Infomax[1]

5 March 2008

Infomax is an algorithm due to Tony Bell and Terry Sejnowski for solving the cocktail party problem. It is interesting both because the cocktail party problem itself is interesting and because the algorithm is very successful but slightly mysterious, hinting of further undiscovered mathematical depths.

I am always reminded of the cocktail party problem at yoga class. During relaxation at the end it is always amazing to realize how much background there is, background noise I was completely unaware of during the earlier part of the class: we say that the brain can solve the cocktail party problem in that we can attend to an acoustic signal even in noisy acoustic environments. We are able to separate a signal from its background. This is a paradigmical example in a large class of important modern problem related to the extraction of salient information from data.

Here, we will look at the simplest example of the cocktail party problem with two sources and two recordings; the challenge is to recover the source signals from the recordings. The two sources are $s_1(t)$ and $s_2(t)$; we ignore temporal structure and assume that at a given time $s_i(t)$ is drawn from a random variable S_i for $i = 1, 2$ with probability distribution p_{S_i}. We also assume the two sources are independent, $p_{S_1,S_2} = p_{S_1}p_{S_2}$. Finally, we assume they are mixed linearly so that we record $r_1(t)$ and $r_2(t)$ with

$$\mathbf{r}(t) = M\mathbf{s}(t) \tag{1}$$

where M is a constant mixing matrix. Now, we want to find an unmixing matrix W so that knowing

$$\mathbf{x}(t) = W\mathbf{r}(t) \tag{2}$$

is as good as knowing the sources; precisely, since $\mathbf{x} = WM\mathbf{s}$ we want WM to be a diagonal matrix multiplied by a permutation matrix, we do not mind if unmixing changes the overall amplitude of the source, or if it reorders the sources. In this two-to-two example, that means

$$MW = \mathrm{diag}\,(d_1, d_2) \tag{3}$$

or

$$MW = \left(\begin{array}{cc} 0 & d_1 \\ d_2 & 0 \end{array}\right) \tag{4}$$

where d_1 and d_2 are real numbers. Hence:

$$\mathbf{s} \overset{\text{mixing}}{\longrightarrow} \mathbf{r} = M\mathbf{s} \overset{\text{unmixing}}{\longrightarrow} \mathbf{x} = W\mathbf{r} \tag{5}$$

[1] Conor Houghton, houghton@maths.tcd.ie please send me any corrections.

One difficulty with looking at this problem is that it involves continuous random variables, whereas the course so far has concentrated on the discrete case: $s_1(t)$ is a continuous variable. For this reason some of the results will just be quoted and this part of the course is not examinable. The main point with continuous random variables is that the probability distribution is now a density, so

$$Pr(a < x < b) = \int_a^b dx p_X(x) \tag{6}$$

with

$$\int_{-\infty}^{\infty} dx p_X(x) = 1 \tag{7}$$

An important difference is demonstrated by the following formula: if $Y = g(X)$ are two random variables related by an invertible function g

$$p_Y(y) = \frac{p_X(x = g^{-1}(y))}{|g'(x)|} \tag{8}$$

The information is defined in the same way

$$H(X) = -\int dx p(x) \log p(x) \tag{9}$$

but it is no longer always positive. Furthermore, by substituting from the formula above, you can see, (C&T Theorem 8.6.4), that

$$H(aX) = H(X) + \log |a| \tag{10}$$

and so the entropy has no maximum or minimum: a similar idea is the maximization of entropy over distributions with the same variance and this would rule out the scaling we see in this formula.

Now, the idea is to solve the problem by using the fact that S_1 and S_2 are independent: we just need to find W so that X_1 and X_2 are also independent. One approach might be to decorrellate the random variables:

$$C(X_1, X_2) = E_{(X_1,X_2)}[(X_1 - EX_1)(X_2 - EX_2)] \tag{11}$$

where the expectation value for continuous random variables has the obvious definition

$$E_X g(X) = \int dx p_X(x) g(x) \tag{12}$$

It is easy to check that the correlation vanishes if X_1 and X_2 are independent, however, the flaw in this approach is that the converse is not true, it is possible to have zero correlation while still having statistical dependence. To see this, imagine, for convenience and without loss of generality, that EX_1 and EX_2 are zero and that we have chosen W so that the correlation matrix is the identity:

$$C_{ab} = C(X_a, X_b) = \mathbf{1} \tag{13}$$

then it is easy to see that rotations

$$\left(\begin{array}{c} X_1' \\ X_2' \end{array}\right) = \left(\begin{array}{cc} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{array}\right)\left(\begin{array}{c} X_1 \\ X_2 \end{array}\right) \tag{14}$$

do not change the correlation matrix. Thus, the decorrelation prescription has a rotational ambiguity and something more is needed. That something, of course, is to require $I(X_1, X_2) = 0$ since this happens if and only if X_1 and X_2 are independent.

The problem is that $I(X_1, X_2)$ is difficult to calculate: the idea behind infomax is to look at $H(X_1, X_2)$:

$$I(X_1, X_2) = H(X_1) + H(X_2) - H(X_1, X_2) \tag{15}$$

The idea is that maximizing the joint entropy, $H(X_1, X_2)$ will give a minimum of the mutual information, in other words, the variations in the individual entropies $H(X_1)$ and $H(X_2)$ can be ignored. However, as state, this will not work because the entropy can be increased by a trivial scaling, $X_a \to \lambda X_a$, changes the joint entropy, $H(X_1, X_2) \to H(X_1, X_2) + \log|\lambda|$ so $H(X_1, X_2)$ can be made arbitrarily large by scaling, something that tells us nothing about mixing and unmixing. Inspired by the behaviour of neurons, Bell and Sejnowski solved this by adding a saturation non-linearity:

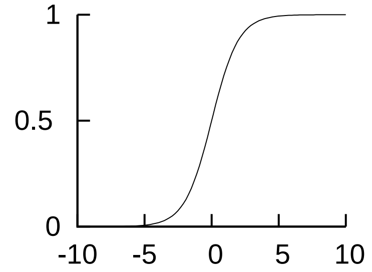
$$\begin{array}{rcl} y_1 & = & g(x_1 + w_1) \\ y_2 & = & g(x_2 + w_2) \end{array} \tag{16}$$

where w_1 and w_2 are parameters and, for example,

$$g(u) = \frac{1}{1 + e^{-u}} \tag{17}$$

so $g : (-\infty, \infty) \to (0, 1)$.

$$\tag{18}$$



Now we have

$$\mathbf{s} \overset{\text{mixing}}{\longrightarrow} \mathbf{r} = M\mathbf{s} \overset{\text{unmixing}}{\longrightarrow} \mathbf{x} = W\mathbf{r} \overset{\text{non-linearity}}{\longrightarrow} \mathbf{y} : y_a = g(x_a + w_a) \tag{19}$$

For later notational convenience, let write $y_a = g(x_a + w_a) = f(r_1, r_2; W, w_a)$ where f is the function, parameterized by W and w_a, mapping from the recording to y. So reason the saturating non-linearity will help is that with a non-linearity like this a large enough scaling will reduce the entropy: consider $H(g(\lambda X))$ where λ is a constant. If λ is very large it will spread X out so that it will take lots of very big positive or negative values; if λx is a big positive number then $g(\lambda x)$ will be near to one, conversely, if it is a large negative number, $g(\lambda x)$ will be near to zero: a distribution that is often zero and often one is not very entropic.

Before considering the source separation problem, lets look at the effect of the non-linearity on its own: we consider the one-to-one case

$$r \overset{\text{multiply}}{\longrightarrow} x = Wr \overset{\text{non-linearity}}{\longrightarrow} y = g(x + w) = f(r; w, W) \tag{20}$$

where W and w are now both scalars and r, x and y are outcomes for random variables R, X and Y. We consider maximizing the entropy $H(Y)$. What does this do; well, it maximizes the information in Y about R:

$$I(R; Y) = H(Y) - H(Y|R) \tag{21}$$

but $H(Y|R)$ is constant since R determines Y. In the discrete case we are familiar with this would be easy to discuss, $H(Y|R)$ would be zero, in the continuous case it is not that simple, it is actually minus infinity, but, the consequence is the same, it does not depend on W and w. To maximize $H(Y)$ we need to calculate its derivative with respect to the parameters W and w; this is a feature of the algorithm, ultimately we need to calculate derivatives and these are calculable and well defined, even if the quantity being differentiated is not. In this case

$$H(Y) = -\int dy p(y) \log p(y) \tag{22}$$

and this is estimated by

$$\tilde{H(y)} = -\log p(y) \tag{23}$$

In other words, if n values y_i is drawn from Y then

$$\frac{1}{n} \sum_i \tilde{H(y_i)} \to H(Y) \tag{24}$$

as n gets large.

Of course we do not have $p_Y(y)$ and it would be difficult to estimate, but, it turns out we do not need it to get the derivative of $\tilde{H(y)}$. We have seen already that since $y = f(r; W, w)$

$$p_Y(y) = \frac{p_R[r = f^{-1}(y)]}{|f'(f^{-1}(y)|} \tag{25}$$

so

$$\tilde{H(y)} = -\log p_R(r) + \log |f'| \tag{26}$$

and $p_R(r)$ is independent of the parameters. Now, for our choice of saturating non-linearity

$$\begin{aligned} g(u) &= \frac{1}{1+\exp(-u)} \\ \frac{dg}{du} &= g(1-g) \end{aligned} \tag{27}$$

and hence

$$\log|f'| = \log W + \log f + \log(1-f) \tag{28}$$

and hence,

$$\frac{dH\tilde{(}y)}{dW} = \frac{1}{W} + \frac{1}{f}rf(1-f) - \frac{1}{1-f}rf(1-f) = \frac{1}{W} + r(1-2y) \tag{29}$$

and

$$\frac{dH\tilde{(}y)}{dw} = 1 - 2y \tag{30}$$

These quantities: s, y and, of course, W, are numbers we have access to, W is a parameter, s is the signal, we can sample $s(t)$ at a set of times to get a set of s's and y is a function of s. This means we can estimate these derivatives, giving the gradient of H at a point (W, w) in the parameter space. This is a common situation in numerical optimization, we don't know the function and, here, it isn't even so easy to define, but we do know its gradient. This means that locally we know which direction brings us in the direction of greater H and so numerical hill-climbing routines can be used, these are well described in, for example, *Numerical Recipes in C++*: steepest ascent, conjugate gradient or metric gradient methods work well here.[2]

The idea then is to chose a starting W and w; estimate the gradient and then change W and w a small amount, repeating until the optimum values are found. What would the optimum value look like, well we know

$$p_Y(y) = \frac{p_R(r)}{|f'(r)|} \tag{31}$$

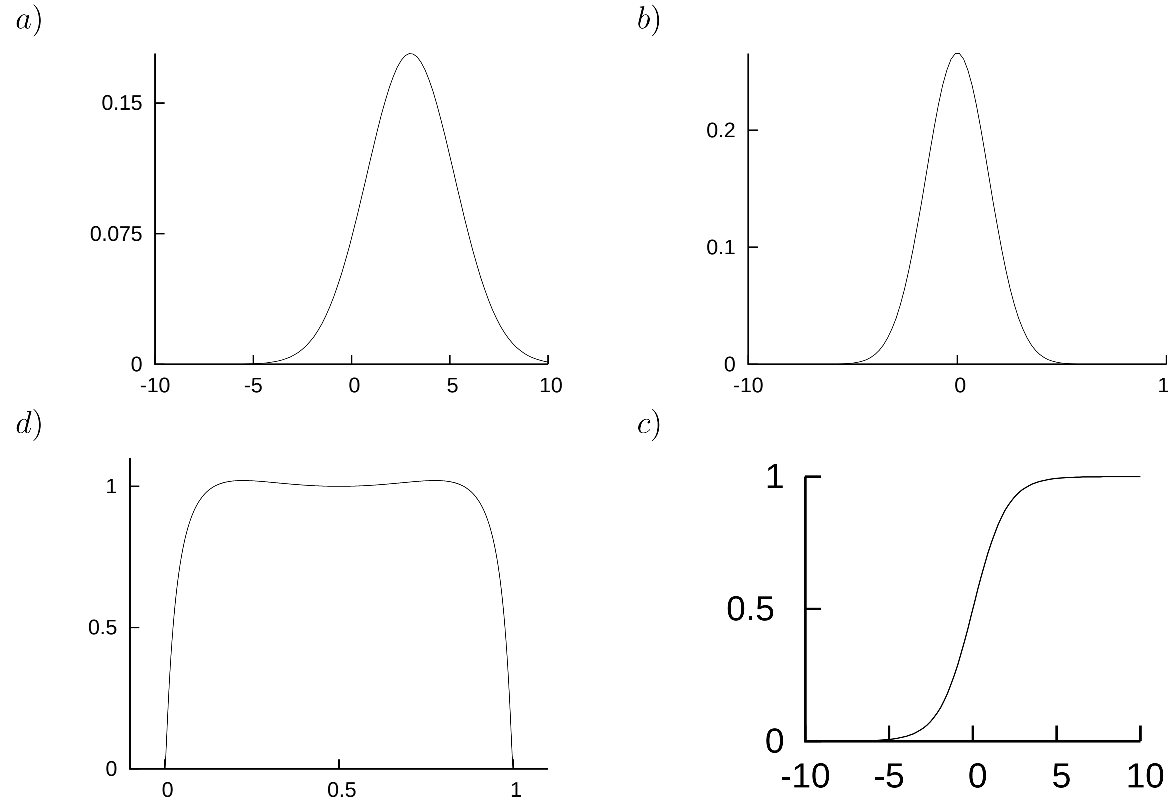
with $r = f^{-1}(y)$. Hence, if Y is evenly distributed on its interval $(0, 1)$ and $f'(r)$ is always positive then

$$f(r) = \int_{-\infty}^{r} p_R(u)du \tag{32}$$

Now this is not what we do, we do not know $p_R(r)$ and we chose $f(r)$ at the start, here it a member of a two-parameter family of functions parameterized by W and w. However, ideally, if the derivative of the saturating non-linearity is somewhat close to the distribution

[2] The metric method uses a slightly surprising, but effective, choice of metric and is due to Amari.

of R then infomax will find the W and w that line everything up so that Y will have something close to an even distribution. Here is an example:



(33)

Figure a) shows an initial distribution

$$p_R(r) = \frac{1}{\sqrt{10\pi}} e^{(r-3)^2/10} \tag{34}$$

b) is a new distribution arrived at by shifting a rescaling r using W and w:

$$p_U(u) = \frac{1}{\sqrt{4.5\pi}} e^{u^2/4.5} \tag{35}$$

where $u = Wr + w$. Now, this distribution is all lined up with the rectifying non-linearity c) so that $p_Y(y)$ where $y = g(u)$ is d).

Now, back to the two-to-two case:

$$\mathbf{s} \xrightarrow{\text{mixing}} \mathbf{r} = M\mathbf{s} \xrightarrow{\text{unmixing}} \mathbf{x} = W\mathbf{r} \xrightarrow{\text{non-linearity}} \mathbf{y} : y_a = g(x_a + w_a) \tag{36}$$

and here we want to maximize $H(Y_1, Y_2)$; the idea being that this should find a matrix W whose eigen-directions give statistically independent Y_a, this is the bit we want since it will also make the X_a independent, and whose eigenvalues, along with the values of w_a make $H(Y_1)$ big by lining the saturating non-linearity up with the underlying distributions: the orientation of W deals with the unmixing, the scale of W and the vector of w_as deals with the one-to-one part. Anyway, doing the calculation gives

$$\begin{aligned} \frac{d\tilde{H(y)}}{dW_{ab}} &= (W^T)^{-1}_{ab} + r_a(1-2y_b) \\ \frac{d\tilde{H(y)}}{dw_a} &= 1-2y_a \end{aligned} \qquad (37)$$

allowing the maximum of $H(Y_1, Y_2)$ to be, hopefully, found and this, again, hopefully, will unmix the signal. Note that this algorithm is not as blind as we might of hoped, the non-linearity needs to be chosen judiciously: however, the algorithm is reasonably robust; reasonably successful and certainly interesting mathematically.