Neural Networks
Lesson 7 -
Independent Component Analysis

Prof. Michele Scarpiniti
INFOCOM Dpt. - “Sapienza” University of Rome

http://ispac.ing.uniroma1.it/scarpiniti/index.htm
michele.scarpiniti@uniroma1.it

Rome, 12 November 2009
1 Introduction
   - Introduction
   - ICA and Gaussianity

2 Different ICA Approaches
   - HOS methods
   - IT methods
Introduction
Independent Component Analysis (ICA) is a computational method for separating a multivariate signal into additive subcomponents supposing the mutual statistical independence of the non-Gaussian source signals.

- The Independent Component Analysis (ICA) of a random vector consists of searching the linear transformation that minimizes the statistical dependence between its components.

- The concept of ICA may be seen as an extension of Principal Component Analysis, which only imposes independence up to the second order and consequently defines directions that are orthogonal. ICA is a much more powerful technique, however, capable of finding the underlying factors or sources when these classic methods fail completely.

- The data analyzed by ICA could originate from many different kinds of application fields, including digital images and document databases, as well as economic indicators and psychometric measurements.

- Applications of ICA include data compression, detection and localization of sources or blind identification, deconvolution and separation.
Introduction

What distinguishes ICA from other methods is that it looks for components that are both statistically independent, and non-gaussian. In practical situations, we cannot in general find a representation where the components are really independent, but we can at least find components that are as independent as possible.

To rigorously define ICA, we can use a statistical “latent variables” model. We observe $n$ random variables $x_1, \ldots, x_n$, which are modeled as linear combinations of $n$ random variables $s_1, \ldots, s_n$:

$$x_i = a_{i1}s_1 + a_{i2}s_2 + \ldots + a_{in}s_n, \quad \forall i = 1, \ldots, n$$

where $a_{ij}, \forall i, j = 1, \ldots, n$ are some real coefficients. By definition the $s_i$ are statistically mutually independent. This is the basic ICA model. The ICA model is a generative model, which means that it describes how the observed data are generated by a process of mixing the components $s_j$. 
The independent components $s_j$ (often abbreviated as ICs) are latent variables, meaning that they cannot be directly observed. Also the mixing coefficients $a_{ij}$ are assumed to be unknown. All we observe are the random variables $x_i$, and we must estimate both the mixing coefficients $a_{ij}$ and the ICs $s_i$ using the $x_i$. This must be done under as general assumptions as possible.

It is usually more convenient to use vector-matrix notation instead of the sums as in the previous equation. Let us denote by $\mathbf{x}$ the random vector whose elements are the mixtures $[x_1, \ldots, x_n]^T$ and likewise by $\mathbf{s}$ the random vector with elements $[s_1, \ldots, s_n]^T$. Let us denote by $\mathbf{A}$ the matrix with elements $a_{ij}$. Using this vector-matrix notation, the mixing model is written as

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

(1)

Sometimes we need the columns of matrix $\mathbf{A}$; if we denote them by $\mathbf{a}_j$ the model can also be written as

$$\mathbf{x} = \sum_{i=1}^{n} a_{ij} s_i$$
Introduction

Different ICA Approaches

Introduction

ICA and Gaussianity

Restriction in ICA

To make sure that the basic ICA model just given can be estimated, we have to make certain assumptions and restrictions.

1. **The independent components are assumed statistically independent.**
   This is the principle on which ICA rests. Surprisingly, not much more than this assumption is needed to ascertain that the model can be estimated.

2. **The independent components must have non-gaussian distributions.**
   The higher order-cumulants are zero for gaussian distributions, but such higher-order information is essential for estimation of the ICA model. Thus, ICA is essentially impossible if the observed variables have gaussian distributions.

3. **For simplicity, we assume that the unknown mixing matrix is square.**
   In other words, the number of independent components is equal to the number of observed mixtures. This assumption can sometimes be relaxed. It is also assumed here that the mixing matrix is invertible. If this is not the case, there are redundant mixtures that could be omitted, in which case the matrix would not be square.
Ambiguities of ICA

In the ICA model in Eq. (1), it is easy to see that the following ambiguities or indeterminacies will necessarily hold:

1. **We cannot determine the variances (energies) of the independent components.** The reason is that, both \( s \) and \( A \) being unknown, any scalar multiplier in one of the sources \( s_i \) could always be canceled by dividing the corresponding column \( a_i \) of \( A \) by the same scalar, say \( \alpha_i \):

\[
x = \sum_{i} \left( \frac{1}{\alpha_i} a_i \right) (\alpha_i s_i)
\]

Note that this still leaves the ambiguity of the sign: we could multiply an independent component by \(-1\) without affecting the model.

2. **We cannot determine the order of the independent components.** The reason is that, again both \( s \) and \( A \) being unknown, we can freely change the order of the terms in the sum in (1), and call any of the independent components the first one. Formally, a permutation matrix \( P \) and its inverse can be substituted in the model to give \( x = AP^{-1}Ps \). The elements of \( Ps \) are the original independent variables \( s_j \), but in another order.
Preprocessing for ICA

Before applying an ICA algorithm on the data, it is usually very useful to do some preprocessing.

1. **Centering**: The most basic and necessary preprocessing is to center \( \mathbf{x} \), subtracting its mean vector \( \mathbf{m} = E \{ \mathbf{x} \} \) so as to make \( \mathbf{x} \) a zero-mean variable. This preprocessing is made solely to simplify the ICA algorithms: after estimating the mixing matrix \( \mathbf{A} \) with centered data, we can complete the estimation by adding the mean vector of \( \mathbf{s} \) back to the centered estimates of \( \mathbf{s} \) given by \( \mathbf{A}^{-1} \mathbf{m} \).

2. **Whitening**: Another useful preprocessing strategy in ICA is to first whiten the observed variables. This means that before the application of the ICA algorithm (and after centering), we transform the observed vector \( \mathbf{x} \) linearly so that we obtain a new vector \( \tilde{\mathbf{x}} \) which is white. In other words, the covariance matrix of \( \tilde{\mathbf{x}} \) equals the identity matrix: \( R_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = E \{ \tilde{\mathbf{x}}\tilde{\mathbf{x}} \} = I \). The whitening transformation is always possible. One popular method for whitening is to use the eigenvalue decomposition (EVD) of the covariance matrix \( E \{ \mathbf{x}\mathbf{x}^T \} = \mathbf{EDE}^T \), 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:

\[
\tilde{\mathbf{x}} = \mathbf{E} \mathbf{D}^{-1/2} \mathbf{E}^T \mathbf{x}
\]
Preprocessing for ICA

Whitening transforms the mixing matrix $A$ into a new one:

$$\tilde{x} = ED^{-1/2}E^T x = ED^{-1/2}E^T As = \tilde{A}s$$

The utility of whitening resides in the fact that the new mixing matrix $\tilde{A}$ is orthogonal. This can be seen from

$$R_{\tilde{x}\tilde{x}} = E \{\tilde{x}\tilde{x}\} = \tilde{A}E \begin{bmatrix} ss^T \end{bmatrix} \tilde{A}^T = \tilde{A}\tilde{A}^T = I$$

Here we see that whitening reduces the number of parameters to be estimated. Instead of having to estimate the $n^2$ parameters that are the elements of the original matrix $A$, we only need to estimate the new, orthogonal mixing matrix $\tilde{A}$ that contains $n(n-1)/2$ parameters. Thus one can say that whitening solves half of the problem of ICA.
Illustration of ICA

To illustrate the ICA model in statistical terms, consider two independent components that have the following uniform distributions:

\[ p(s_j) = \begin{cases} \frac{1}{2\sqrt{3}}, & \text{if } |s_j| \leq \sqrt{3} \\ 0, & \text{otherwise} \end{cases} \]

Now let us mix these two independent components with the following mixing matrix:

\[ A = \begin{pmatrix} 5 & 10 \\ 10 & 2 \end{pmatrix} \]

This gives us two mixed variables, \( x_1 \) and \( x_2 \). It is easily computed that the mixed data has a uniform distribution on a parallelogram, as shown below.
Why Gaussian variable are forbidden

Whitening also helps us understand why gaussian variables are forbidden in ICA. Assume that the joint distribution of two ICs, $s_1$ and $s_2$, is gaussian. This means that their joint pdf is given by

$$p(s_1, s_2) = \frac{1}{2\pi} \exp \left( -\frac{s_1^2 + s_2^2}{2} \right) = \frac{1}{2\pi} \exp \left( -\frac{\|s\|^2}{2} \right)$$

and assume that the mixing matrix $A$ is orthogonal, that is $A^{-1} = A^T$ and $\det A = 1$. So the joint density of the mixtures $x_1$ and $x_2$ is given by

$$p(x_1, x_2) = \frac{1}{2\pi} \exp \left( -\frac{\|A^T x\|^2}{2} \right) \left| \det A^T \right| = \frac{1}{2\pi} \exp \left( -\frac{\|x\|^2}{2} \right)$$

and we see that the orthogonal mixing matrix does not change the pdf. The original and mixed distributions are identical. The phenomenon that the orthogonal mixing matrix cannot be estimated for gaussian variables is related to the property that uncorrelated jointly gaussian variables are necessarily independent.
Why Gaussian variable are forbidden

Graphically, we can see this phenomenon by plotting the distribution of the orthogonal mixtures, which is in fact the *same* as the distribution of the ICs. The following figure shows that the density is *rotationally symmetric*. Therefore, it does not contain any information on the directions of the columns of the mixing matrix $\mathbf{A}$. This is why $\mathbf{A}$ cannot be estimated.

- Thus, in the case of gaussian independent components, we can only estimate the ICA model up to an orthogonal transformation. In other words, the matrix $\mathbf{A}$ is not identifiable for gaussian independent components.
- With gaussian variables, all we can do is whiten the data. We can allow only one gaussian component.
The Darmois-Skitovitch theorem

A connection between gaussian distributions and stochastic independence can be done by the Darmois-Skitovitch theorem. More precisely, it states that if two linear combinations of non-gaussian independent random variables are again independent, then each original random variable can appear only in one of the two linear combinations.

Theorem (Darmois-Skitovitch)

Let \( x_1, \ldots, x_n \) be \( n \) independent random variables and be \( y_1 \) and \( y_2 \) two linear combinations with \( a_j, b_j \in \mathbb{R} \)

\[
\begin{align*}
  y_1 &= a_1 x_1 + a_2 x_2 + \cdots + a_N x_N \\
  y_2 &= b_1 x_1 + b_2 x_2 + \cdots + b_N x_N
\end{align*}
\]

then if \( y_1 \) and \( y_2 \) are independent then all \( x_j \) with \( a_j b_j \neq 0 \) are gaussian.
Identifiability, Separability and Uniqueness

The solution of this problem consists into three different tasks:

1. **identifiability**: the coefficients of the mixing matrix may be determined from the mixtures alone;

2. **separability**: considers the traditional linear ICA model and recovery of the sources;

3. **uniqueness**: the found solution is unique.

All these tasks are up to permutation and scaling of columns.

Moreover, Comon proofed the following theorem as a corollary of the Darmois-Skitovitch theorem:

**Theorem (Comon)**

Let $s \in \mathbb{R}^n$ be an independent random vector with at most one gaussian component, and $A \in \text{Gl}(n)$. Set $x = As$. Then the following statements are equivalent:

1. $x_1, \ldots, x_n$ are pairwise independent.

2. $x$ is independent.

3. $A$ is equivalent to the product of a scaling and a permutation matrix.
Different solution to ICA problem

ICA problem can be solved by two different approaches.

1. As we have noted a central concept in ICA is the non-gaussianity of the input sources. In this way we have to formulate ICA an estimation as the search for directions that are maximally non-gaussian: each local maximum gives one independent component. As a first practical measure of non-gaussianity, one can use the fourth-order cumulant (kurtosis) or other higher-order cumulants. Then, in order to solve some problems associated with kurtosis, one can use the negentropy as an alternative measure of non-gaussianity. This approach is called Higher Order Statistics or HOS approach.

2. As ICA deals with finding independent components, one can adopt some information theoretic quantities (Mutual Information, Kullback-Leibler Divergence, etc.) as independence measurement. This approach is called Information Theoretic approach.
History of ICA

- **Classic ICA**: linear mixtures
  - ICA was born on 1986 (Herault)
    - **Neurometric** approach
  - Organic development by 1994 (Comon, Jutten, Cardoso, etc.)
    - a **Statistical approach (HOS)**
      - Cumulant based method (Amari & al.)
        - Kurtosis maximization (Amari & al., Jutten)
    - b **Information theoretic approach**
      - FAST ICA (Hyvarinen & al.)
      - INFOMAX (Bell & Seynowski)
        - Maximization of Joint Entropy (ME)
        - Minimization of Mutual Information (MMI)
    - c **Different heuristics**
      - Geometric approach (Babaieh-Zadeh)

- **Nonlinear ICA**
  - Organic development by 1999 (Jutten & Taleb)
  - Open problem
Different ICA Approaches
ICA by Maximization of Non-gaussianity: kurtosis

To use non-gaussianity in ICA estimation, we must have a quantitative measure of non-gaussianity of a random variable with zero mean, say $y$. We can use its kurtosis

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

To simplify things, we can further assume that $y$ has been normalized so that its variance is equal to one:

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

Thus, kurtosis is zero for a gaussian random variable. Kurtosis can be both positive or negative:

- Random variables that have a negative kurtosis are called subgaussian or platykurtic and have pdf with heavy tails;
- those with positive kurtosis are called supergaussian or leptokurtic and have typically a “flat” pdf.

Typically non-gaussianity is measured by the absolute value of kurtosis.
ICA by Maximization of Non-gaussianity: kurtosis

Given the ICA model $\mathbf{x} = \mathbf{A}s$, the independent components can be found adapting a matrix $\mathbf{W}$ such that the vector $\mathbf{y} = \mathbf{Wx}$ has independent components. Practically this is done by evaluating the columns $\mathbf{w}$ of the matrix $\mathbf{W}$ that maximize the cost function $|\text{kurt}(\mathbf{w}^T \mathbf{x})|:

$$
\frac{\partial |\text{kurt}(\mathbf{w}^T \mathbf{x})|}{\partial \mathbf{w}} = 4\text{sign}(\text{kurt}(\mathbf{w}^T \mathbf{x})) \left[ E \left\{ \mathbf{x}(\mathbf{w}^T \mathbf{x})^3 \right\} - 3\mathbf{w} \parallel \mathbf{w} \parallel^2 \right]
$$

since for whitened data we have $E \left\{ \mathbf{x}(\mathbf{w}^T \mathbf{x})^2 \right\} = \parallel \mathbf{w} \parallel^2$. Then we can simply divide $\mathbf{w}$ by its norm, obtaining $\parallel \mathbf{w} \parallel^2 = 1$. To further simplify the algorithm, note that since the latter term in brackets would simply be changing the norm of $\mathbf{w}$ in the gradient algorithm, and not its direction, it can be omitted. Thus we obtain the following gradient algorithm:

$$
\Delta \mathbf{w} = \text{sign} \left( \text{kurt} \left( \mathbf{w}^T \mathbf{x} \right) \right) E \left\{ \mathbf{x} \left( \mathbf{w}^T \mathbf{x} \right)^3 \right\}
$$

$$
\mathbf{w} = \mathbf{w} + \eta \Delta \mathbf{w}
$$

$$
\mathbf{w} = \frac{\mathbf{w}}{\parallel \mathbf{w} \parallel}
$$
ICA by Maximization of Non-gaussianity: kurtosis

A gradient method for maximizing non-gaussianity as measured by the absolute value of kurtosis was derived. The advantage of such gradient methods, closely connected to learning in neural networks, is that the inputs $x(t)$ can be used in the algorithm at once, thus enabling fast adaptation in a non-stationary environment. But, a bad choice of the learning rate can, in practice, destroy convergence. Therefore a fixed-point iteration algorithm is a valid alternative.

To derive a more efficient fixed-point iteration, we note that at a stable point of the gradient algorithm, the gradient must point in the direction of $w$, that is the gradient must be equal to $w$ multiplied by some scalar constant. Equating the gradient of kurtosis with $w$, and remembering that $\|w\|^2 = 1$ using a normalization, this means that we should have

$$w = E\{x (w^T x)\} - 3w$$

After every fixed-point iteration, $w$ is divided by its norm to remain on the constraint set. Note that convergence of this fixed-point iteration means that the old and new values of $w$ point in the same direction.
Hyvarinen et al. have shown that such an algorithm works very well, converging very fast and reliably. This algorithm is called FastICA. The FastICA algorithm has a couple of properties that make it clearly superior to the gradient-based algorithms in most cases.

1. It can be shown that the convergence of this algorithm is cubic. This means very fast convergence.

2. Contrary to gradient-based algorithms, there is no learning rate or other adjustable parameters in the algorithm, which makes it easy to use, and more reliable. Gradient algorithms seem to be preferable only in cases where fast adaptation in a changing environment is necessary.

More sophisticated versions of FastICA are introduced later using the negentropy as non-gaussianity measure.
ICA by Maximization of Non-gaussianity: negentropy

However, kurtosis also has some drawbacks in practice, when its value has to be estimated from a measured sample. The main problem is that kurtosis can be very sensitive to outliers. Thus the value of kurtosis may depend on only a few observations in the tails of the distribution, which may be erroneous or irrelevant observations. In other words, kurtosis is not a robust measure of non-gaussianity.

A fundamental result of information theory is that a gaussian variable has the largest entropy among all random variables of equal variance. This means that entropy could be used as a measure of non-gaussianity. In order to obtain a measure of non-gaussianity that is zero for a gaussian variable and always nonnegative, one often uses a normalized version of differential entropy, called negentropy. Negentropy $J$ is defined as follows

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

where $y_{gauss}$ is a Gaussian random vector of the same covariance matrix as $y$. 
ICA by Maximization of Non-gaussianity: negentropy

The advantage of using negentropy, or equivalently, differential entropy, as a measure of non-gaussianity is that it is well justified by statistical theory. The problem in using negentropy is, however, that it is computationally very difficult. Estimating negentropy using the definition would require an estimate (possibly nonparametric) of the pdf.

- The classic method of approximating negentropy

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

is substantially equivalent to the use of the kurtosis, since the first term is zero for symmetric distribution, a quite common situation in practice.

- One useful approach is to generalize the higher-order cumulant approximation so that it uses expectations of general nonquadratic functions, or “nonpolynomial moments”. In general we can replace the polynomial functions \( y^3 \) and \( y^4 \) by any other functions \( G^i \), possibly more than two.
ICA by Maximization of Non-gaussianity: negentropy

- As a simple special case, we can take any two nonquadratic functions $G^1$ and $G^2$ so that $G^1$ is odd and $G^2$ is even, and we obtain the following approximation:

$$J(y) \approx k_1 (E \{ G^1(y) \})^2 + k_2 (E \{ G^2(y) \} - E \{ G^2(\nu) \})^2$$

where $k_1$ and $k_2$ are positive constants and $\nu$ is a Gaussian variable of zero mean and unit variance.

- In the case where we use only one nonquadratic function $G$, the approximation becomes

$$J(y) \propto [E \{ G(y) \} - E \{ G(\nu) \}]^2$$

- The following choices of $G$ have proved very useful:

$$G_1(y) = \frac{1}{a} \ln \cosh ay$$
$$G_2(y) = - \exp (-y^2 / 2)$$

where $1 \leq a \leq 2$ is some suitable constant, often taken equal to one.
ICA by Maximization of Non-gaussianity: negentropy

The resulting FastICA algorithm finds a direction, the unit vector $w$, such that the projection $w^T x$ maximizes non-gaussianity, here measured by the approximation of negentropy $J(w^T x)$. According to the Lagrange conditions, the optima of $E \{ G(w^T x) \}$ under the constraint $E \{ (w^T x)^2 \} = ||w||^2 = 1$ are obtained at points where the gradient of the Lagrangian is zero:

$$F = E \{ xg(w^T x) \} + \beta w = 0$$

where $g$ is the derivative of the function $G$. Evaluating its gradient:

$$\frac{\partial F}{\partial w} = E \{ xx^T g'(w^T x) \} + \beta I$$

Since the data is sphered, a reasonable approximation seems to be $E \{ xx^T g'(w^T x) \} \approx E \{ xx^T \} E \{ g'(w^T x) \} = E \{ g'(w^T x) \} I$. Thus we obtain

$$w = w - \frac{E \{ xg(w^T x) \} + \beta w}{E \{ g'(w^T x) \} + \beta}$$
ICA by Maximization of Non-gaussianity: negentropy

This algorithm can be further simplified by multiplying both sides of the previous equation by $\beta + E \{ g'(w^T x) \}$. This gives, after straightforward algebraic simplification:

$$w = E \left\{ xg(w^T x) \right\} - E \left\{ g'(w^T x) \right\} w$$

(2)

This is the basic fixed-point iteration in FastICA.

Summarizing the algorithm:

1. Center the data to make its mean zero.
2. Whiten the data to give $x$.
3. Choose an initial (e.g., random) vector $w$ of unit norm.
4. Let $w = E \left\{ xg(w^T x) \right\} - E \left\{ g'(w^T x) \right\} w$, for a defined $g$.
5. Let $w = w / \|w\|$.
6. If not converged, go back to step 4.
ICA by Maximum Likelihood Estimation

According to the ICA model $\mathbf{x} = \mathbf{As}$, we want to find a matrix $\mathbf{W}$, such that $\mathbf{s} \approx \mathbf{Wx}$. So the pdf of the vector $\mathbf{x}$ can be expressed as

$$p_{\mathbf{x}}(\mathbf{x}) = |\det \mathbf{W}| \ p_{\mathbf{s}}(\mathbf{s}) = |\det \mathbf{W}| \prod_i p_i(s_i)$$

Then the likelihood can be obtained

$$L(\mathbf{W}) = \ln E \left\{ \prod_{i=1}^{n} p_i(\mathbf{w}_i^T \mathbf{x}) |\det \mathbf{W}| \right\}$$

and using the log-likelihood, for the properties of the logarithmic function, we obtain

$$\mathcal{L}(\mathbf{W}) = \ln L(\mathbf{W}) = E \left\{ \sum_{i=1}^{n} p_i(\mathbf{w}_i^T \mathbf{x}) \right\} + \ln |\det \mathbf{W}|$$
ICA by Maximum Likelihood Estimation

The simplest algorithms for maximizing log-likelihood are obtained by gradient methods:

$$\Delta W = \frac{\partial L}{\partial W} = (W^T)^{-1} + E \left\{ g(Wx)x^T \right\}$$

where $g(y) = [g_1(y_1), \ldots, g_n(y_n)]$ is a component-wise vector function that consists of the so-called (negative) score functions $g_i$ of the distributions of $s_i$, defined as

$$g_i = (\ln p_i)' = \frac{p_i'}{p_i}$$

This algorithm is often called the Bell-Sejnowski algorithm. This algorithm converges very slowly, however, especially due to the inversion of the matrix $W$ that is needed in every step. The convergence can be improved by whitening the data, and especially by using the natural gradient.
ICA by Minimization of Mutual Information

Mutual information (that is the Kullback-Leibler divergence of the joint pdf with the product of its marginal pdfs) is a natural measure of the dependence between random variables. It is always nonnegative, and zero if and only if the variables are statistically independent. Given the estimate of the original vector $y = Wx$, we have:

$$I(y) = \sum_i H(y_i) - H(x) - \ln |\det W|$$

Now, if the $y_i$ are uncorrelated and of unit variance, it is $E \{yy^T\} = WE \{xx^T\} W^T = I$, which implies that $\det W$ must be constant. Moreover, for $y_i$ of unit variance, entropy and negentropy differ only by a constant and the sign. Thus we obtain

$$I(y) = \text{const} - \sum_i J(y_i)$$

This formula shows that ICA estimation by minimization of mutual information is equivalent to maximizing the sum of non-gaussianities of the estimates of the independent components, when the estimates are constrained to be uncorrelated.
ICA by Tensorial Methods

The **cumulant tensor** is a four-dimensional array whose entries are given by the fourth-order cross-cumulants of the data: $\text{cum}(x_i, x_j, x_k, x_l)$ where the indexes $i, j, k, l$ are from 1 to $n$. The **cumulant tensor** is a linear operator defined by the fourth-order cumulants $\text{cum}(x_i, x_j, x_k, x_l)$. This is analogous to the case of the covariance matrix with elements $\text{cov}(x_i, x_j)$, which defines a linear operator just as any matrix defines one. In the case of the tensor we have a linear transformation in the space of $n \times n$ matrices, instead of the space of $n$-dimensional vectors. The space of such matrices is a linear space of dimension $n \times n$, so there is nothing extraordinary in defining the linear transformation. The $i, j$-th element of the matrix given by the transformation is defined as

$$F_{ij}(M) = \sum_{kl} m_{kl} \text{cum}(x_i, x_j, x_k, x_l)$$

where $m_{kl}$ are the elements in the matrix $M$ that is transformed.
ICA by Tensorial Methods

The cumulant tensor of $\mathbf{x}$ has a special structure that can be seen in the eigenvalue decomposition. In fact, given a matrix of the form $\mathbf{M} = \mathbf{w}_m \mathbf{w}_m^T$ for $m = 1, \ldots, n$, we have

$$F_{ij} \left( \mathbf{w}_m \mathbf{w}_m^T \right) = \sum_{klqq'r'r'} w_{mk} w_{ml} w_{qi} w_{q'j} w_{rk} w_{r'l} \text{cum} \left( s_q, s_{q'}, s_r, s_{r'} \right)$$

This expression can be approximated as

$$F_{ij} \left( \mathbf{w}_m \mathbf{w}_m^T \right) = w_{mi} w_{mj} \text{kurt} \left( s_m \right)$$

The optimization problem can be solved obtaining an iteration of the form

$$\mathbf{w} = \mathbf{w}^T \mathbf{F} (\mathbf{w} \mathbf{w}^T)$$
ICA by Tensorial Methods: JADE

Joint approximate diagonalization of eigenmatrices (JADE) refers to one principle of solving the problem of equal eigenvalues of the cumulant tensor. The matrix $W$ diagonalizes $F(M)$ for any $M$. In other words, $WF(M)W^T$ is diagonal. The diagonality of this matrix can be measured, for example, as the sum of the squares of off-diagonal elements. Thus, we could formulate the following measure:

$$J_{JADE}(W) = \sum_i \left\| \text{diag} \left( WF(M)W^T \right) \right\|^2$$

It can be shown that this expression is equivalent to the following ones

$$J_{JADE}(W) = \sum_{ijkl \neq iikl} \text{cum}(y_i, y_j, y_k, y_l)^2$$

in other words, when we minimize $J_{JADE}$ we also minimize a sum of the squared cross-cumulants of the $y_i$. 

M. Scarpiniti  |  Neural Networks Lesson 7 - Independent Component Analysis
The Hérault-Jutten algorithm

Given the ICA model in the simple case of $2 \times 2$:

$$x_1 = a_{11}x_1 + a_{12}x_2$$
$$x_2 = a_{21}x_1 + a_{22}x_2$$

Hérault and Jutten proposed the feedback circuit shown below to solve the problem:

$$y_1 = x_1 - m_{12}y_2$$
$$y_2 = x_2 - m_{21}y_1$$

Defining a matrix $M$ with off-diagonal elements $m_{12}, m_{21}$ and diagonal elements equal to zero, these equations can be compactly written as $y = x - My$, that is

$$y = (I + M)^{-1}x \approx (I - M)x$$
The Hérault-Jutten algorithm

The solution that Jutten and Hérault introduced was to adapt the two feedback coefficients $m_{12}, m_{21}$ so that the outputs of the network $y_1, y_2$ become independent. For independence, they used the criterion of nonlinear correlations. They proposed the following learning rules:

$$\Delta m_{12} = \mu f(y_1)g(y_2)$$
$$\Delta m_{21} = \mu f(y_2)g(y_1)$$

with $\mu$ the learning rate. Both functions $f(\cdot), g(\cdot)$ are odd functions; typically it is used the following ones:

$$f(y) = y^3, \quad g(y) = \arctan(y)$$

Now, if the learning converges, then the right-hand sides must be zero on average, implying

$$E\{f(y_1)g(y_2)\} = E\{f(y_2)g(y_1)\} = 0$$

Thus independence has hopefully been attained for the outputs $y_1$ and $y_2$. 
Noisy ICA

Here we extend the basic ICA model to the situation where noise is present. The noise is assumed to be additive. Thus the model becomes

\[ x = As + n \]

where \( n = [n_1, \ldots, n_n]^T \) is the noise vector. Some further assumptions on the noise are usually made. In particular, it is assumed that

1. the noise is independent from the independent components;
2. the noise is gaussian.

The covariance matrix of the noise, say \( \Sigma \), is often assumed to of the form \( \sigma^2 I \). In this typical case, the noise could be considered as “sensor” noise. This is because the noise variables are separately added on each sensor. Source noise can be modeled with an equation slightly different from the preceding, given by

\[ x = A(s + n) \]

where again the covariance of the noise is diagonal.
Noisy ICA

We could consider the noisy independent components, given by \( \tilde{s}_i = s_i + n_i \), and rewrite the model as

\[ x = A\tilde{s} \]

We see that this is just the basic ICA model, with modified independent components. What is important is that the assumptions of the basic ICA model are still valid: the components of \( \tilde{s} \) are non-gaussian and independent. Thus this model we can be estimated by any method for basic ICA. The estimation of the original independent components from the noisy ones is an additional problem, solved, for example, by a maximum a posteriori estimation, etc.

To recapitulate:

*if the noise is added to the independent components and not to the observed mixtures, or has a particular covariance structure, the mixing matrix can be estimated by ordinary ICA methods. The denoising of the independent components is another problem, though.*
The fixed-point iteration algorithm

In this slides we want to give a look to the fixed-point iteration algorithm.

Definition

A fixed-point \( \bar{x} \) of a function \( f(x) \in \mathbb{R} \) is a point that is mapped to itself by the function:

\[
    f(\bar{x}) = \bar{x}
\]

When we want to numerically solve the equation

\[
    F(x) = 0
\]

it is convenient to rearrange \( F(x) = 0 \) into \( x = K(x) \) and undertake the iteration

\[
    x_{k+1} = K(x_k), \quad \text{for a chosen } x_0
\]

If this series converges to \( x^* \), the by continuity

\[
    x^* = \lim_{k \to \infty} K(x_k) = K(x^*)
\]

Hence \( x^* \) is a fixed-point of \( K(x) \) and thus a zero of \( F(x) \).
The fixed-point iteration algorithm

Let us give an example.

\[ F(x) = x^2 + x - 6 = 0 \quad \Leftrightarrow \quad x = -x^2 + 6 = K(x) \]
References

A. Hyvarinen, J. Karhunen, E. Oja.
Independent Component Analysis.

A. Hyvarinen and E. Oja.
Independent Component Analysis: algorithms and applications.

Fast and Robust Fixed-Point Algorithms for Independent Component Analysis.

A. Cichocki and S.-I. Amari.
Adaptive Blind Signal and Image Processing.

P. Comon.
Self-Organization in a Perceptual Network.

J. Eriksson and V. Koivunen.
Identifiability, separability and uniqueness of linear ICA models.

S. Roberts and R. Everson.
Independent Component Analysis: principles and practice.