

ECE 771

Lecture 13 – Maximum Entropy Estimation

Reading: Sections 11.1, 11.2, 11.4, 11.5, 11.6

The concept of entropy has been applied to estimation problems. Estimation is the art and science of computing a value when incomplete information is available. It is the incompleteness that makes the concept of maximum entropy useful. For many estimation problems, it is necessary to make assumptions about values which are not explicitly available. One possible choice is to assume that the values are such that the entropy is maximized.

Suppose we want to maximize $h(f)$ over all densities with the following constraints:

1. $f(x) \geq 0$
2. $\int_S f(x)dx = 1$
3. $\int_S f(x)r_i(x)dx = \alpha_i, i = 1, 2, \dots, m$. This is some kind of “moment” constraint.

We can “pseudo-solve” this as follows. Let

$$J(f) = - \int f \ln f + \lambda_0 \int f + \sum_{i=1}^m \lambda_i \int f r_i.$$

“Differentiate w.r.t. $f(x)$ ” (this is the part we are skimming on) and equate to zero:

$$\frac{\partial J}{\partial f(x)} = - \ln f(x) - 1 + \lambda_0 + \sum_{i=1}^m \lambda_i r_i(x) = 0$$

which leads to

$$f(x) = e^{\lambda_0 - 1 + \sum_{i=1}^m \lambda_i r_i(x)}$$

where the Lagrange multipliers are chosen to make f satisfy the constraints.

To show that this actually works, we will use an inequality approach. Let g be a density that satisfies the constraints. Then

$$\begin{aligned} h(g) &= - \int_S g \ln g \\ &= - \int_S g \ln \frac{g}{f} \\ &= -D(g||f) - \int_S g \ln f \\ &\leq - \int_S g \ln f \\ &= - \int_S g(\lambda_0 + \sum_{i=1}^m \lambda_i r_i) \\ &= - \int_S f(\lambda_0 + \sum_{i=1}^m \lambda_i r_i) \\ &= - \int f \ln f = h(f) \end{aligned}$$

Example 1 Suppose we have $EX = 0$ and $EX^2 = \sigma^2$. Then the distribution which maximizes the entropy is

$$f(x) = e^{\lambda_0 + \lambda_1 x + \lambda_2 x^2}$$

After finding the constants, we recognize the normal distribution. \square

Example 2 (Weighted dice) Suppose we have a six-sided die with $EX = \sum ip_i = \alpha$. Now suppose we throw n of these dice, and are told that the total number of spots showing is $n\alpha$. What proportion of the dice are showing face i ? (That is, what is the probability of face i ?)

We will find an approximate solution. How many ways are there that n dice can fall so that n_i dice show face i ? There are $\binom{n}{n_1, n_2, \dots, n_6}$ ways. The way that the dice fall collectively is called a *macrostate*, which is denoted by (n_1, n_2, \dots, n_6) . The particular value of each die is called the *microstate*. For each macrostate there are $\binom{n}{n_1, n_2, \dots, n_6}$ microstates. We want to find the most probable macrostate (subject to the constraint), so we wish to maximize $\binom{n}{n_1, n_2, \dots, n_6}$ subject to

$$\sum_{i=1}^6 in_i = n\alpha.$$

Using Stirling's approximation, $n! \approx (n/e)^n$ we find

$$\binom{n}{n_1, n_2, \dots, n_6} \approx \frac{(n/e)^n}{\prod_{i=1}^6 (n_i/e)^{n_i}} = \prod_{i=1}^6 (n/n_i)^{n_i} = e^{nH(n_1/n, n_2/n, \dots, n_6/n)}$$

Thus maximizing $\binom{n}{n_1, n_2, \dots, n_6}$ under the constraint is almost equivalent to maximizing $H(p_1, p_2, \dots, p_6)$ under the constraint $\sum ip_i = \alpha$. Using the previous results we find

$$p_i = \frac{e^{\lambda i}}{\sum_{i=1}^6 e^{\lambda i}}$$

where λ is chosen so that $\sum ip_i = \alpha$. The most probable macrostate is therefore $(np_1, np_2, \dots, np_6)$. \square

Example 3 Let $S = [0, \infty)$ and let $EX = \mu$. Then the entropy-maximizing distribution may be shown to be

$$f(x) = \frac{1}{\mu} e^{-x/\mu} \quad x \geq 0.$$

\square

Spectrum estimation

A problem of ongoing interest in signal processing is to estimate the spectrum of a signal, given its samples (which are often noisy). A large variety of techniques have been developed for this purpose. If the autocorrelation function

$$R(k) = EX_i X_{i+k}$$

is known for all k , then the spectrum (more strictly, the power spectral density) can be computed as the Fourier transform of the autocorrelation function:

$$S(\omega) = \sum_{m=-\infty}^{\infty} R(m) e^{-jm\omega} \quad -\pi < \omega \leq \pi.$$

In practice, we observe only n samples and can only estimate the autocorrelation values by an estimator such as

$$\hat{R}(k) = \frac{1}{n-k} \sum_{i=1}^{n-k} X_i X_{i+k}.$$

This is the *periodogram* method, and it does not converge to the true power spectrum for large n . At large values of k (lags), the estimate has only a few samples to deal with. The inaccuracies can be covered by setting autocorrelations at large lag to zero. However, this abrupt change introduces spectral artifacts. The autocorrelation function could also be windowed, but that can lead to negative power spectrum estimates.

Instead of setting the values to zero, one suggested approach is to set them to values that make the *fewest assumptions about the data*, i.e., which maximize the entropy rate of the process. If the data are assumed to be stationary and Gaussian, this corresponds (as we will see) to an AR process. This approach (due originally to Burg) is of wide application. The model-estimation approach that arises is commonly used, for example, for efficient coding of speech parameters.

We first need to look at the entropy rate of a Gaussian process.

Definition 1 The **differential entropy rate** of a stochastic process is

$$h(\mathcal{X}) = \lim_{n \rightarrow \infty} \frac{h(X_1, X_2, \dots, X_n)}{n},$$

provided that the limit exists. □

We can also write this as

$$h(\mathcal{X}) = \lim_{n \rightarrow \infty} h(X_n | X_{n-1}, \dots, X_1).$$

For a stationary Gaussian process with covariance K we have

$$h(X_1, \dots, X_n) = \frac{1}{2} \log(2\pi e)^n |K^{(n)}|$$

where $K^{(n)}$ is the Toeplitz covariance matrix with entries $R(0), R(1), \dots, R(n-1)$ along the top row, and $K_{ij}^{(n)} = R(|i-j|)$. As $n \rightarrow \infty$ the density of the eigenvalues of the matrix tends to a limit (Szegő's theorem), which is the spectrum of the stochastic process. It has been shown (Kolmogorov) that the entropy rate of a stationary Gaussian stochastic process can be expressed as

$$h(\mathcal{X}) = \frac{1}{2} \log 2\pi e + \frac{1}{4\pi} \int_{-\pi}^{\pi} \log S(\lambda) d\lambda.$$

Using the formulation $h(\mathcal{X}) = \lim_n h(X_n | X^{n-1})$, and the fact that a Gaussian conditioned on Gaussians is Gaussian, we have that $h(\mathcal{X})$ must be the entropy of some Gaussian distribution with entropy $\frac{1}{2} \log 2\pi e \sigma_\infty^2$, where σ_∞^2 is the variance in the error of the best estimate of X_n given the infinite past.

We can now present Burg's result.

Theorem 1 *The maximum entropy rate stochastic process $\{X_i\}$ satisfying the constraints*

$$EX_i X_{i+k} = \alpha_k \quad k = 0, 1, \dots, p$$

is the p th order Gauss-Markov process of the form

$$X_i = - \sum_{k=1}^p a_k X_{i-k} + Z_k$$

where the Z_i are i.i.d. $\mathcal{N}(0, \sigma^2)$ and a_1, a_2, \dots, a_p and σ^2 are chosen to satisfy the correlation constraints.

Note: we have not assumed that X_i is Gaussian, zero-mean, nor stationary.

Proof Let X_1, \dots, X_n be a stochastic process with the given correlation values. Let Z_1, \dots, Z_n be a Gaussian process with the same covariance as X_1, \dots, X_n . We have

$$\begin{aligned} h(X_1, \dots, X_n) &\leq h(Z_1, \dots, Z_n) \\ &= h(Z_1, \dots, Z_p) + \sum_{i=p+1}^n h(Z_i | Z_{i-1}, \dots, Z_1) \\ &\leq h(Z_1, \dots, Z_p) + \sum_{i=p+1}^n h(Z_i | Z_{i-1}, \dots, Z_{i-p}) \end{aligned}$$

Now define Z'_1, Z'_2, \dots, Z'_p as a p th order Gauss-Markov process with the same distribution as Z_1, \dots, Z_p for all orders up to p . Then $h(Z_i | Z_{i-1}, \dots, Z_{i-p}) = h(Z'_i | Z'_{i-1}, \dots, Z'_{i-p})$. We continue the chain of inequalities

$$\begin{aligned} h(X_1, \dots, X_n) &\leq h(Z'_1, \dots, Z'_p) + \sum_{i=p+1}^n h(Z'_i | Z'_{i-1}, \dots, Z'_{i-p}) \\ &= h(Z'_1, Z'_2, \dots, Z'_n). \text{ why?} \end{aligned}$$

Then

$$\lim_{n \rightarrow \infty} \frac{1}{n} h(X_1, \dots, X_n) \leq \lim_{n \rightarrow \infty} \frac{1}{n} h(Z'_1, \dots, Z'_n) = h^*$$

where $h^* = \frac{1}{2} \log 2\pi e \sigma^2$. Hence the maximum entropy rate stochastic process satisfying the constraints is the Gauss-markov process.

□

Summary: the entropy of a finite segment of a stochastic process is bounded above by the entropy of a Gaussian process with the same covariance, which in turn is bounded above by the variance of a minimal order Gauss-Markov process with the given covariance constraints.

Now, how do we select the parameters a_1, \dots, a_p and σ^2 . Multiply the

$$X_i = - \sum_{k=1}^p a_k X_{i-k} + Z_i$$

X_{i-l} and take expectations:

$$R(0) = - \sum_{k=1}^p a_k R(k) + \sigma^2$$

$$R(l) = - \sum_{k=1}^p a_k R(l-k), \quad l = 1, 2, \dots,$$

This gives rise to the Yule-Walker equations.

Having determined the values of a_i the spectrum is

$$S(\omega) = \frac{\sigma^2}{|1 + \sum_{k=1}^p a_k e^{-jk\omega}|^2}.$$

Show this!