Course 495: Advanced Statistical Machine Learning/Pattern Recognition

Deterministic Component Analysis

- Goal (Lecture): To present two more component analysis algorithms, Independent Component Analysis (ICA).

- Goal (Tutorials): To provide the students the necessary mathematical tools for deeply understanding the CA techniques.
Materials

- Pattern Recognition & Machine Learning by C. Bishop Chapter 12
‘Cocktail party’

Sources $y_1$ and $y_2$

Mixing matrix $W$

Observations

$x_1 = w_{11}y_1 + w_{12}y_2$

$x_2 = w_{21}y_1 + w_{22}y_2$

$x = Wy$

$d$ sources (latent space), $N$ observations
‘Cocktail party’

- Observing signals
- Original source signal
- ICA

\[ x \]

\[ y \]
‘Cocktail party’

• Two Independent Sources

\[ x_1(t) = w_{11} y_1(t) + w_{12} y_2(t) \]

\[ x_2(t) = w_{21} y_1(t) + w_{22} y_2(t) \]

• Mixture at two Mics
‘Cocktail party’

• Get the Independent Signals out of the Mixture
‘Cocktail party’
‘Cocktail party’

Independent Components of natural images
Formulating the problem

\[ x_{i1} = \sum_{k=1}^{d} w_{ik} y_{k1} \]
\[ \vdots \]
\[ x_{iN} = \sum_{k=1}^{d} w_{ik} y_{kN} \]

\[ X = WY \]

Find the mixing matrix \( W \) and the independent components \( Y \)

If \( W \) is a square and invertible then \( Y = W^{-1}X \)
Formulating the problem

Sources of ambiguity:

1. We cannot determine the variances (energies) of the independent components.

\[ X = WL^{-1}LY \]

\[ L = \begin{bmatrix} l_1 & 0 & 0 \\ 0 & l_2 & 0 \\ 0 & 0 & l_3 \end{bmatrix} \quad L^{-1} = \begin{bmatrix} \frac{1}{l_1} & 0 & 0 \\ 0 & \frac{1}{l_2} & 0 \\ 0 & 0 & \frac{1}{l_3} \end{bmatrix} \]

It can be partially resolved by

\[ E[y] = 1 \]

Though sign ambiguity cannot be resolved
Formulating the problem

2. We cannot determine the order of the independent components.

\[ X = WP^{-1}PY \]

\( P \) can be a permutation matrix

\( PY \) are the independent components but in another order

\( WP^{-1} \) is the new mixing matrix
A first example

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

\[ y_1 \sim p(y) \quad y_2 \sim p(y) \]

\[ E[y_1] = 0 \quad E[y_2] = 0 \]

\[ \sigma_1^2 = E[y_1^2] = 1 \]

\[ \sigma_2^2 = E[y_2^2] = 1 \]
A first example

\[ p(y_1, y_2) = p(y_1)p(y_2) \]

\[ y_1 \sim p(y) \quad y_2 \sim p(y) \]
A first example

\[ x_1 = 2y_1 + 3y_2 \]
\[ x_2 = 2y_1 + 1y_2 \]

\[ W = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \]

\[ w_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \]

\[ w_2 = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \]

\[ W^{-1} = \begin{bmatrix} -0.25 & 0.5 \\ 0.75 & -0.5 \end{bmatrix} \]
Definition of independence

\( y_1 \) and \( y_2 \) are independent iff \( p(y_1, y_2) = p(y_1)p(y_2) \)

Alternative definition

\[
E(h_1(y_1), h_2(y_2)) = E(h_1(y_1))E(h_2(y_2))
\]

\[
E(h_1(y_1), h_2(y_2)) = \int_{-\infty}^{+\infty} h_1(y_1)h_2(y_2)p(y_1, y_2)dy_1dy_2
\]

\[
= \int_{-\infty}^{+\infty} h_1(y_1)p(y_1)dy_1 \int_{-\infty}^{+\infty} h_2(y_2)p(y_2)dy_2
\]

\[
= E(h_1(y_1))E(h_2(y_2))
\]
Uncorrelated variables are only partly independent

\[ h_1(y) = y \quad h_2(y) = y \]

\[ E(y_1, y_2) = E(y_1)E(y_2) \]

To show that

\[
\begin{align*}
(y_1, y_2) &= (0, 1) & E(y_1, y_2) &= E(y_1)E(y_2) \\
(y_1, y_2) &= (0, -1) & E(y_1^2 y_2^2) &= 0 \\
(y_1, y_2) &= (1, 0) & \neq E(y_1^2)E(y_2^2) &= \frac{1}{4} \\
(y_1, y_2) &= (-1, 0)
\end{align*}
\]
Non Gaussianity

\[ p(y_1) = \frac{1}{\sqrt{2\pi}} e^{-y_1^2} \]

\[ p(y_2) = \frac{1}{\sqrt{2\pi}} e^{-y_2^2} \]

When \( W \) is orthogonal it gives

\[ p(y_1, y_2) = \frac{1}{\sqrt{2\pi}} e^{-(y_2^2 + y_1^2)} \]
Non Gaussianity

The joint density of unit variance $y_1$ & $y_2$ is symmetric. So it doesn‘t contain any information about the directions of the cols of the mixing matrix $W$.

So $W$ can‘t be estimated.

We need non-gaussianity for the independent components (IC).

If only one IC is gaussian, the estimation is still possible.
Non Gaussianity

• Key element is non-gaussianity

\[ A = W^{-1} \quad \tilde{y} = a^T X \]

• If \( a \) was one of the rows of the inverse of \( W \), this linear combination \( \tilde{y} \) would actually equal to one of the independent components.

• How could we use the Central Limit Theorem to determine \( W \) so that it would equal to one of the rows of the inverse of \( A \)?
Non Gaussianity

• Let us make a change of variables

\[ \mathbf{z} = \mathbf{W}^T \mathbf{a} \Rightarrow \]

\[ \tilde{y} = \mathbf{a}^T \mathbf{X} = \mathbf{a}^T \mathbf{WY} = \mathbf{z}^T \mathbf{Y} \]

\( \tilde{y} \) is thus a linear combination of \( \mathbf{Y} \), with weights given by \( \mathbf{z} \)

Even the sum of two independent random variables is more Gaussian than the original variables
Non Gaussianity

• Maximize the non-Gaussianity of $\mathbf{a}^T \mathbf{x}$. This means that $\mathbf{a}^T \mathbf{x} \hat{=} \mathbf{a}$ equals to one of the independent components!

• Maximizing the non-Gaussianity of $\mathbf{a}^T \mathbf{x}$ thus gives us one of the independent components.

• In fact, non-Gaussianity in the $n$-dimensional space of vectors $\mathbf{a}_i$ has $2n$ local maxima, two for each independent component, corresponding to $\mathbf{y}_i$ and $-\mathbf{y}_i$.

• To find several independent components, we exploit the fact that different independent components are uncorrelated.
Measures of Non Gaussianity

- Assuming a random variable $y$ such that

$$E[y] = 0 \quad \sigma^2 = E[y^2] = 1$$

- The classical measure of non-Gaussianity is kurtosis or the fourth-order cumulant

$$kurt(y) = E[y^4] - 3(E[y^2])^2 \Rightarrow kurt(y) = E[y^4] - 3$$

Kurtosis is zero for a Gaussian random variable. For most (but not quite all) non-Gaussian random variables, kurtosis is non-zero.
Measures of Non Gaussianity - Kurtosis

- $kurt(y) < 0$ are called sub-Gaussian,
- $kurt(y) > 0$ are called super-Gaussian.

Super-Gaussian random variables have typically a “spiky” pdf with heavy tails, i.e.

Sub-Gaussian random variables, on the other hand, have typically a “flat” pdf,

$$p(y) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|y|}$$
Measures of Non Gaussianity - Kurtosis

• Kurtosis, or rather its absolute value, has been widely used as a measure of non-Gaussianity in ICA and related fields.

• The main reason is its simplicity, both computational and theoretical.

• Computationally, kurtosis can be estimated simply by using the fourth moment of the sample data.
Measures of Non Gaussianity - Kurtosis

- Kurtosis has problems when its value is estimated from a measured sample.
  1. It is very sensitive to outliers.
  2. Its value depends on only a few observations in the tails of the distribution, which may be erroneous or irrelevant observations.

- Thus, other measures of non-Gaussianity might be better than kurtosis in some situations, i.e. negentropy that more or less combine the good properties of both measures.
Measures of Non Gaussianity - Negentropyp

• A very important measure of non-Gaussianity is given by
negentropy. Negentropy is based on the information-
theoretic quantity of (differential) entropy.

• Entropy is the basic concept of information theory
(measure of “randomness” of a variable)

• Entropy $H$ is defined for discrete/continuous random
variable $y$ as

$$H(y) = - \sum p(y_i) \log p(y_i) \quad H(y) = - \int p(y) \log p(y) dy$$
A fundamental result of information theory is that: a Gaussian variable has the largest entropy among all random variables of equal variance.

Gaussian distribution is the “most random” or the least structured of all distributions.

How entropy could be used as a measure of non-Gaussianity?

Negentropy $J$ is defined as follows:

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

where $y_{Gauss}$ is a Gaussian random variable of the same covariance matrix as $y$. 
Measures of Non Gaussianity - Negentropy

• Negentropy is always non-negative, and is zero if and only if $y$ has a Gaussian distribution.

• Negentropy is in some sense the optimal estimator of non-Gaussianity, as far as statistical properties are concerned.

• The problem in using negentropy is, however, that it is computationally very difficult.

• Estimating negentropy using the definition would require an estimate (possibly non-parametric) of the pdf.
Measures of Non Gaussianity - Negentropy

• The classical method of approximating negentropy is using higher-order moments, for example as follows (zero mean and unit variance)

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

• However, the validity of such approximations may be rather limited.

• In particular, these approximations suffer from the non-robustness encountered with kurtosis
Measures of Non Gaussianity - Negentropy

• Assuming a zero mean and unit variance $\mathbf{y}$, a more useful approximation is the following

$$J(\mathbf{y}) \approx c[E(G(\mathbf{y})) - E(G(\mathbf{v}))]^2$$

where $G$ is practically any non-quadratic function (higher order than 2), $c$ is an irrelevant constant, and $\mathbf{v}$ is a Gaussian variable of zero mean and unit variance (i.e., standardized).

• If we set $\mathbf{y} = \mathbf{a}^T \mathbf{X}$ then negentropy is reformulated as

$$J(\mathbf{a}) = [E(G(\mathbf{a}^T \mathbf{X})) - E(G(\mathbf{v}))]^2$$
Measures of Non Gaussianity - Negentropy

Examples of function $G$

\[ G_1(y) = \frac{1}{4} y^4 \quad g_1(y) = y^3 \]

\[ G_2(y) = -\frac{1}{c_1} e^{-\frac{c_1}{2} y^2} \quad g_2(y) = y e^{-\frac{c_1}{2} y^2} \]

\[ G_3(y) = \frac{1}{c_2} \log \cosh c_2 y \quad g_3(y) = \tanh c_2 y \]

\[ 1 \leq c_2 \leq 2, \quad c_1 \approx 1 \]
Measures of Non Gaussianity - Negentropy

\[ A = \arg \max_A J(A) = \sum_{k=1}^{d} J(a_k) \]

\[ \text{s.t. } A^T XX^T A = I \]

Assuming whitened data, i.e. \( XX^T = I \)

\[ A = \arg \max_A J(A) = \sum_{k=1}^{d} J(a_k) \]

\[ \text{s.t. } A^T A = I \]
Measures of Non Gaussianity - Negentropy

Let that we want to find one $\mathbf{a}$

$$\mathbf{a} = \text{argmax}_a J(\mathbf{a}) \quad \text{s.t.} \quad \mathbf{a}^T \mathbf{a} = 1$$

Lagrangian

$$L(\mathbf{a}, \lambda) = J(\mathbf{a}) - \lambda(\mathbf{a}^T \mathbf{a} - 1)$$

$$\frac{\partial L(\mathbf{a}, \lambda)}{\partial \mathbf{a}} = E[\mathbf{x}_i g(\mathbf{a}^T \mathbf{x}_i)] - \lambda \mathbf{a} \Rightarrow \lambda = E[\mathbf{a}^T \mathbf{x}_i g(\mathbf{a}^T \mathbf{x}_i)]$$

$$\frac{\partial^2 L}{\partial a_i \partial a_j} = E[\mathbf{x}_i \mathbf{x}_i^T g'(\mathbf{a}^T \mathbf{x}_i)] - \lambda \mathbf{I}$$
Measures of Non Gaussianity - Negentropy

Assuming the approximation

\[ E[x_i x_i^T g'(a^T x_i)] \approx E[x_i x_i^T] E[g'(a^T x_i)] = E[g'(a^T x_i)] I \]

We get the Newton updates

\[ \alpha_+^{(t)} = \alpha^{(t-1)} - \frac{E[x_i g(a^T x_i)] - \lambda \alpha}{E[g'(a^T x_i)] - \lambda} \]

\[ \alpha^{(t)} = \frac{\alpha_+^{(t)}}{||\alpha_+^{(t)}||} \]

Also setting that \( \lambda = E[a^T x_i g(a^T x_i)] \)

We get the fix point updates

\[ \alpha_+^{(t)} = E[x_i g(a^T x_i)] - E[g'(a^T x_i)] \alpha^{(t-1)} \]
Measures of Non Gaussianity - Negentropy

• A simple way of achieving decorrelation is a deflation scheme based on a Gram–Schmidt-like decorrelation.

• We estimate the independent components one by one, i.e. to estimate $d$ independent components, or $d$ vectors $a_1, a_2, \ldots a_d$ we run the one-unit fixed point algorithm for $a_{i+1}$; and after every iteration step we subtract from $a_{i+1}$ the “projections” of the previously estimated $i$ vectors, and then renormalize as:

$$a_{i+1} = a_{i+1} - \sum_{j=1}^{i} a_{i+1}^T a_j a_j$$

$$a_{i+1} = \frac{a_{i+1}}{||a_{i+1}||}$$
Preprocessing

• Centering
\[ x_i = x_i - \mu \quad \text{where} \quad \mu = \frac{1}{N} \sum_{i=1}^{N} x_i \]

• Sphering
\[ \tilde{X} = U \Lambda^{-\frac{1}{2}} U^T X \]

where \( \Lambda \) is the diagonal matrix of the positive eigenvalues and \( U \) is the matrix with the corresponding eigenvectors.
An overview

PCA: Maximize the global variance

LDA: Minimize the class variance while maximizing the mean variance

LPP: Minimize the local variance

ICA: Maximize independence by maximizing non-Gaussianity

All are deterministic!!