APPENDIX C
ESTIMATING AMPLITUDE DISTRIBUTIONS

In Appendices A and B I assumed specific prior signal and noise pdf's for the MAP objective function (A.3). If the noise is Gaussian, the standard deviation can be chosen pessimistically as equal to that of the data. The standard deviation of Gaussian signal is important to the stability of the linearized MAP estimate defined by equation (3.10). Let us first choose some physically unrestrictive upper bound for the variance, and if the perturbation is unstable, let us reduce the value.

Much more critical are the choices for the pdf's of transformed parameters, those pdf's used in equations (3.14) and (3.15) for the Bayesian estimate and reliability. By observing statistics of the residual data before and after linear transformation, we can estimate these pdf's directly from the data.

C.1. Stationarity

Let us first assume that the signal has a stationary dimension, so that enough redundancy exists for histograms to approximate pdf's. Prior pdf's should reflect regional possibilities. An inversion that was told where a physical structure was likely to appear could create a model that only reinforced the interpreter's prejudices. If a structure appears often in one location then one should assume that it can also appear nearby. Thus, one not only expects but desires that estimated pdf's change slowly over spatial dimensions and time. Because of this stationarity, a histogram prepared from a great many samples with identical pdf's will describe the possibilities open to them all.

Likewise we can usually find some dimension over which we expect noise parameters to have similar statistics. When I later drop subscripts from signal and noise pdf's, I am assuming stationarity over some dimension.

C.2. Pessimistic estimates of distributions

We shall need the following relations between random variables (z's and y's) and their corresponding pdf's. a is a constant, and x a dummy variable.

$$z = y + a \quad \text{implies} \quad p_z(x) = p_y(x - a) \ ; \tag{C.1}$$

$$z = a \cdot y \quad \text{implies} \quad p_z(x) = \frac{1}{a} p_y(x/a) \ ;$$

$$z = y_1 + y_2 \quad \text{implies} \quad p_z(x) = p_{y_1}(x) * p_{y_2}(x)$$

The star indicates convolution. Because we assumed that the samples of $\mathbf{n}$ are

statistically independent, equation (B.3) requires that

$$p_{n_i'}(x) = \prod_j * \left[\frac{1}{F_{ij}} p_{n\,res}\left(\frac{x}{F_{ij}}\right)\right] \quad . \tag{C.2}$$

The $\prod *$ indicates multiple convolutions. Many convolutions eventually produce a Gaussian distribution, according to the central limit theorem. In many applications, including that of the vertical seismic profile, the transform F will change slowly enough that local stationarity is preserved. We shall then be able to suppress subscripts on all pdf's.

Define a pessimistic estimate of $p_{n_i'}(x)$ as the distribution that would result if all residual data were noise. Ignore the coherence of any signal.

$$\hat{p}_{n_i'}(x) \equiv \prod_j * \left[\frac{1}{F_{ij}} p_{d\,res}\left(\frac{x - c_i}{F_{ij}}\right)\right] = p_{n_i'}(x) * \prod_j * \left[\frac{1}{F_{ij}} p_{s\,res}\left(\frac{x - c_i}{F_{ij}}\right)\right] \tag{C.3}$$

This pdf must overestimate the transformed noise and all positive moments. If the data contain no signal, then the estimate is perfect (the signal pdf becomes a delta function).

To find the pessimistic estimate (C.3), first generate a random array whose statistically independent samples have the same pdf's as do the data. Use the same transform and take histograms over stationary dimensions. Because the signal and noise samples remain statistically independent and additive after a linear transformation is performed, choices of their pdf's determine pdf's for the data.

$$p_{d'}(x) = p_{s'}(x) * p_{n'}(x) \tag{C.4}$$

Using the assumption of local stationarity, estimate $p_{d'}(\cdot)$ from local histograms of the transformed data. To estimate the signal pdf $p_{s'}(\cdot)$ we must now "deconvolve" the data pdf with the noise pdf.

C.3. Estimating distributions with cross entropy

Let $\{\hat{p}_i\}$ be a histogram; let $\{p_i\}$ be the discrete pdf approximating it. Subscript i indexes a narrow range of amplitudes (a bin). $\hat{p}_i$ is the frequency of the bin in the data, and p_i its assumed probability. If the parameters sampled by a histogram are statistically independent, then the probability of the ensemble is equal to the product of the probabilities of the individual samples. A histogram of N samples will have the probability shown within brackets below.

$$\max_{\{p_i\}} \left[\prod_i C_i (p_i)^{\hat{p}_i N}\right] \quad ; \quad C_i = \frac{(\hat{p}_i N)!}{i\,!(\hat{p}_i N - i)!} \tag{C.5}$$

(Exclamation points indicate factorials.) The optimum p_i will maximize this probability. Let us take the natural logarithm, reverse the sign, add and subtract constants; we may equivalently minimize

$$\min_{\{p_i\}} \sum_i \hat{p}_i \log(\hat{p}_i / p_i) \quad . \tag{C.6}$$

We discover, in the continuous limit, Kullback's directed divergence (1959), or cross-entropy.

$$\min_{p(x)} \int \hat{p}(x) \log[\hat{p}(x)/p(x)]dx \tag{C.7}$$

x is now the index of amplitude.

Assuming that the pessimistic noise pdf is correct, I define the MAP estimate as the signal pdf that maximizes the probability of the data histogram. To find $p_{s'}(x)$, given $p_{d'}(x)$ and $p_{n'}(x)$, minimize the following (suppressing primes).

$$J_4[p_s(x)] = \int p_d(x) \ln[p_d(x)/\int p_s(x-y)p_n(y)dy]dx$$

$$+ \frac{\lambda_1}{2}[\int p_s(x)dx - 1]^2 + \frac{\lambda_2}{4}\int [p_s(x) - |p_s(x)|]^2 dx \tag{C.8}$$

Add two Lagrange multipliers for the constraints of unit area and of positivity.

Many non-linear methods will minimize this distribution. I use a steepest descent algorithm and select optimum Lagrangian multipliers during the necessary line search. The histograms contain fewer than a hundred samples, so the iterations are fast and take a negligible percentage of the inversion's total run time.

To calculate the gradient of J_4 with respect to each point of the function $p_s(x)$, perturb the previous estimate with a delta function, $p_s(x) + \epsilon\delta(x - x^0)$, differentiate, and set ϵ to zero.

$$\frac{\partial}{\partial \epsilon} J_4[p_s(x) + \epsilon\delta(x - x^0)] \; |_{\epsilon=0}$$

$$= -\int \frac{p_d(x)}{\int p_s(y)p_n(x-y)dy} p_n(x - x^0)dx$$

$$+ \lambda_1[\int p_s(x)dx - 1] + \lambda_2[p_s(x^0) - |p_s(x^0)|] \tag{C.9}$$

Iteratively perturb $p_s(x)$ with the negative of this gradient; an inexpensive line search finds the correct magnitude. The constraints easily determine the values of λ_1 and λ_2 for any magnitude of perturbation. The second term raises or lowers all points of $p_s(x)$ equally until the constraint of unit area is satisfied. The third term moves each

point a large enough positive distance to remove any negative excursions. The first term divides the estimate $p_d(x)$ by the convolved value and cross correlates with a shifted noise distribution (contributed to the data distribution by the perturbation of $p_s[x^0]$). The cross correlation identifies the points at which the divergence is not uniform and compensates with appropriate perturbations.

APPENDIX D
DERIVATION OF A FOCUSING MEASURE

A number of authors have proposed measures of "simplicity," "sparseness," and "parsimony" that indirectly estimate the non-Gaussianity of data. Wiggins (1978) proposed a Varimax norm ratio that compared the probability that data resulted from a Gaussian rather than a generalized Gaussian distribution: $p(x) = C\exp(-|x|^4)$. Gray (1979) adapted a result by Hogg (1972) to define a "variable norm" ratio that compared the likelihood of two different generalized Gaussian distributions (see equation [3.1]). Mehta (1977) first proposed a more direct measure of non-Gaussianity using histograms of the data. He used quantiles to estimate the Gaussian distribution that fit the data best and then calculated the divergence of the histogram from this Gaussian distribution using the cross entropy function. I shall here derive a similar result that simplifies the calculation of the best fitting Gaussian.

Cross entropy (defined first by Kullback as directed divergence) measures the unpredictability of a given $p_1(x)$ with respect to some reference $p_2(x)$.

$$I\ [p_1(x) : p_2(x)] \equiv \int p_1(x)\ \log[p_1(x)/p_2(x)]\ dx \tag{D.1}$$

If p_2 is uniform, then cross entropy becomes the negative of Shannon's statistical entropy. I approaches the minimum value of 0 when p_1 is most like p_2. Define a measure of non-Gaussianity, F, as the minimum cross entropy of the data pdf's with respect to Gaussian distributions of all variances σ^2. F increases with non-Gaussianity and thereby with the focusing of the data. (Assume zero-mean processes for simplicity in notation. A mean is easily estimated and subtracted.) For a single pdf, define

$$F[p(x)] \equiv \min_{\sigma} I\ [p(x) : \text{Gaussian}(\sigma,x)] \tag{D.2}$$

$$= \min_{\sigma} \int p(x)\ \log[p(x)/(\frac{1}{\sqrt{2\pi}\sigma}\ e^{-\frac{x^2}{2\sigma^2}})]\ dx$$

$$= \min_{\sigma} \left\{ \int p(x)\ \log p(x)\ dx\ + \frac{1}{2\sigma^2}\int x^2 p(x)\ dx\ + \log\sigma + \log\sqrt{2\pi} \right\}$$

Not surprisingly, F attains this minimum value when the Gaussian distribution possesses the same standard deviation as $p(x)$:

$$\frac{d}{d\sigma}\ I[p(x) : Gaussian(\sigma,x)] = -\frac{1}{\sigma^3}\int x^2\ p(x)\ dx\ + \frac{1}{\sigma} = 0 \tag{D.3}$$

$$\rightarrow \sigma^2 = \int x^2\ p(x)\ dx\ .$$

Substitute this result into (D.2).

$$F[p(x)] \equiv \int p(x) \log p(x)\, dx + \frac{1}{2} \log \int x^2 p(x)\, dx + C \qquad \text{(D.4)}$$

$$\text{where } C = \log\sqrt{2\pi} + \frac{1}{2}$$

(D.4) provides a simpler, working definition of F. Notice that (D.4) is scale invariant: multiplying the random variable x by a constant a does not affect F.

$$F[\frac{1}{a}\, p(\frac{x}{a})] = F[p(x)] \qquad \text{(D.5)}$$

Finally, we may prove *a posteriori* that a Gaussian distribution minimizes (D.4). First replace $p(x)$ in (D.4) by a perturbed $(1-\epsilon)p(x) + \epsilon\delta(x-x_0)$. $\delta(x)$ is the Dirac delta function, with unit area. Then, setting the ϵ derivative equal to zero yields

$$-\int p(x) \log p(x)\, dx + \log p(x_0)\, dx + \frac{x_0^2}{2\sigma^2} - \frac{1}{2} = 0$$

$$\rightarrow \quad p(x_0) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x_0}{2\sigma 2}} \quad \text{and } F[p(x)] = 0\,. \qquad \text{(D.6)}$$

(D.6) gives the equation of a Gaussian. (D.3) again defines the variance. The constraint of unit area requires that the measure attain a minimum value of 0.

In practice, evaluate (D.4) from discrete histograms functioning as pdf's. Represent the sampled versions as $\{p_i\}$, defining each sample (indexed with i) as an average of $p(x)$ over a short interval of x. Assume that the $\{p_i\}$ are sampled N times per standard deviation.

$$p_i = \frac{1}{(\sigma/N)} \int \Pi\left(\frac{x - i\,\sigma/N}{\sigma/N}\right) p(x)\, dx$$

$$\text{where } \Pi(x) \equiv \begin{cases} 1 & -0.5 \leq x \leq +0.5 \\ 0 & \text{else} \end{cases} \qquad \text{(D.7)}$$

The sampling reduces $F[p(x)]$ by some ϵ made arbitrarily small by large N.

$$F[p(x)] = \int p(x) \log p(x)\, dx + \log\sigma + C \qquad \text{(D.8a)}$$

$$= \sum_i \frac{\sigma}{N} p_i \log p_i + \log\sigma + C + \epsilon$$

$$= \sum_i (\frac{\sigma}{N} p_i) \log(\frac{\sigma}{N} p_i) + \log N + C + \epsilon$$

Let $\{q_i = \sigma\, p_i / N\}$ be the probabilities of the amplitude bins. If a histogram selects a

fixed N *bins per standard deviation,* then the focusing measure equals Shannon's statistical entropy plus constants.

$$F\left[\{q_i\}\right] = \sum_i q_i \log q_i + \log N + C \tag{D.8b}$$

The focusing measure requires two inexpensive passes through the data: one, to find the standard deviation from a sum of squares; second, to calculate a coarse histogram scaled accordingly.