# Independent Component Analysis and FastICA



June 2016

*last update: April 5, 2020*

TABLE OF CONTENTS

This note is to introduce independent component analysis and FastICA method developed by Hyvarinen et al [1].

## 1. INTRODUCTION

Independent component analysis (ICA) is a powerful technique for separating an observed multivariate signal into statistically independent non-Gaussian components. For example, let's have $n$ independent (non-Gaussian) sound sources installed at different locations in a room. Assuming no time delay and echoes, a microphone placed somewhere in the room can pick up an audio signal that is a linear mixture of the sound sources due to different proximities to the sound sources from the microphone. By sampling the audio signals from $n$ microphones located differently, the independent sound sources can be retrieved using ICA on the sampled audio signals.

ICA shares certain similarity with the well-known principle component analysis (PCA) method. But unlike the PCA that imposes strong assumption of Gaussian features and seek variance maximization and ensure uncorrelatedness only by the first and second moments of the signal, the ICA exploits inherently non-Gaussian features for independence and employs information from higher moments. For ICA to function effectively, the independent components of source signal must be non-gaussian. The non-gaussianity is crucial, because a *whitened* multivariate Gaussian signal, in which the components are uncorrelated and of unit variance, has a completely symmetric multivariate density. The symmetricity provides no information on the directions of the columns of the mixing matrix and hence makes the matrix estimation impossible. Although the ICA model can still be applied to Gaussian signals, it can only be estimated up to an orthogonal transformation and the mixing matrix cannot be identified.

## 2. INDEPENDENCE BY NON-GAUSSIANITY

As mentioned, the key to estimating ICA model is non-gaussianity. From the central limit theorem, we know that the distribution of a sum of independent random variables tends toward a Gaussian distribution. In other words, a sum of two independent random variables usually has a distribution closer to Gaussian than any of the two original random variables. Hence identifying the independent source

signal $s$ from the observed mixture $v$ is equivalent to finding a transformation matrix $B$ such that the non-gaussianity of $Bv$ is maximized.

## 2.1. Differential Entropy

To find a good measure of the objective, i.e. non-gaussianity, we must introduce a quantity, known as entropy, for a probability distribution. In information theory, entropy is a measure of randomness. For a discrete random variable $X$, it is defined by

$$\mathbb{H}[X] = -\sum_i P(X = \xi_i) \log P(X = \xi_i) \tag{1}$$

and for continuous random variables, in which case it is often called differential entropy, is defined by

$$\mathbb{H}[X] = -\int_\Omega p_X(\xi) \log p_X(\xi)\, d\xi = -\mathbb{E}[\log p_X(X)] \tag{2}$$

Here we want to show one property of the differential entropy that is of special interest to us when we come to the ICA method. Suppose $Y = f(X)$ is an invertible transformation from random vector $X$ to random vector $Y$, we want to find the connection between the entropy $\mathbb{H}[Y]$ and $\mathbb{H}[X]$. Firstly, the density of $Y$ can be formulated as

$$p_Y(y) = \frac{p_X\left(f^{-1}(y)\right)}{\left|\det J\left(f^{-1}(y)\right)\right|} \tag{3}$$

where $J(\xi)$ is the Jacobian matrix of function $f(X)$ evaluated at $\xi$. Hence the entropy $\mathbb{H}[Y]$ becomes

$$\mathbb{H}[Y] = -\mathbb{E}[\log p_Y(Y)] = -\mathbb{E}\left[\log \frac{p_X\left(f^{-1}(Y)\right)}{\left|\det J\left(f^{-1}(Y)\right)\right|}\right] = \mathbb{H}[X] + \mathbb{E}[\log|\det J(X)|] \tag{4}$$

It shows the entropy is changed in the transformation by $\mathbb{E}[\log|\det J(X)|]$ amount. In a special case where the transformation is linear, e.g. $Y = MX$, we obtain

$$\mathbb{H}[Y] = \mathbb{H}[X] + \log|\det M| \tag{5}$$

## 2.2. Maximum Entropy Probability Distribution

Assume that $X$ is a continuous random variable with density $p_X(\xi)$ and the information available on the density $p_X(\xi)$ is of the form (e.g. if $F_i(\xi) = \xi$, it gives the mean of $X$)

$$\int_{\Omega} p_X(\xi) F_i(\xi) d\xi = \mathbb{E}[F_i(X)] = c_i \;\; \forall \;\; i = 1, \cdots, n \tag{6}$$

we want to find a density $p_X(\xi)$ that maximizes the entropy subject to the constraints in (6). To do so, we may consider the following functional

$$\begin{aligned} \mathcal{J}[p_X(\xi)] = &- \int_{\Omega} p_X(\xi) \log p_X(\xi)\, d\xi + a_0 \left( \int_{\Omega} p_X(\xi) d\xi - 1 \right) \\ &+ \sum_{i=1}^{n} a_i \left( \int_{\Omega} p_X(\xi) F_i(\xi) d\xi - c_i \right) \end{aligned} \tag{7}$$

where the $a_i \; \forall \; i = 0, \cdots, n$ are the Lagrange multipliers. The zeroth constraint ensures the total probability sums to 1. The entropy attains a maximum when the density function satisfies *Euler-Lagrange equation*, i.e. the functional derivative must equal zero

$$\frac{\delta \mathcal{J}[f(\xi)]}{\delta f(\xi)} = \frac{\partial L}{\partial f} - \frac{d}{d\xi} \frac{\partial L}{\partial \dot{f}} = 0 \quad \text{for} \quad \mathcal{J}[f(\xi)] = \int_{\Omega} L\left(\xi, f(\xi), \dot{f}(\xi)\right) dx \tag{8}$$

where $\dot{f}$ stands for the first derivative of function $f$. Since the $\dot{f}$ does not appear explicitly in integrands of $\mathcal{J}[p_X(\xi)]$, the second term in *Euler-Lagrange equation* vanishes for all function $f$ and thus

$$\frac{\delta \mathcal{J}[p_X(\xi)]}{\delta p_X(\xi)} = -\log p_X(\xi) - 1 + a_0 + \sum_{i=1}^{n} a_i F_i(\xi) = 0 \tag{9}$$

Hence the density $\tilde{p}_X(\xi)$ with maximum entropy under the constraints in (6) must be of the form

$$\tilde{p}_X(\xi) = \exp\left( -1 + a_0 + \sum_{i=1}^{n} a_i F_i(\xi) \right) = A \exp\left( \sum_{i=1}^{n} a_i F_i(\xi) \right) \tag{10}$$

In the case that the constraints are limited to the mean $\mathbb{E}[X] = \mu$ and the variance $\mathbb{E}[(X-\mu)^2] = \sigma^2$, the (univarate) density of maximum entropy in (10) can be further derived into

$$\tilde{p}_X(\xi) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left( -\frac{(\xi - \mu)^2}{2\sigma^2} \right) \tag{11}$$

which is the well-known Gaussian density.

## 2.3. Measure of Non-gaussianity: Negentropy

A fundamental result from previous discussion is that a Gaussian variable has the largest entropy among all random variables of equal mean and variance. This means that entropy could be used as a measure of non-gaussianity. To obtain such a measure that is zero for Gaussian and positive for all other random variables, one often uses a slightly modified version of the differential entropy, called negentropy $\Theta$, which is defined as

$$\Theta[X] = \mathbb{H}[\mathcal{Z}] - \mathbb{H}[X] \tag{12}$$

where $X$ is an arbitrary (multivariate) random variable and $\mathcal{Z}$ is a (multivariate) Gaussian of the same covariance matrix as $X$. It is evident that the negentropy is always non-negative and becomes zero if and only if $X$ is a Gaussian. An interesting property arises from (5) is that the negentropy (12) is invariant for an invertible linear transformation $Y = MX$, that is

$$\Theta[Y] = \Theta[MX] = \mathbb{H}[\mathcal{Z}] + \log|\det M| - \mathbb{H}[X] - \log|\det M| = \Theta[X] \tag{13}$$

Negentropy appears to be a good measure of the non-gaussianity. However, the use of negentropy turns out to be computationally challenging. This is because, practically it is quite difficult to accurately and efficiently estimate the density function, which is the key ingredient in computation of the integral in (2). Therefore, some approximations [2], though possibly rather coarse, have to be used. In the following, we will introduce two of such approximation methods.

### 2.4. Approximation of Negentropy by Cumulants

The first approximation is based on cumulants of random variables. The cumulants [3] of a random variable $X$, denoted by $\kappa_n[X]$, are given by cumulant generating function $C_X(t)$, which is defined as logarithm of characteristic function $\varphi_X(t)$ of $X$, that is $C_X(t) = \log \varphi_X(t)$. The characteristic function $\varphi_X(t)$ is a Fourier transform of the density function and it can be used to generate raw moments $\nu_n[X]$

$$\varphi_X(t) = \mathbb{E}[\exp(itX)] = 1 + \sum_{n=1}^{\infty} \nu_n \frac{(it)^n}{n!} \qquad \text{and} \qquad \nu_n[X] = \mathbb{E}[X^n] = i^{-n}\varphi_X^{(n)}(0) \tag{14}$$

where $\varphi_X^{(n)}(0)$ denotes the $n$-th derivative of the characteristic function evaluated at $0$. Since the cumulant generating function $C_X(t)$ is a composition of two functions, which have power series expansions itself,

it also has a power series expansion (in Maclaurin series, i.e. a Taylor series centered at zero). The cumulant $\kappa_n[X]$ can be derived from the $C_X(t)$ by

$$C_X(t) = \log \varphi_X(t) = \sum_{n=1}^{\infty} \kappa_n[X] \frac{(it)^n}{n!} \qquad \text{and} \qquad \kappa_n[X] = i^{-n} C_X^{(n)}(0) \tag{15}$$

For example, when written in terms of the raw moments, the first 4 cumulants appear like

$$\begin{aligned} &\kappa_1 = \nu_1, \qquad \kappa_2 = \nu_2 - \nu_1^2, \qquad \kappa_3 = \nu_3 - 3\nu_2\nu_1 + 2\nu_1^3 \\ &\kappa_4 = \nu_4 - 3\nu_2^2 - 4\nu_3\nu_1 + 12\nu_2\nu_1^2 - 6\nu_1^4 \end{aligned} \tag{16}$$

The approximation imposes an assumption that the random variable $X$ is close to a standard Gaussian. Let $\phi(\xi)$ be the density of a (univariate) standard Gaussian

$$\phi(\xi) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\xi^2}{2}\right) \tag{17}$$

Derivatives of the density $\phi(\xi)$ defines a series of functions $h_n(\xi)$, known as Hermite polynomials

$$\phi^{(n)}(\xi) = (-1)^n h_n(\xi)\phi(\xi) \tag{18}$$

For example, the first a few Hermite polynomials are as follows

$$\begin{aligned} &h_0(\xi) = 1, \qquad h_1(\xi) = \xi, \qquad h_2(\xi) = \xi^2 - 1, \qquad h_3(\xi) = \xi^3 - 3\xi \\ &h_4(\xi) = \xi^4 - 6\xi^2 + 3, \qquad h_5(\xi) = \xi^5 - 10\xi^3 + 15\xi, \qquad \cdots \end{aligned} \tag{19}$$

The Hermite polynomials form an orthogonal system, such that

$$\mathbb{E}[h_m(Z)h_n(Z)] = \int_{\mathbb{R}} \phi(\xi)h_m(\xi)h_n(\xi)d\xi = \begin{cases} n! & \text{if } m = n \\ 0 & \text{if } m \neq n \end{cases} \tag{20}$$

where $Z$ is a standard Gaussian. Because the density $P_X(\xi)$ of random variable $X$ is close to $\phi(\xi)$, it can be approximated by *Gram-Charlier* expansion [4] truncated to include first two non-constant terms. The approximation of $P_X(\xi)$ is defined in terms of cumulants $\kappa_n[X]$ of random variable $X$ as follows

$$p_X(\xi) \approx \hat{p}_X(\xi) = \phi(\xi)\left(1 + \frac{\kappa_3[X]h_3(\xi)}{3!} + \frac{\kappa_4[X]h_4(\xi)}{4!} + \cdots\right) \tag{21}$$

The next step is to estimate the entropy using the $\hat{p}_X(\xi)$ in (21). Since $P_X(\xi)$ is close to $\phi(\xi)$, the cumulant $\kappa_3[X]$ and $\kappa_4[X]$ are small, we are able to apply the approximation $\log(1+\epsilon) = \epsilon - \epsilon^2/2 + o(\epsilon^3)$ to get

$$
\begin{aligned}
\mathbb{H}[X] &\approx -\int_\Omega \hat{p}_X(\xi)\log\hat{p}_X(\xi)\,d\xi \\
&\approx -\int_\Omega \phi(\xi)\left(1+\frac{\kappa_3[X]h_3(\xi)}{3!}+\frac{\kappa_4[X]h_4(\xi)}{4!}\right)\left(\log\phi(\xi)+\frac{\kappa_3[X]h_3(\xi)}{3!}+\frac{\kappa_4[X]h_4(\xi)}{4!}\right. \\
&\qquad\left.-\frac{1}{2}\left(\frac{\kappa_3[X]h_3(\xi)}{3!}+\frac{\kappa_4[X]h_4(\xi)}{4!}\right)^2\right)d\xi \\
&= -\int_\Omega \phi(\xi)\log\phi(\xi)\,d\xi + \underbrace{\int_\Omega \phi(\xi)\log\phi(\xi)\left(\frac{\kappa_3[X]h_3(\xi)}{3!}+\frac{\kappa_4[X]h_4(\xi)}{4!}\right)d\xi}_{=0} \\
&\quad \underbrace{-\int_\Omega \phi(\xi)\left(\frac{\kappa_3[X]h_3(\xi)}{3!}+\frac{\kappa_4[X]h_4(\xi)}{4!}\right)d\xi}_{=0} - \frac{1}{2}\int_\Omega \phi(\xi)\left(\frac{\kappa_3[X]h_3(\xi)}{3!}+\frac{\kappa_4[X]h_4(\xi)}{4!}\right)^2 d\xi \\
&\quad \underbrace{+\frac{1}{2}\int_\Omega \phi(\xi)\left(\frac{\kappa_3[X]h_3(\xi)}{3!}+\frac{\kappa_4[X]h_4(\xi)}{4!}\right)^3 d\xi}_{\approx 0} \\
&\approx \mathbb{H}[Z] - \frac{1}{2}\left(\frac{\kappa_3[X]^2}{3!}+\frac{\kappa_4[X]^2}{4!}\right)
\end{aligned}
\tag{22}
$$

Here we have used the orthogonality in (20), and in particular the fact that $h_3(\xi)$ and $h_4(\xi)$ are orthogonal to any second-order polynomials. Furthermore, because both $\kappa_3[X]$ and $\kappa_4[X]$ are small, their third order monomials are much smaller than terms involving only second order monomials. Given the results in (22), we may approximate the negentropy of random variable $X$ that is close to a Gaussian by

$$
\Theta[X] = \mathbb{H}[Z] - \mathbb{H}[X] = \frac{\kappa_3[X]^2}{12} + \frac{\kappa_4[X]^2}{48} = \frac{\mathbb{E}[X^3]^2}{12} + \frac{\mathrm{kurt}[X]^2}{48}
\tag{23}
$$

where $\quad \mathrm{kurt}[X] = \mathbb{E}[X^4] - 3\mathbb{E}[X^2]^2$

It is obvious that this approximation of negentropy is computationally very efficient. However, it possesses another issue that it is very sensitive to outliers. It mainly measures the tails and is largely

unaffected by structure near the center of the distribution. In order to mitigate this issue, another more robust approximation of negentropy is developed as follows.

## 2.5. Approximation of Negentropy by Nonpolynomial Functions

Let's assume again that we have observed (or estimated) a number of expectations of different functions of $X$ as shown in (6). The function $F_i(\xi)$ are in general not polynomials. Similarly we make a simple approximation $\tilde{p}_X(\xi)$ of the maximum entropy density based on the assumption that the true density $p_X(\xi)$ is not far from the Gaussian density of the same mean and variance. To make it simpler, let's assume that the $X$ has zero mean and unit variance so that we can put two additional constraints into (6), defined by

$$F_{n+1}(\xi) = \xi, \qquad c_{n+1} = \mathbb{E}[X] = 0 \qquad \text{and} \qquad F_{n+2}(\xi) = \xi^2, \qquad c_{n+2} = \mathbb{E}[X^2] = 1 \tag{24}$$

which makes the (10) into

$$\tilde{p}_X(\xi) = A \exp\left(\sum_{i=1}^{n+2} a_i F_i(\xi)\right) \tag{25}$$

To further simplify the calculations, let us make another, purely technical assumption: The functions $F_i(\xi)$, form an orthonormal system by the measure $\phi(\xi)$ and are orthogonal to all polynomials up to second degree (similar to the orthogonality of Hermite polynomials). In other words, we have

$$\begin{aligned} &\mathbb{E}\big[F_i(Z)F_j(Z)\big] = \int_\Omega \phi(\xi)F_i(\xi)F_j(\xi)d\xi = \delta(i,j) \;\; \forall \;\; i = 1, \cdots, n \quad \text{and} \quad \delta(i,j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \\ &\mathbb{E}[F_i(Z)Z^k] = \int_\Omega \phi(\xi)F_i(\xi)\xi^k d\xi = 0 \;\; \forall \;\; k = 0, 1, 2 \end{aligned} \tag{26}$$

where $Z$ is a standard Gaussian. Due to the assumption of near-gaussianity, the exponential in (25) is not far from $\exp(-\xi^2/2)$, thus all the other $a_i$ in (25) are small compared to $a_{n+2} \approx -1/2$. Hence we can make a first-order approximation of the exponential function using expansion $\exp(\epsilon) \approx 1 + \epsilon + o(\epsilon^2)$

$$\tilde{p}_X(\xi) = A \exp\left(\sum_{i=1}^{n+2} a_i F_i(\xi)\right) = A \exp\left(-\frac{\xi^2}{2} + a_{n+1}\xi + \frac{2a_{n+2}+1}{2}\xi^2 + \sum_{i=1}^{n} a_i F_i(\xi)\right) \tag{27}$$

$$\approx \mathcal{A}\phi(\xi)\left(1 + a_{n+1}\xi + \frac{2a_{n+2}+1}{2}\xi^2 + \sum_{i=1}^{n} a_i F_i(\xi)\right) \qquad \text{where} \qquad \mathcal{A}\phi(\xi) = A\exp\left(-\frac{\xi^2}{2}\right)$$

By orthogonality in (26) and use the following expectation of powers of standard Gaussian

$$\mathbb{E}[Z] = 0, \qquad \mathbb{E}[Z^2] = 1, \qquad \mathbb{E}[Z^3] = 0, \qquad \mathbb{E}[Z^4] = 3 \tag{28}$$

we can derive a series of equations as follows

$$\begin{aligned}
&\int_\Omega \hat{p}_X(\xi)d\xi = \mathcal{A}\mathbb{E}\left[1 + a_{n+1}Z + \frac{2a_{n+2}+1}{2}Z^2 + \sum_{i=1}^{n} a_i F_i(Z)\right] = \mathcal{A}\left(\frac{3}{2} + a_{n+2}\right) = 1 \\
&\int_\Omega \hat{p}_X(\xi)\xi d\xi = \mathcal{A}\mathbb{E}\left[Z + a_{n+1}Z^2 + \frac{2a_{n+2}+1}{2}Z^3 + \sum_{i=1}^{n} a_i F_i(Z)Z\right] = \mathcal{A}a_{n+1} = 0 \\
&\int_\Omega \hat{p}_X(\xi)\xi^2 d\xi = \mathcal{A}\mathbb{E}\left[Z^2 + a_{n+1}Z^3 + \frac{2a_{n+2}+1}{2}Z^4 + \sum_{i=1}^{n} a_i F_i(Z)Z^2\right] = \mathcal{A}\left(\frac{5}{2} + 3a_{n+2}\right) = 1 \\
&\int_\Omega \hat{p}_X(\xi)F_i(\xi)d\xi = \mathcal{A}\mathbb{E}\left[F_i(Z)\left(1 + a_{n+1}Z + \frac{2a_{n+2}+1}{2}Z^2 + \sum_{j=1}^{n} a_j F_j(Z)\right)\right] = \mathcal{A}a_i = c_i
\end{aligned} \tag{29}$$

These equations can be solved to give

$$\mathcal{A} = 1, \qquad a_i = c_i \;\forall\, i = 1, \cdots, n, \qquad a_{n+1} = 0, \qquad a_{n+2} = -\frac{1}{2} \tag{30}$$

The approximative maximum entropy density, denoted by $\hat{p}_X(\xi)$, can then be obtained from (27) using the solved constants, that is

$$\hat{p}_X(\xi) = \phi(\xi)\left(1 + \sum_{i=1}^{n} c_i F_i(\xi)\right) \quad \text{with} \quad c_i = \mathbb{E}[F_i(X)] \tag{31}$$

Given the density above, we can approximate the differential entropy as

$$\begin{aligned}
&\mathbb{H}[X] \approx -\int_\Omega \hat{p}_X(\xi)\log\hat{p}_X(\xi)\,d\xi \\
&\approx -\int_\Omega \phi(\xi)\left(1 + \sum_{i=1}^{n} c_i F_i(\xi)\right)\left(\log\phi(\xi) + \log\left(\sum_{i=1}^{n} c_i F_i(\xi) - \frac{1}{2}\left(\sum_{i=1}^{n} c_i F_i(\xi)\right)^2\right)\right)d\xi
\end{aligned} \tag{32}$$

$$\approx \mathbb{H}[Z] - \sum_{i=1}^{n} c_i \underbrace{\mathbb{E}[\log \phi(Z)\, F_i(Z)]}_{=0} - \sum_{i=1}^{n} c_i \underbrace{\mathbb{E}[F_i(Z)]}_{=0} - \frac{1}{2}\sum_{i,j=1}^{n} c_i c_j \mathbb{E}\big[F_i(Z)F_j(Z)\big]$$

$$= \mathbb{H}[Z] - \frac{1}{2}\sum_{i=1}^{n} c_i^2$$

where another approximative expansion $(1+\epsilon)\log(1+\epsilon) \approx \epsilon + \epsilon^2/2 + o(\epsilon^3)$ has been applied considering the quantity $\epsilon$ is much smaller than 1. The negentropy approximation is then given by

$$\Theta[X] = \frac{1}{2}\sum_{i=1}^{n} c_i^2 = \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}[F_i(X)]^2 \tag{33}$$

There is a special and yet simple case of (31) can be obtained by using two functions: an odd function $G_1(\xi)$ and an even function $G_2(\xi)$. The odd function measures the asymmetry and the even function measures the dimension of bimodality vs. peak at zero, closely related to sub- vs. super-gaussianity [5]. First we define $F_1(\xi)$ and $F_2(\xi)$ as

$$F_1(\xi) = \frac{G_1(\xi) + \alpha\xi}{\lambda_1}, \qquad F_2(\xi) = \frac{G_2(\xi) + \beta_1 + \beta_2\xi^2}{\lambda_2} \tag{34}$$

Since $G_1(\xi)$ and $G_2(\xi)$ are odd and even function respectively, and so are the $F_1(\xi)$ and $F_2(\xi)$, the orthogonality $\mathbb{E}[F_1(Z)F_2(Z)] = 0$ is automatically satisfied. To determine constants $\alpha, \beta_1, \beta_2, \lambda_1, \lambda_2$, the functions $F_1(\xi)$ and $F_2(\xi)$ must satisfy the orthonormality in (26) by measure $\phi(\xi)$, which defines the following equations

$$\begin{aligned}
&\mathbb{E}[F_1(Z)^2] = \frac{\mathbb{E}[G_1(Z)^2] + 2\alpha\mathbb{E}[G_1(Z)Z] + \alpha^2\mathbb{E}[Z^2]}{\lambda_1^2} = \frac{\mathbb{E}[G_1(Z)^2] + 2\alpha\mathbb{E}[G_1(Z)Z] + \alpha^2}{\lambda_1^2} = 1 \\
&\mathbb{E}[F_1(Z)Z] = \frac{\mathbb{E}[G_1(Z)Z] + \alpha\mathbb{E}[Z^2]}{\lambda_1} = \frac{\mathbb{E}[G_1(Z)Z] + \alpha}{\lambda_1} = 0 \\
&\mathbb{E}[F_2(Z)^2] = \frac{\mathbb{E}[G_2(Z)^2] + \beta_1^2 + \beta_2^2\mathbb{E}[Z^4] + 2\beta_1\mathbb{E}[G_2(Z)] + 2\beta_1\beta_2\mathbb{E}[Z^2] + 2\beta_2\mathbb{E}[G_2(Z)Z^2]}{\lambda_2^2} \\
&\qquad = \frac{\mathbb{E}[G_2(Z)^2] + \beta_1^2 + 3\beta_2^2 + 2\beta_1\mathbb{E}[G_2(Z)] + 2\beta_1\beta_2 + 2\beta_2\mathbb{E}[G_2(Z)Z^2]}{\lambda_2^2} = 1
\end{aligned} \tag{35}$$

$$\mathbb{E}[F_2(Z)] = \frac{\mathbb{E}[G_2(Z)] + \beta_1 + \beta_2\mathbb{E}[Z^2]}{\lambda_2} = \frac{\mathbb{E}[G_2(Z)] + \beta_1 + \beta_2}{\lambda_2} = 0$$

$$\mathbb{E}[F_2(Z)Z^2] = \frac{\mathbb{E}[G_2(Z)Z^2] + \beta_1\mathbb{E}[Z^2] + \beta_2\mathbb{E}[Z^4]}{\lambda_2} = \frac{\mathbb{E}[G_2(Z)Z^2] + \beta_1 + 3\beta_2}{\lambda_2} = 0$$

The constants can be estimated from the equations above by

$$\begin{aligned} &\mathbb{E}[G_1(Z)Z] = -\alpha, \qquad \mathbb{E}[G_1(Z)^2] = \lambda_1^2 + \alpha^2, \qquad \mathbb{E}[G_2(Z)] = -\beta_1 - \beta_2 \\ &\mathbb{E}[G_2(Z)Z^2] = -\beta_1 - 3\beta_2, \qquad \mathbb{E}[G_2(Z)^2] = \lambda_2^2 + \beta_1^2 + 2\beta_1\beta_2 + 3\beta_2^2 \end{aligned} \tag{36}$$

Given the assumption (24) that $X$ has zero mean and unit variance, we can further write

$$\begin{aligned} c_1 &= \mathbb{E}[F_1(X)] = \frac{\mathbb{E}[G_1(X)] + \alpha\mathbb{E}[X]}{\lambda_1} = \frac{\mathbb{E}[G_1(X)]}{\lambda_1} \\ c_2 &= \mathbb{E}[F_2(X)] = \frac{\mathbb{E}[G_2(X)] + \beta_1 + \beta_2\mathbb{E}[X^2]}{\lambda_2} = \frac{\mathbb{E}[G_2(X)] + \beta_1 + \beta_2}{\lambda_2} = \frac{\mathbb{E}[G_2(X)] - \mathbb{E}[G_2(Z)]}{\lambda_2} \end{aligned} \tag{37}$$

where the denominators are estimated as follows

$$\begin{aligned} \lambda_1^2 &= \mathbb{E}[G_1(Z)^2] - \mathbb{E}[G_1(Z)Z]^2 \\ \lambda_2^2 &= \mathbb{E}[G_2(Z)^2] - \mathbb{E}[G_2(Z)]^2 - \frac{1}{2}(\mathbb{E}[G_2(Z)] - \mathbb{E}[G_2(Z)Z^2])^2 \end{aligned} \tag{38}$$

Hence the negentropy approximation in (33) can be expressed as

$$\Theta[X] = \frac{c_1^2 + c_2^2}{2} = k_1\mathbb{E}[G_1(X)]^2 + k_2(\mathbb{E}[G_2(X)] - \mathbb{E}[G_2(Z)])^2, \qquad k_1 = \frac{1}{2\lambda_1^2}, \qquad k_2 = \frac{1}{2\lambda_2^2} \tag{39}$$

Note that in real applications, the assumption that the true density of $X$ is not too far from the Gaussian can be easily violated. However, even in the case that the approximation is not very accurate, the (39) can still be used to construct a non-gaussianity measure that is consistent in the sense that it is always nonnegative, and equal to zero if $X$ is a Gaussian.

To make the approximation more robust, the functions $G_i(\xi)$ must be chosen wisely. The selective criteria should consider: 1). estimation of $\mathbb{E}[G_i(X)]$ should be statistically easy and insensitive to outliers. 2). the $G_i(\xi)$ must not grow faster than quadratically to ensure that the density in (31) is integrable and therefore to ensure that the maximum entropy distribution exists in the first place. 3). the $G_i(\xi)$ must

capture aspects of the distribution of $X$ that are pertinent in the computation of entropy. For example, if we use simple polynomials (which apparently are bad choices), we would end up with something very similar to what we have in the preceding section. This can be seen by letting $G_1(\xi) = \xi^3$ and $G_2(\xi) = \xi^4$, then from (38) (40)we find

$$\begin{aligned} \lambda_1^2 &= \mathbb{E}[Z^6] - \mathbb{E}[Z^4]^2 = 15 - 9 = 6 \\ \lambda_2^2 &= \mathbb{E}[Z^8] - \mathbb{E}[Z^4]^2 - \frac{1}{2}(\mathbb{E}[Z^4] - \mathbb{E}[Z^6])^2 = 105 - 9 - \frac{1}{2}(3 - 15)^2 = 24 \end{aligned} \tag{40}$$

and hence the negentropy approximation in (39) becomes

$$\Theta[X] = \frac{\mathbb{E}[X^3]^2}{12} + \frac{(\mathbb{E}[X^4] - 3)^2}{48} \tag{41}$$

which is identical to that of (23).

When using only one non-quadratic even function, an even simpler approximation of negentropy can be obtained. This amounts to omitting the term associated with the odd function in (39) to give

$$\Theta[X] \propto (\mathbb{E}[G(X)] - \mathbb{E}[G(Z)])^2 \tag{42}$$

It has been suggested [6] that the following two functions exhibit very good properties

$$G(\xi) = \frac{1}{a}\log\cosh(a\xi) \quad \forall\, 1 \le a \le 2 \qquad \text{and} \qquad G(\xi) = -\exp\left(-\frac{\xi^2}{2}\right) \tag{43}$$

The first one is useful for general purposes, while the second on may be highly robust.

## 3. FastICA using Negentropy

In this section, we are going to summarize a numerical method, known as FastICA method [7], for ICA estimation. Suppose we are able to observe signal $V$ such that

$$V = MS \tag{44}$$

are linearly mixed by an $n \times n$ invertible matrix $M$ of independent source signal $S$. Both $V$ and $S$ are $n \times 1$ vectors. We want to the independent components by maximizing the statistical independence of the estimated components.

### 3.1. Preprocessing for ICA

Preprocessing the observed signal $V$ for ICA generally involves two steps: centering and whitening. Centering means to subtract the mean from the signal. Whitening is to linearly transform the observed signal so that the transformed signal is white, i.e. its components are uncorrelated and their variances equal unity. While centering appears simple, there can be many ways to perform whitening. One of the methods we want to discuss is the PCA whitening, which performs eigenvalue decomposition on covariance matrix $\Sigma$ of the observed signal $V$ such that

$$\Sigma[V] = E\Lambda E' \tag{45}$$

Whitening can then be done on the centered signal by

$$X = \Lambda^{-\frac{1}{2}} E'(V - \mathbb{E}[V]) \tag{46}$$

where $X$ is the whitened signal that will be subject to FastICA estimation. After centering and whitening, we can show that the signal $X$ has the following properties

$$\mathbb{E}[X] = 0 \quad \text{and} \quad \mathbb{E}[XX'] = I \tag{47}$$

where $I$ is an identity matrix.

## 3.2. Maximization of Negentropy

Suppose we want to find an invertible matrix $B$ such that the linear transformation $Y = BX$ produces a signal $Y$ whose components are mutually independent with unit variance. The independence of the components requires that they are uncorrelated, and in the whitened space we must have

$$I = \mathbb{E}[YY'] = \mathbb{E}[BXX'B'] = BB' \tag{48}$$

This means after whitening, the matrix $B$ can be taken to be orthonomal (i.e. orthogonal and normalized). Hence we want to maximize the negentropy of each component subject to the constraint that components are mutually independent.

Let $b$ be the transpose of a row vector of matrix $B$. We want to find the $b$ such that it maximizes the negentropy of the component on the unit sphere, i.e. $b'b = 1$. The constrained maximization can be done by using Lagrange multiplier, that is

$$\mathcal{L}(b,\lambda) = (\mathbb{E}[G(b'X)] - \mathbb{E}[G(Z)])^2 + \lambda(b'b - 1) \tag{49}$$

The first term on the right hand side comes from the approximation of negentropy and the second is the optimization constraint. The optimality condition requires

$$\begin{aligned} &\frac{\partial \mathcal{L}(b,\lambda)}{\partial b} = 2\gamma\mathbb{E}\left[\dot{G}(b'X)X\right] + 2\lambda b = 0 \quad \text{where} \quad \gamma = \mathbb{E}[G(b'X)] - \mathbb{E}[G(Z)] \\ &\Longrightarrow \mathbb{E}\left[\dot{G}(b'X)X\right] + \frac{\lambda}{\gamma}b = 0 \end{aligned} \tag{50}$$

Since $\gamma$ is a scalar, the vector $\mathbb{E}\left[\dot{G}(b'X)X\right]$ at maximum must be equal to vector $b$ multiplied by a scalar constant. In order words, the vector $\mathbb{E}\left[\dot{G}(b'X)X\right]$ must point in the same direction of $b$. Because in every iteration, the vector $b$ is normalized, this leads to a fixed point scheme

$$b \leftarrow \text{Normalize}\left(\mathbb{E}\left[\dot{G}(b'X)X\right]\right) \tag{51}$$

However, the fixed point iteration is often accompanied by slow convergence rate. Instead, the solution is sought by solving a nonlinear equation for an arbitrary scalar $\beta$

$$\mathbb{E}\left[\dot{G}(b'X)X\right] + \beta b = 0 \tag{52}$$

using Newton's method. Denoting the left hand side of (52) by

$$F(b) = \mathbb{E}\left[\dot{G}(b'X)X\right] + \beta b \tag{53}$$

the Newton's method defines another fixed point iteration, such that

$$b \leftarrow \text{Normalize}\left(b - \left(\dot{F}(b)\right)^{-1} F(b)\right) = \text{Normalize}\left(b - \frac{\mathbb{E}\left[\dot{G}(b'X)X\right] + \beta b}{\mathbb{E}\left[\ddot{G}(b'X)\right] + \beta}\right) \tag{54}$$

where the gradient matrix $\dot{F}(b)$ can be approximated by

$$\dot{F}(b) = \mathbb{E}\left[\ddot{G}(b'X)XX'\right] + \beta I \approx \mathbb{E}\left[\ddot{G}(b'X)\right]\mathbb{E}[XX'] + \beta I = \left(\mathbb{E}\left[\ddot{G}(b'X)\right] + \beta\right)I \tag{55}$$

given that the components of whitened $X$ are uncorrelated (though unnecessarily independent). Since the quantity $\mathbb{E}\left[\ddot{G}(b'X)\right] + \beta$ is a scalar and would be eventually eliminated by the normalization anyway, we can multiply both sides of (54) by the scalar and simplify the fixed point iteration to

$$b \leftarrow \text{Normalize}\left(\mathbb{E}\left[\dot{G}(b'X)X\right] - \mathbb{E}\left[\ddot{G}(b'X)\right]b\right) \tag{56}$$

Once the iteration converges (e.g. the $\|\Delta b\| < \epsilon$ for a small number $\epsilon$), then the $b'X$ is one of the independent components that we want to estimate. As a summary, we list the non-quadratic nonlinear candidate functions in (43) and their first and second derivatives as follows ($1 \le a \le 2$)

$$\begin{aligned} &G(\xi) = \frac{1}{a}\log\cosh(a\xi), \qquad \dot{G}(\xi) = \tanh(a\xi), \qquad \ddot{G}(\xi) = a(1-\tanh(a\xi)^2) \\ &G(\xi) = -\exp\left(-\frac{\xi^2}{2}\right), \qquad \dot{G}(\xi) = \xi\exp\left(-\frac{\xi^2}{2}\right), \qquad \ddot{G}(\xi) = (1-\xi^2)\exp\left(-\frac{\xi^2}{2}\right) \end{aligned} \tag{57}$$

3.3. Symmetric Orthogonalization

In certain applications, it may be desirable to use a symmetric decorrelation, in which no vectors are "privileged" over others. This means that the vectors $b$'s are not estimated one by one; instead, they are estimated in parallel by a fixed point iteration derived from (56)

$$B \leftarrow \text{SymmetricOrthonomalize}\left(\mathbb{E}\left[\dot{G}(BX)X'\right] - \text{Diag}\left(\mathbb{E}\left[\ddot{G}(BX)\right]\right)B\right) \tag{58}$$

where the symmetric orthonomalization of matrix $B$ is given by

$$\text{SymmetricOrthonomalize}(B) = (BB')^{-\frac{1}{2}}B \tag{59}$$

with the inverse square root of matrix $BB'$ estimated from eigenvalue decomposition

$$(BB')^{-\frac{1}{2}} = E\Lambda^{-\frac{1}{2}}E' \quad \text{for} \quad BB' = E\Lambda E' \tag{60}$$

Alternatively, another iteration based symmetric orthonomalization can be done by first normalizing $B \leftarrow B/\|B\|$ and then running the iteration

$$B \leftarrow \frac{3}{2}B - \frac{1}{2}BB'B \tag{61}$$

until the matrix $BB'$ is sufficiently close to an identity.

## REFERENCES

1. Hyvärinen, A; Oja, E., *Independent Component Analysis: Algorithms and Applications*, Neural Networks, London, 13(4-5),411-430, 2000

2. Hyvärinen, A; Karhunen, J.; Oja, E., *Independent Component Analysis*, Wiley, 2001, pp. 113-120

3. Wikipedia: https://en.wikipedia.org/wiki/Cumulant

4. Wikipedia: https://en.wikipedia.org/wiki/Edgeworth_series

5. Hyvärinen, A; Karhunen, J.; Oja, E., *Independent Component Analysis*, Wiley, 2001, pp. 118

6. Hyvärinen, A; Karhunen, J.; Oja, E., *Independent Component Analysis*, Wiley, 2001, pp. 184

7. Hyvärinen, A; Karhunen, J.; Oja, E., *Independent Component Analysis*, Wiley, 2001, pp. 188-196