

Maximum Likelihood Derivation for Informax ICA/PCA

Tingfan Wu

1 Probability Relationship

$$y = f(x) \quad (1)$$

$$P_y(y) = \frac{1}{\det[\nabla f(x)]} P_x(x) \quad (2)$$

2 Maximum Likelihood

In our case,

$$y = Wx \quad (3)$$

$$P_x(x) = \prod_i P_x(x_i) \quad (4)$$

$$p(x_i) \text{ known.} \quad (5)$$

then

$$P_y(y) = \frac{P_x(W^{-1}y)}{\det[W]} \quad (6)$$

$$\log P_y(y) = -\log \det[W] + \sum_i \log P_x(c_i^T W^{-1}x) \quad (7)$$

where c_i is a vector of zeros except that the i -th element is 1, which is used to extract the i -th element of vector $W^{-1}x$.

For convenience, let $Q = W^{-1}$,

$$\log P_y(y) = \log \det[Q] + \sum_i \log P_x(c_i^T Qx) \quad (8)$$

$$\frac{\partial \log P_y(y)}{\partial Q} = Q^{-T} + \sum_i \nabla_Q \log P_x(c_i^T Qx) \quad (9)$$

2.1 ICA

$$P_x(x) = \frac{\partial \text{logistic}(x)}{\partial x} = \frac{e^{-x}}{(1 + e^{-1})^2} \quad (10)$$

$$\log P_x(x) = -x - 2 \log(1 + e^{-x}) \quad (11)$$

$$d \log P_x(x) = - \left(1 - 2 \frac{1}{1 + e^x} \right) dx \quad (12)$$

Plugin back into (9), we obtain

$$\frac{\partial \log P_y(y)}{\partial Q} = Q^{-T} - \sum_i \left(1 - 2 \frac{1}{1 + e^{c_i^T Qy}} \right) c_i y^T \quad (13)$$

Let $u = -y$, we get Tony Bell's ICA update rule

$$\frac{\partial \log P_y(y)}{\partial Q} = Q^{-T} + \sum_i \left(1 - 2 \frac{1}{1 + e^{-c_i^T Q u}}\right) c_i u^T \quad (14)$$

$$= Q^{-T} + (1 - 2 \text{logistic}(x)) u^T \quad (15)$$

2.2 PCA

$$P_x(x) = \frac{1}{z} e^{-x^2} \quad (16)$$

$$\log P_x(x) = -x^2 - \log z \quad (17)$$

$$d \log P_x(x) = -2x dx \quad (18)$$

Plug in back into (9), we obtain

$$\frac{\partial \log P_y(y)}{\partial Q} = Q^{-T} - 2 \sum_i Q y c_i y^T \quad (19)$$

$$= Q^{-T} \left(I - 2 \sum_i Q y c_i (Q y)^T \right) \quad (20)$$

At the maximum, we have zero gradient. The problem is equivalent (hand waving) to PCA up to a scale. The rows of Q are eigenvectors divided by corresponding eigenvalues.

3 Infomax ICA

A linear neural network

$$x = \sigma(Qy) \quad (21)$$

$$\frac{\partial x}{\partial y} = \sigma(Qy)(1 - \sigma(Qy))Q \quad (22)$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is a logistic activation function.

In the infomax setting, we want to maximize the mutual information between x and y .

$$I(x, y) = H[x] - H[x|y], \quad (23)$$

where the second term $H[x|y]$ contains no information. Therefore, we need only maximize the $H[x]$.

$$H[x] = - \int p(x) \log p(x) dx \quad (24)$$

$$= - \int p(y) \log \frac{p(y)}{\left| \frac{\partial x}{\partial y} \right|} dy \quad (25)$$

$$= -E[\log p(y)] + E\left[\log \left| \frac{\partial x}{\partial y} \right| \right] \quad (26)$$

The first term is independent of Q . Therefore, we only need to maximize the second term. In an online manner fashion, we need to maximize $\log \left| \frac{\partial x}{\partial y} \right|$.

$$\frac{\partial}{\partial Q} \log \left| \frac{\partial x}{\partial y} \right| = \frac{\frac{\partial}{\partial w} \left| \frac{\partial x}{\partial y} \right|}{\left| \frac{\partial x}{\partial y} \right|} = Q^{-1} + [1 - 2\sigma(Qy)]y^T \quad (27)$$

$$= Q^{-1} + [1 - 2\sigma(x)]y^T \quad (28)$$