Probability and Entropy of Seismic Data

by Jon F. Claerbout

We model the reflectivity in a sedimentary column as a realization of an independent, non-Gaussian, random process. The means by which we observe the reflectivity involve time averaging (filtering) and space averaging (diffraction). Thus, the central limit theorem implies that our observations will be more Gaussian than the underlying random process. Conceivably the unknown filter (related to shot waveform) and diffraction (related to RMS velocity) could be based on maximizing the non-Gaussian-ness of inverse filtered and inverse diffracted (migrated) data.

The present chapter will focus on estimating an unknown probability density function $p(x)$ given n samples (X_i , $i=1,n$) drawn from $p(x)$. We defer the task of measuring departure from normality. However, before we go into details of estimation of probability functions let us review some basic facts.

Entropy S is defined as

$$S(p) = \int_{-\infty}^{+\infty} p(x) \ \ln p(x) \ dx \tag{1}$$

As our first review exercise let us show the well known result that the Gaussian probability density maximizes S provided that the variance σ^2 is known. Being known means that a constraint to the maximization of S is

$$\sigma^2 = \int_{-\infty}^{+\infty} x^2 \ p(x) \ dx \tag{2}$$

The first requirement every probability density function must satisfy is the unit area constraint

$$1 = \int_{-\infty}^{+\infty} p(x)\, dx \tag{3}$$

Introducing two Lagrange multipliers λ_0 and λ_1 for the constraints (3) and (2) we maximize (1) by maximizing

$$H = \int_{-\infty}^{+\infty} p \,\ell n\, p\, dx + \lambda_0 \left(1 - \int_{-\infty}^{+\infty} p(x)\, dx\right) + \lambda_1 \left(\sigma^2 - \int_{-\infty}^{+\infty} x^2\, p(x)\, dx\right) \tag{4}$$

Setting to zero the variation of H with respect to P gives

$$0 = \delta H = \int_{-\infty}^{+\infty} (\ell n\, p(x) + 1 - \lambda_0 - \lambda_1 x^2)\, \delta P\, dx \tag{5}$$

Since the variation of P is arbitrary the parenthesized expression in (5) must vanish identically. Hence,

$$p(x) = \exp(-1 + \lambda_0 + \lambda_1 x^2) \tag{6}$$

It remains to choose numerical values of λ_1 and λ_2 to ensure satisfaction of the constraints, however it is already clear that (6) is a Gaussian function.

If instead of the expectation of x^2 in (2) we had been given the expectation of $|x|$ then the probability density would turn out to be $\exp(-\text{const}\,|x|)$.

Now suppose that instead of being given moments we are given various quantiles. A quantile x_α is defined for any $0 \leq \alpha \leq 1$ as follows

$$\alpha \;=\; P\,(\,X < x_\alpha\,) \;=\; \int_{-\infty}^{x_\alpha} p(x)\;dx \tag{7a}$$

If $\alpha = .5$ then x_α is called the median or the 50^{th} percentile. We shall have occasion to work with a group of n quantiles. They will be ordered from smallest to largest and indexed by the integer i . Rather than subscript x_α with the index i we will write the defining equation (7a) as

$$\begin{aligned} \alpha(i,n) \;&=\; P\,(\,X < x(i,n)\,) \\ &=\; \int_{-\infty}^{x(i,n)} p(x)\;dx \\ &=\; \int_{-\infty}^{+\infty} \mathrm{step}\,[\,x - x(i,n)\,]\,p(x)\;dx \end{aligned} \tag{7b}$$

With this notation we can easily write a set of n Lagrange constraints for each quantile as

$$\sum_{i=1}^{n} \lambda_i \left\{ \alpha(i,n) - \int_{-\infty}^{+\infty} \mathrm{step}\,[\,x - x(i,n)\,]\,p(x)\;dx \right\} \tag{8}$$

Now the maximum entropy probability function turns out to be

$$p(x) \;=\; \exp\left\{ \lambda_0 + \sum_{i=1}^{n} \lambda_i\,\mathrm{step}\,[\,x - x(i,n)\,] \right\} \tag{9}$$

This probability function is a constant function of x except at the given percentiles $x(i,n)$ where the function jumps. The result is

$$p(x) = \frac{\alpha(i+1,n) - \alpha(i,n)}{x(i+1,n) - x(i,n)} \qquad \text{for} \quad x(i+1,n) > x > x(i,n) \quad (10)$$

Let us briefly consider a few other possibilities. Suppose instead of making the entropy $p \ln p$ stationary we considered making the integral of p stationary. Such an integral is called the information. Then the variance constraint would give

$$p(x) = \frac{1}{\lambda_1 + \lambda_2 x^2} \qquad (11)$$

which is a Cauchy density and has infinite variance (so we can't choose λ_2 unless the integrals terminate before infinite x). Again, minimizing $\ln p$ now under the quantile constraints we get

$$p(x) = \frac{1}{\lambda_0 + \Sigma \lambda_i \text{ step } [x-x(i,n)]} \qquad (12)$$

which although it looks different than (9) is in fact a constant function between the given percentiles, and hence, reduces exactly to (10).

Now we return to our central theme, given N data points X_i , $i=1, N$ find some estimated probability density function $\hat{p}(x)$. Later we will look for means to test $\hat{p}$ against the Gaussian. One possible procedure for estimating $\hat{p}$ would be to estimate the variance of the data and other higher moments. We could then determine the probability function which maximizes entropy. Another possible procedure would be to take the observations X_i and reorder them from smallest to largest. Let the reordered data points be denoted $X(i,N)$. If N is odd then $X[(i+1)/2, N]$ is the sample median and it would seem to be a good estimate of the true median. Likewise, $X(i/4, N)$ would seem to be a reasonable estimate of the 25^{th} percentile, etc.

The traditional approach seems to be moment expansions. That, however, seems particularly hazardous because of the bursty nature of seismic reflectivity. Furthermore, some probability functions like the Cauchy density arise in perfectly valid applications yet they have second and higher moments which are divergent. Thus, we will turn to the estimation of quantiles.

Suppose we have only one data point, X_1 . Then we have a 50% probability that this data point exceeds the median $x_{.5}$ of the unknown probability function $p(x)$. Thus, with only one data point our best estimate $\hat{x}_{.5}$ of the median $x_{.5}$ would be the given data point X_1 . Hence, even without knowledge of $p(x)$ we may assert that for this estimate

$$P(\hat{x}_{.5} < x_{.5}) = \frac{1}{2} \tag{13}$$

Likewise, suppose that we had three independent samples X_1 , X_2 , and X_3 from the unknown probability function. Then these could be reordered from smallest to largest, namely $X(1,3)$, $X(2,3)$, $X(3,3)$ and we could take the middle one as our median estimate $\hat{x}_{.5} = X(2,3)$.

It is still true regardless of the true $p(x)$ and the true unknown median $x_{.5}$ that $P(\hat{x}_{.5} < x_{.5}) = 1/2$.

Now suppose we have two data points X_1 and X_2 . Let us estimate some quantiles by the formulas

$$\hat{x}(2,2) = \max(X_1, X_2) \tag{12a}$$

$$\hat{x}(1,2) = \min(X_1, X_2) \tag{12b}$$

What quantiles are these? The answer is that $\max(X_1, X_2)$ is a good estimator of something like the 71^{st} percentile of the unknown probability function $p(x)$.

Rephrasing the basic idea, suppose every person in the world had two random numbers X_1 and X_2 from some $p(x)$. Then if each person asserts that the maximum of his two numbers is the 71^{st} percentile of $p(x)$ then (about) half of the people have underestimated the percentile and half of the people have overestimated it.

Let us now get the precise percentile. Recall equation (7a)

$$\alpha = P(X < x_\alpha) = \int_{-\infty}^{x_\alpha} p(x)\, dx \tag{7a}$$

For the maximum of two numbers to be less than x_α we must have both of the numbers less than x_α . The probability of two independent events is the product of the individual probabilities, so

$$P[\max(X_1, X_2) < x_\alpha] = P(X_1 < x_\alpha)\ P(X_2 < x_\alpha) = \alpha^2 \tag{13}$$

If half of the people in the world are supposed to underestimate while the other half overestimate, then we want to set (13) equal one half. Hence,

$$\alpha \;=\; (.5)^{.5} \;=\; .7071 \ldots \qquad (14)$$

A similar analysis with 3 data points shows that the ordered data points X(i,3) are good estimators of the quantiles

$$\alpha(i,3) \;=\; (\, 1 - 2^{-1/3} \,,\, 1/2 \,,\, 2^{-1/3} \,) \qquad (15)$$

Likewise, with n points the maximum X(n,n) gives the α value

$$\alpha \;=\; (\tfrac{1}{2})^{1/n} \;=\; \exp \frac{\ell n(1/2)}{n} \qquad (16a)$$

$$\approx\; 1 - \frac{.69}{N} \qquad (16b)$$

Thus, for one hundred points, the minimum estimates something smaller than the first percentile; it corresponds to an $\alpha = .0069$. You might say that the minimum is in a bin which goes from $\alpha = 0$ to $\alpha = .01$.

Now let us find the quantile levels $\alpha(i,n)$ which correspond to the ordered data points X(i,n) . Define the "upper tail probability" β as

$$\beta \;=\; P\,(\,X > x_{\alpha}\,) \;=\; 1 - \alpha \qquad (17)$$

Consider the expression

$$(\alpha + \beta)^2 \;=\; \alpha^2 + \alpha\,\beta \;+\; \beta\,\alpha \;+\; \beta^2 \qquad (18)$$

Each of the terms on the right side of (18) corresponds to a possible result of a two number draw. The term α^2 means both X_1 and X_2 were less than x_α . The term $\alpha\beta$ means X_1 was less but X_2 was more, etc. Next consider expansion of the expression $(\alpha+\beta)^n$ where like terms such as $\alpha\beta + \beta\alpha$ are not combined to $2\alpha\beta$ as they were not confined in (18). There will be 2^n terms, one for each possible result in an n number draw. To find the probability that just one number was in the upper tail area and the rest were in the lower tail area we gather all the terms with the combination $\alpha^{n-1}\beta$. To find the probability that one or none were in the upper tail area we gather the two terms

$$\alpha^n + n\,\alpha^{n-1}\beta$$

For the probability that 2 or fewer were in the upper tail area we have

$$\alpha^n + n\,\alpha^{n-1}\beta + \frac{n(n-1)}{2}\,\alpha^{n-2}\beta^2$$

The probability that $i-1$ or fewer are in the upper tail area defines the so-called incomplete beta function

$$\sum_{s=n-i+1}^{n} \binom{n}{s} \alpha^s (1-\alpha)^s = I_\alpha(n-i+1,\, i) \qquad (19)$$

We can set (19) equal one half and solve the resulting polynomial equation for α .