1 Results

1.1 Notations

Cocktail party problem :

$$S \text{ sources} \Rightarrow \mathcal{F}(X) \text{ observations} \Rightarrow \mathcal{H}(y) \text{ estimations}$$

Where $\mathcal{F}$ is the mixing function, and $\mathcal{H}$ is the separating function. $S = (s_1, \cdots, s_p)^T$ is the sources in $\mathbb{R}^p$, X is a random variable in $\mathbb{R}^n$. We assume $p = n$.
Linear case :

$$y = WX = WAS$$

We note $G = WA$ We consider the times series as matrices :

$$X_T = \begin{pmatrix} x_1(1) & \cdots & x_1(T) \\ \vdots & & \vdots \\ x_n(1) & \cdots & x_n(T) \end{pmatrix}$$

We note the i-th rox of a matrix W as w_i.

1.2 PCA

We note V_x the covariance matrix of x, and

$$V_x = FDF^T$$

its diagonal reduction in an orthonormal basis. **Result** If we denote w_1 the eigen vector corresponding to the greatest eigen value of V_x, then the data projection on w_1 is called the *principal component*, and it is the one coding the most variance : $w_1 = \underset{|w|=1}{\text{argmax}}\ \mathbb{E}[(w^T x)^2]$. We can compute w using methods described in @articleoja1992principal, title=Principal components, minor components, and linear neural networks, author=Oja, Erkki, journal=Neural Networks, volume=5, number=6, pages=927–935, year=1992, publisher=Elsevier
We then have $W_{PCA} = D^{-1/2}F^T$. We call this operation *spectral whitening* of the data. This basicly nullifies the data variances.

We can also have $W_{ZCA} = \mathbb{E}[x^T x]^{-1/2}$. Then the covariance matrix of $y = W_{ZCA}x$ is diagonal and our datas are uncorrelated.

2 ICA

The major hypothesis of ICA is the statistical independance of our data. In the second hypothesis according to which the sources are linearly mixed,

the first hypothesis is sufficient to provide a good source separation.
let V_x be the covariance matrix of x. We factorize V_x as:

$$V_x = AD^2A^T$$

Where $D \preccurlyeq 0$ is diagonal, and $A \in \mathcal{M}_{pn}$ is of rank n. We can put observations as $x = As$, where D^2 is the covariance matrix of y, and where s is a vector of $\mathbb{R}^n$ which is maximizing a certain *contrast function* or *independance measure* which can be found in @articlecomon1994independent, title=Independent component analysis, a new concept?, author=Comon, Pierre, journal=Signal processing, volume=36, number=3, pages=287–314, year=1994, publisher=Elsevier
Optimization algorithm found in @articlehyvarinen1999survey, title=Survey on independent component analysis, author=Hyvarinen, Aapo, journal=Neural computing surveys, volume=2, number=4, pages=94–128, year=1999 For ICA we need the following conditions :

- At most one source can follow a gaussian distribution
- the rank of A must be the number of sources

First condition is due to the fact that a gaussian distribution as all this moment of order greater than 2 equal to 0. Thus the independance hypothesis is just a simple decorrelation such as the one done in PCA. The second condition can be lifted as shown in several papers (P51 du cours). With dimensional reduction techniques, we can always assume the the number of observations is equal to the number of sources (which implies tha A is squared).
Two last problems arise from ICA : it is insensible to a permutation of sources and a change in amplitude of a source. We can modelize this by a multiplication by a diagonal matrix of weight D for the amplitude and a multiplication by the inverse of a permutation matrix P.

Let p_y be the density of the estimated sources.

Contrast function The contrast function Ψ should be invariant to permutations : $\Psi(P.p_y) = \psi(p_y)$, scale invariant: $\Psi(p_{\Delta y})$ for any Δdiagonal, and finaly discriminant for y with mutualy independant composants : $\Psi(p_{My}) \geq \Psi(p_y)$ with equality only when M is of the form ΔP. The mutual information is great for this problem, but hard to compute. Thus we use numerical approximation such as Gram-Charlier series.

2.1 Neuronal approach

The idea is to try to find y as $y = (I + W)^{-1}x$, where the diagonal of W is 0, and where $w_{ij} = f(y_i)g(y_j)$ elsewhere. f and g are supposed to be odd functions, several choices are given in the quoted articles (P53). It is in the necessity to resort to statistics of higher order that lies the contribution of ICA.

2.2 Learning approach

Using the contrast function, we define the ACI problem as an optimization one, and an iterative learning of the separation matrix. This kind of algorithms are dependant to the mixing matrix. Thus we use invariant estimator : $\hat{A}_{MX_T} = M\hat{A}_{X_T}$ ($\hat{A}_{X_T}$ is the estimation of A obtain with T samples X. We can see that with such an estimator of the sources, we have : $\hat{s}(t) = (\hat{A}_{S_T})^{-1}$. We define the relative gradient :

$$W_{t+1} = (I - \lambda_t \nabla J_{\psi}(y_t))W_t$$

Where $\nabla J_{\psi}(y_t)$ is the gradient of a cost function depending on the contrast function ψ computed on y_t. Then the serial update of the global source matrix $G = WA$ verifies :

$$G_{t+1} = (I - \lambda_t \nabla J_{\psi}(G_t s))G_t$$

Results For $\psi(y) = \sum_{i=1}^{n} |y_i|^4$ and $J_{\psi}(y) = \mathbb{E}[\psi(y)]$, we have $W_{t+1} = W_t - (y_t y_t^T - I + g(y_t)y_t^T - y_t g(y_t)^T)W_t$.
See @inproceedingscardoso1996independent, title=Independent component analysis, a survey of some algebraic methods, author=Cardoso, Jean-Frcsnçois and Comon, Pierre, booktitle=Circuits and Systems, 1996. ISCAS'96., Connecting the World., 1996 IEEE International Symposium on, volume=2, pages=93–96, year=1996, organization=IEEE This idea is at the base of the HJ algorithm.

2.3 Tensorial approach

The idea is to do tensorial diagonalization in order to optimize the contrast function, giving the *JADE* algorithm, which follows the *FOBI* one. The cumulant tensor of the 4-th order is a 4D matrix containing all the 4-th order cumulant :$N = (Cum(x_i, x_j, x_k, x_l))_{i,j,k,l}$. We can also view N as a linear application of $\mathcal{M}_{nn}$ to $\mathcal{M}_{nn}$, and see N as a matrix of size $n^2 \times n^2$. The aim of this algorithm is then to diagonalize N which, as N is normalized, is equivalent to try to maximize the diagonal elements of N. It can be shown that the contrast function is then $c(x) = \sum_{i,k,l} |Cum(x_i, x_i*, x_k, x_l*)|^2$.
This algorithm is of the monstruous complexity of $\mathcal{O}(n^4)$, and thus unusable for real data…

2.4 Maximum likelihood

Likelihood of the observations given our mixing matrix :

$$p_{x|A}(y) = \int p_s(A^{-1}u)|\det(A)|^{-1}du$$

If we note $\Phi_i = [\log(p_{s_i})]'$, then the maximum likelihood estimator is obtained by solving :

$$\hat{\mathbb{E}}[\Phi_i(e_i^T A^{-1}x)e_j^T A^{-1}x] = 0$$

With $\hat{s}_i = e_i^T A^{-1}x$ the estimation of our sources, we get the simple condition :

$$\hat{\mathbb{E}}[\Phi_i(\hat{s}_i)\hat{s}_j] = 0$$

This becomes an optimization problem, we can use what we want to solve it... This approach as been proved to be equivalent to the *Infomax* method by Cardoso.

2.5 ICA as non linear PCA

We can derive an andaptation rule for neurons network learning :

$$W_{t+1} = W_t + \lambda_t[x_t g_1(e_t^T)W_t G_2(x_t^T W_t) + g_1(e_t)f_2(x_t^T W_t)]$$

Where f_1, f_2 are two non-linear functions, $g_1 = f_1', g_2 = f_2'$, $G_2 = \text{diag}(g_2(.))$ and $e_t = x_t - W_t g_2(W_t^T x_t)$ For stability reasons, g_1 needs to be a impaire croissante function. With f_1 quadratic and f_2 lineal, this is the standard PCA.
With W orthogonal and $y = Wx$, we have :

$$\|x - W^T g(Wx)\|^2 = \sum_{i=1}^{n}[y_i - g(y_i)]^2$$

2.6 Infomax

Theorticaly equivalent to the learning approach. From neurons network theory :

$$\frac{\partial I}{\partial w}(x,y) = \frac{\partial H}{\partial w}(y)$$

Where I(x,y) is the mutual information between inputs x and outputs y of the neurones network, $H(y)$ is the ouput entropy and w the network parameters. Thus the update rule :

$$\Delta W = (W^{-T}) + \frac{\partial}{\partial w}\ln\prod_i |y_i'|$$

This rule can be upgraded for faster convergence and better results into:

$$\Delta W = [I - K \tanh(y) y^T - y y^T] W$$

Where K is a diagonal matrix with 1 for an over gaussian source, and -1 for a sub gaussian one.

2.7 Non gaussianity measure

Accordiang to CLT, a sum of independant variables converges toward a gaussiand distribution. Thus we try to have estimations as much non-gaussian as possible. A robust measure of non-gaussianity is the negentropy : $J(y) = H(y_{gauss}) - H(y)$ where H is the classical entropy, and y_{gauss} is a gaussian variable of same mean and variance as y. To obtain even more robust estimators, we approach negentropy by : $J(y) \propto [\mathbb{E}[G(y)] - \mathbb{E}[G(\mu)]^2$. Where G is a non quadratic function. The algorithm (which is based on the last section update rule) needs uncorrelated and centered data at each iteration, thus an orthogonalization step.

The orthonormalization step can be done either column by column (Deflation) or all matrix directly (symmetric).

3 Algorithms

3.1 HJ

Described in section 2.2. One matlab version found.

3.2 CoM1-2

Using contrast functions and cumulants. Introuvable...

3.3 Jade

Described in section 2.3. One matlab version (Cardoso's ma gueule) found.

3.4 Fast-ICA

Described the last four sections. Available in matlab and python packages.

3.5 BS

Based on infomax and neurones networks. Extract matlab from `ftp://ftp.cnl.salk.edu/pub/tony/sep96.public` .

4 Applications

- Speech signals
- Medical imagery : EEG and MEG, **MRI**
- **Finance : look for independant factors explaining the actions market structure**
- **Analyze of the money flux of 40 stores of a same chain over 3 years to see temporal events or relation to other chains.**
- **Temporal series (Stephaaaaane Mallaaaaaaaat) construct a predictor**
- **Use on natural images to mimique the brain**
- **Image recognition and classification by extraction of patterns**
- **Multimedia data modelization**
- **Audio-Video fusion**
- **Image compression ***
- **Image denoizing ***
- **Transparence separation (photo through galss) ***

* Not studied enough, or not an improvement enough at the time.