Information Theory and Predictability
Lecture 7: Gaussian Case

1 Gaussian distribution

The general multivariate form of this density for a random n dimensional vector $\mathbf{x}$ can be written as

$$p(\mathbf{x}) = \left[(2\pi)^n \det(\boldsymbol{\sigma})\right]^{-1/2} \exp\left[-\frac{1}{2}\left(\mathbf{x}-\overline{\mathbf{x}}\right)^t \boldsymbol{\sigma}^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}\right)\right]$$

where $\boldsymbol{\sigma}$ is the $n \times n$ covariance matrix for the random variables which is given by:

$$\sigma_{ij} \equiv \overline{x_i x_j} - (\overline{x_i})(\overline{x_j})$$

Notice that such a distribution is specified exactly by its mean vector and covariance matrix.

Such distributions are encountered frequently in systems with many degrees of freedom where random variables often have such distributions to a good approximation. One possible reason why this occurs is the central limit theorem of statistics. In classical form this states that given n iid random variables with a general distribution of finite variance then the average of these random variables has a distribution that approaches a Gaussian as $n \to \infty$. The condition that the random variables be iid can be weakened considerably and the result still holds. Since such distributions are ubiquitous it is of value as an analysis tool to consider their entropic functionals. This is further facilitated by the ease with which analytical expressions may be obtained and elementary linear algebra techniques used in their analysis.

2 Entropic functionals

Consider firstly the differential entropy. This is easily evaluated because the logarithm of the distribution is a quadratic form plus another constant piece. The expectation of this quadratic form actually reduced to a constant and so only the normalization of the distribution contributes anything of interest:

$$h(\mathbf{X}) = \frac{1}{2}\log\left[(2\pi e)^n \det(\boldsymbol{\sigma})\right]$$

Consider now a linear transformation of our vector random variable that diagonalizes the covariance matrix (this is always possible since it is symmetric). These transformed random variables are commonly called EOFs or principal components in the geophysical and statistical literature respectively. In such a basis it is clear that the differential entropy is simply the sum of the logarithms of all the standard deviations of the principal components plus a constant. This corresponds nicely to our intuition of entropy as total uncertainty.

The relative entropy can be evaluated almost as easily although now the different means of the two Gaussian distributions comes into play and we obtain the following analytical expression

$$\begin{array}{rcll} D(p||q) & = & \frac{1}{2}\left[\log\left(\det\left(\boldsymbol{\sigma}_y\right)/\det\left(\boldsymbol{\sigma}_x\right)\right)+tr\left(\boldsymbol{\sigma}_x\left(\boldsymbol{\sigma}_y\right)^{-1}\right)-n\right] & \textit{Dispersion} \\ & + & \frac{1}{2}(\overline{\mathbf{x}}-\overline{\mathbf{y}})^t\boldsymbol{\sigma}_y^{-1}(\overline{\mathbf{x}}-\overline{\mathbf{y}}) & \textit{Signal} \end{array} \tag{1}$$

where the subscripts x and y refer to the distributions p and q respectively while $\mathbf{x}$ and $\mathbf{y}$ are the respective values of the random vectors. We have deliberately separated this expression into two pieces one depending only on the covariance of the p distribution while the other piece only depends on the means of this distribution. Notice that the first piece of the "dispersion" terms is simply the difference of the differential entropies of p and q i.e. it is the difference in the uncertainties of the two distributions. The second term is often small in the applications we shall consider later. The "signal" term can be understood by transforming to the principal component basis for the q distribution. In that basis this term is proportional to the sum of the squares of the difference in means normalized by the variances of the principal components.

Finally it is interesting to evaluate the mutual information of two Gaussian distributions since it is a measure of their independence as random variables. Since this is the relative entropy of the joint distribution with the product of the marginal distributions we can use equation (1) for its evaulation. The means of these two distributions are the same so the signal term vanishes. Additionally it is easily shown by elementary matrix manipulation that the second and third terms of the dispersion cancel. We are left therefore with just the first dispersion term and we can reduce this to

$$I(X;Y)=\frac{1}{2}\log\left(\frac{\det(\sigma_X)\det(\sigma_Y)}{\det(\sigma)}\right)$$

where the subscripted covariance matrices are those for the appropriate marginal distributions i.e. for X and Y. The unsubscripted matrix is for the full $2n\times 2n$ dimensional covariance matrix. Note that when $n=1$ this reduces to

$$I(X;Y)=\log\sqrt{\frac{1}{1-r^2}}$$

where r is the usual correlation coefficient. We could use this form and the previous equation to define a multivariate correlation coeï¬cient but this is not yet common in statistical circles.