Maximum Entropy Priciple

Jayanti Prasad
Inter-University Centre for Astronomy & Astrophysics
Pune, INDIA, 411007

February 17, 2016

1 Entropy

The standard definition of entropy, as given bu Shannon is:

$$S = -\sum_{i=1}^{N} p_i \log p_i, \tag{1}$$

where $0 < 1 < p_i$ is the probability and so $S > 0$. Note that this entropy is just -ve of the Shannon information I, i.e., $I = -S$.

In order to understand Entropy we can consider a four cell case and the three probability distribution $P_A = (1/4, 1/4, 1/4, 1/4), P_B = (1, 0, 0, 0), P_C = (1/2, 1/4/, 1/4, 0)$ for which we can compute entropy in the following way:

$$\begin{aligned} S_A &= -4 \times \frac{1}{4} \log \left(\frac{1}{4}\right) = \log 4 = 2 \log 2 \\ S_B &= -1 \times \log 1 = 0 \\ S_C &= -\frac{1}{2} \log \left(\frac{1}{2}\right) - 2 \times \frac{1}{4} \log \left(\frac{1}{4}\right) = \frac{3}{2} \log 2. \end{aligned} \tag{2}$$

From this it is clear that the probability distribution A which has least structural information has the maximum entropy and B which has the maximum information (least uncertainty) has the minimum entropy and case C is between these two extremes.

It is extremely important to keep in mind that the entropy/information as defined above is a pure number and should not have any dimension. In fact in most cases it is expressed in terms of log and can be converted to bits/bytes also [1].

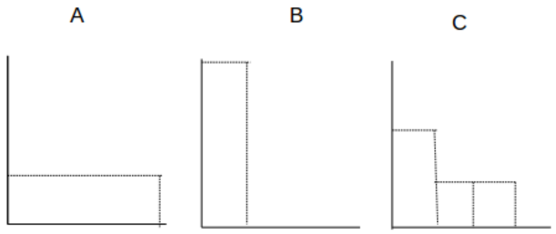

Figure 1: Entropy depends on the probability distribution.

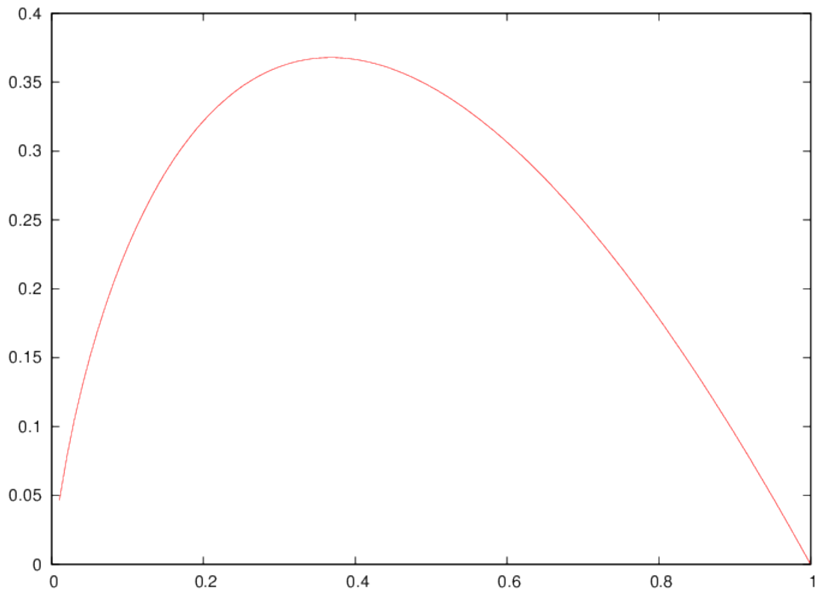

Figure 2: The argument $-x \log x$ of the entropy is a convex function for $0 < x < 1$.

Entropy as defined above can be used for regularization in the following way. Let us consider $C(f)$ is the χ^2 and λ is a regularization parameter so we try to maximize $M(f)$ as defined:

$$Q(f) = S(f) - \lambda C(f). \tag{3}$$

If we define $S(f) = -\sum p_i \log p_i$ with $p_i = f_i / \sum f_I$ then we can

$$\frac{\partial S}{\partial f_i} = \frac{1}{\sum f_i} \left[\log A - \log f_i \right], \tag{4}$$

with

$$\log A = \log(\sum f_i) - 1; \quad \text{or} \quad \sum f_i = Ae. \tag{5}$$

Note that here we have introduced a parameter A which is necessary to convert any distribution to a probability distribution and its value is decided by the distribution itself and not given from outside. In other words A is a normalization parameter which makes entropy positive. In terms of A we can write:

$$S(f) = -\sum \frac{f_i}{Ae} \log \left(\frac{f_i}{Ae} \right) \tag{6}$$

Note that $Ae = \sum f_i$ is always greater than any f_i so the entropy given by the above expression is always positive as expected. Note that if we start giving the value of A by hand then there is no guarantee that entropy will be positive ! For example for $A < f_i/e$ we get negative entropy.

We can find out the maximum entropy solution i.e., which will minimize chi-square and maximize entropy by:

$$\frac{\partial M}{\partial f_i} = 0, \tag{7}$$

we get :

$$\log A - \log f_i = \lambda(\sum f_i) \frac{\partial C}{\partial f_i} \tag{8}$$

and by solving this we can find the maximum entropy solution. Since the above equation is non-linear so we must solve it iteratively:

$$f_i^{n+1} = A \exp \left[-\lambda(\sum f_i) \frac{\partial C(f^n)}{\partial f_i} \right] \tag{9}$$

Although exponential guarantees that the solution will always remain positive but it also introduces instability which can be fixed by modifying the solution as:

$$f_i^{n+1} = (1-p)f_i^n + pA\exp\left[-\lambda(\sum f_i)\frac{\partial C(f^n)}{\partial f_i}\right], \tag{10}$$

where p is a parameter which can be tuned.

2 Some other solutions

Our goal is to solve :

$$\frac{\partial Q}{\partial f_i} = 0, \;\text{ with }\; Q(f) = S(f) - \lambda C(f). \tag{11}$$

One of the common methods to solve such optimization problem is to use gradient based methods line steepest descent, Newton-Raphon or conjugate gradient, however, there are some issue with these methods.

Steepest descent solution can be written as:

$$f_i^{n+1} = f_i^n + x\frac{\partial Q(f_i^n)}{\partial f_i}, \tag{12}$$

which may give negative values of f_i when the gradient term is large and negative. The same is true for Newton-Raphson also.

In conjugate gradient method at every step we move along the direction which is orthogonal to the directions we have moved in earlier steps and so this insures that we will find solutions in n steps for a problem in n dimensions. Anyway, this method also does not guarantee that the solutions will remain positive.

3 Skilling and Bryan method

In place of searching along one direction as is done in conjugate gradient Skilling and Bryan [2] suggested to construct a subspace at every step and carry out the search within that. The subspace can be constructed from the derivatives of entropy S and chi-square C in the way discussed below.

Within the subspace a quadratic model is constructed for Q

$$\tilde{Q}(x) = Q_0 + Q_\mu x^\mu + \frac{1}{2}H_{\mu\nu}x^\mu x^\nu. \tag{13}$$

If the basis of the subspace are given by e^i with $i = 1, .., r$ then

$$\begin{aligned} \delta f &= x^\mu e_\mu \\ Q_\mu &= e_\mu^T.\nabla Q \\ H_{\mu\nu} &= e_\mu^T.\nabla\nabla Q.e_\nu, \end{aligned} \tag{14}$$

with Q_μ and $H_{\mu\nu}$ agreeing with the local gradient and curvature of Q. Within the subspace $\tilde{Q}$ is maximized with:

$$x_\mu = -(H_{\mu\nu}^{-1})Q_\nu. \tag{15}$$

In general we can always approximate S and C some qudratic function (at least localy in some region):

$$\begin{aligned} C(x) &= C_0 + C_\mu x^\mu + \frac{1}{2}h_{\mu\nu}x^\mu x^\nu \\ S(x) &= S_0 + S_\mu x^\mu - \frac{1}{2}g_{\mu\nu}x^\mu x^\nu. \end{aligned} \tag{16}$$

where $C_\mu, S_\mu, h_{\mu\nu}$ and $g_{\mu\nu}$ are the gradiants and Hessian's of C and S respectively.

According to [2] $g_{\mu\nu} = \mathbf{e}_{\mu}^{T}.\mathbf{e}_{\nu}$ is an identity matrix (see equation (23)).

If the expansion is done in subspace (as is the case for Skilling & Bryan) then C_μ and S_ν are the components of the ∇C and ∇S along the axes of the subsapce:

$$\begin{aligned} S_\mu &= \mathbf{e}^T.\nabla S \\ C_\mu &= \mathbf{e}^T.\nabla C, \end{aligned} \tag{17}$$

where e represents the basis vectors of the subspace.

$$\begin{aligned} S &= -\sum_{i=1}^{N} \frac{f_i}{Ae} \log\left(\frac{f_i}{Ae}\right) \quad \text{with} \;\; F = Ae \\ \frac{\partial S}{\partial f_i} &= \frac{1}{Ae} \log\left(\frac{A}{f_i}\right) \\ \frac{\partial^2 S}{\partial f_i \partial f_j} &= -\frac{1}{Ae}\left(\frac{\delta_{ij}}{f_i}\right) \end{aligned} \tag{18}$$

References

[1] Joy A. Thomas, Thomas M. Cover, *Elements of Information Theory*, Wiley Interscience (1991).

[2] Skilling J. and Bryan R. K., MNRAS **211**, 111-124 (1984).

[3] Skiling J. (1984), *Maximum Entropy and Bayesian Method in Applied Science*, Page 129 (Cambridge University Press).

[4] S. H. Suyu, P. J. Marshall,M. P. Hobson, and R. D. Blandford, MNRAS **371**, 983-998 (2006).

[5] Golub G. H. and Van Loan C. F., *Matrix Computation, 4th Ed* (The John Hopkins University Press).

[6] Gaurav Goswami and Jayanti Prasad, Phys. Rev. D **88**, 023522 (2013).