Relative Entropy Spectral Analysis slide notes from SEP-35 invited lecture by John Burg

Stewart A. Levin

I. What is relative entropy?
   A. Problem it solves
   B. Axiomatic derivation
   C. Properties

II. Specialization to spectral analysis
   A. Derivation of more general forms of maximum entropy
   B. Signal processing application

Relative entropy is a general method of inference about an unknown probability density, q, when there is an initial estimate of the probability density, p, and new information, I, in the form of expected values.

We make a few definitions:

x random variable

D domain of x

$\mathfrak{D}$ collection of all possible probability densities, $q(x)$, on D, i.e., $q \geq 0$ for $x \in D$ and

$$\int_D q(x)\,dx = 1 \quad .$$

$q^{\dagger}$ the true but unknown density

$p(x)$ the initial estimate of $q^{\dagger}$

I new information in the form of expected values

$\mathcal{I}$ all probability densities agreeing with I. $\mathcal{I} \subset \mathfrak{D}$, $q^{\dagger} \in \mathcal{I}$.

$q(x)$ the final density

$\circ$ the "information operation", used as $q = p \circ I$. This operator takes two arguments.

If $a_k(x)$, $k = 1, \ldots K$, are functions of the random variable x, then the true expected value of $a_k(x)$, $\bar{a}_k$, is give by

$$\bar{a}_k = \int_D a_k(x)\, q^{\dagger}(x)\, dx \qquad k = 1, \ldots, K \quad ,$$

where $q^{\dagger}(x)$ is the true density.

Information given in the form of some expected values, $\bar{a}_k$, thus allows us to place the linear equality constraints

$$\int_D a_k(x)\, q(x)\, dx = \overline{a}_k \qquad k = 1, \ldots, K$$

on the final probability density, $q(x) \in \mathcal{I}$.

Problem and solution

Given the initial probability density p, how do you choose q? There is only one logically consistent way of doing this. q must be chosen to minimize the relative entropy $H[q,p]$ given by

$$H[q,p] = \int q(x) \log \frac{q(x)}{p(x)}\, dx$$

subject to the constraint that q agrees with the expected values $\overline{a}_k$. Our notation $q = p \circ I$ is shorthand for this minimization.

This claim follows from four Consistency Axioms. These four axioms are based on the fundamental principle that if a problem can be solved in more than one way, the results should be consistent.

I. Uniqueness
Result should be unique

II. Invariance
The choice of coordinate system should not matter

III. System Independence
It should not matter whether one accounts for independent information about independent systems separately in terms of different densities or together in terms of a joint density. In terms of the $\circ$ operator:

$$(p_1 p_2) \circ (I_1 \wedge I_2) = (p_1 \circ I_1)(p_2 \circ I_2)$$

IV. Subset Independence
It should not matter whether one accounts for independent information about a subset of D in terms of a separate conditional density or in terms of the full probability density.

The $q \in \mathcal{I}$ that minimizes the relative entropy

$$H[q,p] = \int q \log \frac{q}{p} dx$$

satisfies these axioms uniquely.

In the references, relative entropy is termed *cross-entropy*, and the initial and final probability densities are called *prior* and *posterior* respectively.

Some properties of $\circ$:

1) $p = p \circ I$ if and only if $p \in \mathcal{I}$

2) $(p \circ I) \circ I = p \circ I$

3) Triangle equality: For any $r \in \mathcal{I}$

$$H[r,p] = H[r,q] + H[q,p]$$

where $q = p \circ I$. The special case $r = q^{\dagger}$ shows q is closer to $q^{\dagger}$ than p, i.e., $H[q^{\dagger},q] \leq H[q^{\dagger},p]$ with equality if and only if $q = p$.

4) Sequential new information

$$(p \circ I_1) \circ (I_1 \Lambda I_2) = p \circ (I_1 \Lambda I_2)$$

5) Remeasured information

$$(p \circ I) \circ I' = p \circ I'$$

where I' is a later measurement of the expected value given by I, e.g. I' wipes out I.

Application to spectral analysis

We shall derive the way to estimate the power density spectrum of a stationary Gaussian time series, given a prior spectral density $P(f)$ and exact autocorrelation information $R(n)$, $|n| \leq N$.

A simplistic way of looking at a stationary Gaussian time series, $y(t)$, is

$$y(t) = \sum_{m=1}^{M} \left(a_m \sin[2\pi f_m t] + b_m \cos[2\pi f_m t] \right)$$

where a_m and b_m are independent, zero mean, normally distributed with variance σ_m^2:

$$P(a_m) = \frac{1}{\sqrt{2\pi\sigma_m^2}} \exp[-\frac{a_m^2}{2\sigma_m^2}]$$

Since the average power of sin or cos is ½ of the peak-squared amplitude, the power at frequency f_m is σ_m^2.

Continuing with the simple-minded analysis, we see our power spectrum at f_m is proportional to σ_m^2. Thus our prior power spectrum determines the prior probability density for the a_m and b_m.

Now our autocorrelation sequence forms the expected values of

$$\sum \frac{a_m^2 + b_m^2}{2} \cos[2\pi f_m \tau]$$

This becomes

$$R(\tau) = \sum_1^M \sigma_m^2 \cos[2\pi f_m \tau]$$

Starting with an initial Gaussian distribution and adding the expected value information from the autocorrelation measurements, the relative entropy principle gives us a final probability density, also zero mean, independent Gaussian with variance at f_m of

$$\frac{1}{\dfrac{1}{\sigma_m^2} + \sum\limits_{-N}^{N} \lambda_n Z^n}$$

where the λ_n are Lagrange multipliers adjusted to match the given autocorrelation values and Z is the unit delay operator.

Thus our final estimate is

$$Q(f) = \frac{1}{\dfrac{1}{P(f)} + \sum\limits_{-N}^{N} \lambda_n Z^n} .$$

This looks very much like the normal maximum entropy solution. (If P is white it is.)

If the prior is N^{th} order autoregressive, then

$$Q(f) = \frac{1}{\sum\limits_{-N}^{N} \beta_n Z^n + \sum\limits_{-N}^{N} \lambda_n Z^n}$$

$$= \frac{1}{\sum\limits_{-N}^{N} (\beta_n + \lambda_n) Z^n}$$

For an arbitrary prior, $P(f)$, $Q(f)$ has no special form such as AR or ARMA.

Actually, we are really assuming a χ^2 distribution with two degrees of freedom on the initial power spectrum $P(f)$. Instead of assuming only 2 degrees of freedom per frequency, suppose you weight the spectral estimate by assuming $W(f)$ degrees of freedom at f. Then the resulting final spectrum is

$$Q(f) = \frac{1}{\dfrac{1}{P(f)} + \dfrac{2}{W(f)} \sum_{-N}^{N} \lambda_n Z^n} .$$

Suppose we have prior probability densities $P_S(f)$ and $P_N(f)$ of the signal and noise spectra. Let our new information be the autocorrelation $R(\tau)$ of the combined signal + noise. Relative entropy yields the final densities

$$Q_S(f) = \frac{1}{\dfrac{1}{P_S(f)} + \sum_{-N}^{N} \lambda_n Z^n}$$

$$Q_N(f) = \frac{1}{\dfrac{1}{P_N(f)} + \sum_{-N}^{N} \lambda_n Z^n}$$

with the *same* Lagrange multipliers λ_n.

Conclusion

Use of the relative entropy principle leads to the only logically consistent approach to spectral analysis.

References

Shore, J.E., 1981, Minimum cross-entropy spectral analysis, IEEE Trans. Acous., Speech, Signal Processing, vol. 29, no. 2, p. 230-237.

Shore, J.E., and Gray, R.M., 1982, Minimum cross-entropy pattern classification and cluster analysis, IEEE Trans. Pattern Anal. Machine Intell., vol. 4, p. 11-17.

Shore, J.E., and Johnson, R.W., 1980, Axiomatic derivation of the principle of the maximum entropy and the principle of minimum cross-entropy, IEEE Trans. Info. Theory, no. 1, p. 26-37.

Shore, J.E., and Johnson, R.W., 1981, Properties of cross-entropy minimization, IEEE Trans. Info. Theory, no. 2, p. 472-482.

Total number of words = 387243

Number of words of length 1 = 70659	Percentage = 18.25%
Number of words of length 2 = 76294	Percentage = 19.70%
Number of words of length 3 = 69817	Percentage = 18.03%
Number of words of length 4 = 42866	Percentage = 11.07%
Number of words of length 5 = 34370	Percentage = 8.88%
Number of words of length 6 = 32915	Percentage = 8.50%
Number of words of length 7 = 23717	Percentage = 6.12%
Number of words of length 8 = 17014	Percentage = 4.39%
Number of words of length 9 = 9819	Percentage = 2.54%
Number of words of length 10 = 5957	Percentage = 1.54%
Number of words of length 11 = 1960	Percentage = 0.51%
Number of words of length 12 = 1096	Percentage = 0.28%
Number of words of length 13 = 530	Percentage = 0.14%
Number of words of length 14 = 185	Percentage = 0.05%
Number of words of length 15 = 28	Percentage = 0.01%
Number of words of length 16 = 13	Percentage = 0.00%
Number of words of length 17 = 0	Percentage = 0.00%
Number of words of length 18 = 1	Percentage = 0.00%