Maximizing mutual information between input and neural response

S. Marzen

March 9, 2021

This is basically ICA in a certain limit with added noise.
We will deal with the one-dimensional case first:

$$z = f(Ax) \tag{1}$$
$$n = z + B\eta \tag{2}$$

where η is Gaussian noise, but we assume that such noise is small enough to ignore, and where z is the mean firing rate of the neuron. Then we have

$$I[x;n] = H[n] - H[n|x] \tag{3}$$
$$= H[n] + \frac{1}{2}\log(2\pi e B^2) \tag{4}$$

and with

$$\rho(n) = \int_{-\infty}^{\infty} \rho_z(n-y) \times \frac{1}{\sqrt{2\pi B^2}} e^{-y^2/2B^2} dy. \tag{5}$$

Since B is small, we simply Taylor expand and find

$$\rho_z(n-y) = \rho_z(n) - \frac{\partial \rho_z}{\partial z}|_n y + \frac{1}{2}\frac{\partial^2 \rho_z}{\partial z^2}|_n y^2 + O(y^3) \tag{6}$$

and thus

$$\rho(n) = \rho_z(n) + \frac{1}{2}\rho_z''(n)B^2 + O(B^4). \tag{7}$$

To find its differential entropy, we simply have

$$H[n] = -\int \rho(n)\log\rho(n) dn \tag{8}$$
$$= -\int \left(\rho_z(n) + \frac{1}{2}\rho_z''(n)B^2\right)\log\left(\rho_z(n) + \frac{1}{2}\rho_z''(n)B^2\right) + O(B^4) \tag{9}$$
$$= -\int \rho_z(n)\log\rho_z(n) dn - \frac{1}{2}B^2\int \rho_z''(n)\log\rho_z(n) dn - \frac{1}{2}B^2\int \rho_z''(n) dn + O(B^4). \tag{10}$$

Although we have the noise terms in there, we will simply ignore them in order to understand the predictions that come from maximizing mutual information between stimulus and neural response. We have (and this is in Cover and Thomas)

$$\rho_z(z) = \frac{\rho_x(x)}{|dz/dx|} \tag{11}$$
$$H[n] = -\int \rho_z(z)\log\frac{\rho_x(x)}{|dz/dx|} dz + O(B^2) \tag{12}$$
$$= -\int \rho_x(x)\log\rho_x(x) + \int \rho_x(x)|\frac{dz}{dx}| dx + O(B^2) \tag{13}$$
$$= H[x] + \int \rho_x(x)\log|\frac{dz}{dx}| dx + O(B^2). \tag{14}$$

We make yet another small noise approximation– that x is highly peaked at some value x^* and can be approximated as a Gaussian around that point.

$$\begin{aligned}
\frac{dz}{dx} &= Af'(Ax) && (15)\\
&= Af'(Ax^*) + A^2 f''(Ax^*)(x-x^*) + \frac{1}{2}A^3 f'''(Ax^*)(x-x^*)^2 + O((x-x^*)^3) && (16)\\
\log|\frac{dz}{dx}| &= \log|Af'(Ax^*) + A^2 f''(Ax^*)(x-x^*) + \frac{1}{2}A^3 f'''(Ax^*)(x-x^*)^2 + O((x-x^*)^3)| && (17)\\
&= \log|Af'(Ax^*)| + \frac{Af''(Ax^*)}{f'(Ax^*)}(x-x^*) + \frac{A^2 f'''(Ax^*)}{2f'(Ax^*)}(x-x^*)^2 + O((x-x^*)^3) && (18)
\end{aligned}$$

and so

$$\begin{aligned}
H[n] &= H[x] + \int \rho_x(x)\left(\log|Af'(Ax^*)| + \frac{Af''(Ax^*)}{f'(Ax^*)}(x-x^*) + \frac{A^2 f'''(Ax^*)}{2f'(Ax^*)}(x-x^*)^2 + O((x-x^*)^3)\right)dx\\
&\quad + O(B^2) && (19)\\
&= H[x] + \log|A| + \log|f'(Ax^*)| + \frac{A^2}{2}\frac{f'''(Ax^*)}{f'(Ax^*)}\sigma^2 + O(\sigma^4) && (20)
\end{aligned}$$

where σ is $\langle (x-x^*)^2\rangle$. Thus,

$$I[x;n] = H[x] + \frac{1}{2}\log(2\pi e B^2) + \log|A| + \log|f'(Ax^*)| + \frac{A^2}{2}\frac{f'''(Ax^*)}{f'(Ax^*)}\sigma^2 + O(B^2,\sigma^4). \quad (21)$$

To maximize this, we want to make A as large as possible in magnitude, we want $|f'(Ax^*)|$ to be as close to zero as possible but positive, and we want $f'''(Ax^*)$ to be as large as possible. The last of these leads to the weird conclusion that if we have a maximum in firing rate at x^*, then $f'''(Ax^*)$ being large and positive will cause the downturn in firing rate to swing into an upturn. The firing rate might only have a local max at the most likely value of the input.

One can do this for the multi-dimensional case. For ease (although this can be done more generally) we assume that the dimensionality of input and neural responses match. Then, given that

$$n = \vec{f}(Ax) + B\eta \quad (22)$$

where A is now a matrix and η is now a vector of ignorable noise, one finds that

$$I[x;n] = \log\left|\det\left(A^\top \mathrm{diag}(\vec{f}(x^*))\right)\right| + \frac{1}{2}\mathrm{tr}\left(\mathrm{diag}(\vec{f}(x^*)^{\odot -1})A^{-\top}\mathrm{diag}(\vec{f}'''(x^*))A^\top \mathrm{diag}\left((A\odot A)\vec{\sigma^2}\right)\right),$$

where x^* is the most likely input and $\vec{\sigma^2}$ is $\langle (x-x^*)(x-x^*)^\top\rangle$. The qualitative conclusions from the single neuron case seem to hold up.

The derivation goes as follows. We again ignore to $O(B^2)$ and to $O(\sigma^4)$ and find that, as is typical for ICA, the only thing that matters is $\langle \log|\det \nabla_x \vec{f}|\rangle$. The strategy again is to Taylor expand $\nabla_x \vec{f}$ around x^*:

$$\begin{aligned}
\left(\nabla_x \vec{f}\right)_{i,j} &= \frac{\partial f_j}{\partial x_i} && (23)\\
&= \frac{\partial f_j}{\partial x_i}|_{x^*} + \sum_k \frac{\partial^2 f_j}{\partial x_i \partial x_k}|_{x^*}(x_k - x_k^*) + \frac{1}{2}\sum_{k,k'}\frac{\partial^3 f_j}{\partial x_i \partial x_k \partial x_{k'}}|_{x^*}(x_k - x_k^*)(x_{k'} - x_{k'}^*) && (24)
\end{aligned}$$

and so we can write

$$\nabla_x \vec{f} = M(x^*) + \delta M \quad (25)$$

where δM is "small" in a way to be described. Then

$$\langle \log |\det(M(x^*) + \delta M)| \rangle = \langle \log |\det M(x^*)| \rangle + \langle \mathrm{tr}(M(x^*)^{-1} \delta M) \rangle. \quad (26)$$

We have

$$\begin{aligned} \langle \log |\det M(x^*)| \rangle &= \log |\det M(x^*)| & (27) \\ &= \log |\det(A^\top \mathrm{diag}(\vec{f}'(Ax^*)))| & (28) \\ &= \log |\det A| + \sum_i \log f_i'(Ax^*) & (29) \end{aligned}$$

and then we have

$$\langle \mathrm{tr}(M(x^*)^{-1} \delta M) \rangle = \mathrm{tr}(M(x^*)^{-1} \langle \delta M \rangle) \quad (30)$$

with

$$\begin{aligned} \langle \delta M_{ij} \rangle &= \frac{1}{2} \sum_{k,k'} \frac{\partial^3 f_j}{\partial x_i \partial x_k \partial x_{k'}} |_{x^*} \delta_{k,k'} \sigma_k^2 & (31) \\ &= \frac{1}{2} \sum_{k} \frac{\partial^3 f_j}{\partial x_i \partial x_k \partial x_k} |_{x^*} \sigma_k^2 & (32) \end{aligned}$$

where the components of x are assumed to be independent. (It is rather easy to transform your data so that this is roughly true around the peak.) This is unfortunately ugly, but hopefully useful.