ICA Tutorial

Mika Inki

September 28, 2000

Based on the article:
"Independent component analysis: algorithms and applications"
by A. Hyvärinen and E. Oja

Background

ICA or Independent Component Analysis is closely related to BSS, or blind signal separation. Blind means that we have no knowledge (or only some very general assumptions) about the signals and about the way they are mixed. Still we would like to find these signals from the mixtures. ICA is one method for performing BSS.

Another closely related consept is projection pursuit, in which interesting (linear) projections are searched from a data set. Projection pursuit utilizes indices of interestingness, which are in many cases closely related to the measures of independence (non-Gaussianity) used in ICA. But unlike in projection pursuit, which is mostly used as an exploratory data analysis tool, in ICA one assumes a model for the data.

In many cases, BSS problems can be (at least approximately) written in a form in which the observed variables are linear combinations of underlying signals, which are also called sources. A classic example of a BSS problem is the coctail-party problem. Assume that we have two speakers (emitting signals s_1 and s_2), and two microphones (observing signals x_1 and x_2). The relation between the observed signals and the underlaying sources can approximately be written as

$$x_1 = a_{11}s_1 + a_{12}s_2$$

$$x_2 = a_{21}s_1 + a_{22}s_2$$

Where a_{ij} are unknown mixing coefficients. One would then search for separating coefficients (or weights) w_{ij} for which

$$s_1 = w_{11}x_1 + w_{12}x_2$$

$$s_2 = w_{21}x_1 + w_{22}x_2$$

In reality, there are time lags involved in the problem (i.e. $x_1(i) = a_{11}s_1(i - d_{11}) + a_{12}s_2(i - d_{12})$) as well as echos, there may be noise in the observed signals and the speakers may be moving, so this problem cannot be exactly formulated in such a simple manner as an ICA problem.

Independent Component Analysis

ICA can be described as a statistical model in which the observed variables $x_i, i = 1 \ldots n$ are obtained as a linear transformation ($\mathbf{A}$) of latent, independent variables $s_i, i = 1 \ldots n$:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{1}$$

Where $\mathbf{x} = [x_1 \quad x_2 \quad \ldots \quad x_n]^T$, $\mathbf{s} = [s_1 \quad s_2 \quad \ldots \quad s_n]^T$ and $\mathbf{A}$ is the matrix with elements a_{ij}. Several things are assumed in the model:

- only the variables $x_i, i = 1 \ldots n$ can be observed. The mixing $\mathbf{A}$ is unknown.
- $\mathbf{A}$ a matrix (i.e. a linear mapping)
- $\mathbf{A}$ static (i.e. $\mathrm{d}a_{ij}(t)/\mathrm{d}t = 0 \quad \forall t, \forall i, j$) and instantaneous
- the sources are independent
- usually there are as many independent components as there are observed variables (The case where there are less IC:s than dimensions can be solved by using deflation (if possible) or by reducing the dimension (e.g. with PCA), but the case where there are more independent components than observed variables is considerably harder.)
- usually noise is omitted from the model (or the level of noise is assumed to be relatively small)

Independence

Two variables x and y are considered to be independent when knowledge about one yields no information about the other. This can be written with the help of the conditional probability densities:

$$p(x|y) = p(x), \quad \text{and} \quad p(y|x) = p(y) \tag{2}$$

or equivalently in an other way using the joint probability density function:

$$p(x, y) = p(x)p(y) \tag{3}$$

Independence means that for any two functions h_1 and h_2 we have

$$E\{h_1(x)h_2(y)\} = \int\int h_1(x)h_2(y)p(x,y)\mathrm{d}x\mathrm{d}y =$$

$$\int h_1(x)p(x)\mathrm{d}x \int h_2(y)p(y)\mathrm{d}y = E\{h_1(x)\}E\{h_2(y)\} \tag{4}$$

And therefore the signals are also uncorrelated:

$$E\{xy\} - E\{x\}E\{y\} = E\{x\}E\{y\} - E\{x\}E\{y\} = 0. \tag{5}$$

We can find a decorrelating transformation for a set of variables $\mathbf{x}$ by using the eigenvalues and eigenvectors of its covariance matrix.

$$E\{\mathbf{x}\mathbf{x}^T\} = \mathbf{E}\mathbf{D}\mathbf{E}^T, \mathbf{V} = \mathbf{E}\mathbf{D}^{-\frac{1}{2}}\mathbf{E}^T \tag{6}$$

And it is easily proved that the new signals $\mathbf{V}\mathbf{x}$ are indeed white:

(HOMEWORK)

By keeping the dimensions corresponding to the highest eigenvalues, we do PCA (Principal Component Analysis) dimension reduction to the data.

Reconstruction of the data using only the principal components is optimal in the sense of mean square error. In this way one can drop the dimension of the data to the number of independent components (and remove only (measurement) noise). Reducing the dimensionality also reduces the risk of overlearning. (Also any algorithm should converge faster in a lower dimensional space.)

Examples of the independence / ICA model

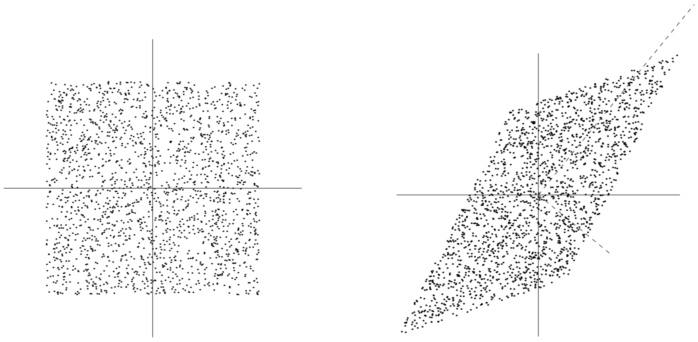

Figure 1: Left: x and y uniformly distributed and independent. Center: After linear mixing x and y are no longer independent. (Dashed: PCA directions). Right: After whitening.

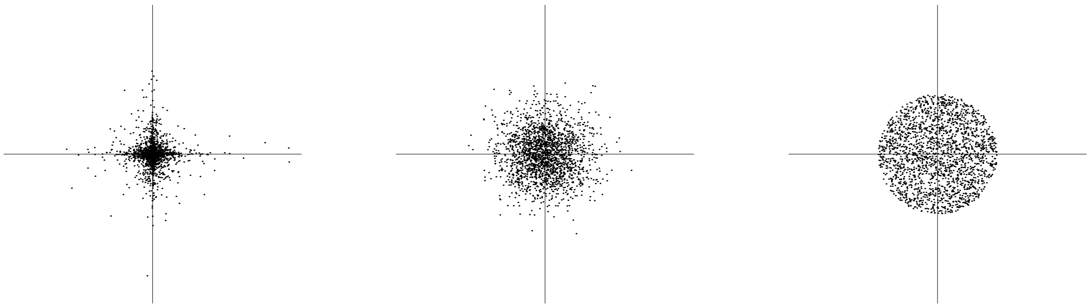

Figure 2: Left: x and y Laplacian and independent. Center: x and y Gaussian and independent, $p(x, y) = \frac{\exp(-\frac{1}{2}(x^2+y^2))}{2\pi}$. Right: x and y not independent.

Properties of ICA

- We cannot determine the order of the independent components.
- We cannot determine the scales or the signs of the independent components.

$$\mathbf{x} = \mathbf{A}\mathbf{s} = (\mathbf{A}\mathbf{K}^{-1}\mathbf{P}^{-1})\mathbf{P}\mathbf{K}\mathbf{s} \tag{7}$$

Where $\mathbf{K}$ is a diagonal scaling (and sign) matrix, and $\mathbf{P}$ is a permutation matrix. Note that the signals $\mathbf{PKs}$ are still independent.

- For simplicity the components are usually centered, i.e. the mean is removed.(Does not affect their independence / dependence)
- As the scales of the components are undetermined and the components are uncorrelated due to independence, constraining the estimates to be uncorrelated and of unit variance, i.e. whitening them, is sensible. In a whitened space ICA means that we are searching for an orthogonal matrix:

$$\begin{aligned} E\{\mathbf{s}\mathbf{s}^T\} &= E\{\mathbf{W}\mathbf{V}\mathbf{x}\mathbf{x}^T\mathbf{V}^T\mathbf{W}^T\} = \\ \mathbf{W}E\{\mathbf{V}\mathbf{x}\mathbf{x}^T\mathbf{V}^T\}\mathbf{W}^T &= \mathbf{W}\mathbf{W}^T = \mathbf{I}. \end{aligned} \tag{8}$$

As an orthogonal (n by n) matrix has $n(n-1)/2$ degrees of freedom compared to the n^2 degrees of a general (n by n) matrix, we could say that this solves half the problem (!).

- At most one of the components can be Gaussian. One can determine only the directions of variance from a Gaussian distribution (in a whitened space we can say therefore nothing about the directions of the independent components).

Independence criteria

By using ICA methods one attempts to find the separating matrix $\mathbf{W}$, i.e. the estimate for $\mathbf{A}^{-1}$, for which the separated signals $\mathbf{y} = \mathbf{W}\mathbf{x}$ are the most independent. Essential to ICA estimation are the criteria by which one can measure the independence of a particular rotation. All these measures are, in one way or another, measures of non-Gaussianity.

Non-Gaussianity

The reason behind the usability of non-Gaussianity as a criterion for independence is related to the central limit theorem. Asymptotically by summing independent variables, one gets a variable with a gaussian distribution under certain conditions. Therefore, if the original signals are not gaussian to begin with, (intuitively) one should find them from the mixtures by maximizing non-Gaussianity.

Kurtosis

The classical measure for non-Gaussianity is kurtosis or the fourth order cumulant.

$$\mathrm{kurt}(y) = E\{y^4\} - 3(E\{y^2\})^2 \tag{9}$$

The kurtosis of a gaussian variable is zero, for super-Gaussian distributions it is positive and for sub-Gaussian distributions it is negative. Super-Gaussian distributions have characteristically a "spiky" pdf with heavy tails. (for example the Laplace distribution $p(y) = \frac{1}{\sqrt{2}} \exp(\sqrt{2}|y|))$. Sub-Gaussian distributions have typically a "flat" pdf (for example the uniform distribution).

The estimates of kurtosis from a finite sample can be very sensitive to outliers, and there are non-Gaussian distributions for which kurtosis is zero.

Negentropy

Another measure of non-Gaussianity is given by negentropy. Negentropy J is defined using differential entropy H as:

$$J(\mathbf{y}) = H(\mathbf{y}_{gauss}) - H(\mathbf{y}) \tag{10}$$

here $\mathbf{y}_{gauss}$ is a Gaussian random variable with the same covariance matrix as $\mathbf{y}$. As a Gaussian variable has the largest entropy among all variables of equal variance, this measure is always non-negative and zero only for Gaussian variables. By assuming the y_i to be independent, this can be decomposed as the sum of the negentropies of the y_i:

$$J(\mathbf{y}) = \sum_i H(y_{gauss}) - H(y_i) \tag{11}$$

Negentropy as a measure of non-Gaussianity is well justified by statistical theory, and is in one sense the optimal estimator of non-Gaussianity. However, negentropy requires estimates for the probability distributions of the y_i.

Negentropy can, however, be approximated by using nonquadratic functions G_i as:

$$J(y) \approx \sum_i^p k_i [E\{G_i(y)\} - E\{G_i(y_{Gauss})\}]^2 \tag{12}$$

When carefully chosen (so that it is not sensitive to outliers), one can use just one function in the approximation. For example the following choices have proven useful:

$$G_1(u) = \frac{1}{a_1} \log\cosh a_1 u, \qquad G_2(u) = -exp(-u^2/2). \tag{13}$$

Minimization of mutual information

Mutual information is defined using differential entropy as:

$$I(\mathbf{y}) = \sum_{i=1}^m H(y_i) - H(\mathbf{y}) \tag{14}$$

Mutual information is a natural measure of the dependence between random variables. It is equivalent to the Kullback-Leibler divergence between the joint density $f(\mathbf{y})$ and the product of its marginal densities, a very intuitive measure for independence. Further, it can be (easily) shown that (for uncorrelated y_i with unit variance)

$$I(\mathbf{y}) = C - \sum_i J(y_i) \tag{15}$$

so, minimizing mutual information is closely related to maximizing negentropy.

Maximum likelihood

In the maximum likelihood approach to ICA, we maximize (with respect to $\mathbf{W}$) the likelihood of the observed data $\mathbf{x}(\mathbf{t}), t = 1..N$ assuming that the model $\mathbf{s} = \mathbf{W}\mathbf{x}$ holds. In the maximum, it should hold that $\mathbf{W}^{-1} = \mathbf{A}$. The logarithm of the likelihood function takes the form:

$$L = \sum_{t=1}^{N} \sum_{i=1}^{n} \log f_i(\mathbf{w}_i^T \mathbf{x}(t)) + T \log |\det \mathbf{W}| \tag{16}$$

Assuming that the f_i are the true distributions of y_i, and that y_i are independent, it can be shown that

$$E\{L(\mathbf{x}, \mathbf{W})\} = -I(\mathbf{y}) - H(\mathbf{x}) \tag{17}$$

where $H(\mathbf{x})$ is constant.

Infomax

Infomax is based on maximizing the output entropy of a neural network with nonlinear outputs.

$$L_2 = H(\phi_1(\mathbf{w}_1^T \mathbf{x}), \ldots , \phi_n(\mathbf{w}_n^T \mathbf{x})) \tag{18}$$

If the ϕ_i are chosen as the cumulative distribution functions corresponding to the independent components, Infomax is equivalent to maximum likelihood estimation.

FastICA

One, of cource, needs an algorithm for finding the minimum (maximum) of the objective functions we have just presented. Classically one has used simple gradient descent (ascent) for this purpose. However, the FastICA algorithm is in many ways much more efficient. The FastICA algorithm can be derived as an approximate Newton's method.

Starting with a function G (in this case an approximation of negentropy), one attempts to maximize $E\{G(\mathbf{w}^T\mathbf{x})\}$, thus finding an estimate for one independent component.

The problem can be written as

$$\begin{aligned} &\max E\{G(\mathbf{w}^T\mathbf{x})\} \\ s.t. \qquad &\mathbf{w}^T\mathbf{w} - 1 = 0 \end{aligned} \tag{19}$$

The Kuhn-Tucker conditions for this problem can be written as $(g = G')$

$$E\{\mathbf{x}g(\mathbf{w}^T\mathbf{x})\} - \beta\mathbf{w} = 0 \tag{20}$$

The Newton's method iteration formula for $f(\mathbf{x}) = 0$ can be written as

$$\mathbf{x}_{n=1} = \mathbf{x}_n - J(\mathbf{x_n})^{-1} f(\mathbf{x}_n), \tag{21}$$

where J is the Jacobian matrix of f. The jacobian in this case would be

$$J(\mathbf{w}) = E\{\mathbf{x}\mathbf{x}^T g'(\mathbf{w}^T\mathbf{x})\} - \beta\mathbf{I}. \tag{22}$$

To simplify this inversion the we make the approximation

$$E\{\mathbf{x}\mathbf{x}^T g'(\mathbf{w}^T\mathbf{x})\} \approx E\{\mathbf{x}\mathbf{x}^T\}E\{g'(\mathbf{w}^T\mathbf{x})\} = E\{g'(\mathbf{w}^T\mathbf{x})\}\mathbf{I}. \tag{23}$$

Now the iteration (for one unit) can be written as

$$\mathbf{w}_{n+1} = \mathbf{w}_n - [E\{\mathbf{x}g(\mathbf{w}_n^T\mathbf{x})\} - \beta\mathbf{w}_n]/[E\{g'(\mathbf{w}_n^T\mathbf{x})\} + \beta] \tag{24}$$

Now, by multiplying both sides by $[E\{g'(\mathbf{w}^T\mathbf{x})\} + \beta]$ and using the constraint, one gets the final algorithm:

1. Initialize $\mathbf{w}$ to a possibly random initial value $\mathbf{w}_0$
2. Let $\mathbf{w}_{n+1}^+ = E\{\mathbf{x}g(\mathbf{w}_n^T\mathbf{x})\} - E\{g(\mathbf{w}_n^T\mathbf{x})\}\mathbf{w}_n$
3. Let $\mathbf{w}_{n+1} = \mathbf{w}_{n+1}^+/||\mathbf{w}_{n+1}^+||$
4. If not converged, go to step 2.

FastICA for several units uses basically the same algorithm, new values are calculated for several weight vectors $\mathbf{w}_i, i = 1 \ldots \mathbf{n}$ while decorrelating the $\mathbf{w}_i^T\mathbf{x}$. This decorrelation can be done either nonsymmetrically by using Gram-Schmidt-like decorrelation, or symmetrically by using matrix square roots .

In reality one must naturally use sample averages instead of expected values. One does not have to use the whole data for the calculation, a smaller sample that is chosen randomly for each step will usually suffice.

The FastICA algorithm has some appealing qualities:

- The convergence is cubic (or at least quadratic), under the assumption of the ICA data model. Remember that the convergence of gradient descent methods is only linear.
- There are no step size parameters.
- The algorithm gives reliable estimates for the indipendent components of practically any non-Gaussian distribution using only a single function g. Usually in ICA estimation one needs estimates for the probability densities of the components.
- One can optimize the algorithm by choosing a suitable nonlinearity g.
- The independent components can be estimated one by one, roughly equivalent to doing projection pursuit.
- The algorithm is parallel, distributed, computationally simple and requires little memory space.