12. Estimating the Noise Parameters

In Chapter 3 of GIT we saw that the concept of an adequate fit of model predictions to the measured values depends on having a quantitative estimate of the measurement uncertainty, most conveniently, an estimate of the noise **variance**. This is often very difficult to do objectively, and in seismology in particular, people often say it's too difficult to do. In ordinary parameter estimation in statistics, fitting a straight line for example, we have many more data than parameters, and the misfit to the model gives a measure of uncertainty all on its own. A standard result for linear parameter estimation, used to estimate the noise, is

$$\mathcal{E}[\sum_{j=1}^{N} (d_j - \Theta_j)^2] = (N-P)\,\sigma^2 \tag{1}$$

where Θ_j are the predictions of the linear model, P is the number of parameters in the model, and σ is the standard error of the noise in the measurements d_j. The number $N-P$ is often called the number of degrees of freedom in the data. In the linear inverse problem, this result presents us with a difficulty: in principle P is infinite! With linearly independent representers (the normal situation) we can always reduce the misfit to zero if necessary.

How can we get a value for σ? In some inverse problems there data themselves are often composites estimated by averaging over large numbers of actual measurements. The best example of this electromagnetic sounding, in which the response of the Earth is obtained by time-series

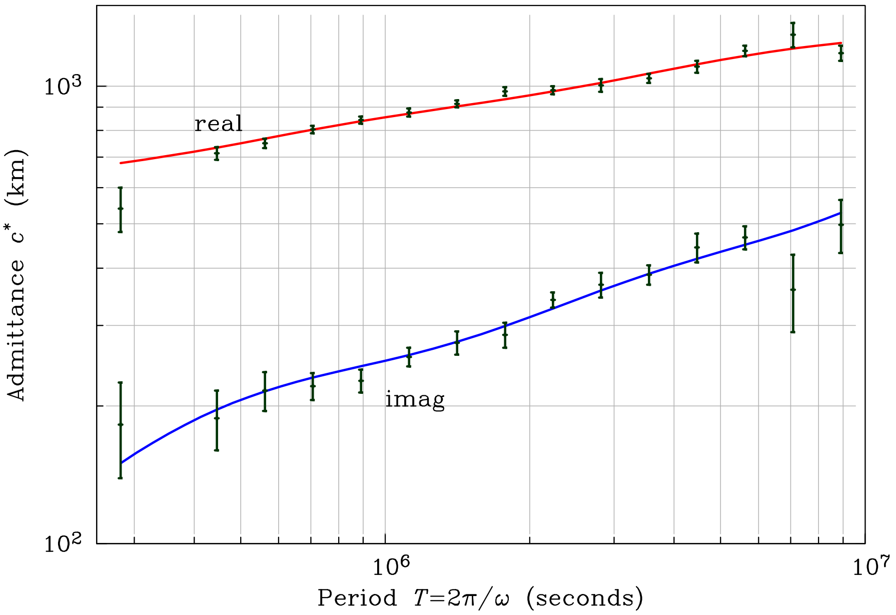

Figure 12.1: Global electromagnetic response with one-standard-deviation error bars.

analysis of long data series of electric and magnetic fields. Those field measurements themselves are never used directly. In this subject a complex function of frequency (called an *admittance* or a *transfer function*) is found statistically, and comes with error bars already; consult Egbert, G. D., and Booker, J. R., Robust estimation of geomagnetic transfer functions, *Geophys. J. R. Astron. Soc.,* 87, 173-94, 1986. Alternatively, when Steve Constable wanted to calculate a global transfer function, he averaged together responses from several independent studies made in scattered locations, and came up with response whose uncertainties were estimated by their deviation from the mean; see Figure 12.1 and Constable, S., Constraints on mantle electrical conductivity from field and laboratory measurements, *J. Geomag. Geoelectr.,* 45, 707-9, 1993.

If we are to identify and quantify noise in data, we need a characteristic that separates it from signal. When the noise is uncorrelated from point to point, it will have a flat power spectral density; even if it is not completely uncorrelated, the noise spectrum is usually much flatter than

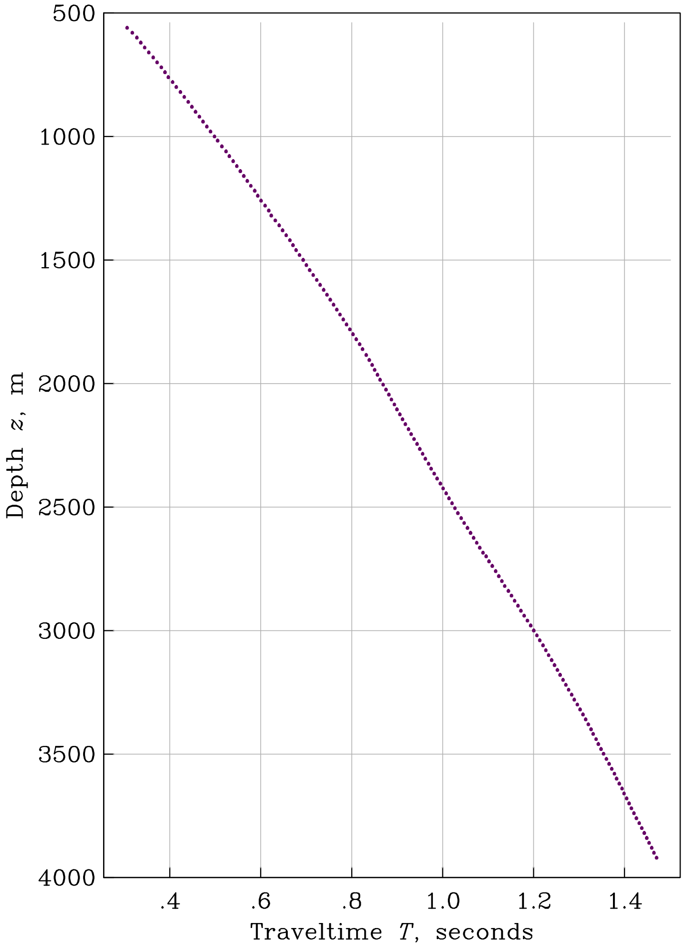

Figure 12.2: Travel-time picks from geophones in a well: the check-shot data.

that of the signal. When the measurements are made serially in time, or in space (along a profile, for example) it is usually possible to estimate the power spectrum, or power spectral density (PSD), from the data set. There is no space to go into how PSD are calculated here; that will be covered in the course on geophysical data analysis in the Spring Quarter. For a good reference see: Priestley, M. B., Spectral Analysis and Time Series, Academic Press, New York, 1981.

On the previous page we see an example of seismic “check-shot” data. These are first-arrival travel time picks from geophones in an oil well, from a charge fired at the surface. If we want to invert this record we will need an estimate of uncertainties. One way would be to look at the original traces, and to take a guess at how accurately the first pulse emerges from the ambient noise, but I don’t have the original records, just the time picks. So we take the power spectrum, from the 169 data. The result, show in Figure 12.3 is remarkably revealing. The two curves belong to different estimation methods for the PSD. What we see is a steeply falling part (a **red spectrum**), out to a wavenumber of around 0.006 m^{-1} , and then a plateau. The flat portion is characteristic of **white noise**, or an uncorrelated random signal. A very compelling interpretation of this spectrum is that the red part comes from the geophysical signal, and the white spectrum is the result of noise in the data. We will assume the noise spectrum continues to the smallest wavenumbers, but is

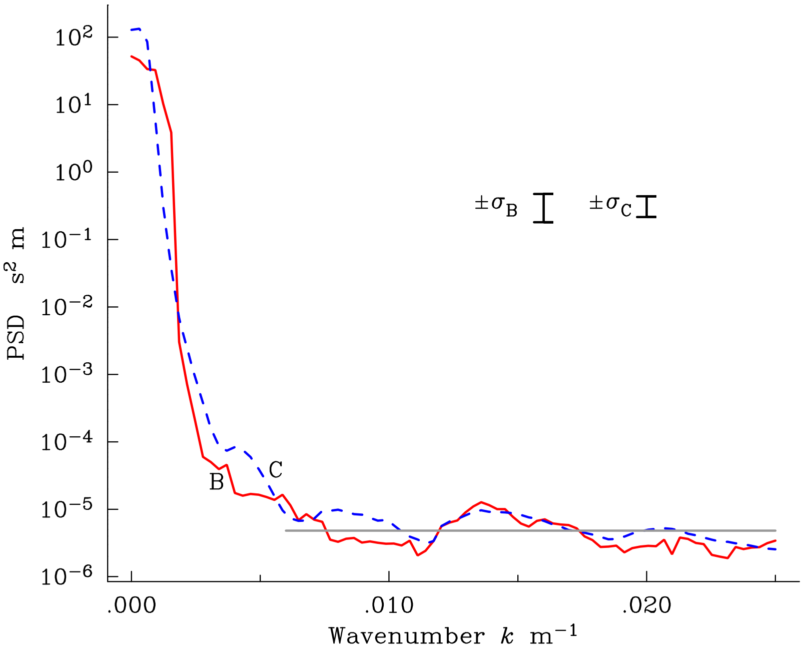

Figure 12.3: Power spectral density of checkshot data.

completely overwhelmed by signal below 0.006 m^{-1}. Then we use the famous result that

$$\mathrm{var}\,[X] = \int_0^{k_{\max}} P_X(k)\, dk \qquad (2)$$

The variance is just the area under the PSD curve. This gives us approximately $\sigma^2 = 0.03 \times 4 \times 10^{-6} = 1.2 \times 10^{-7}\mathrm{s}^2$; hence $\sigma = 0.00035$ s, or 0.35 milliseconds. The uncertainty in these data would plot as an error bar too small to see on Figure 12.2. A completely different approach leads to almost exactly the same error estimate: see Malinverno, A., and Parker, R. L., Two ways to quantify uncertainty in geophysical inverse problems, *Geophysics,* 71, 15-27, 2005.

We will use the same technique on the near seafloor magnetic data, since it is part of a long serial record. The PSD of our magnetic anomaly is shown as the solid curve in Figure 12.4. At first the picture is not as convincing as it is for the checkshot seismic data, where there is a completely clear division between a red and a flat spectrum, but in this case there is a bit of theory to guide us in what to expect; see Parker, R. L., and O'Brien, M. S., Spectral analysis of vector magnetic field profiles, *J. Geophys. Res.,* 102, 24815-24824, 1997. If the track were level and the basement were flat and contained a random magnetization, we would expect

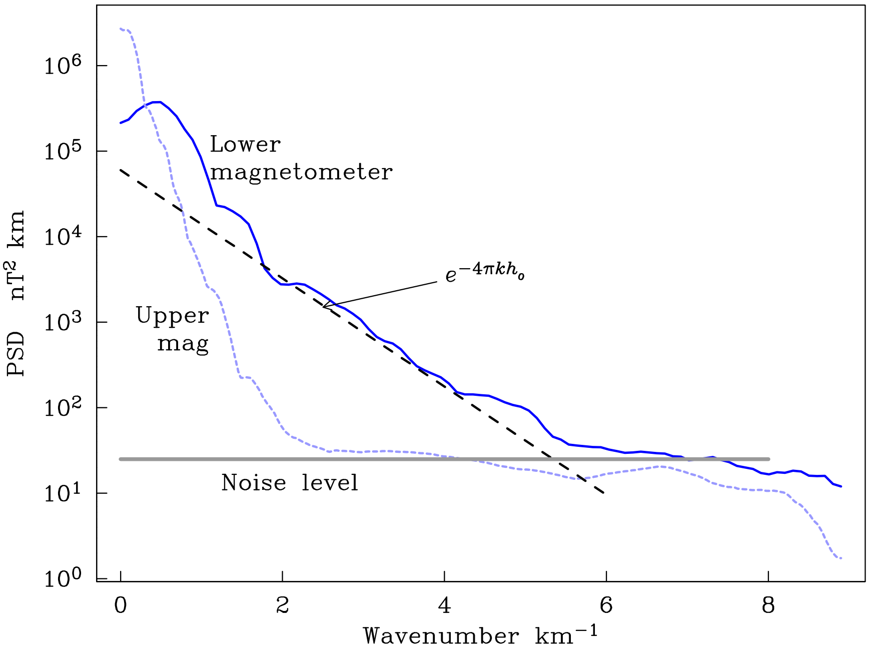

Figure 12.4: PSD of a segment of the near-bottom magnetic anomaly profile.

the magnetic anomaly PSD to fall approximately exponentially, like $\exp(-4\pi kh)$. The elevation of the magnetometer above the basement lies roughly between 116 m and 200 m, and the lower level will dominate at the higher wavenumbers; the long dashed line shows the exponential fall-off for the lower value of h, and it fits the spectrum rather well. There is a break at about $6\,\mathrm{km}^{-1}$, which I will interpret as the point where noise begins to exceed signal. In marine magnetic surveys the noise is caused by other environmental magnetic fields, such as time-varying fields due to currents in the ionosphere, and also those from electric currents in the water caused by induction. I will assume the PSD of the noise is approximately white, and follows the horizontal grey line. The corresponding variance from (2) is $8.5 \times 25 = 210\,\mathrm{nT}^2$ which yields a value for $\sigma = 14.5\,\mathrm{nT}$. Remarkable confirmation for this model comes from the second PSD, shown with short dashes. Unusually in a survey of this kind, a second magnetometer was tethered to the cable 300 m higher up than the one we have been using. By the theory I mentioned, its spectrum falls much faster, so the PSD hits the noise level at a lower wavenumber, and as we see in the figure, it levels out at the same value, because it is in essentially the same noise environment as the lower instrument. Thanks to the second magnetometer we can be confident in our estimate for the uncertainty.

The spectral approach requires a number of additional assumptions about the noise, primarily, that it is *statistically stationary.* This means that the random process responsible for the noise is the same everywhere in the data series. That is often a plausible assumption, and would be accepted for the marine data. The message I want to leave you with is that the power spectrum usually gives an important clue about the noise, because the noise persists out to the highest frequencies, while most natural processes have a red spectrum, and the components of the signal at the high frequencies (or wavenumbers) are usually attenuated, thus permitted the noise spectrum to show itself. If this doesn't happen the data are not being sampled at a high enough rate, and there is danger of aliasing, which means the signal is being sampled too slowly to capture its true behavior.

References

Constable, S., Constraints on mantle electrical conductivity from field and laboratory measurements, *J. Geomag. Geoelectr.,* 45, 707-9, 1993.

Egbert, G. D., and Booker, J. R., Robust estimation of geomagnetic transfer functions, *Geophys. J. R. Astron. Soc.,* 87, 173-94, 1986.

Malinverno, A., and Parker, R. L., Two ways to quantify uncertainty in geophysical inverse problems, *Geophysics,* 71, 15-27, 2005.

Parker, R. L., Coherence of signals from magnetometers on parallel paths, *J. Geophys. Res.,* 102, 5111-7, 1997.

Parker, R. L., and O'Brien, M. S., Spectral analysis of vector magnetic field profiles, *J. Geophys. Res.,* 102, 24815-24824, 1997.

Priestley, M. B.,*Spectral Analysis and Time Series,* Academic Press, New York, 1981.

13. Fitting within a Tolerance

The statistical theory of the early sections of Chapter 3 in GIT tell us that a good fit to the observations can be expressed in the form

$$\| \Sigma^{-1}(d - \Theta(m)) \| \leq T \tag{1}$$

where $d, \Theta \in \mathbb{R}^M$ are vectors containing the measurements and the predictions of the model $m \in \mathbb{R}^N$; $\Sigma \in \mathbb{R}^{M \times M}$ is usually a diagonal matrix of standard errors (for the rare case of correlated errors something else replaces the diagonal matrix); and T is tolerance that we arrive at by a subjective decision about what we regard as acceptable odds of being wrong. For mere model building, a loose 50% level is just fine. We will almost always use the 2-norm on the data space, and thus the chi-squared statistic will be our guide.

From here on in our discussion we will take the purely practical road, and so the vector of theory, Θ will not be a collection of inner products in a Hilbert space, but instead

$$\Theta(m) = GWm \tag{2}$$

where $m \in \mathbb{R}^N$ is a vector representing the model itself, and $G \in \mathbb{R}^{M \times N}$ is a matrix with rows sampling the representers, and $W \in \mathbb{R}^{N \times N}$ is the quadrature matrix, another diagonal matrix.

We seek the smallest model, or the simplest solution, and for now that idea will be encapsulated in the minimization of **penalty function**, a norm or seminorm:

$$\| Rm \| \tag{3}$$

where $R \in \mathbb{R}^{L \times N}$ is a regularizing matrix, which might not be of full rank, and might penalize only part of the solution, so that $L < N$ as in 11(20); recall the brief discussion on pp 51-52 in the Notes, where we differenced m. A welcome feature of the numerical approach (as opposed to the analytic one) is that the treatment is indifferent to which of the two choices, norm or seminorm, is made. At first glance the normal strategy of calling in a Lagrange multiplier to handle the constraint (1) is inapplicable because of the inequality. But as GIT demonstrates at great length, we can still use this useful tool after all, with the caveat that we must first check that a model satisfying

$$Rm = 0 \tag{4}$$

cannot satisfy (1). If such a model does exist, then clearly zero is the minimum of the penalty function. In practice, this will almost never happen, and so then the problem to be solved is to find the stationary value of the unconstrained function

$$u(m, \nu) = \| R\, m \|^2 + \nu[\| \Sigma^{-1}(d - GW\, m) \|^2] - T^2] \tag{5}$$

where ν is the Lagrange multiplier accompanying the constraint:

$$||\Sigma^{-1}(d - GW\,m)|| = T\,. \tag{6}$$

To reduce the clutter I introduce a couple of abbreviations: let

$$\hat{d} = \Sigma^{-1}d, \quad \text{and} \quad B = \Sigma^{-1}GW\,. \tag{7}$$

This gives us the new unconstrained function

$$u(m, \nu) = \,||R\,m\,||^2 + \nu[\,||\hat{d} - B\,m\,||^2] - T^2]\,. \tag{8}$$

Notice that, for a fixed value of T, minimization of this expression can be regarded as seeking a compromise between two undesirable properties of the solution: the first term represents model complexity, which we wish to keep small; the second measures model misfit, also a quantity to be suppressed as far as possible. By making $\nu > 0$ but small we pay attention to the penalty function at the expense of data misfit, while making ν large works in the other direction, and allows large penalty values to secure a good match to observation. Let us continue.

We can differentiate (8) with respect to m by writing out the expression in terms of components; I will spare you the intermediate steps which we have seen several times in slightly different contexts. At a stationary point of (8) the gradient of u vanishes and we find the vector m_0 obeys

$$R^T R\,m_0 + \nu B^T B\,m_0 - \nu B^T \hat{d} = 0\,. \tag{9}$$

Or, equivalently, m_0 satisfies the linear system

$$(B^T B + \frac{1}{\nu}\,R^T R)\,m_0 = B^T \hat{d}\,. \tag{10}$$

Differentiating with ν returns the constraint, now written as

$$||\hat{d} - B\,m_0\,|| = T\,. \tag{11}$$

If we knew the value of ν, we could find the model by solving (10). So the tactic for solving (10) and (11) together, as we must, requires solving (10) for a sequence of νs seeking the vector m_0 that gives (11). We need to show that as ν increases, the misfit norm in (11) decreases. This result is intuitive from our discussion after (8), but it also is useful to have the derivative itself. Consider the squared misfit in (11) to be purely a function of ν:

$$F(\nu) = \,||\hat{d} - B\,m_0(\nu)\,||^2\,. \tag{12}$$

Then differentiating on ν

$$\frac{dF}{d\nu} = -2\,(\hat{d} \; - B\,m_0(\nu))^T B\,\frac{dm_0}{d\nu} = -2\,(B^T(\hat{d} - B\,m_0))^T\,\frac{dm_0}{d\nu}\,. \tag{13}$$

By rearranging (9) we see that

$$B^T(\hat{d} - B\,m_0) = \frac{1}{\nu}\,R^T R m_0 \tag{14}$$

and hence

$$\frac{dF}{d\nu} = -\frac{2}{\nu}(R^T R m_0)^T \frac{dm_0}{d\nu} \,. \tag{15}$$

Now to get $dm_0/d\nu$, differentiate both sides of (10):

$$(B^T B + \frac{1}{\nu} R^T R)\frac{dm_0}{d\nu} - \frac{1}{\nu^2} R^T R m_0 = 0 \,. \tag{16}$$

Solving for $dm_0/d\nu$ and plugging the answer into (15) gives the glorious result

$$\frac{dF}{d\nu} = -\frac{2}{\nu^3}(R^T R m_0)^T (B^T B + \frac{1}{\nu} R^T R)^{-1} (R^T R m_0) \,. \tag{17}$$

The inverse matrix in the middle is positive definite, because it is composed of the sum of positive definite pieces; then, since $\nu > 0$ the whole thing must be negative.

This simplifies the strategy for solving the pair (10)-(11), because now we know that when a guess for ν yields a value of F that is too high, we must increase ν, and conversely. Better yet we can even use **Newton's method,** which you will recall can be used for solving equations in a single variable: in this case the equation is

$$F(\nu_0) = T^2 \,. \tag{18}$$

We begin with an initial value $\nu_1 > 0$, and we perform a one-term Taylor expansion on (18) as follows:

$$T^2 = F(\nu_1 + \nu_0 - \nu_1) = F(\nu_1) + (\nu_0 - \nu_1)F'(\nu_1) + \varepsilon \tag{19}$$

where F' denotes the derivative, and ε is error due to the neglect of higher order terms in the series. Rearranging this expression gives

$$\nu_0 = \nu_1 - \frac{F(\nu_1) - T^2}{F'(\nu_1)} + \frac{\varepsilon}{F'(\nu_1)} \,. \tag{20}$$

If the ε, the second order term in the Taylor expansion is neglected, (20) gives a recipe for the next step in an iterative process which we write

$$\nu_{n+1} = \nu_n - \frac{F(\nu_n) - T^2}{F'(\nu_n)} \,, \quad n = 1, 2, \cdots \,. \tag{21}$$

It is shown in GIT that this procedure always converges, provided the initial guess obeys $\nu_1 < \nu_0$. But surprisingly perhaps, a faster rate of convergence is usually obtained by writing (18) as

$$\ln(F(\nu_0)) = 2 \ln T \tag{22}$$

and solving this equation with Newton's method.

Let us summarize the process. Recall the abbreviations introduced in (7). We wish to discover the solution vector m_0 and the Lagrange multiplier ν_0 which solve simultaneously the linear system (10) and the misfit constraint (11). We make an initial guess for ν which we call ν_1, and with it we solve (10). We take the resultant model vector and put it into (12) which gives us F. We also solve the system (17), which provides us with $dF/d\nu$. If F is close enough to T^2 we stop. Otherwise we use (21) to obtain a revised version of ν, with which we can begin the cycle again.

There are a number of points, before we examine this process in an illustration. Equation (10) can be written as the solution to an *overdetermined* least-squares approximation problem:

$$\begin{bmatrix} B \\ \nu^{-½}R \end{bmatrix} m_0 \sim \begin{bmatrix} \hat{d} \\ 0 \end{bmatrix} \tag{23}$$

you will easily verify that the normal equations for this overdetermined least squares problem is exactly (10). We can write a similar equation for $dm_0/d\nu$

$$\begin{bmatrix} B \\ \nu^{-½}R \end{bmatrix} \frac{dm_0}{d\nu} \sim \begin{bmatrix} 0 \\ \nu^{-3/2}Rm_0 \end{bmatrix} \tag{24}$$

There are two possible reasons for treating (10) as the solution of (23). First, when the size of the system is modest, we can solve (23) by QR factorization and avoid poor numerical stability. Second, R is almost always sparse, so when the system is large we can take advantage of the sparse LS form of the solution, 6(13) in our notes on linear algebra. And if the system is very large, that form can be conveniently solved by the method of conjugate gradients, as we will see later.

Let us now apply this approach to our near-bottom magnetic anomaly problem. You may recall that we looked for the smallest magnetization model in L_2 that fit the measured values exactly and obtained a mess, plotted in Figure 11.4. The model is one hundred times larger than a reasonable solution should be, yet it is the smallest model. The fault is the demand of an exact fit. In Section 11 of these Notes I obtained a noise estimate of $\sigma = 14.5\,\mathrm{nT}$, a quantity too small to be distinguished in a plot showing the full range of the data; can such a small misfit reduce the size of the solution to a reasonable level? We decide in advance what will be acceptable as a plausible size of misfit. I suggest that we take the expected value of error norm as a trial value. Thus, as shown in GIT on p 124

$$\mathcal{E}[\,\|\, d - \Theta \,\|\, /\sigma] = \sqrt{N}\left[1 - \frac{1}{4N} + \cdots\right] \tag{25}$$

where we have used the fact that noise will be treated as iid. Then since $N = 100$, we calculate that the misfit tolerance in 11(6) is $T = 9.975$. The target misfit for Newton's method is T^2. We will minimize the L_2 norm as before, using trapezoidal rule for all the integrations.

In the figure on the next page we see the progress of the iterative solution, which starts with $\nu = 0.001$ and then goes down in F and up in ν. The starting guess lies below the solution ν_0, and so Newton is

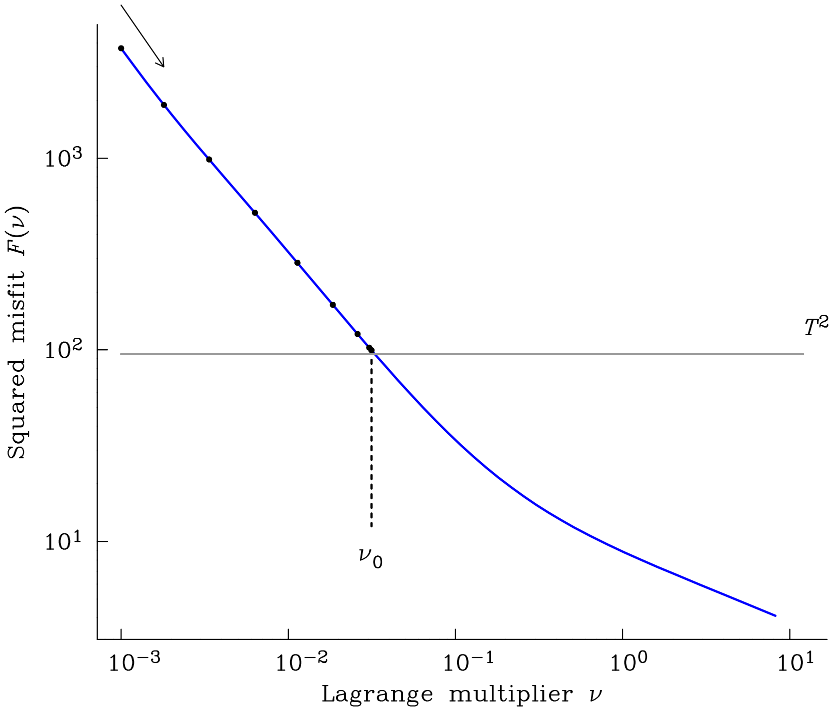

Figure 13.1: Squared misfit vs Lagrange multiplier.

guaranteed to converge. If the guess had been too high, we could not reliably use the Newton iterate, because it can ask for negative values, which are forbidden; so in those circumstances we just divide the guess by ten and try again. In this example the procedure took 10 steps to converge to about 4 significant figures. It is obvious we could get much more rapid convergence if logarithmic values (both $\ln F$ and $\ln \nu$) were used, because the curve is nearly straight in these variables, and Newton's method is based on a linear approximation. I leave the details for a homework exercise.

The norm of the new model m_0 is considerably smaller than the one obtained by an exact fit: now $|| m_0 || = 6.26$, while a precise match yields a norm of 697. The new model is considerably more reasonable in size, as we had hoped. And this is confirmed in Figure 13.2, where the solid line is the L_2 norm minimizer. This solution is spiky but keeps its magnetization in a range of perfectly acceptable numbers. Notice the sign is predominantly positive, which we might perhaps expect as the profile is the Bruhnes normal magnetic period. The strongly reversed section between 1.6 and 2.5 km is interesting, because it is not a well recognized brief reversal. Are any of the model's reversed magnetization sections real, or can they be dispensed with while still matching the measurements? This is a question we must wait to answer.

In the same figure shown dashed is the minimizer of the 2-norm of dm/dx; it is noticeably smoother, and a little larger. The nasty spike in m_0 near 3.2 km has been greatly reduced, but that is hard to see in this

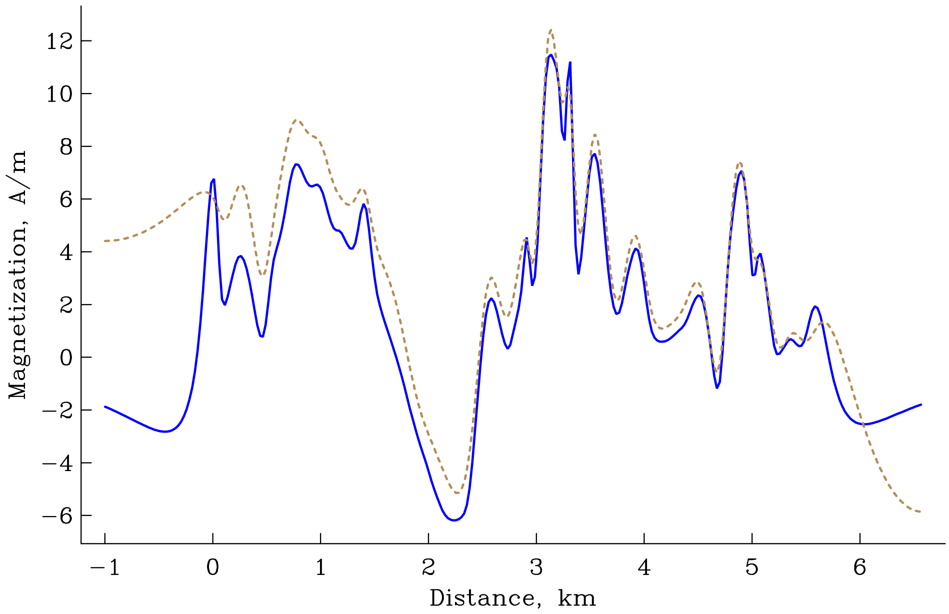

Figure 13.2: Minimum norm and seminorm magnetizations with plausible misfits.

graph. We probably can conclude that the crustal magnetization is far from constant along this profile, and that big swings in the original field are not due to effects of topography (changes in range of the magnetometer from the sources), but are a genuine reflection of variable magnetic intensity in the basement. But whether or not reversed magnetization is required has not been established; it certainly looks like it on the present evidence.