Introduction to Independent Component Analysis

Formulation of the Problem

A typical example is the "cocktail party problem". Given the signals $x_j(t)$ from m microphones recording n speakers in the room, one wants to recover the voice $s_i(t)$ of each speaker. The problem can be formulated as

- **Given**

  $$x_i(t) = \sum_{j=1}^{n} a_{ij} y_j(t) \qquad (i = 1, \cdots, m)$$

  or in matrix form

  $$X = AS$$

- **Find**
  - the source variables $y_j(t)$ *the independent components*: $S = WX$
  - and the linear combination matrix $A = W^{-1}$.

This is a blind source separation (BSS) problem, i.e., to separate linearly mixed source signals. The word "blind" means that we do not assume any prior knowledge about sources S or the mixing process A except that the source signals s_i are statistically independent.

Although this BSS problem seems severely under constrained, the independent component analysis (ICA) can find nearly unique solutions satisfying certain properties.

Methods of ICA Estimations

Non-Gaussianity

One theoretical foundation of ICA is the *Central Limit Theorem*, which states that the distribution of the sum of N independent random variables approaches normal as $N \longrightarrow \infty$. For example, the face number of a dice has a uniform distribution from 1 to 6. But the distribution of a pair of dice is no longer uniform, it has a maximum probability at the mean of 7. The distribution better approximates Gaussian as the number of dice increases.

Assume x_i are random variables independently drawn from an arbitrary distribution with mean μ and variance σ^2. Then the distribution of the mean $x = \sum_{i=1}^{N} x_i/N$ approaches Gaussian with mean μ and variance σ^2/N.

Thus, in the BSS problem, the distribution of $x_i = \sum_{j=1}^{m} a_{ij} s_i(t)$ as the sum of m random variables, is closer to Gaussian than any of the variables $s_i(t)$. We want to find W so that $S = WX$ is less Gaussian than X, in other words, we want to maximize the **non-Gaussianality** of WS. Obviously if all source variables are normally distributed, the ICA method will not work.

The ICA method has the following constrains:

- The independent components have to be non-Gaussian.

- The estimation of A and S is up to a scaling factor c_i. Let $C = diag(c_1, \cdots, c_n)$ and $C^{-1} = diag(1/c_1, \cdots, 1/c_n)$, and $A' = AC^{-1}$ and $S' = CS$, we have

$$X = AS = [AC^{-1}][CS] = A'S'$$

  Also the scaling factor c_i could be either positive or negative. For this reason, we will always assume the independent components have unit variance $E\{s_i^2\} = 1$. Since They are also uncorrelated (all independent variables are uncorreclated), we have $E\{s_i s_j\} = \delta_{ij}$, i.e.,

$$E\{SS^T\} = I$$

- The estimated independent components are not in any particular order. When the order of the corresponding elements in both S and A is rearranged, $X = AS$ still holds.

Measures of Non-Gaussianality

The ICA method depends on certain measurement of the non-Gaussianity:

- **Kurtosis**

Kurtosis is defined as the normalized form of the fourth central moment of a distribution:

$$kurt(x) = E\{x^4\} - 3(E\{x^2\})^2$$

If we assume x to have zero mean $\mu_x = E\{x\} = 0$ and unit variance $\sigma_x^2 = D\{x\} = E\{x^2\} - \mu_x^2 = 1$, then $E\{x^2\} = 1$ and $kurt(x) = E\{x^4\} - 3$. Kurtosis measures the degree of peakedness (spikiness) of a distribution and it is zero only for Gaussian distribution. Any other distribution's kurtosis is either positive if it is supergaussian (spikier than Gaussian) or negative if it is subgaussian (flatter than Gaussian). Therefore the absolute value of the kurtosis or kurtosis squared can be used to measure the non-Gaussianity of a distribution. However, kurtosis is very sensitive to outliers, and it is not a robust measurement of non-Gaussianity.

- **Negentropy**

  Another unique property of Gaussian distribution is that it has maximum entropy among all distributions over the entire real axis $(-\infty, \infty)$. Based on this property, we define differential entropy, called *negentropy*, as

  $$J(y) = H(y_G) - H(y) \geq 0$$

  where $Y = [y_1, \cdots, y_m]^T$ is a random vector and y_G is a Gaussian vector with the same covariance as y. As $J(y)$ is always greater than zero unless Y is Gaussian, it is a good measurement of non-Gaussianity. We want to find W so that $Y = WX$ has the maximum negentropy $J(y)$, i.e., Y is most non-Gaussian.

  However, exact $J(y)$ is difficult to get as its calculation requires the specific density distribution function $p(y)$.

- **Approximations of Negentropy**

  The negentropy can be approximated by

  $$J(y) \approx \frac{1}{12}E\{y^3\}^2 + \frac{1}{48}kurt(y)^2$$

However, this approximation also suffers from the non-robustness due to the kurtosis function. A better approximation is

$$J(y) \approx \sum_{i=1}^{p} k_i [E\{G_i(y)\} - E\{G_i(g)\}]^2$$

where k_i are some positive constants, y is assumed to have zero mean and unit variance, and g is a Gaussian variable also with zero mean and unit variance. G_i are some non-quadratic functions such as

$$G_1(y) = \frac{1}{a} log\ cosh\ a\ x, \qquad G_2(y) = -exp(-y^2/2)$$

where $1 \leq a \leq 2$ is some suitable constant. Although this approximation may not be an accurate, it is always greater than zero except when x is Gaussian.

Minimization of Mutual Information

The mutual information $I(X, Y)$ of two random variables X and Y is defined as

$$I(X, Y) = H(X) + H(y) - H(X, Y)$$

Similarly the mutual information $I(y_1, \cdots, y_m)$ of a set of m variables y_i $(i = 1, \cdots, m)$ is defined as

$$I(y_1, \cdots, y_m) = \sum_{i=1}^{m} H(y_i) - H(y_1, \cdots, y_m)$$

If random vector $Y = [y_1, \cdots, y_m]^T$ is a linear transform of another random vector $x = [x_1, \cdots, x_m]^T$:

$$y_i = \sum_{j=1}^{m} w_{ij} x_j, \qquad \text{or} \qquad Y = WX$$

the above can be written as

$$\begin{aligned} I(y_1, \cdots, y_m) &= \sum_{i=1}^{m} H(y_i) - H(y_1, \cdots, y_m) \\ &= \sum_{i=1}^{m} H(y_i) - H(x_1, \cdots, x_m) - log\ det(W) \end{aligned}$$

We further assume y_i to be uncorrelated and of unit variance, i.e., the covariance matrix of Y is

$$E\{YY^T\} = WE\{XX^T\}W^T = I$$

and its determinant is

$$det\ I = 1 = (det\ W)\ (det\ E\{XX^T\})\ (det\ W^T)$$

This means $det\ W$ is a constant (same for any W). Moreover, since all y_i have the same unit variance, their negentropy becomes

$$J(y_i) = H(y_G) - H(y_i) = C - H(y_i)$$

where $C = H(y_G)$ is the entropy of a Gaussian with unit variance, same for all y_i. Substituting $H(y_i) = C - j(y_i)$ into the expression of mutual information, and realizing the other two terms $H(X)$ and $log\ det\ W$ are both constant (same for any W), we get

$$I(y_1, \cdots, y_m) = Const - \sum_{i=1}^{m} J(y_i)$$

where $Const$ is a constant (including all terms C, $H(X)$ and $log\ det\ W$) which is the same for any linear transform matrix W. This is the fundamental relation between mutual information and negentropy of the variables y_1. If the mutual information of a set of variables is decreased (indicating the variables are less dependent) then the negentropy will be increased, and y_i are less Gaussian. We want to find a linear transform matrix W to minimize mutual information $I(y_1, \cdots, y_m)$, or equivalently maximize negentropy (under the assumption that y_i are uncorrelated).

Preprocessing for ICA

To simplify the ICA algorithms, the following preprocessing steps are usually taken:

- **Centering**

  Subtract the mean $E\{X\}$ from the observed variable $X = AS$ so it has zero mean. By doing so, the sources S also become zero mean because $E\{X\} = A\ E\{S\} = 0$. After the mixing matrix A is estimated, $E\{S\}$ can be estimated to be $A^1 E\{X\}$.

- **Whitening**

  Transform observed variables X so that they are uncorrelated and have unit variance. We first obtain the eigenvalues λ_i and their corresponding eigenvectors ϕ_i of the covariance matrix $E\{XX^T\}$, and form the diagonal eigenvalue matrix $\Lambda = diag(\lambda_1, \cdots, \lambda_m)$ and orthogonal eigenvector matrix $\Phi = [\phi_1, \cdots, \phi_m]$. We have

  $$E\{XX^T\}\Phi = \Phi\Lambda, \quad \text{i.e.,} \quad \Lambda^{-1/2}\Phi^T E\{XX^T\}\Phi\Lambda^{-1/2} = I$$

  It is now clear that if we carry out transform $X' = TX = \Lambda^{-1/2}\Phi^T X$, the covariance matrix of X' becomes

  $$\begin{aligned} E\{X'X'^T\} &= E\{(\Lambda^{-1/2}\Phi^T X)(\Lambda^{-1/2}\Phi^T X)^T\} = E\{\Lambda^{-1/2}\Phi^T XX^T\Phi\Lambda^{-1/2}\} \\ &= \Lambda^{-1/2}\Phi^T E\{XX^T\}\Phi\Lambda^{-1/2} = I \end{aligned}$$

  After the transform $T = \Lambda^{-1/2}\Phi^T$, the mixing process becomes

  $$X' = TX = TAS = A'S$$

  where $A' = TA$ is the new mixing matrix, which is orthogonal, since

  $$E\{X'X'^T\} = I = A'E\{SS^T\}A'^T = A'A'^T$$

  (recall that $E\{SS^T\} = I$).

  The whitening process can also reduce the dimensionality of the problem by ignoring the components corresponding to very small eigenvalues (PCA).

Appendix

- **Entropy**

  The entropy of a distribution $p_x(x)$ is defined as

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

  Entropy represents the uncertainty of the random variable. Among all distributions, uniform distribution has maximum entropy over a finite region $[a, b]$, while Gaussian distribution has maximum entropy over the entire real axis.

The joint entropy of two random variables X and Y is defined as

$$H(X,Y) = -\int p(x,y)\ log\ p(x,y)dx\ dy = E\{-log\ p(X,Y)\}$$

The conditional entropy of X given y is

$$H(X|y) = -\int p(x|y)\ log\ p(x|y)dx = E\{-log\ p(X|Y)\mid Y=y\}$$

and the conditional entropy of X given Y is

$$\begin{aligned} H(X|Y) &= \int p(y)H(X|y)dy = -\int p(y)\int p(x|y)\ log\ p(x|y)dxdy \\ &= -\int\int p(x,y)\ log\ p(x|y)dxdy = E\{E\{-log\ p(X|Y)\mid Y\}\} \end{aligned}$$

- **Mutual information**

  Mutual information is defined as

$$\begin{aligned} I(X,Y) &= H(X) + H(Y) - H(X,Y) \\ &= E\{-log\ \ p(X)\} + E\{-log\ \ p(Y)\} + E\{-log\ \ p(X,Y)\} \\ &= E\{log\ \frac{p(X,Y)}{p(X)\ p(Y)}\} \end{aligned}$$

  Mutual information measures the amount of information shared between the two random variables X and Y. Since

$$p(X,Y) = p(X|Y)\ p(Y) = p(Y|X)\ p(X)$$

  we have

$$\begin{aligned} I(X,Y) &= E\{log\ \frac{p(X,Y)}{p(X)\ p(Y)}\} \\ &= E\{log\ \frac{p(X|Y)}{p(X)}\} = H(X) - H(X|Y) \\ &= E\{log\ \frac{p(Y|X)}{p(Y)}\} = H(Y) - H(Y|X) \end{aligned}$$

- **Functions of random variables**

  Assume X is a random variable with distribution $p_x(x)$, then its function $Y = \phi(X)$ is also a random variable. We have $X = \phi^{-1}(Y) = \psi(Y)$ and $dy/dx = \phi'$ and $dx/dy = \psi' = 1/\phi'$.

  – Distribution $p_y(y)$ of Y is related to distribution $p_x(x)$ of X:

    * If $Y = \phi(X)$ monotonically increases ($\phi' > 0$ and $\psi' > 0$), then

    $$P_y(y) = P(Y < y) = P(X < x) = \int_{-\infty}^{x} p_x(x)dx = \int_{-\infty}^{\psi(y)} p_x(\psi(y))dx$$

    $$p_y(y) = dP_y(y)/dy = p_x(\psi(y))\psi'(y) = p_x(x)/\phi'(x)$$

    * If $Y = \phi(X)$ monotonically decreases ($\phi' < 0$ and $\psi' < 0$), then

    $$P_y(y) = P(Y < y) = P(X > x) = \int_{x}^{\infty} p_x(x)dx = \int_{\psi(y)}^{x} p_x(\psi(y))dx$$

    $$p_y(y) = dP_y(y)/dy = -p_x(\psi(y))\psi'(y) = p_x(\psi(y))|\psi'(y)| = p_x(x)/|\phi'(x)|$$

    In general, we have

    $$p_y(y) = \sum_i p_x(x_i)/|\phi'(x_i)|$$

    where x_i are solutions for equation $y = \phi(x)$.

  – Entropy $H(Y)$ of Y is related to entropy $H(X)$ of $X = \phi^{-1}(Y) = \psi(Y)$:

$$\begin{aligned} H(Y) &= -\int p_y(y)\, log\, p_y(y) dy = -\int \frac{p_x(x)}{|\phi'(x)|} log\, \frac{p_x(x)}{|\phi'(x)|} dy \\ &= -\int p_x(x)\, log\, \frac{p_x(x)}{|\phi'(x)|} dx = -\int p_x(x)\, log\, p_x(x)\, dx + \int p_x(x)\, log\, |\phi'(x)| dx \\ &= H(X) + E\left\{log\, |\phi'(X)|\right\} \end{aligned}$$

If the inverse function $X = \psi^{-1}(Y)$ is not unique, than

$$H(Y) < H(X) + E\ \{log\ |\phi'(x)|\}$$

This result can be generalized to multi-variables. If

$$Y_i = \phi_i(X_1, \cdots, X_n), \qquad (i = 1, \cdots, n)$$

then

$$H(Y_1, \cdots, Y_n) \leq H(X_1, \cdots, X_n) + E\ \{log\ J(X_1, \cdots, X_n)\}$$

where $J(X_1, \cdots, X_n)$ is the Jacobian of the above transformation. In particular, if the functions are linear

$$Y_i = \sum_{j=1}^{n} a_{ij} X_j, \qquad (i = 1, \cdots, n)$$

then

$$H(Y_1, \cdots, Y_n) \leq H(X_1, \cdots, X_n) + log\ det(A)$$

where $det(A)$ is the determinant of the transform matrix $A = [a_{ij}]_{n \times n}$.