheless, there is clearly some value for the ratio that indicates a significant difference between the two fits.

To compute this critical value, we consider the theoretical distribution for the ratio of two variance estimates derived from two different realizations of the *same* data set. Of course, the ratio of the true variance with itself always has the value unity; but the ratio of two estimates of the true variance will fluctuate randomly about unity. We therefore determine whether or not ratios greater than or equal to the observed ratio occur less than, say, 5% of the time. If they do, then there is a 95% probability that the two estimates are derived from data sets with different true variances. We are justified in concluding that the second model is a significant improvement over the first.

To handle data with nonuniform variance, we form a ratio, not of estimated variances, but of the related quantity

$$\chi_\nu^2 = \frac{\nu(\sigma_d^{\text{est}})^2}{(\sigma_d^{\text{true}})^2} = \frac{\sum_{i=1}^{N} e_i^2}{(\sigma_{di}^{\text{true}})^2} \tag{5.59}$$

This quantity is chosen because it has a χ_ν^2 distribution with ν degrees of freedom. The ratio of the χ_ν^2 for the two models is given by

$$F(\nu_A, \nu_B) = \frac{\chi_{\nu A}^2/\nu_A}{\chi_{\nu B}^2/\nu_B} = \frac{(\sigma_{dA}^{\text{est}})^2/(\sigma_{dA}^{\text{true}})^2}{(\sigma_{dB}^{\text{est}})^2/(\sigma_{dB}^{\text{true}})^2} \tag{5.60}$$

Note that the true variance cancels out of the equation, as long as it is the same for the two models. Thus, the F ratio is not a function of the overall amplitude of the total error but only of the relative error between the two models. It is, however, a function of the number of degrees of freedom of the two models.

The probability density function of the F ratio is known, but it cannot be written in terms of elementary functions, so we omit it here. It is a unimodal distribution with mean and variance given by

$$\langle F \rangle = \frac{\nu_B}{\nu_B - 2} \qquad \sigma_F^2 = \frac{2\nu_B^2(\nu_A + \nu_B - 2)}{\nu_A(\nu_B - 2)^2(\nu_B - 4)} \tag{5.61}$$

Note that for large degrees of freedom, $\langle F \rangle \approx 1$. Note, also, that F^{est} and $1/F^{\text{est}}$ play symmetrical roles, in the sense that the first quantifies the improvement of fit of Model B with respect to Model A, and the latter, the improvement of fit of Model A with respect to Model B. Thus, in testing the *Null Hypothesis* that any difference between the two estimated variances is due to random variation, we should compute the probability that F is smaller than $1/F^{\text{est}}$ or larger than F^{est}:

$$P\left(F < \frac{1}{F^{\text{est}}} \quad \text{or} \quad F > F^{\text{est}}\right) \tag{5.62}$$

The Null Hypothesis can be rejected if this probability is less than 5%. The *MatLab* function `fcdf()` computes the cumulative probability of F:

```
Fobs = (EA/vA) / (EB/vB);
if( Fobs<1 )
   Fobs=1/Fobs;
end
P = 1 - (fcdf(Fobs,vA,vB)-fcdf(1/Fobs,vA,vB));
```

(*MatLab* script gda05_17)

Here `EA` and `EB` are E_A and E_B, respectively, and `vA` and `vB` are ν_A and ν_B, respectively. An example is shown in Figure 5.17.

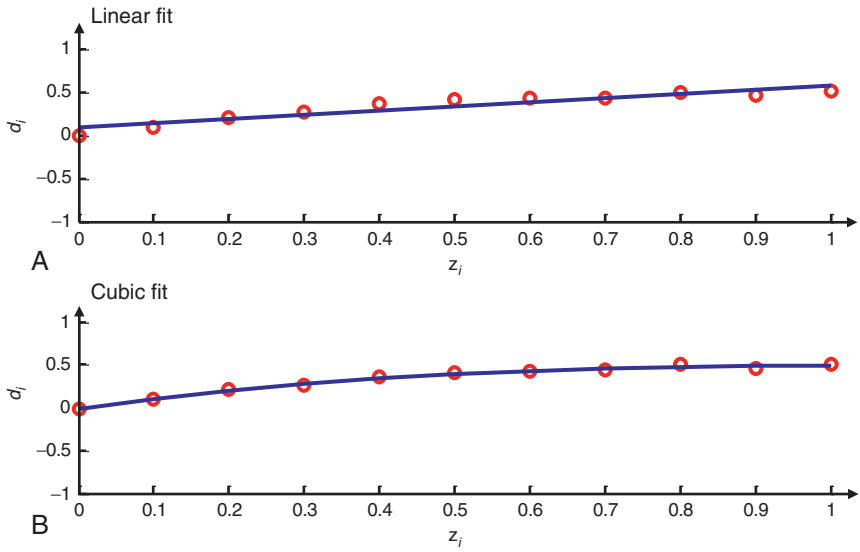

FIGURE 5.17 Hypothetical data set (red circles) fit (blue curve) with (A) a straight line and (B) a cubic polynomial. Although the cubic fit appears superior, an F-test reveals that this level of improvement of fit will be obtained 6.4% of the time under the Null Hypothesis that the improvement is due to random variation. The improvement of fit is not significant at the 95% level. *MatLab* script gda05_17.

5.6 PROBLEMS

5.1 Suppose that a random variable m is defined on the interval $[-1,1]$. A reasonable choice for the null probability density function is $p_N(\mathbf{m}) = 1/2$, meaning that m can be anywhere within the interval with equal probability. (A) Calculate (either analytically or numerically) the information gain S of the *a priori* probability density function $p_A(\mathbf{m}) = 1/2 + cm$, where $0 < c < 1/2$. Note that the larger the constant c, the higher the probability that m will fall in the positive half of the interval. (B) Make and interpret a plot of $S(c)$.

5.2 Suppose that a Gaussian probability density function with mean $\langle m \rangle$ and variance σ_m^2 is used to represent *a priori* information about the following types of model parameters: (A) the density of sea water; (B) the shear velocity at 100-km depth in the earth; (C) the O^{18} to O^{16} ratio in glacial ice. Propose plausible values of $\langle m \rangle$ and σ_m^2 in each case, citing references that justify your values.

5.3 Suppose that you are fitting a cubic polynomial to data, $d_i = m_1 + m_2 z_i + m_3 z_i^2 + m_4 z_i^3$, but have *a priori* information that $m_1 = 2m_2 = 4m_3 = 8m_4$. Write a *MatLab* script to solve this problem using Equation (3.55). Set up the problem so that $N = 50$, $0 < z_i < 1$, and $m_1^{\text{true}} = 1$, and generate synthetic data with $\sigma_d^2 = (0.1)^2$. How well do the data alone (that is, without the *a priori* information) constrain the ratios between the ms?

5.4 This problem builds upon Problem 5.3. Suppose that you are fitting a cubic polynomial to data, $d_i = m_1 + m_2 z_i + m_3 z_i^2 + m_4 z_i^3$, but have *a priori* information that $m_1 = 2m_2 = 4m_3 = 8m_4$. Write a *MatLab* script to solve this problem using Equation (5.55). Use a range of values for the variance σ_m^2 of the *a priori* information, from very uncertain to very certain. Set up the problem so that $N = 50$, $0 < z_i < 1$, and $m_1^{\text{true}} = 1$ and generate synthetic data with $\sigma_d^2 = (0.1)^2$. How well do the data alone (that is, without the *a priori* information) constrain the ratios between the *m*s?

5.5 This problem builds upon Problems 5.3 and 5.4. Modify your script in Problem 5.4 in the case where the model of the data fitting a cubic polynomial is thought to be inexact, with a variance $\sigma_d^2 = (0.05)^2$. How does this modification change your results?

5.6 Write a *MatLab* function to empirically generate the $p(F)$ probability density function with $\nu_1 = \nu_2 = 20$ by (A) using the `random('Normal', ...)` function to generate batches of 20 Gaussian-distributed random numbers with zero mean and unit variance. (B) Calculating χ_{20}^2 by summing the squares of each batch. (C) Calculating F for pairs of batches. (D) Repeat many times, creating a histogram of the resulting Fs, and normalize to unit area to produce an empirical estimate of $p(F)$. (E) Compare your result with *MatLab*'*s* `fpdf()` function.

5.7 This problem expands upon Problem 5.5. Suppose that the random numbers in step (A) are drawn from a uniform distribution with zero mean and unit variance, but that F is calculated the same as before (let us call it F'). (A) How different is $p(F')$ from $p(F)$? (B) How would treating data with uniformly distributed error as if they were Gaussian-distributed affect the results of an F-test? Hint: A distribution that is uniform between a and b has a variance of $(b-a)^2/12$.

REFERENCES

Kapur, J.N., 1989. Maximum Entropy Models in Science and Engineering. Wiley, New York 636pp.

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

Tarantola, A., Valette, B., 1982b. Inverse problems = quest for information. J. Geophys. 50, 159–170.

Woodbury, A.D., 2011. Minimum relative entropy, Bayes and Kapur. Geophys. J. Int. 185, 181–189.