INFORMATION THEORETIC APPROACHES TO RINCIPAL AND INDEPENDENT COMPONENT ANALYSIS: A SIMPLIFIED VIEW

[1]Dinesh Kumar, [2]C.S.Rai, [3]Shakti Kumar

[1]Department of Computer Science & Engineering, G J U S & T, Hisar, Haryana, India-125001

[2]University School of Information Technology, GGS I P U, Delhi, India -110006

[3]Institute of Science & Technology Klawad, Distt. Yamuna Nagar, Haryana, India- 135001

E-mail: [1]dinesh_chutani@yahoo.com , [2]csrai_ipu@yahoo.com , [3]shakti@istk.org

ABSTRACT

Principal Component Analysis (PCA) and Independent Component Analysis (ICA) are the techniques that deal with extracting the independent components from linear mixtures of Gaussian and non-Gaussian data at the input respectively. PCA is a classical method that deals with the second order statistics of data. It is also known as Karhunen-Loeve Transform or the Hotelling Transform in some application areas. ICA is a generalization of PCA that takes into account the higher order statistics also. This paper presents a simplified view of information theoretic approaches to the problem of Principal Component and Independent Component Analysis. With the help of these techniques we find a linear representation of multivariate data so that the components are as statistically independent as possible. Such representations capture essential features of the data in many applications. This paper summarizes major approaches that are based on information theoretic concepts and includes the applications of ICA.

Keywords: Information Theory, Principal Component Analysis, Independent Component Analysis

1. INTRODUCTION

Information theory has been used and developed extensively by the communication engineers ever since Shannon first formulated his ‘Mathematical Theory of Computation’ [9]. The information theory provides a framework for the study of fundamental issues such as efficiency of information representation and the limitations involved in the reliable transmission of information over a communication channel. It has proved particularly useful in the development of unsupervised learning algorithms.

One of the basic problems in Information Theory is the measurement of degree of independence or interdependence. This is the reason that the concepts of Information Theory are widely used whenever we talk of Principal and Independent Component Analysis.

Principal Component Analysis (PCA), a statistical method, has widely been used in signal processing and neural computing to find a set of basis vectors by rotating the data such that maximum variabilities are projected onto the axes. The principal components are orthogonal and projections of data onto them are linearly decorrelated and consider only second order characteristics of data. Whereas ICA seeks a transformation to coordinates in which data are statistically independent to a maximum possible extent rather than merely decorrelated. Independent Component Analysis (ICA) is a computational method for finding the components from multivariate statistical data. ICA was originally developed to deal with problems that are closely related to the cocktail party problem most commonly known as Blind Source Separation (BSS). Due to the recent increase of interest in ICA, it is widely being used in variety of other applications also such as signal processing, pattern recognition, telecommunications and medical signal processing. ICA not only decorrelates the signals (second order statistics) but also reduces higher order statistical dependencies, with an attempt to make the signals as independent as possible. The analysis of independent components constituted the study of separating mixed sources observed in an array of sensors [10-13]. A cost function related to the approximate minimization of mutual information between the sensors was proposed by Comon [14]. Linsker [15] proposed unsupervised learning rules based on information theory with the goal to maximize the mutual information between the inputs and the outputs of a neural network. Bell and Sejnowski [6] also put the blind source separation problem into information theoretic framework and demonstrated the separation and deconvolution of mixed sources. A similar adaptive method for source separation was proposed by Cardoso and Laheld [17]. Hyvärinen proposed a measure of non-Gaussianity as Negentropy, which is based on the information theoretic quantity of (differential) entropy for independent component analysis [7, 8, 20, 21]. Section 2 includes basics of PCA , ICA and information theoretic concepts. Section 3 reviews different information theoretic approaches followed by some potential applications in section 4.

2. PCA, ICA AND INFORMATION THEORY

2.1. PCA

In many real world problems, reducing dimensionality of a problem is an essential step before any analysis of data is performed. The general criterion for reducing the dimensions is the desire to preserve most of the relevant information of the original data according to some optimality criteria. PCA is concerned with explaining the variance-covariance structure of a set variable through linear combinations of these variables. Consider a random vector $X = (x_1, x_2, ..., x_p)$ .The covariance matrix of X is *C* and the eigenvalues are $\lambda_1, \lambda_2, ..., \lambda_p$ such that $\lambda_1 \geq \lambda_2 \geq, ..., \geq \lambda_p$ and the eigenvector-eigenvalue pairs are $(V_1, \lambda_1), (V_2, \lambda_2), ..., (V_p, \lambda_p)$ .Then Principal Components (PCs) are linear combinations $y_1, y_2, ..., y_p$, with the changed coordinate system, of the *p* number of random variables $x_1, x_2, ..., x_p$. The first principal component, y_1 is a linear combination of $x_1, x_2, ..., x_p$, that is

$$y_1 = b_{11}x_1 + b_{12}x_2 + ... + b_{1p}x_p = \sum_{i=1}^{p} b_{1i}x_i = b_1' X \quad (1)$$

The first principal component y_1 is such that its variance is maximized given the constraint that $b_1' b_1 = 1$. Principal components analysis finds the optimal weights vectors $(b_{11}, b_{12}, ..., b_{1p})$ and associated variance of y_1 which is usually denoted by λ_1 . The second principal component, involves finding a second weights vectors $(b_{21}, b_{22}, ..., b_{2p})$ such that the variance of y_2 is maximized subject to the constraints that $b_2' b_2 = 1$ and the associated variance value is denoted by λ_2 . This process can be continued until as many components as variables have been calculated. The sum of variance of principal components is equal to the sum of the variance of original variables such that $\sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \sigma_i^2$ where λ_i is the variance of the *i*th principal component. This way there are *p* linear transformations (PCs) of the original p variables. These are $y_1 = \sum_{i=1}^{p} b_{1i}x_i, \ldots, y_p = \sum_{i=1}^{p} b_{pi}x_i$. These can be expressed as $Y = B'X$ where $Y = (y_1, y_2, ..., y_p)$. B' is $p \times p$ matrix. Using the eigenvector-eigenvalue pair, the *i*th principal component may be written as

$$y_i = V_i' X = v_{i1}x_1 + \cdots + v_{ip}x_p, \, i = 1, 2, \ldots p \quad (2)$$

So $Var(y_i) = V_i' C V_i = \lambda_i \quad Cov(y_i, y_k) = 0, i \neq k$, we have

$$Y = VX = (V_1, V_2, \ldots V_p)' X \quad (3)$$

$$Var(Y) = VCV' = diag(\lambda_1, \lambda_2, \ldots \lambda_p) \quad (4)$$

We can retain the maximum information by retaining the coordinate axes that have largest eigenvalues and delete those that have less information. The success of PCA is due to the following important properties

1. Principal components sequentially capture the maximum variability among the data, thus

guaranteeing minimal information loss when lesser components are discarded.
2. Principal components are uncorrelated, so one can talk about each of the Principal components without referring to the others, each one makes an independent contribution to accounting for the variance of the original variables.

2.2. Basic ICA Model

Consider a random vector $\mathbf{x} = (x_1, x_2, \cdots, x_m)^T$ and the components as the random vector as $\mathbf{s} = (s_1, s_2, \cdots, s_n)^T$. The aim is to find the components s_i as independent as possible in the sense of maximizing some function $F(s_1, s_2, \cdots, s_n)$ that measures independence. The independent component analysis of the observed data $\mathbf{x}$ consists of finding a linear transformation

$$\mathbf{s} = \mathbf{W}\mathbf{x} \qquad (5)$$

so as to get the components s_i as independent as possible. This is the most general definition of independent component analysis. There are two more definitions for ICA, noisy and noiseless ICA. The noisy ICA model is

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n} \qquad (6)$$

where s_i in $\mathbf{s}$ are assumed independent. The matrix $\mathbf{A}$ is a constant $m \times n$ mixing matrix and $\mathbf{n}$ is a m-dimensional random noise vector. In order to make (6) and hence the model more simple, we consider another simplified model in which $\mathbf{n}$ is zero and the model is known as noise free model and it is given as

$$\mathbf{x} = \mathbf{A}\mathbf{s} \qquad (7)$$

where $\mathbf{A}$ and $\mathbf{s}$ are as defined above. The linear form of ICA has been considered, though nonlinear forms of ICA also exit. The linear functions make the interpretation and the computation of the representation much simpler.

So the basic ICA model consists of observing n number of random variables $x_1, x_2, \cdots, x_n$ and these are modeled as linear combination of n random variables $s_1, s_2, \cdots, s_n$

$$x_i = a_{i1}s_1 + a_{i2}s_2 + \cdots + a_{in}s_n \; for\; all\; i = 1, \cdots, n \qquad (8)$$

where $a_{ij}, i, j = 1, \cdots n$ are some real coefficients. This model is termed as generative model due to the fact that it describes how the observed data are generated by the process of mixing the components s_i. The independent components s_i (statistically mutually independent) are latent variables, meaning that they cannot be directly observed.

For the basic ICA model the independent components are assumed to be statistically independent and the components must have non-Gaussian distribution. This is due to the fact that Gaussian distribution gives us information up to second order only whereas all higher order cumulants ate zero. It is further assumed that the unknown mixing matrix is square meaning thereby that the number of independent components is equal to the number of observed mixtures. The only reason behind it is that it simplifies the estimation to a larger extent.

2.3. Information Theoretic Concepts

Consider a discrete random variable X whose k possible values are $x_1, x_2, \ldots, x_k$. The probability of each value is $p_1, p_2, \ldots, p_k$ .The probability space can be expressed as follows

$$\mathrm{X} = (x_1, x_2, \ldots, x_k) \quad P = (p_1, p_2, \ldots, p_k) \quad (9)$$

where $P(x_i) = p_i$ is the probability of occurrence of an event $X = x_i$ with the requirement that $0 \le p_i \le 1$ and $\sum_{i=1}^{k} p_i = 1$ .Suppose that the event $X = x_i$ occurs with the probability $p_i = 1$,means that $p_i = 0$ for all except the one for which it is equal to one. It means that there is no surprise or uncertainty and therefore no "information" conveyed by the occurrence of the event $X = x_i$ .On the other hand, if the probability of occurrence is low, then there is more surprise and hence the more information. The amount of information is related to the inverse of the probability of occurrence.

In general, the amount of information gained after the occurrence of the event $X = x_i$ with probability p_i is defined as the logarithmic function

$$I(x_i) = \log(\tfrac{1}{p_i}) = -\log p_i \tag{10}$$

$I(x_i)$ is called information function. It is also known as self information

For a random vector $\mathbf{x}$ consisting of n number of random variables $x_1, x_2, \cdots, x_n$, the (differential) entropy of $\mathbf{x}$ is

$$\begin{aligned} H(\mathbf{x}) &= -\int_{-\infty}^{\infty} p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} \\ &= -E[\log p(\mathbf{x})] \end{aligned} \tag{11}$$

where $p(\mathbf{x})$ is the probability density function (pdf) of $\mathbf{x}$. The joint entropy $H(\mathbf{x},\mathbf{y})$ of two random vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$H(\mathbf{x},\mathbf{y}) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(\mathbf{x},\mathbf{y}) \log p(\mathbf{x},\mathbf{y}) d\mathbf{x}d\mathbf{y} \tag{12}$$

and the conditional entropy of $\mathbf{x}$ and $\mathbf{y}$ is

$$H(\mathbf{x}\,|\,\mathbf{y}) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(\mathbf{x},\mathbf{y}) \log p(\mathbf{x}\,|\,\mathbf{y}) d\mathbf{x}d\mathbf{y} \tag{13}$$

$H(\mathbf{x}\,|\,\mathbf{y})$ can also be written as[24]

$$H(\mathbf{x}\,|\,\mathbf{y}) = H(\mathbf{x},\mathbf{y}) - H(\mathbf{y}) \tag{14}$$

with the property that

$$0 \le H(\mathbf{x}\,|\,\mathbf{y}) \le H(\mathbf{x}) \tag{15}$$

The conditional entropy $H(\mathbf{x}\,|\,\mathbf{y})$ represents the amount of uncertainty remaining about the system input $\mathbf{x}$ after the system output $\mathbf{y}$ has been observed.

Since the entropy $H(\mathbf{x})$ represents our uncertainty about the system input before observing the system output and the conditional entropy $H(\mathbf{x}\,|\,\mathbf{y})$ represents our uncertainty about the system input after observing the system output, the difference $H(\mathbf{x}) - H(\mathbf{x}\,|\,\mathbf{y})$ must represent our uncertainty about the system input that is resolved by observing the system output. This quantity is called mutual information between $\mathbf{x}$ and $\mathbf{y}$ and is denoted as $I(\mathbf{x};\mathbf{y})$ and has the following properties

$$\begin{aligned} I(\mathbf{x};\mathbf{y}) &= H(\mathbf{x}) - H(\mathbf{x}\,|\,\mathbf{y}) \\ &= H(\mathbf{y}) - H(\mathbf{y}\,|\,\mathbf{x}) \\ &= H(\mathbf{x}) + H(\mathbf{y}) - H(\mathbf{x},\mathbf{y}) \end{aligned} \tag{16}$$

$$I(\mathbf{x};\mathbf{y}) = I(\mathbf{y};\mathbf{x}) \tag{17}$$

$$I(\mathbf{x};\mathbf{y}) \ge 0 \tag{18}$$

$I(\mathbf{x};\mathbf{y})$ may also be written as

$$I(\mathbf{x};\mathbf{y}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(\mathbf{x},\mathbf{y}) \log\left(\frac{p(\mathbf{x}\,|\,\mathbf{y})}{p(\mathbf{x})}\right) d\mathbf{x}d\mathbf{y} \tag{19}$$

Consider that $\mathbf{x}$ and $\mathbf{y}$ are independent, then

$$p(\mathbf{x},\mathbf{y}) = p(\mathbf{x})p(\mathbf{y}) \tag{20}$$

And we may write $p(\mathbf{x}\,|\,\mathbf{y}) = p(\mathbf{x})$ which means that the knowledge of outcome of $\mathbf{y}$ does not affect the distribution of $\mathbf{x}$ .If we apply this to (19), we get

$$I(\mathbf{x};\mathbf{y}) = 0 \tag{21}$$

This shows that the mutual information is zero if and only if $\mathbf{x}$ and $\mathbf{y}$ are independent.

The basic idea of ICA is to minimize the dependency among the output components. The dependency is measured by Kulback-Leibler (KL) divergence between the joint and the product of the marginal distributions of the outputs. The KL divergence between the two different probability density functions $f(\mathbf{x})$ and $g(\mathbf{x})$ of a random vector $\mathbf{x}$ is as follows

$$D_{f\|g} = \int_{-\infty}^{\infty} f(\mathbf{x}) \log\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) d\mathbf{x} \tag{22}$$

It can be considered as a kind of distance between the two probability distributions, though it is not a real distance measure because it is not symmetric. If we have a perfect match between the two distributions i.e., if $f(\mathbf{x}) = g(\mathbf{x})$, then $D_{f\|g}$ is exactly zero.

We know that

$$p(\mathbf{x},\mathbf{y}) = p(\mathbf{x}\,|\,\mathbf{y})p(\mathbf{y}) = p(\mathbf{y}\,|\,\mathbf{x})p(\mathbf{x}) \tag{23}$$

The (19) can be rewritten as

$$I(\mathbf{x};\mathbf{y}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(\mathbf{x},\mathbf{y}) \log\left(\frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}\right) d\mathbf{x}d\mathbf{y} \tag{24}$$

Comparing the two (22) and (24) we deduce that

$$I(\mathbf{x};\mathbf{y}) = D_{p(\mathbf{x},\mathbf{y})\|p(\mathbf{x})p(\mathbf{y})} \tag{25}$$

In other words, the mutual information $I(\mathbf{x};\mathbf{y})$ between $\mathbf{x}$ and $\mathbf{y}$ is equal to the KL divergence between the joint probability density function $p(\mathbf{x};\mathbf{y})$ and the product of probability density functions $p(\mathbf{x})$ and $p(\mathbf{y})$.

3. INFORMATION THEORETIC APPROACHES

3.1. PCA Using Information

For performing the dimensionality reduction on input data, we need to compute the eigenvalues and eigenvectors of the covariance matrix of the input data vector and then project the data orthogonally onto the subspace spanned by the eigenvectors belonging to the dominant eigenvalues and leaving those that possess less information. From the discussion as above, it has become very clear that the information can be compressed using Principal Component Analysis by selecting only a few large eigenvalues and ignoring the other. The large eigenvaues are selected with the understanding that they contribute more information In order to estimate the degree of information compression, the authors [45] made full use of the concept of Shannon information theory and gave the new concept called possibility information function (PIF)

For the *n*-dimensional random variable $\boldsymbol{x} = (x_1, x_2, \dots, x_n)^T$ composed of *n* features and $\mathrm{C} = \mathrm{E}[(\mathbf{x} - \overline{\mathbf{x}})(\mathbf{x} - \overline{\mathbf{x}})^{\mathrm{T}}]$ being the covariance matrix of $\boldsymbol{x}$. All eigenvalues $[\lambda_1, \lambda_2, \cdots, \lambda_n]$ of C were obtained and transformation of λ_i was carried out and written as follows [45]

$$\rho_i = 1 - \lambda_i / \sum_{i=1}^{n} \lambda_i \qquad (26)$$

From the above formula it is clear that $0 \le \rho_i \le 1$ and therefore ρ_i has the numerical properties of probability. Similar to the definition of information function, the possibility information function *(PIF*) and possibility information entropy (*PIE*) were defined as [45]

$$I(\lambda_i) = \log(\tfrac{1}{\rho_i}) = -\log \rho_i \; (i = 1, 2, \dots n) \qquad (27)$$

$$H(T) = H(\rho_1, \rho_2, \dots, \rho_n) = -\sum_{i=1}^{n} \rho_i \log \rho_i \qquad (28)$$

Based upon *PIF* two new concepts information rate *(IR*) and accumulated information rate (*AIR*) were defined that are as follows

$$IR(\lambda_i) = \frac{I(\lambda_i)}{\sum_{i=1}^{n} I(\lambda_i)}, i = 1, 2, \dots n \qquad (29)$$

$$AIR(\lambda_{1,} \lambda_2, \dots \lambda_m) = \frac{\sum_{i=1}^{m} I(\lambda_i)}{\sum_{i=1}^{n} I(\lambda_i)} \qquad (30)$$

These two concepts are in accordance with variance contribution rate (*CR*) and total variance contribution rate (*TCR*) for standard *PCA* method. *CR* and *TCR* are defined as

$$CR(\lambda_i) = \frac{\lambda_i}{\sum_{i=1}^{n} (\lambda_i)}, i = 1, 2, \dots n \qquad (31)$$

$$TCR(\lambda_{1,} \lambda_2, \dots \lambda_m) = \frac{\sum_{i=1}^{m} (\lambda_i)}{\sum_{i=1}^{n} (\lambda_i)} \qquad (32)$$

3.2. ICA Using Negentropy

Nongaussianity is the parameter to estimate Independent Component Analysis. Kurtosis also known as fourth order cummulant gives us the measure of nongaussianity. The kurtosis of random variable say y is defined as

$$kurt(y) = E\{y^4\} - 3\left(E\{y^2\}\right)^2 \qquad (33)$$

If we assume that y has been normalized so as to have its variance equal to one i.e., $E\left(y^2\right) = 1$ then

$$kurt(y) = E\{y^4\} - 3 \qquad (34)$$

From the above equation we can say that kurtosis is simply normalized version of the fourth moment. For a Gaussian random variable y the fourth order moment is equal to $3\left(E\{y^2\}\right)^2$, making kurtosis of the variable equal to zero and for most of the nongaussian random variables it is nonzero. This is the reason kurtosis is most widely used a measure of nongaussianity in Independent Component Analysis. Theoretically it is quite simple and it has computational simplicity also. It has drawback also. It is quite sensitive to outliers. Its value may depend on only a few observations in the tails of the distribution, which may be erroneous or irrelevant observations. So we can say that it is not the robust measure of nongaussianity. Another important measure is negentropy, which is robust but computationally complicated. It is based on information theoretic quantity of differential entropy or simply entropy. The (differential) entropy of a random variable y is

$$H(y) = \int p(y) \log p(y) dy \qquad (35)$$

A fundamental result of information theory is that a Gaussian variable has the largest entropy among all random variables of equal variance meaning thereby that entropy could be used as a measure of nongaussianity. The value of the entropy being largest implies that the Gaussian distribution is the most random or least structured of all distributions.

To obtain a measure of nongaussianity that is zero for a Gaussian random variable and always nonnegative, a slightly modified version of definition of differential entropy called Negentropy is used. The Negentropy is defined as follows

$$J(y) = H(y_{gauss}) - H(y) \qquad (36)$$

Negentropy is also defined as the KL divergence between probability density function $p(y)$ and the Gaussian distribution $p_{gauss}(y)$ with the same mean and covariance as $p(y)$ [24] and written as

$$J(y) = D(p(y) \,||\, p_{gauss}(y)) \qquad (37)$$

where y_{gauss} a Gaussian random variable of same covariance matrix as y. Negentropy is always has nonnegative value and it is zero if and only if y has Gaussian distribution. Because of computational difficulty, Negentropy would require an estimate (possibly nonparametric) of the probability density function. Therefore the simple approximations of Negentropy are used and the classical method of approximating Negentropy is using higher order moments.

$$J(y) \approx \tfrac{1}{12} E\left\{y^3\right\}^2 + \tfrac{1}{48} kurt(y)^2 \qquad (38)$$

The random variable is assumed to be of zero mean and unit variance. But these approximations also suffer from non-robustness. To avoid this, new approximations were developed and proposed by [20,21]. In general

$$J(y) \propto \left[E\left\{G(y)\right\} - E\left\{G(\upsilon)\right\}\right]^2 \qquad (39)$$

where G is any non-quadratic function. By choosing G carefully, we can obtain the approximations of Negentropy that are better than the one given by (38). The following choices of G have proved very useful

$$G_1(y) = \frac{1}{a_1} \log\cosh a_1 y \quad G_2(y) = -\exp\left(-y^2/2\right)$$

where $1 \le a_1 \le 2$ is some suitable constant. υ is a standardized (zero mean and unit variance) Gaussian random variable.

A method for maximizing the negentropy can be found using a fixed-point algorithm. The algorithm is known as FastICA algorithm [25] that finds a direction i.e. a unit vector $\mathbf{w}$, such that projection $\mathbf{w}^T\mathbf{z}$ maximizes nongaussianity, which is measured by the approximation of negentropy $J(\mathbf{w}^T\mathbf{z})$ where $\mathbf{z} = \mathbf{V}\mathbf{x}$ is a new vector that is white and $\mathbf{V}$ is the whitening transformation matrix and $\mathbf{V} = \mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^T$ where $\mathbf{E}$ is the orthogonal matrix of eigenvectors of $E\{\mathbf{x}\mathbf{x}^T\}$ and $\mathbf{D}$ is the diagonal matrix of its eigenvalues $\mathbf{D} = diag(d_1, d_2, \ldots, d_n)$ .The basic fixed point iteration in FastICA is given as [25]

$$\mathbf{w} \leftarrow E\{zg(\mathbf{w}^T\mathbf{z})\} - E\{g'(\mathbf{w}^T\mathbf{z})\}\mathbf{w} \qquad (40)$$

Iteration in (40) is used and is followed by normalization. Here the nonlinearity g is chosen, which is the derivative of the nonquadratic function G. Thus we can use the derivatives of the functions

$$G_1(y) = \frac{1}{a_1} \log\cosh a_1 y \quad G_2(y) = -\exp\left(-y^2/2\right)$$

that gives robust approximation of negentropy. We can choose

$$\begin{array}{ll} g_1(y) = \tanh(a_1 y) & g_1^{'}(y) = a_1(1 - \tanh^2(a_1 y)) \\ g_2(y) = y e^{-y^2/2} & g_2^{'}(y) = (1 - y^2) e^{-y^2/2} \\ g_3(y) = y^3 & g_3^{'}(y) = 3y^2 \end{array} \qquad (41)$$

The above-mentioned algorithm estimates only one independent component. To estimate more independent components either Deflationary orthogonolization (one by one estimation) or Symmetric orthogonolization (estimation in parallel) method is used.

3.3. ICA Using Minimization of Mutual Information

The KL divergence between the pdf $p(\mathbf{x})$ of random vector $\mathbf{x}$ and the product of its marginal pdfs and hence the mutual information $I(\mathbf{x})$ of the observed vector, is

$$I(\mathbf{x}) \quad = D_{p(\mathbf{x})\|\prod_{i=1}^{n} p(x_i)}$$

$$= \int_{-\infty}^{\infty} p(\mathbf{x}) \log \left(\frac{p(\mathbf{x})}{\prod_{i=1}^{n} p(x_i)} \right) d\mathbf{x} \qquad (42)$$

It may also be written as

$$\int_{-\infty}^{\infty} p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} - \sum_{i=1}^{n} \int_{-\infty}^{\infty} p(\mathbf{x}) \log p(x_i) d\mathbf{x} \qquad (43)$$

Since $d\mathbf{x} = d\mathbf{x}^{(i)} dx_i$ we may write

$$\int_{-\infty}^{\infty} p(\mathbf{x}) \log p(x_i) d\mathbf{x} = \int_{-\infty}^{\infty} \log p(x_i) \int_{-\infty}^{\infty} p(\mathbf{x}) d\mathbf{x}^{i} dx_i \quad (44)$$

where the inner integral is with respect to $(n-1)$-by-1 vector $\mathbf{x}^{(i)}$ and the outer integral is with respect to the scalar integral x_i. Let $p(x_i)$ denotes the *i*th marginal pdf of element x_i, which is defined as

$$p(x_i) = \int_{-\infty}^{\infty} p(\mathbf{x}) d\mathbf{x}^{i} \quad i = 1, 2, \ldots, n \qquad (45)$$

where $\mathbf{x}^{(i)}$ is the $(n-1)$-by-1 vector left after removing the *i*th element from vector $\mathbf{x}$. Equation (44) becomes

$$\int_{-\infty}^{\infty} p(\mathbf{x}) \log p(x_i) d\mathbf{x} = \int_{-\infty}^{\infty} \log p(x_i) p(x_i) dx_i \quad (46)$$

$$\int_{-\infty}^{\infty} \log p(x_i) p(x_i) dx_i = -H_{mar}(x_i) \qquad (47)$$

where $H_{mar}(x_i)$ is the marginal entropy based on the marginal probability density function $p(x_i)$. Using (43) and (47) we may write

$$D = -H(\mathbf{x}) + \sum_{i=1}^{n} H_{mar}(x_i) \qquad (48)$$

where

$$H(\mathbf{x}) \ = -\int_{-\infty}^{\infty} p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} \qquad (49)$$

$$H_{mar}(x_i) = -\int_{-\infty}^{\infty} p(x_i) \log p(x_i) dx_i \qquad (50)$$

Assume that the ICA model is given by (7), which is reproduced as

$$\mathbf{x} = \mathbf{As}$$

without knowing the source signals and the mixing matrix, we want to recover the original signals from the observations $\mathbf{x}$ by the following linear transformation

$$\mathbf{y} = \mathbf{Wx} \qquad (51)$$

where $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ and $\mathbf{W}$ is a demixing matrix. It is impossible to obtain original sources s_i because they are not identifiable in the statistical sense. However except for a permutation of indices, it is possible to obtain $c_i s_i$ where the constants c_i are nonzero scalar functions. The source signals are identifiable in this sense. So here the aim is to find the matrix $\mathbf{W}$ such that $(y_1, y_2, \cdots, y_n)$ coincides with a permutation of $(s_1, s_2, \cdots, s_n)$ except for the scalar functions. The solution $\mathbf{W}$ is the matrix, which finds all independent components in the output.

Here the aim is to have components of output vector $\mathbf{y}$ as statistically independent as possible. We have chosen the mutual information $I(y_i; y_j)$ between the random variables y_i and y_j constituting any two components of the output vector $\mathbf{y}$. Ideally $I(y_i; y_j)$ is zero when the components y_i and y_j are statistically independent. So, this suggests minimizing the mutual information between every pair of the random variables constituting the output vector $\mathbf{y}$. This objective is equivalent to minimizing the KL divergence between the two distributions, one the probability density function $p(\mathbf{y}, \mathbf{W})$ parameterized by $\mathbf{W}$ and second the corresponding factorial distributions defined by

$$p_{mar}(\mathbf{y}, \mathbf{W}) = \prod_{i=1}^{n} p(y_i, \mathbf{W}) \qquad (52)$$

where $p(y_i, \mathbf{W})$ is the marginal probability density function of y_i. So the problem statement is

Given an n-by-1 vector $\mathbf{x}$ representing a linear combination of n independent source signals, the transformation of the observation vector $\mathbf{x}$ by a neural system into a new vector $\mathbf{y}$ should be carried out in such a way that the KL divergence between the parameterized probability denoting function $p(\mathbf{y}, \mathbf{W})$ and the corresponding factorial distribution $p_{mar}(\mathbf{y}, \mathbf{W})$

is minimized with respect to the unknown parameter matrix $\mathbf{W}$.
So

$$D(\mathbf{W}) = -H(\mathbf{y}) + \sum_{i=1}^{n} H_{mar}(y_i) \qquad (53)$$

Since $\mathbf{y} = \mathbf{Wx}$, we may write $H(\mathbf{y}) = H(\mathbf{Wx})$ and in case of linear transformation, we have

$$H(\mathbf{y}) = H(\mathbf{x}) + \log|\det(\mathbf{W})| \qquad (54)$$

where $\det(\mathbf{W})$ is the determinant of $\mathbf{W}$. To determine marginal entropy $H_{mar}(y_i)$ we require the knowledge of marginal distribution of y_i. For a vector of high dimensionality it is usually more difficult to calculate $H_{mar}(y_i)$ than $H(\mathbf{y})$. This difficulty is overcome by deriving an approximate formula for $H_{mar}(y_i)$ in terms of higher order moments of random variable y_i. This as accomplished by properly truncating one of the two expansions, Edgeworth series or Gram-Charlier series.

3.3.1. Edgeworth Series

If the observed vector has a covariance matrix $\left\langle \mathbf{xx}^T \right\rangle = E\left\{\mathbf{xx}^T\right\}$ then the mutual information in (42) can be expressed as [14]

$$I(\mathbf{x}) = J(\mathbf{x}) - \sum_{i=1}^{n} J(x_i) + \frac{1}{2}\log\frac{\left(\prod_{i=1}^{n}\left\langle x_i^2\right\rangle\right)}{\det(E\{\mathbf{xx}^T\})} \qquad (55)$$

where $\left\langle x_i^2 \right\rangle$ are the diagonal elements of the covariance matrix. $J(\mathbf{x})$ is the multivariate negentropy as in (37)and $J(x_i)$ are the marginal negentropies.

$$J(x_i) = \int p(x_i)\log\frac{p(x_i)}{p_{gauss}(x_i)}dx_i \qquad (56)$$

There is standardization of $\mathbf{x}$ by using whitening transformation, which results in removal of second order redundancy. The observed vector $\mathbf{x}$ is transformed linearly so that a new vector $\tilde{\mathbf{x}}$ is obtained which is white i.e. its components are uncorrelated and their variances equal unity. In other words $E\{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T\} = \mathbf{I}$, $\mathbf{I}$ being identity matrix, and $\det(E\{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T\}) = 1$ thus making the third term of (55) always equal to zero and

$$\tilde{\mathbf{x}} = \mathbf{Vx} \qquad (57)$$

where $\mathbf{V}$ is the whitening transformation matrix. The mutual information of the spatially white data can be written as

$$I(\tilde{\mathbf{x}}) = J(\tilde{\mathbf{x}}) - \sum_{i=1}^{n} J(\tilde{x}_i) \qquad (58)$$

A further transformation $\mathbf{y} = \mathbf{W}\tilde{\mathbf{x}}$ using higher order correlations is required to reduce the remaining redundancy within the vector for non Gaussian sources. This transformation seeks an orthogonal matrix that accounts for the correct rotation of data. The first term in the above equation is constant because of orthogonal transform. We only need to minimize the second term, the sum of marginal negentropies. Comon [14] minimized the degree of dependence among outputs using contrast functions in terms of higher order moments using the Edgeworth expansion. The key advantage of using the Edgeworth expansion over Gram-Charlier expansion lies in the ordering of terms according to their decreasing significance as a function of $m^{-1/2}$ where y is modeled as being made up of sum of m independent random variables. The truncated Edgeworth expansion upto order 4, of $p(y_i)$ in terms of n^{th} order cumulant and Hermite polynomial, denoted as k_n and h_n respectively is (zero mean and unit variance)

$$\begin{aligned}\frac{p(y_i)}{p_{gauss}(y_i)} &= 1 + \frac{1}{3!}k_3h_3(y_i) + \frac{1}{4!}k_4h_4(y_i) \\ &+ \frac{10}{6!}k_4^2h_6(y_i) + \frac{1}{5!}k_5h_5(y_i) + \frac{35}{7!}k_3k_4h_7(y_i) \\ &+ \frac{280}{9!}k_3^3h_9(y_i) + \frac{1}{6!}k_6h_6(y_i) + \frac{56}{8!}k_3k_5h_8(y_i) \\ &+ \frac{35}{8!}k_4^2h_8(y_i) + \frac{2100}{10!}k_3^2k_4h_{10}(y_i) \\ &+ \frac{15400}{12!}k_3^4h_{12}(y_i)\end{aligned} \qquad (59)$$

The cumulants k_n are the coefficients and they can be expressed in terms of moments. The terms $h_n(y_i)$ are the orthogonal Hermite polynomials defined as

$$(-1)^k \frac{\partial^k p_{gauss}(y_i)}{\partial y^k} = h_k(y_i)\,p_{gauss}(y_i) \qquad (60)$$

Using (56) and (59) we can write [14]

$$J(y_i) \cong \frac{1}{12}k_3^2(i) + \frac{1}{48}k_4^2(i) + \frac{7}{48}k_3^4(i) - \frac{1}{8}k_3^2(i)k_4(i) \qquad (61)$$

Here an assumption has been made that the pdf of the signals under consideration are approximately symmetric, then the third order cumulants will have negligible contribution in the above equation. The mutual information in (55) of the transformed data $\mathbf{y}$ is now approximated by

$$I(\mathbf{y}) \cong J(\mathbf{y}) - \frac{1}{48}\sum_{i=1}^{n} k_4^2(i) \qquad (62)$$

$J(\mathbf{y})$ is invariant under an orthogonal transformation

$$\begin{aligned} J(\mathbf{y}) &= \int p(\mathbf{y}) \log \frac{p(\mathbf{y})}{p_{gauss}(\mathbf{y})} d\mathbf{y} \\ &= H(\mathbf{y}) - \frac{1}{2}\log((2\pi e)^N \det(E\{\mathbf{y}\mathbf{y}^T\})) \\ &= H(\tilde{\mathbf{x}}) + \log|\det(\mathbf{W})| \\ &\quad - \frac{1}{2}\log((2\pi e)^N \det(\mathbf{W}E\{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T\}\mathbf{W}^T)) \\ &= H(\tilde{\mathbf{x}}) - \frac{1}{2}\log((2\pi e)^N \det(E\{\mathbf{x}\mathbf{x}^T\})) \\ &= H(\tilde{\mathbf{x}}) - H_G(\tilde{\mathbf{x}}) = J(\tilde{\mathbf{x}}) \end{aligned} \qquad (63)$$

where $H_G(\tilde{\mathbf{x}})$ is the entropy of a normal density and

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

$$\det(\mathbf{W}^T) = \det(\mathbf{W})$$

So the approximation can be written as

$$I(\mathbf{y}) \cong J(\tilde{\mathbf{x}}) - \frac{1}{48}\sum_{i=1}^{n} k_4^2(i) \qquad (64)$$

Thus maximizing the contrast function is approximately equivalent to maximizing the marginal negentropies and maximizing the marginal negentropies with respect to $\mathbf{W}$ minimizes the mutual information.

$$\frac{\partial}{\partial \mathbf{W}} I(\mathbf{y}) \cong \frac{\partial}{\partial \mathbf{W}}(-\frac{1}{48}\sum_{i=1}^{n} k_4^2(i)) \qquad (65)$$

Therefore the following contrast function was proposed [14]

$$\varphi = \sum_{i=1}^{n} k_4^2(i) \qquad (66)$$

3.3.2. Gram-Charlier Series

The Gram-Charlier expansion of the parameterized marginal pdf $p(y_i, \mathbf{W})$ is

$$= \alpha(y_i)\left[1 + \sum_{k=3}^{\infty} c_k H_k(y_i)\right] \qquad (67)$$

where $\alpha(y_i)$, the multiplying factor, is the pdf of a normalized Gaussian random variable with zero mean and unit variance i.e.,

$$\alpha(y_i) = \tfrac{1}{\sqrt{2\pi}} e^{-y_i^2/2}$$

and $H_k(y_i)$ are Hermite polynomials. The coefficients of expansion $c_k, k = 3,4,\ldots$ are defined in terms of the cumulants of the random variable y_i. The natural order of the terms is not best for the Gram-Charlier series. Rather, the terms are listed in groups as given below

$$k = (0),(3),(4,6),(5,7,9),\ldots$$

So, truncating the series we may write,

$$p(y_i) = \alpha(y_i)\left(1 + \frac{k_{i,3}}{3!}H_3(y_i) + \frac{k_{i,4}}{4!}\right) \qquad (68)$$

$$c_1 = 0, c_2 = 0, c_3 = \frac{k_3}{3!}, c_4 = \frac{k_4}{4!}$$

where $k_{i,k}$ is the $k\,th$ order cumulant of y_i. Let $m_{i,k}$ denote the $k\,th$ order moment of y_i defined by $m_{i,k} = E\left[{y_i}^k\right] = E\left[(\sum_{k=1}^{n} w_{i,k}x_i)^k\right]$ where x_i is the $i\,th$ element of observation vector $\mathbf{x}$ and $w_{i,k}$ is the $i\,k\,th$ element of weight matrix $\mathbf{W}$. We further assume that y_i has zero mean value and $\sigma_i^2 = m_{i,2}$ with

$$k_{i,3} = m_{i,3} \quad and \quad k_{i,4} = m_{i,4} - 3{m_{i,2}}^2$$

Taking log of (68), we get

$$\log p(y_i) = \log\alpha(y_i) + \log\left(1 + \frac{k_{i,3}}{3\cdot 2\cdot 1}H_3(y_i) + \frac{k_{i,4}}{4\cdot 3\cdot 2\cdot 1}\right) \qquad (69)$$

We use the expansion of a logarithm

$$\log(1+y) \cong y - \frac{y^2}{2} \qquad (70)$$

where all the terms of order three and higher are ignored. The Chebyshev –Hermite polynomials are defined by the identity

$$(-1)^k \frac{d^k\alpha(y)}{dy^k} = H_k(y)\alpha(y)$$
$$\int \alpha(y)(H_3(y))^2 H_4(y)\alpha(y)dy = (3!)^3$$
$$\int \alpha(y)(H_4(y))^3 dy = (12)^3$$

and

$$\int y^{2k+1}\alpha(y)dy = 0$$
$$\int y^{2k}\alpha(y)dy = 1\cdot 3\cdot\ldots\cdot(2k-1)$$
$$-\int \alpha(y)\log\alpha(y)dy = \tfrac{1}{2}\log(2\pi e)$$

Using (47), (69), (70) and the above identities, we may write [22]

$$\begin{aligned} H_{mar}(y_i) = \frac{1}{2}\log(2\pi e) - \frac{k_{i,3}^2}{12} - \frac{k_{i,4}^2}{48} \\ + \frac{5}{8}k_{i,3}^2 k_{i,4} + \frac{1}{16}k_{i,4}^3 \end{aligned} \qquad (71)$$

Equation (53) may now be written as

$$\begin{aligned} D(\mathbf{W}) \approx -H(\mathbf{x}) - \log|\det(\mathbf{W})| + \frac{n}{2}\log(2\pi e) \\ -\sum_{i=1}^{n} \frac{k_{i,3}^2}{12} + \frac{k_{i,4}^2}{48} - \frac{5}{8}k_{i,3}^2 k_{i,4} - \frac{1}{16}k_{i,4}^3 \end{aligned} \qquad (72)$$

The derivation for $D(\mathbf{W})$ is based on Gram-Charlier expansion, assuming that random variable y_i has zero mean and unit variance. In order to develop a learning algorithm for computing $\mathbf{W}$, we need to differentiate $D(\mathbf{W})$ with respect to $\mathbf{W}$. Let A_{ik} denote the $i\,k\,th$ cofactor of matrix $\mathbf{W}$. Using Laplacian expansion of $\det(\mathbf{W})$ by the $i\,th$ row, we may write

$$\det(\mathbf{W}) = \sum_{k=1}^{n} w_{ik}A_{ik} \quad i = 1, 2, \ldots, n \qquad (73)$$

w_{ik} is the $i\,k\,th$ element of matrix $\mathbf{W}$.

$$\begin{aligned} \frac{\partial}{\partial w_{ik}}\log(\det(\mathbf{W})) &= \frac{1}{\det(\mathbf{W})}\frac{\partial}{\partial w_{ik}}\det(\mathbf{W}) \\ &= \frac{A_{ik}}{\det(\mathbf{W})} \\ &= (\mathbf{W}^{-T})_{ik} \end{aligned} \qquad (74)$$

where $\mathbf{W}^{-T}$ is the inverse of transposed matrix $\mathbf{W}^T$.The partial derivatives of other terms that depend on $\mathbf{W}$ with respect to w_{ik} are

$$\frac{\partial k_{i,3}}{\partial w_{ik}} = 3E[y_i^2 x_k] \qquad \frac{\partial k_{i,4}}{\partial w_{ik}} = 4E[y_i^3 x_k] \quad (75)$$

In deriving an adaptive algorithm, we usually replace expectations with their instantaneous values. So the above values can be written as

$$\frac{\partial k_{i,3}}{\partial w_{ik}} \cong 3y_i^2 x_k \qquad \frac{\partial k_{i,4}}{\partial w_{ik}} \cong 4y_i^3 x_k \qquad (76)$$

Substituting (74) and (76) in the (72) we get

$$\frac{\partial}{\partial w_{ik}}D(\mathbf{W}) = -(\mathbf{W}^{-T})_{ik} + \psi(y_i)x_k \qquad (77)$$

where $\psi(y_i)$ is the activation function of the learning algorithm, defined by [22]

$$\begin{aligned} \psi(y_i) = \frac{3}{4}y^{11} + \frac{25}{4}y^9 - \frac{14}{3}y^7 \\ -\frac{47}{4}y^5 + \frac{29}{4}y^3 \end{aligned} \qquad (78)$$

The objective of the learning algorithm is to minimize KL divergence between the probability density function of $\mathbf{y}$ and the factorial distribution of y_i $i = 1,2,\ldots,n$. This minimization may be implemented using the method of gradient descent whereby the adjustment applied to the weight w_{ik} is defined by

$$\Delta w_{ik} = -\eta\frac{\partial}{\partial w_{ik}}D(\mathbf{W}) \qquad (79)$$

$$\Delta w_{ik} = \eta\left((\mathbf{W}^{-T})_{ik} - \psi(y_i)x_k\right) \qquad (80)$$

where η is learning rate parameter. The formula of (80) can be extended to the entire weight matrix $\mathbf{W}$ and the adjustment $\Delta\mathbf{W}$ applied to $\mathbf{W}$ may be expressed as follows

$$\Delta\mathbf{W} = \eta\left(\mathbf{W}^{-T} - \boldsymbol{\psi}(\mathbf{y})\mathbf{x}^T\right) \qquad (81)$$

where $\mathbf{x}^T$ is the transpose of n-by-1 observation vector $\mathbf{x}$ and $\boldsymbol{\psi}(\mathbf{y}) = [\psi(y_1), \psi(y_2), \cdots, \psi(y_n)]$

Equation (81) can be rewritten as

$$\Delta\mathbf{W} = \eta\left(\mathbf{I} - \boldsymbol{\psi}(\mathbf{y})\mathbf{x}^T\mathbf{W}^T\right)\mathbf{W}^{-T} \qquad (82)$$

where $\mathbf{I}$ is the identity matrix. Since $\mathbf{y}^T = \mathbf{x}^T\mathbf{W}^T$, we may write the above equation as

$$\Delta\mathbf{W} = \eta\left(\mathbf{I} - \boldsymbol{\psi}(\mathbf{y})\mathbf{y}^T\right)\mathbf{W}^{-T} \qquad (83)$$

It is better to replace the above algorithm by the natural gradient of the objective function $D(\mathbf{W})$. The natural gradient of the objective function $D(\mathbf{W})$, defined in terms of usual gradient ∇D is as

$$\nabla_{nat} D(\mathbf{W}) = (\nabla D(\mathbf{W}))\mathbf{W}^T\mathbf{W} \qquad (84)$$

The gradient $\nabla D(\mathbf{W})$ is the optimum direction for descent only when the parameter space {**W**} is Euclidean with an orthonormal coordinate system. In neural networks, however, the parameter space has a coordinate system that is nonorthonormal. In such situations, the natural gradient $\nabla_{nat} D(\mathbf{W})$ will provide the steepest descent. For the natural gradient space the parameter space must be *Riemannian* and the matrix $\mathbf{W}$ must be nonsingular (i.e., invertible) Using natural gradient we get

$$\begin{aligned}\Delta\mathbf{W} &= \eta\left(\mathbf{I} - \boldsymbol{\psi}(\mathbf{y})\mathbf{y}^T\right)\mathbf{W}\mathbf{W}^T\mathbf{W}^{-T} \\ &= \eta\left(\mathbf{I} - \boldsymbol{\psi}(\mathbf{y})\mathbf{y}^T\right)\mathbf{W}\end{aligned} \qquad (85)$$

Hence we may write the weight update as

$$\begin{aligned}\mathbf{W}(n+1) &= \mathbf{W}(n) \\ &\quad + \eta(n)[\mathbf{I} - \boldsymbol{\psi}(\mathbf{y}(n))\mathbf{y}^T(n)]\mathbf{W}(n)\end{aligned} \qquad (86)$$

3.4. Maximum Entropy Method

Consider the following figure. The (differential) entropy of the random vector $\mathbf{z}$ at the output of the nonlinearirty $\mathbf{G}$ is

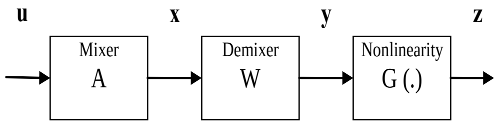

Figure 1. Block diagram of maximum entropy method for Independent Component Analysis.

$$H(\mathbf{z}) = -E\left[\log p(\mathbf{z})\right] \qquad (87)$$

$$\mathbf{z} = \mathbf{G}(\mathbf{y}) = \mathbf{G}(\mathbf{Wx}) = \mathbf{G}(\mathbf{WAu})$$

and the original source vector may be expressed as

$$\mathbf{u} = \mathbf{A}^{-1}\mathbf{W}^{-1}\mathbf{G}^{-1}(\mathbf{z})$$

where $\mathbf{G}(\cdot)$ is invertible.

The probability density function of the output vector $\mathbf{z}$ in terms of source vector $\mathbf{u}$ is defined as [23]

$$p(\mathbf{z}) = \frac{f(\mathbf{u})}{|\det(\mathbf{J}(\mathbf{u}))|} \qquad (88)$$

where $\det(\mathbf{J}(\mathbf{u}))$ is the determinant of the Jacobian matrix $\mathbf{J}(\mathbf{u})$ and

$$\mathbf{J} = \begin{bmatrix} \frac{\partial z_1}{\partial u_1} & \cdots & \frac{\partial z_1}{\partial u_n} \\ \vdots & & \vdots \\ \frac{\partial z_n}{\partial u_1} & \cdots & \frac{\partial z_n}{\partial u_n} \end{bmatrix} \qquad (89)$$

The ij-th element of the matrix is defined as

$$J_{ij} = \frac{\partial z_i}{\partial u_j} \qquad (90)$$

Putting (88) into(87), we get

$$H(\mathbf{z}) = -E\left[\log\left(\frac{f(\mathbf{u})}{|\det(\mathbf{J}(\mathbf{u}))}\right)\right] \qquad (91)$$

Using chain rule of calculus, (90) can be rewritten as

$$\begin{aligned}J_{ij} &= \sum_{k=1}^{n} \frac{\partial z_i}{\partial y_i}\frac{\partial y_i}{\partial x_k}\frac{\partial x_k}{\partial u_j} \\ &= \sum_{k=1}^{n} \frac{\partial z_i}{\partial y_i} w_{ik}\alpha_{kj}\end{aligned} \qquad (92)$$

The Jacobian matrix can therefore be written as

$$\mathbf{J} = \mathbf{DWA} \qquad (93)$$

where **D** is the diagonal matrix

$$D = diag\left(\frac{\partial z_1}{\partial y_1}, \frac{\partial z_2}{\partial y_2}, \cdots, \frac{\partial z_n}{\partial y_n}\right)$$

Hence

$$|\det(J)| = |\det(\mathbf{WA})| \prod_{i=1}^{n} \frac{\partial z_i}{\partial y_i} \qquad (94)$$

The maximization of entropy $H(\mathbf{z})$ requires the maximization of expectation of the denominator term in (91) that is $\log|\det(\mathbf{J}(\mathbf{u}))|$ with respect to the weight matrix $\mathbf{W}$. So we may consider the objective function as

$$\Phi = \log|\det(\mathbf{J})| \qquad (95)$$

Putting (94) into (95) yields

$$\Phi = \log|\det(\mathbf{A})| + \log|\det(\mathbf{W})| + \sum_{i=1}^{n} \log\left(\frac{\partial z_i}{\partial y_i}\right) \qquad (96)$$

Differentiating Φ with respect to the weight matrix $\mathbf{W}$ gives

$$\frac{\partial\Phi}{\partial\mathbf{W}} = \mathbf{W}^{-\mathbf{T}} + \sum_{i=1}^{n} \frac{\partial}{\partial\mathbf{W}} \log\left(\frac{\partial z_i}{\partial y_i}\right) \qquad (97)$$

The nonlinearity used was the logistic function given as

$$z_i = g\left(y_i\right)$$
$$z_i = \frac{1}{1+e^{-y_i}} \qquad i = 1, 2, \ldots n \tag{98}$$

Substituting (98) into (97), we get

$$\frac{\partial \Phi}{\partial \mathbf{W}} = \mathbf{W}^{-\mathbf{T}} + \left(\mathbf{1} - 2\mathbf{z}\right)\mathbf{x}^{\mathbf{T}} \tag{99}$$

The objective of learning algorithm is to maximize the entropy $H(\mathbf{z})$. Using the method of steepest ascent, the change applied to the weight matrix $\mathbf{W}$ is [6]

$$\Delta\mathbf{W} = \eta \frac{\partial \Phi}{\partial \mathbf{W}} = \eta\left(\mathbf{W}^{-\mathbf{T}} + \left(\mathbf{1} - 2\mathbf{z}\right)\mathbf{x}^{\mathbf{T}}\right) \tag{100}$$

where η is the learning rate parameter. Using natural gradient we get

$$\begin{aligned}\Delta\mathbf{W} &= \eta\left(\mathbf{W}^{-\mathbf{T}} + \left(\mathbf{1} - 2\mathbf{z}\right)\mathbf{x}^{\mathbf{T}}\right)\mathbf{W}^T\mathbf{W} \\ &= \eta\left(\mathbf{I} + \left(1 - 2\mathbf{z}\right)\left(\mathbf{W}\mathbf{x}\right)^T\right)\mathbf{W} \\ &= \eta\left(\mathbf{I} + \left(1 - 2\mathbf{z}\right)\mathbf{y}^T\right)\mathbf{W}\end{aligned} \tag{101}$$

Hence the weight update rule is

$$\Delta\mathbf{W} = \eta\left(\mathbf{I} + \left(1 - 2\mathbf{z}\right)\mathbf{y}^{\mathbf{T}}\right)\mathbf{W} \tag{102}$$

3.5. Maximum Likelihood Estimation

Maximum likelihood is a well-established procedure for statistical estimation with some nice properties. In this we first formulate a log-likelihood function and then optimize it with respect to the parameter vector of the probabilistic model under consideration. The likelihood function is the probability density function of a data set in a given model, but viewed as a function of the unknown parameters of the model.

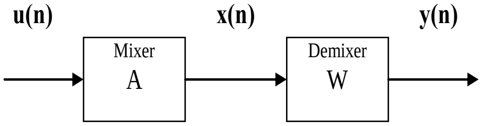

Figure 2. Block diagram of maximum likelihood estimation method for independent component analysis.

Let $f(\cdot)$ denote the probability density function of the random source vector $\mathbf{u}$. Then the pdf of the observation vector $\mathbf{x} = \mathbf{A}\mathbf{u}$ at the output of the mixer is defined by [23]

$$q(\mathbf{x}, \mathbf{A}) = |\det(\mathbf{A})|^{-1} f(\mathbf{A}^{-1}\mathbf{x}) \tag{103}$$

where $\det(\mathbf{A})$ is the determinant of the mixing matrix $\mathbf{A}$.Let us assume that we have N samples of $\mathbf{x}$ denoted by $\mathbf{x}(1), \mathbf{x}(2), \cdots, \mathbf{x}(N)$, and $\Gamma = \{\mathbf{x}(k)\}_{k=1}^{N}$. We may write

$$q(\Gamma, \mathbf{A}) = \prod_{k=1}^{N} q(\mathbf{x}(k), \mathbf{A}) \tag{104}$$

The log-likelihood function is written as

$$\log q(\Gamma, \mathbf{A}) = \sum_{k=1}^{N} \log(q(\mathbf{x}(k), \mathbf{A}) \tag{105}$$

It is convenient to work with normalized version of the log-likelihood function

$$\begin{aligned}\frac{1}{N}\log q(\Gamma, \mathbf{A}) &= \frac{1}{N}\sum_{k=1}^{N}\log(q(\mathbf{x}(k), \mathbf{A}) \\ &= \frac{1}{N}\sum_{k=1}^{N}\log(|\det(\mathbf{A})|^{-1} f(\mathbf{A}^{-1}\mathbf{x}(k))) \\ &= \frac{1}{N}\sum_{k=1}^{N}\log(|\det(\mathbf{A}^{-1})|) + \frac{1}{N}\sum_{k=1}^{N}\log(f(\mathbf{A}^{-1}\mathbf{x}(k))) \\ &= -\log|\det(\mathbf{A})| + \frac{1}{N}\sum_{k=1}^{N}\log(f(\mathbf{A}^{-1}\mathbf{x}(k)))\end{aligned}$$

Let $\mathbf{y} = \mathbf{A}^{-1}\mathbf{x}$ be a realization of the random vector $\mathbf{y}$ at the demixer output, thus we may write

$$\frac{1}{N}\log q(\Gamma, \mathbf{A}) = -\log|\det(\mathbf{A})| + \frac{1}{N}\sum_{k=1}^{N}\log(f(\mathbf{y}(k))) \tag{106}$$

Let $\mathbf{A}^{-1} = \mathbf{W}$ and let $p(\mathbf{y}, \mathbf{W})$ denote the pdf of $\mathbf{y}$ parameterized by $\mathbf{W}$. Since $\frac{1}{N}\sum_{k=1}^{N}\log(f(\mathbf{y}(k)))$ is the sample average of $\log f(\mathbf{y}(k))$ and as $N \to \infty$, we may write

$$\begin{aligned}L(\mathbf{W}) &= \lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}\log(f(\mathbf{y}(k))) + \log|\det(\mathbf{W})| \\ &= E[\log(f(\mathbf{y}(k)))] + \log|\det(\mathbf{W})| \\ &= \int_{-\infty}^{\infty} p(\mathbf{y}, \mathbf{W})\log f(\mathbf{y})d\mathbf{y} + \log|\det(\mathbf{W})|\end{aligned} \tag{107}$$

The quantity $L(\mathbf{W})$ is the desired log-likelihood function.

We know that $f(\mathbf{y}) = \left(\frac{f(\mathbf{y})}{p(\mathbf{y}, \mathbf{W})}\right)p(\mathbf{y}, \mathbf{W})$, we may express the above equation as

$$L(\mathbf{W}) = \int_{-\infty}^{\infty} p(\mathbf{y},\mathbf{W}) \log\left(\frac{f(\mathbf{y})}{p(\mathbf{y},\mathbf{W})}\right) d\mathbf{y}$$
$$+\int_{-\infty}^{\infty} p(\mathbf{y},\mathbf{W}) \log p(\mathbf{y},\mathbf{W}) d\mathbf{y} + \log|\det(\mathbf{W})|$$

$$L(\mathbf{W}) = -D_{p\|f} - H(\mathbf{y},\mathbf{W}) + \log|\det(\mathbf{W})| \quad (108)$$

where $H(\mathbf{y},\mathbf{W})$ is the (differential) entropy of random vector $\mathbf{y}$ parameterized by $\mathbf{W}$ and $D_{p\|f}$ is the KL divergence between $p(\mathbf{y},\mathbf{W})$ and $f(\mathbf{y})$. Since

$$H(\mathbf{y}) = H(\mathbf{Wx}) = H(\mathbf{x}) + \log|\det(\mathbf{W})|$$

We may write [19]

$$L(\mathbf{W}) = -D_{p\|f} - H(\mathbf{x}) \quad (109)$$

where $H(\mathbf{x})$ is the (differential) entropy of the vector $\mathbf{x}$ at the demixer input. From the above equation it is clear that the KL divergence $D_{p\|f}$ is the only quantity that depends upon $\mathbf{W}$, the weight vector. From (109), we can, therefore, conclude that maximizing the log-likelihood function $L(\mathbf{W})$ is equivalent to minimizing the KL divergence $D_{p\|f}$, which is, matching the probability distribution of the demixer output $\mathbf{y}$ to that of the original source vector $\mathbf{u}$.

Let $p_{mar}(x_i)$ denote the marginal probability density function of each x_i. Then using Pythagorean decomposition we may write

$$D_{p\|f} = D_{p\|p_{mar}} + D_{p_{mar}\|f} \quad (110)$$

Since $D_{p\|p_{mar}}$ does not depend on f, the pdf of input source vector $\mathbf{u}$, the above equation shows that $D_{p\|f}$ is minimized in f by minimizing its second term i.e. $D_{p_{mar}\|f}$. This is simply achieved by taking $f = p_{mar}$ for which $D_{p_{mar}\|f} = 0$, so that $\min_f D_{p\|f} = D_{p\|p_{mar}}$. So the objective now is to minimize $D_{p\|p_{mar}}$, which is the KL divergence between a distribution and the closest distribution with independent entries and is called as mutual information between the entries of $\mathbf{y}$. It satisfies $D_{p\|p_{mar}} \geq 0$ and is equal if and only if $\mathbf{y}$ is distributed as p_{mar}. By the definition of p_{mar}, this happens when the entries of $\mathbf{y}$ are independent. In other words $D_{p\|p_{mar}}$ measures the independence between the entries of $\mathbf{y}$. Thus, the mutual information appears as the quantitative measure of independence associated to the maximum likelihood principle.

The first KL divergence $D_{p\|p_{mar}}$ in (110) is a measure of structural mismatch that characterizes the method of independent component analysis. The second KL divergence $D_{p_{mar}\|f}$ is a measure of marginal mismatch between the marginal distributions of the demixer output $\mathbf{y}$ and the distribution of original source vector $\mathbf{u}$. The global distribution-matching criterion for maximum likelihood may be expressed as [19]

$$\begin{pmatrix} Total \\ Mismatch \end{pmatrix} = \begin{pmatrix} Structural \\ Mismatch \end{pmatrix} + \begin{pmatrix} M\arg inal \\ Mismatch \end{pmatrix}$$

Structural Mismatch refers to the structure of a distribution pertaining to a set of independent variables whereas Marginal Mismatch refers to the mismatch between the individual marginal distributions. Therefore, maximizing the likelihood with fixed assumptions about the distributions of the sources amounts to minimize a sum of two terms: the first one is the true objective i.e. mutual information as a measure of independence while the second term measures how far the (marginal) distributions of the outputs $y_1, y_2, \cdots, y_n$ are from the assumed distributions. Under the ideal conditions $\mathbf{W} = \mathbf{A}^{-1}$, both the structural mismatch and marginal mismatch vanish. At that point, maximum likelihood and independent component analysis yield exactly the same solution.

4. APPLICATIONS

ICA has many potential applications and a few of these are mentioned here

4.1. ICA in Biomedical Signal Analysis

The information theoretic methods were used in multi feature analysis of human chromosome images. The architectures based on information theory were proposed for the prediction of metastases in early breast cancer patients. These algorithms were also applied to real world problems such as analyzing electroencephalographic (EEG) data, functional magnetic resonance imaging (fMRI) data etc [26-29]

4.2. ICA for Feature Extraction

In multispectral/hyperspectral imagery the independent components can be associated with features present in the image. Algorithms based on information theory can separate these features [30, 31].

4.3. ICA for Signal and Speech Processing

Neural networks are being used in signal and speech processing for design and implementation of filters for noise reduction and separation of signals. Independent Component Analysis and Blind Source Separation techniques based on information theoretic approaches are being successfully used for this purpose [32-34].

4.4. ICA in Communications

Signal mixing occurs in radio channels. Such type of problem exists in mobile communication applications such as Code Division Multiple Access (CDMA) systems. Blind source separation techniques have been used to unmix radio signals in fading channels [35, 36].

4.5. ICA for Image Processing

The ICA algorithms based on information theoretic approaches are well suited for image processing where the objective is the discovery of properties of a noisy sensory input exhibiting coherence across both space and time and in applications such as dual image processing where the objective is to maximize the spatial differentiation between the corresponding regions of two separate images (views) of an environment of interest. These techniques also find applications in satellite image analysis, radar images etc. ICA based filters have also been proposed for removing noise from images corrupted with additive Gaussian noise [37-42].

4.6. ICA in Financial Market Data Analysis and Data Mining

In such types of applications the input consists of a set of different stock market data and the requirement is to extract the underlying set of dominant independent components. Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with a great potential for helping companies focus on the most important information in their data warehouse. ICA has been explored for financial data modeling and ICA based projection pursuit networks have been suggested for data clustering and data mining [43,44].

4.7. ICA for Face Recognition

Face recognition has always been a fascinating research area because of its various potential applications. ICA has successfully been used to extract local features for face recognition systems [1-4].

5. CONCLUSIONS

Principal Component Analysis is a classical second order unsupervised statistical method. It is widely used in signal processing, statistics and neural computing. Independent Component Analysis, an extension of PCA, is a general-purpose statistical technique that deals with higher order statistics of observed random data and transforms the data linearly into components that are as independent as possible from each other. In this paper a brief overview of information theory based prominent techniques such as ICA based on maximum entropy method, minimization of mutual information, maximization of negentropy, maximum likelihood estimation has been covered. In addition a number of potential applications such as biomedical signal processing, audio signal processing, image processing, pattern recognition and telecommunications etc. have also been described.

References

[1] M.S.Bartlett and T.J.Sejnowski, “Independent Components of Face Images: A Representation for Face Recognition,” Proc. of the 4th Annual Joint Symposium on Neural Computation, Pasadena, 1997

[2] M.S.Bartlett, H.M.Lades, T.J.Sejnowski, “Independent Component Representations for Face Recognition,” Proc. of SPIE Symposium on Electronic Imaging: Science and Technology; Conference on Human Vision and Electronic Imaging III, California, 1998

[3] M.S.Bartlett, “Face Image Analysis by Unsupervised Learning and Redundancy

Reduction," Ph.D. Dissertation, University of California, San Diego, 1998

[4] M.S.Bartlett, J.R.Movellan and T.J.Sejnowski, "Face Recognition by Independent Component Analysis," IEEE Trans. on Neural Networks, Vol.13, No.6, pp.1450-1464, 2002

[5] Simon Haykin, "Neural Networks – A Comprehensive Foundation," 2nd Edition, Pearson Education, 1999

[6] A.J.Bell and T.J.Sejnowski, "An Information-maximization approach to blind separation and blind deconvolution," Neural computation, Vol.7, No.6, pp. 1129-1159, 1995

[7] A.Hyvärinen, "Survey on Independent Component Analysis," Neural Computing Surveys, Vol.2, pp.94-128, 1999

[8] A.Hyvärinen and E.Oja, "Independent Component Analysis: Algorithms and Applications," Neural Networks, Vol.3, No. 4-5, pp.411- 430, 200

[9] C.E.Shannon, "A Mathematical Theory of Communication," Bell System Technical Journal, Vol. 27, pp 379-423, 623-656, 1948.

[10] J.Herault and J.Jutten, "Space or time adaptive signal processing by neural network models," AIP Conference proceedings 151, New York, 1986.

[11] C.Jutten and J. Herault, "Blind separation of sources, part 1: An adaptive algorithm based on neuromimetic architecture," Signal Processing, Vol. 24,pp 1-10,1991.

[12] J.Karhunen and J. Joutsensalo, "Representation and separation of signals using nonlinear PCA type learning,"Neural networks, Vol. 7, pp 113-127, 1994.

[13] A. Cichocki, R. Unbehauen and E. Rummert, "Robust learning algorithm for blind separation of signals," Electronics Letters, Vol. 30, No. 17, pp 1386-1387,1994

[14] P. Comon, "Independent Component Analysis-a new concept?," Signal processing, Vol. 36, No.3, pp 287-314, 1994.

[15] R. Linsker, "Local Synaptic learning rules suffice to maximize mutual information in a linear network," Neural Computation, Vol. 4, pp 691-702, 1992

[16] R. Linsker, " A local learning rule that enables information maximization for arbitrary input distributions," Neural Computation, Vol. 9, pp 1661-1665, 1997.

[17] J-F. Cardoso and B. Laheld, "Equivariant adaptive source separation," IEEE Trans. On Signal Processing," Vol.45 No.2, pp 434-444, 1996.

[18] J-F. Cardoso, " Infomax and maximum likelihood for source separation," IEEE Letters on Signal Processing, Vol. 4, pp 112-114, 1997

[19] J-F. Cardoso, "Blind signal separation: statistical principles," Proceedings of IEEE, Vol. 9, No.10, pp 2009-2025, 1998.

[20] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," IEEE Trans. On Neural Networks, Vol.10, No.3, pp 626-634,1999.

[21] A. Hyvärinen and E. Oja, "A fast fixed point algorithm for independent component analysis," Neural Computation, Vol. 9, No.7, pp 1483-1492, 1997.

[22] S.Amari, A.Cichocki and H.Yang, "A new learning algorithm for blind signal separation," In Advances in Neural information Processing Systems 8, pp 757-763,1996

[23] A.Papoulis, "Probability, random variablesand stochastic processes," Second Edition, McGraw-Hill, 1984.

[24] T.Cover and J.Thomas, "Elements of information theory,"John Wiley and Sons, New York, 1991.

[25] A.Hyvarinen, "Fast and Robust fixed-point algorithms for independent component analysis," IEEE Trans. on Neural Networks, Vol.10, No.3, pp 626-634,1999.

[26] P.L.Choong, C.J.S deSilva, H.J.S.Dawkins and G.F. Sterrett, "Entropy maximization networks: an application to breast cancer prognosis," IEEE Trans. on Neural Networks, No.3, Vol 7, pp 568-577, 1996

[27] Yasuo Matsuyama and Shuichiro Imahara, " Independent component analysis by conves divergence minimization: application to brain FMRI analysis," Proc. of IJCNN'01, Vol.1, pp 412-417, 2001

[28] S. Makeig, T-P. Jung, A. Bell, D. Ghahremani and T. Sejnowski, "Independent Component analysis of electroencephalographic data," Procs. of National Academy of Sciences, 94, pp 10979-10984, 1997

[29] M. McKeown, T-P. Jung, S. Makeig, G. Brown, S. Kindermann, T-W. Lee, T. Sejnowski, "Transiently Time-locked fMRI Activations revealed by independent component analysis," Procs. of National Academy of Sciences, 95, pp 803-810, 1997

[30] S.A. Robila and P.K. Varshney, "A fast source sepatration algorithm for hyperspectral image processing," IEEE International Geoscience and Remote

Sensing symposium, Vol.6, pp 3516-3518, 2002

[31] P.O. Hoyer and Aapo Hyvarinen, “Feature extraction from colour and stereo images using ICA,”Proc. of IJCNN’02, Vol.3, pp 369-374, 2000

[32] R. Lambert and A. Bell, “Blind separation of multiple speakers in a multipath environment,” Proc. of ICASSP, pp 423-426, 1997

[33] T-W. Lee, A.J. Bell and R. Orglmeister, “Blind source separation of real world signals,” IEEE Proc. ICNN pp 2129-2135, 1997

[34] A. Cichocki, S. Amari, J. Cao, “Neural network models for blind separation of time delayed and convolved signals,” IECE Trans. Fundamentals, E82-A.

[35] K. Torkkola, “Blind separation of radio signals in fading channels,” Advances in Neural Information Processing Systems 10, MIT Press

[36] Y.Yao and H.Vinnet Poor, “Blind detection of synchronous CDMA in non-Gaussian channels,” IEEE Trans. on Signal processing, Vol 52, No.1, pp 271-279, 2004.

[37] M.Ukrainec and S.Haykin, “A modular neural network for enhancement of cross polar radar targets,” Neural Networks, Vol. 9, pp 143-168, 1992

[38] M.Ukrainec and S.Haykin, “Enhancement of radar images using mutual information based unsupervised neural networks,” Canadean Conference of Electrical and Computer Engineering, pp M.A.6.9.1-M.A.6.9.4, 1996

[39] J.Liu,J.Sun,Z.Du, Y.Wan, “Embodying information into images by an MMI based independent component analysis algorithm,”International Conference on Signal Processing, Vol.2, pp 1600-1603, 2002

[40] I.R.Farah, M.B.Ahmed, M.S.Naceur and M.R.Boussema, “Satellite image analysis based on the method of blind separation of sources for the extraction of information,” Proc. of IGARSS’02, Vol.2, pp 919-921, 2002

[41] A. Bell and T. sejnowski, “The independent components of natural scenes are edge filters,” Vision Research, 37,pp 3327-3338,1997

[42] A. Hyvarinen, “Sparse code shrinkage: denoising of nonGaussian data by maximum likelihood estimation, Neural Computation, Vol.11, No.7, pp 1739-1768, 1999

[43] A.D. Back and A.S. Weigend, “A first application of independent component analysis to extracting structure from stock returns,” International Journal on Neural Systems, Vol.8, No.4, pp 483-484, 1998

[44] M. Girolami, A. Cichocki, and S.Amari, “A common neural network model for exploratory data analysis and independent component analysis,” Technical Report, bip-97-001,Brain Information processing Group, RIKEN, 1997

[45] Shi-Fei Ding,Zhong-Zhi She, Yong Liang, Feng-Xiang Jin, “Information Feature Analysis and Improved Algorithm of PCA,” Proceedings of the Fourth International Conference on Machine Learning and Cybernetics , Guangzhou, pp.1756-1761, August 2005.