# Likelihood ICA

Mike Mayer

October 27, 2015

# 2nd oldest ICA method: Infomax, Likelihood

- Original paper idea from A. Bell and T. Sejnowski
- Gives insight to the information theoretic framework
- Disadvantage: It requires assumptions about the probability distributions of the sources...

# Transforming pdfs from random variable $a$ to $b$ (1st part)

Task: We have a real random variable $a$ (scalar for the moment) with a probability density function $p_b(b)$, we want to express the random variable by

$$b = f(a),$$

where $f(a)$ may be any differentiable, monotonously increasing function: Question is how we can calculate the pdf $p_b(b)$: It is not: $p_a(a) = p_b(b)$! Instead, we have to go via the condensed probability density function

$$P_a(a) = \int_{-\infty}^{a} p_a(a)da,$$

which is monotonously increasing from 0 to 1.

# Transforming pdfs from random variable $a$ to $b$ (2nd part)

In the case of cdf we really have

$$P_a(a) = P_b(f(a)) = P_b(b).$$

Thus,

$$p_a(a)da = p_b(f(a))db$$

and since

$$\frac{db}{da} = f'(a),$$

we get

$$p_a(a) = p_b(f(a)) \cdot f'(a).$$

# Transforming multivariate random variables...

If $a$ and $b$ are (equally dimensional) $\mathcal{R}^n$ vectors and consequently $f(.)$ is a vector function $\mathcal{R}^n \rightarrow \mathcal{R}^n$ the formula becomes:

$$p_a(a) = p_b(f(a)) \cdot |\mathcal{J}(f(a))|,$$

where $|\mathcal{J}(f)|$ is the determinant of the Jacobian of $f(.)$.

# Back to ICA: We want to know the sources!

Thus, again we say:

$$p(s) = \prod_{i=1}^{n} p_s(s_i)$$

We assume again linear mixing, thus

$$\mathbf{x} = A\mathbf{s}$$

We search for $W = A^{-1}$, so

$$\mathbf{s} = W\mathbf{x}.$$

Thus,

$$s = f(x) = W\mathbf{x},$$

which means the Jacobian $\mathcal{J}(f(x)) = W$.

# Transforming the pdf from **s** to **x**

We write down how the pdf transfers from a function of $s$ to a function of $x$.

$$p_x(\mathbf{x}) = p_s(\mathbf{s}) \cdot |W|$$

For every component $s_i$ of $\mathbf{s}$ we can write:

$$s_i = \mathbf{w}_i^T \mathbf{x},$$

where $\mathbf{w}_i$ are the line vectors of the matrix $W$. We can calculate

$$p_s(\mathbf{s}) = \prod p_s(s_i) = \prod p_s(\mathbf{w}_i^T \mathbf{x}),$$

and so:

$$p_x(\mathbf{x}) = \prod p_s(\mathbf{w}_i^T \mathbf{x}) \cdot |W|$$

which is a conditional probability $p_x(x|W)$.

# Maximum likelihood estimate

We get in every iteration of ICA a sample of $x$ we would like to get a optimal $W$, i.e. that gives on average the highest $p_x$ for our samples, so our guess should be:

$$\tilde{W} = argmax_W p_x(x|W),$$

which is a maximum likelihood estimate. The problem is: We do **not** know $p_s(s_i)$. The Bell and Sejnowski approach now just **assumes** a $p_s(s_i) = g'(s_i)$:

$$g(z) = \frac{1}{1 + e^{-z}}$$

($g(z)$ is the condensed probability density function: It is monotonously increasing from 0 to 1.)

# Adaptation rule

We have then:

$$\mathcal{L}(W) = \sum_{i=1}^{m} \left( \sum_{j=1}^{n} \log\ g'(w^T x^{(j)}) + log(|W|) \right),$$

where $\mathcal{L}$ is the logarithm of the likelihood. We can find the minimum point by stochastic gradient descent:

$$W := W + \alpha \left( \begin{bmatrix} 1 - 2g(w_1^T x^{(i)}) \\ 1 - 2g(w_2^T x^{(i)}) \\ 1 - 2g(w_3^T x^{(i)}) \\ \dots \end{bmatrix} x^{(i),T} + (W^{-1})^T \right),$$

where we can use: $\nabla_W |W| = |W|(W^{-1})^T$ in order to get the derivative. [1]

[1] The arguments here are taken from the Stanford scripts 2011. The formula is identical the one in this script if you account for the identity $(A^T)^{-1} = (A^{-1})^T$.

# Literature

The arguments are taken from Stanford University lectures available in the internet:
Andrew Ng, Computer Science: Machine Learning Autumn 2011, Part XII, Independent Component Analysis.
The original paper is (with different set of arguments but ending up with the same learning rule) is:
A. Bell, T. Sejnowski, 'An information-maximation approach to blind separation and blind deconvolution', Neural Comput. 1995 Nov;7(6):1129-59.