Independent Component Analysis

This introduction is based on the tutorial paperhttp://www.cis.hut.fi/aapo/papers/IJCNN99

Formulation of the Problem

A typical example is the "cocktail party problem". (See a demohttp://www.cis.hut.fi/projects/ic
Given the signals $x_j(t)$ from m microphones recording n speakers in the room ($m \geq n$), one wants to recover the voice $s_i(t)$ of each speaker. The problem can be formulated as

- **Given**

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

  or in matrix form

$$\begin{bmatrix} x_1 \\ \cdots \\ x_m \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} s_1 \\ \cdots \\ s_n \end{bmatrix} \qquad \text{or} \qquad \mathbf{x} = \mathbf{A}\mathbf{s}$$

- **Find**
  - the estimation $\mathbf{y} = \mathbf{W}\mathbf{x}$ of the source variables $s_j(t)$ *the independent components*,
  - and the linear combination matrix $\mathbf{A}$.

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 $\mathbf{s}$ or the mixing process $\mathbf{A}$ 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.

ICA can be compared with *principal component analysis* (PCA)../pca/index.html for decorrelation. Given a set of variables $\mathbf{x}$, PCA finds a matrix $\mathbf{W}$ so that the components of $\mathbf{y} = \mathbf{W}\mathbf{x}$ are uncorrelated. Only under the special case when $\mathbf{y} = [y_1, \cdots, y_n]$ are gaussian, are they also independent. In comparison,

ICA is a more powerful method in the senese that it satisfies a stronger requirement of finding $\mathbf{W}$ so that the components of $\mathbf{y} = \mathbf{Wx}$ are independent (and therefore are also necessarily uncorrelated).

Methods of ICA Estimations

Non-Gaussianity is Independence

The theoretical foundation of ICA is the *Central Limit Theorem*http://en.wikipedia.org/wiki/Ce
∞. For example, the face value of a dice has a uniform distribution from 1 to 6. But the distribution of the sum of a pair of dice is no longer uniform. It has a maximum probability at the mean of 7. The distribution of the sum of the face values will be better approximated by a Gaussian as the number of dice increases.

Specifically, if 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.

To solve the BSS problem, we want to find a matrix $\mathbf{W}$ so that $\mathbf{y} = \mathbf{Wx} = \mathbf{WAs} \approx \mathbf{s}$ is as close to the independent sources $\mathbf{s}$ as possible. This can be seen as the reverse process of the central limit theorem above.

Consider one component $y_i = \mathbf{w_i^T As}$ of $\mathbf{y}$, where $\mathbf{w_i^T}$ is the ith row of $\mathbf{W}$. As a linear combination of all components of $\mathbf{s}$, y_i is necessarily more Gaussian than any of the components unless it is equal to one of them (i.e., $\mathbf{w_i^T A}$ has only one non-zero component. In other words, the goal $\mathbf{y} \approx \mathbf{s}$ can be achieved by finding $\mathbf{W}$ that maximizes the **non-Gaussianity** of $\mathbf{y} = \mathbf{Wx} = \mathbf{WAs}$ (so that $\mathbf{y}$ is least Gaussian). This is the essence of all ICA methods. Obviously if all source variables are Gaussian, the ICA method will not work.

Based on the above discussion, we get requirements and constrains for the ICA methods:

- The number of observed variables m must be no fewer than the number of independent sources n (assume $m = n$ in the following).
- The source components are stochastically independent, and have to be non-Gaussian (with possible exception of at most one Gaussian).

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

  $$\mathbf{x} = \mathbf{A}\mathbf{s} = [\mathbf{A}\mathbf{C}^{-\mathbf{1}}][\mathbf{C}\mathbf{s}] = \mathbf{A}'\mathbf{s}'$$

  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 $E\{s_i^2\} = 1$. As 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 $\mathbf{s}$ and $\mathbf{A}$ is rearranged, $\mathbf{x} = \mathbf{A}\mathbf{s}$ still holds.

All ICA methods are based on the same fundamental approach of finding a matrix $\mathbf{W}$ that maximizes the non-Gaussianity of $\mathbf{s} = \mathbf{W}\mathbf{x}$ thereby minimizing the independence of S, and they can be formulated as:

$$\textbf{ICA method} = \textbf{objective function} + \textbf{optimization algorithm}$$

All ICA methods are an optimization process (always iterative) to find a matrix $\mathbf{W}$ to maximize some objective function that measures the degree of non-Gaussianity or independence of the estimated components $\mathbf{s} = \mathbf{W}\mathbf{x}$. In the following, we will discuss some common objective functions.

Measures of Non-Gaussianity

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 $\sigma_x^2 = E\{x^2\} - \mu_x^2 = 1$, 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.

../figures/kurtosis.gif

- **Differential Entropy – Negentropy**

  The entropy of a random variable y with density function $p(y)$ is defined as

  $$H(y) = -\int_{-\infty}^{\infty} p(y) log\ p(y) dy = -E\{log\ p_i(y)\}$$

  An important property of Gaussian distribution is that it has the maximum entropy among all distributions over the entire real axis $(-\infty, \infty)$. (And uniform distribution has the maximum entropy among all distributions over a finite range.) Based on this property, the differential entropy, also called *negentropy*, is defined as

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

  where y_G is a Gaussian variable with the same variance as y. As $J(y)$ is always greater than zero unless y is Gaussian, it is a good measurement of non-Gaussianity.

  This result can be generalized from random variables to random vectors, such as $\mathbf{y} = [y_1, \cdots, y_m]^T$, and we want to find a matrix $\mathbf{W}$ so that $\mathbf{y} = \mathbf{Wx}$ has the maximum negentropy $J(\mathbf{y}) = H(\mathbf{y_G}) - H(\mathbf{y})$, i.e., $\mathbf{y}$ is most non-Gaussian. However, exact $J(\mathbf{y})$ is difficult to get as its calculation requires the specific density distribution function $p(\mathbf{y})$.

- **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

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

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

$$J(y) = [E\{G(y)\} - E\{G(g)\}]^2$$

Since the second term is a constant, we want to maximize $E\{G(y)\}$ to maximize $J(y)$.

../figures/G$_f unctions.gif$

Minimization of Mutual Information

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

$$I(x, y) = H(x) + H(y) - H(x, y) = H(x) - H(x|y) = H(y) - H(y|x)$$

Obviously when x and y are independnent, i.e., $H(y|x) = H(y)$ and $H(x|y) = H(x)$, their mutual information $I(x, y)$ is zero.

../figures/mutual$_i nfo.gif$

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

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

If random vector $\mathbf{y} = [y_1, \cdots, y_n]^T$ is a linear transform of another random vector $\mathbf{x} = [x_1, \cdots, x_n]^T$:

$$y_i = \sum_{j=1}^{n} w_{ij} x_j, \quad \text{or} \quad \mathbf{y} = \mathbf{W}\mathbf{x}$$

then the entropy of $\mathbf{y}$ is related to that of $\mathbf{x}$ by:

$$\begin{aligned} H(y_1, \cdots, y_n) &= H(x_1, \cdots, x_n) + E\left\{log\ J(x_1, \cdots, x_n)\right\} \\ &= H(x_1, \cdots, x_n) + log\ det\ \mathbf{W} \end{aligned}$$

where $J(x_1, \cdots, x_n)$ is the Jacobian of the above transformation:

$$J(x_1, \cdots, x_n) = \begin{vmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \cdots & \cdots & \cdots \\ \frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_n} \end{vmatrix} = det\ \mathbf{W}$$

The mutual information above can be written as

$$\begin{aligned} I(y_1, \cdots, y_n) &= \sum_{i=1}^{n} H(y_i) - H(y_1, \cdots, y_n) \\ &= \sum_{i=1}^{n} H(y_i) - H(x_1, \cdots, x_n) - log\ det\ \mathbf{W} \end{aligned}$$

We further assume y_i to be uncorrelated and of unit variance, i.e., the covariance matrix of $\mathbf{y}$ is

$$E\{\mathbf{y}\mathbf{y}^\mathbf{T}\} = \mathbf{W}E\{\mathbf{x}\mathbf{x}^\mathbf{T}\}\mathbf{W}^\mathbf{T} = \mathbf{I}$$

and its determinant is

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

This means $det\ \mathbf{W}$ is a constant (same for any $\mathbf{W}$). Also, as the second term in the mutual information expression $H(x_1, \cdots, x_n)$ is also a constant (invariant with respect to $\mathbf{W}$), we have

$$I(y_1, \cdots, y_n) = \sum_{i=1}^{n} H(y_i) + \text{Constant}$$

i.e., minimization of mutual information $I(y_1, \cdots, y_n)$ is achieved by minimizing the entropies

$$H(y_i) = -\int p_i(y_i) log\ p_i(y_i) dy_i = -E\{log\ p_i(y_i)\}$$

As Gaussian density has maximal entropy, minimizing entropy is equivalent to minimizing Gaussianity.

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 $H(\mathbf{x})$ and $log\ det\ \mathbf{W}$ are both constant (same for any $\mathbf{W}$), we get

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

where $Const$ is a constant (including all terms C, $H(\mathbf{x})$ and $log\ det\ \mathbf{W}$) 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 $I(y_1, \cdots, y_n)$, or, equivalently, to 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 $E\{\mathbf{x}\}$ from the observed variable $\mathbf{x} = \mathbf{As}$ so it has zero mean. By doing so, the sources $\mathbf{s}$ also become zero mean because $E\{\mathbf{x}\} = \mathbf{A}\ E\{\mathbf{s}\} = 0$. When the mixing matrix $\mathbf{A}$ is available, $E\{\mathbf{s}\}$ can be estimated to be $\mathbf{A}^{-1}E\{\mathbf{x}\}$.

- **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 $E\{\mathbf{x}\mathbf{x}^T\}$, and form the diagonal eigenvalue matrix $\mathbf{\Lambda} = diag(\lambda_1, \cdots, \lambda_m)$ and orthogonal eigenvector matrix $\mathbf{\Phi} = [\phi_1, \cdots, \phi_m]$ ($\mathbf{\Phi}^{-1} = \mathbf{\Phi}^T$). We have

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

  If we carry out a linear transform $\mathbf{T} = \mathbf{\Lambda}^{-1/2}\mathbf{\Phi}^T$ so that $\mathbf{x}' = \mathbf{T}\mathbf{X} = (\mathbf{\Lambda}^{-1/2}\mathbf{\Phi}^T)X$, the covariance matrix of $\mathbf{x}'$ becomes

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

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

$$\mathbf{x}' = \mathbf{T}\mathbf{x} = \mathbf{T}\mathbf{A}\mathbf{s} = \mathbf{A}'\mathbf{s}$$

  Here the new mixing matrix $\mathbf{A}' = \mathbf{T}\mathbf{A}$ is orthogonal, as it satisfies:

$$E\{\mathbf{x}'\mathbf{x}'^T\} = \mathbf{I} = \mathbf{A}'E\{\mathbf{s}\mathbf{s}^T\}\mathbf{A}'^T = \mathbf{A}'\mathbf{A}'^T$$

  (recall that we assume $E\{\mathbf{s}\mathbf{s}^T\} = \mathbf{I}$). Similarly we also know that the matrix $\mathbf{W}$ we seek is also orthogonal, as

$$E\{\mathbf{s}\mathbf{s}^T\} = \mathbf{I} = \mathbf{W}E\{\mathbf{x}'\mathbf{x}'^T\}\mathbf{W}^T = \mathbf{W}\mathbf{W}^T$$

  The whitening process reduces the independent variables, the n^2 components of the mixing matrix A' to half ($n(n-1)/2$) due to the constraint that $\mathbf{A}'$ is orthogonal. Moreover, the whitening can also reduce the dimensionality of the problem by ignoring the components corresponding to very small eigenvalues (PCA).

FastICA Algorithm

Summarizing the objective functions discussed above, we see a common goal of maximizing a function $\sum_i E\{G(y_i)\}$, where $y_i = \mathbf{w}_i^T\mathbf{x}$ is a component of $\mathbf{y} = \mathbf{W}\mathbf{x}$

$$\sum_i E\{G(y_i)\} = \sum_i E\{G(\mathbf{w}_i^T\mathbf{x})\}$$

where $\mathbf{w}_i^T$ is the ith row vector in matrix $\mathbf{W}$. We first consider one particular component (with the subscript i dropped). This is an optimization problem which can be solved by Lagrange multiplier method with the objective function

$$O(\mathbf{w}) = E\{G(\mathbf{w}^T\mathbf{x})\} - \beta(\mathbf{w}^T\mathbf{w} - 1)/2$$

The second term is the constraint representing the fact that the rows and columns of the orthogonal matrix $\mathbf{W}$ are normalized, i.e., $\mathbf{w}^T\mathbf{w} = 1$. We set the derivative of $O(\mathbf{w})$ with respect to $\mathbf{w}$ to zero and get

$$F(\mathbf{w}) \triangleq \frac{\partial O(\mathbf{w})}{\partial \mathbf{w}} = E\{\mathbf{x}g(\mathbf{w}^T\mathbf{x})\} - \beta\mathbf{w} = 0$$

where $g(z) = dG(z)/dz$ is the derivative of function $G(z)$. This algebraic equation system can be solved iteratively by Newton-Raphson method:

$$\mathbf{w} \Leftarrow \mathbf{w} - J_F^{-1}(\mathbf{w})F(\mathbf{w})$$

where $J_F(\mathbf{w})$ is the Jacobian of function $F(\mathbf{w})$:

$$J_F(\mathbf{w}) = \frac{\partial F}{\partial \mathbf{w}} = E\{\mathbf{x}\mathbf{x}^T g'(\mathbf{w}^T\mathbf{x})\} - \beta\mathbf{I}$$

The first term on the right can be approximated as

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

and the Jacobian becomes diagonal

$$J_F(\mathbf{w}) = [E\{g'(\mathbf{w}^T\mathbf{x})\} - \beta]\mathbf{I}$$

and the Newton-Raphson iteration becomes:

$$\mathbf{w} \Leftarrow \mathbf{w} - \frac{1}{E\{g'(\mathbf{w}^T\mathbf{x})\} - \beta}[E\{\mathbf{x}g(\mathbf{w}^T\mathbf{x})\} - \beta\mathbf{w}]$$

Multiplying both sides by the scaler $\beta - E\{g'(\mathbf{w}^T\mathbf{x})\}$, we get

$$\mathbf{w} \Leftarrow E\{\mathbf{x}g(\mathbf{w}^T\mathbf{x})\} - E\{g'(\mathbf{w}^T\mathbf{x})\}\mathbf{w}$$

Note that we still use the same representation $\mathbf{w}$ for the left-hand side, while its value is actually multiplied by a scaler. This is taken care of by renormalization, as shown in the following FastICA algorithm:

1. Choose an initial random guess for $\mathbf{w}$
2. Iterate:
$$\mathbf{w} \Leftarrow E\{\mathbf{x}g(\mathbf{w}^T\mathbf{x})\} - E\{g'(\mathbf{w}^T\mathbf{x})\}\mathbf{w}$$
3. Normalize:
$$\mathbf{w} \Leftarrow \mathbf{w}/||\mathbf{w}||$$
4. If not converged, go back to step 2.

This is a demohttp://www.cis.hut.fi/projects/ica/icademo/ of the FastICA algorithm.

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 $X = \phi^{-1}(Y) = \psi(Y)$:

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

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:

$$J(X_1, \cdots, X_n) = \begin{vmatrix} \frac{\partial \phi_1}{\partial X_1} & \cdots & \frac{\partial \phi_1}{\partial X_n} \\ \cdots & \cdots & \cdots \\ \frac{\partial \phi_n}{\partial X_1} & \cdots & \frac{\partial \phi_n}{\partial X_n} \end{vmatrix}$$

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 $A = [a_{ij}]_{n \times n}$. Again, the equation holds if the transform is unique.

Newton-Raphson Method (Uni-Variate)

To solve an algebraic equation $f(x) = 0$, select a random initial guess x_0 and follow the iteration:

$$x \Leftarrow x - \frac{f(x)}{f'(x)}$$

This Newton-Raphson formula can be derived below. The equation of the tangent of $f(x)$ at $x = x_0$ is

$$f'(x_0) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

If we let $f(x_1) = 0$, i.e., x_1 is the zero crossing of the tangent, we get

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

which is closer to the desired solution than x_0. Repeating the process we will get $x_2, x_3, \cdots$, which approach the actual solution.

../figures/newtonraphson.gif

Newton-Raphson Method (Multi-Variate)

The above method can be generalized to multi-variate case to solve n simultaneous algebraic equations

$$\begin{cases} f_1(x_1, \cdots, x_n) = f_1(\mathbf{x}) = 0 \\ \cdots\cdots\cdots \\ f_n(x_1, \cdots, x_n) = f_n(\mathbf{x}) = 0 \end{cases}$$

where $\mathbf{x} = [x_1, \cdots, x_n]^T$ is an n-dimensional vector. This equation system can be more concisely represented in vector form as $\mathbf{f}(\mathbf{x}) = 0$. The Newton-Raphson formula for multi-variate problem is

$$\mathbf{x} \Leftarrow \mathbf{x} - J_f^{-1}(\mathbf{x}) f(\mathbf{x})$$

where $J_f(\mathbf{x})$ is the Jacobian of function $\mathbf{f}(\mathbf{x})$:

$$J_f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \cdots & \cdots & \cdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}$$

To derive this iteration, consider Taylor series

$$f_i(\mathbf{x} + \delta\mathbf{x}) = f_i(\mathbf{x}) + \sum_j \frac{\partial f_i}{\partial x_j} \delta x_j + O(\delta\mathbf{x}^2) \qquad (i = 1, \cdots, n)$$

We ignore the terms of $\delta\mathbf{x}^2$ and higher and let $f_i(\mathbf{x} + \delta\mathbf{x})$ be zero (i.e., $\mathbf{x} + \delta\mathbf{x}$ is the zero-crossing of the tangent), and get

$$\sum_j \frac{\partial f_i}{\partial x_j} \delta x_j = -f_i(\mathbf{x}) \qquad (i = 1, \cdots, n)$$

Solving this linear equation system for δx_j, we get

$$\delta\mathbf{x} = -J_f^{-1}(\mathbf{x}) f(\mathbf{x})$$

and the Newton-Raphson formula:

$$\mathbf{x} \Leftarrow \mathbf{x} + \delta\mathbf{x} = \mathbf{x} - J_f^{-1}(\mathbf{x}) f(\mathbf{x})$$