INDEPENDENT COMPONENT ANALYSIS

Independent Component Analysis is a statistical generative model which aims to reveal hidden independent additive components from multi-dimensional data.

1 Mathematical Preliminaries

1.1 Single Random Variable

Let the Probability Mass Function of a discrete random variable X be defined as follows.

	$x = 1$	$x = 2$	$x = 3$	$x = 4$
$P(X = x)$	1/6	1/3	1/3	1/6

The expectation $E(X)$ of this random variable is defined as follows.

$$\begin{aligned} E(X) &= \sum_k x_k P(X = x_k) \\ &= 1 \times \frac{1}{6} + 2 \times \frac{1}{3} + 3 \times \frac{1}{3} + 4 \times \frac{1}{6} \\ &= \frac{15}{6} \end{aligned}$$

If a new variable $Y = 2X + 1$ is defined, the PMF of Y becomes

	$y = 2 \times 1 + 1$	$y = 2 \times 2 + 1$	$y = 2 \times 3 + 1$	$y = 2 \times 4 + 1$
$P(Y = y = 2x + 1)$	1/6	1/3	1/3	1/6

$$\begin{aligned}E(Y) &= \sum_k y_k P(Y = y_k) \\ &= \sum_k (2x_k + 1)P(Y = y_k) \\ &= \sum_k (2x_k + 1)P(X = x_k) \\ &= 2E(X) + 1\end{aligned}$$

Hence, expectation is a linear operator which satisfies $E(aX + b) = aE(X) + b$.

The variance of a discrete random variable X is defined as

$$\begin{aligned}Var(X) &= E((X - E(X))(X - E(X))) \\ &= E(X^2 - 2E(X)X + E(X)E(X)) \\ &= E(X^2) - 2E(X)E(X) + E(X)E(X) \\ &= E(X^2) - E(X)^2\end{aligned}$$

In case of continuous random variables, instead of Probability Mass Function a Probability Density Function is defined because the sample space is continuous and infinite as opposed to discrete and finite.

The expectation and variance of a continuous random variable are defined as

$$\begin{aligned}E(X) &= \int xP(X = x)dx \\ Var(X) &= \int (x - E(X))^2 P(X = x)dx\end{aligned}$$

Just as in the case of discrete random variables where the sum of probabilities of all events in sample space must sum to 1, the area under PDF curve in the case of continuous random variable must also integrate to 1.

An example of a well known continuous random variable is the Gaussian random variable whose PDF is defined as follows.

$$P(X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

with the following expectation and variance.

$$\begin{aligned}E(X) &= \mu \\ Var(X) &= \sigma^2\end{aligned}$$

1.2 A Pair of Random Variables

Let the joint Probability Mass Function of a pair of discrete random variables X and Y be defined as follows.

P(X=x, Y=y)	$x = 1$	$x = 2$	$x = 3$
$y = 1$	3/8	3/16	3/16
$y = 2$	1/8	1/16	1/16

As usual, the sum of all probabilities is equal to 1.

Given the joint distribution of a set of random variables, one can also find out the distribution of a subset of those random variables and the distribution of such a subset is called the marginal distribution. In the above example, the marginal distribution of X and Ycan be computed by summing up each column and row respectively in the table above.

$$P(X = x_k) = \sum_{y_l} P(X = x_k,\ y = y_l)$$
$$P(Y = y_k) = \sum_{x_l} P(X = x_l,\ y = y_k)$$

In the case of a pair of continuous random variables, their joint Probability Distribution Function becomes a surface in three dimensions. The marginal distributions take the following formulae.

$$P(X = x) = \int_y P(X = x,\ y = y)dy$$
$$P(Y = y) = \int_x P(X = x,\ y = y)dx$$

Two random variables are said to statistically independent if and only if

$$P(X = x,\ Y = y) = P(X = x)P(Y = y)$$

For independent random variables, the following result holds.

$$\begin{aligned}
E(f(X)g(Y)) &= \int\int f(x)g(y)P(X=x,\ Y=y)dxdy \\
&= \int\int f(x)g(y)P(X=x)P(Y=y)dxdy \\
&= \int f(x)P(X=x)dx \int g(y)P(Y=y)dy \\
&= E(f(X))E(g(Y))
\end{aligned}$$

A weaker form of independence is uncorrelatedness. If $f(X)$ and $g(Y)$ are defined such that $f(X) = X$ and $g(Y) = Y$ and $E(XY) = E(X)E(Y)$ holds true, then two variables are said to be uncorrelated.

1.3 Sum of Random Variables

If X denotes the outcome of roll of a dice and Y denotes the outcome of roll of another dice, then $Z = X + Y$ denotes a random variable whose value is equal to the sum of the two rolls. If PMFs of X and Y are known, what is the PMF of $Z = X + Y$?

$$\begin{aligned}
P(Z = X + Y = z) &= \sum_x P(X=x,\ Y=z-x) \\
&= \sum_y P(X=z-y,\ Y=y)
\end{aligned}$$

It is easy to see why the above equation holds. To find $P(Z = z)$, one has to sum the probabilities $P(X = x,\ Y = y)$ of all possible scenarios where $x + y = z$. In the case of continuous random variables the formulae become

$$\begin{aligned}
P(Z = X + Y = z) &= \int P(X=x,\ Y=z-x)dx \\
&= \int P(X=z-y,\ Y=y)dy
\end{aligned}$$

Random variable sums are commutative and associative.

$$\begin{aligned}
X + Y &= Y + Z \\
X + Y + Z &= (X + Y) + Z \\
&= X + (Y + Z)
\end{aligned}$$

The expectation of a sum of two random variables is the sum of expectations of individual random variables.

$$
\begin{aligned}
E(X+Y) &= \int\int (x+y)P(X=x,\ Y=y)dxdy \\
&= \int\int xP(X=x,\ Y=y)dxdy + \int yP(X=x,\ Y=y)dxdy \\
&= \int x \ \int P(X=x,\ Y=y)dy\ dx + \int y \ \int P(X=x,\ Y=y)dx\ dy \\
&= \int xP(X=x)dx + \int yP(Y=y)dy \\
&= E(X)+E(Y)
\end{aligned}
$$

Interestingly, If X and Y are independent Gaussian random variables then the above result holds for variances too.

$$
\begin{gathered}
E(X+Y) = E(X)+E(Y) \\
Var(X+Y) = Var(X)+Var(Y)
\end{gathered}
$$

2 Motivation

Suppose there are N people speaking at a party and there are N microphones placed at different locations recording sound. Each microphone is recording a simultaneous mix of all speakers' sounds depending on the relative position of the microphone from the speakers. Looking at the microphone recordings alone, is it possible to decipher what each speaker said?

3 Problem Statement

Denote each speaker's sound as a random variable S_i (S stands for signal). A square mixing matrix A is assumed to be responsible for linearly transforming N signals into another set of N random variables observed one each by a microphone.

$$
\begin{pmatrix} X_1 \\ X_2 \\ . \\ . \\ . \\ X_N \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1N} \\ a_{21} & a_{22} & ... & a_{2N} \\ . & . & ... & . \\ . & . & ... & . \\ . & . & ... & . \\ a_{N1} & a_{N2} & ... & a_{NN} \end{pmatrix} \begin{pmatrix} S_1 \\ S_2 \\ . \\ . \\ . \\ S_N \end{pmatrix}
$$

$$
X = AS
$$

If the jth column of the matrix A is denoted by a_j, then the equation becomes

$$X = \sum_j a_j S_j$$

A set of observations corresponding to X are given and the goal is to recover S.

$$\begin{aligned} X &= AS \\ S &= A^{-1}X \\ S &= WX \end{aligned}$$

The goal is to find this un-mixing matrix W.

4 Assumptions

- The signals to be recovered must be mutually independent. This assumption gives basis to any ICA algorithm. Find a linear combination W of the observed random variables to yield “signal” random variables that are as independent as possible.
- The signals must be non-Gaussian. The reason for this restriction will be explained later.
- The signals have zero expectation. This is for mathematical convenience. The observed random variables can be centered to impose this restriction. How?

$$\begin{aligned} if\ E(X_i) &= 0 \ \forall\ i \\ then\ E(S_i) &= E(\sum_j w_{ij} S_j) \\ &= \sum_j w_{ij} E(S_j) \\ &= 0 \ \forall\ i \end{aligned}$$

5 Ambiguities

- Variances of the signals can’t be calculated.

$$\begin{aligned} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} &= \begin{pmatrix} 2 & 3 \\ 4 & 5 \end{pmatrix} \begin{pmatrix} S_1 \\ S_2 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 3 \\ 2 & 5 \end{pmatrix} \begin{pmatrix} 2S_1 \\ S_2 \end{pmatrix} \end{aligned}$$

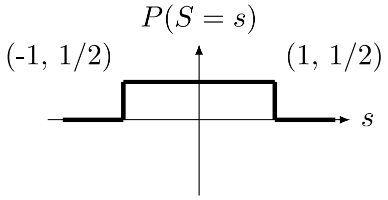
Hence the variances of all signals are assumed to be equal to unity.

- The order of signals can't be recovered.

$$\begin{aligned}\begin{pmatrix}X_1\\X_2\end{pmatrix} &= \begin{pmatrix}2 & 3\\4 & 5\end{pmatrix}\begin{pmatrix}S_1\\S_2\end{pmatrix}\\ &= \begin{pmatrix}3 & 2\\5 & 4\end{pmatrix}\begin{pmatrix}S_2\\S_1\end{pmatrix}\end{aligned}$$

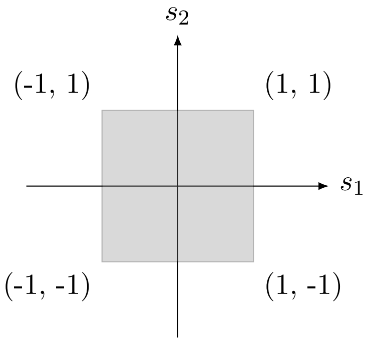
6 Example

Let S_1 and S_2 be two independent signals with a uniform density of 1/2 from -1 to 1. The PDF of each signal looks like



Their joint distribution takes the the following form.

$$\begin{aligned}P(S_1 = s_1,\ S_2 = s_2) &= \frac{1}{4} \quad if\ |s_1| <= 1\ and\ |s_2| <= 1\\ &= 0 \quad otherwise\end{aligned}$$

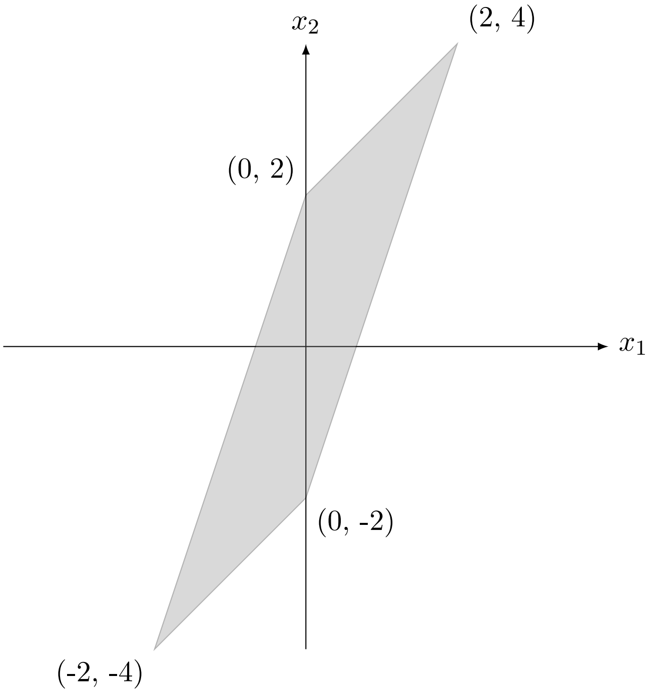


Consider the following linear transformation of these signals.

$$\begin{pmatrix}X_1\\X_2\end{pmatrix} = \begin{pmatrix}1 & 1\\3 & 1\end{pmatrix}\begin{pmatrix}S_1\\S_2\end{pmatrix}$$

The joint density of the observed signals becomes

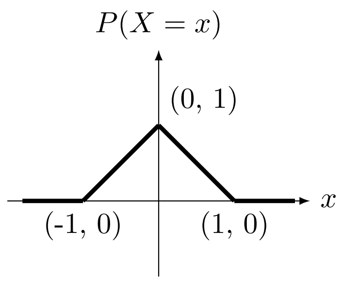
$$
\begin{aligned}
P(X_1 = x_1,\ X_2 = x_2) &= \frac{1}{2} \times \frac{1}{4} = \frac{1}{\begin{vmatrix} 1 & 1 \\ 3 & 1 \end{vmatrix}} \times \frac{1}{4} \quad && \textit{if } (x_1, x_2) \textit{ lie in parallelogram below} \\
&= 0 && \textit{otherwise}
\end{aligned}
$$



One edge of the parallelogram points in the direction $(1, 1)$ and the other in $(1, 3)$. These are the same as the columns in the mixing matrix. This gives some hope that ICA could exploit some properties of the joint distribution to recover the original signals like in this case where the edges of the joint density give the mixing matrix away.

7 Density Under Linear Transformation

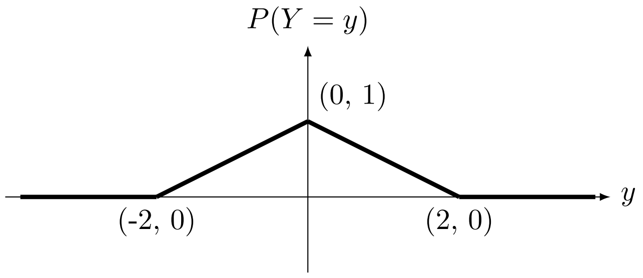
Consider a random vector X with the following PDF.



Consider another random vector $Y = 2X$. How is the PDF of Y related to that of X? One can proceed in the following manner.

$$\begin{aligned} P(Y = y) &= P(2X = y) \\ &= P(X = \frac{y}{2})) \end{aligned}$$

The calculated PDF is drawn below.



On closer inspection, the calculated PDF is wrong as the area under it is 2. Thus a division by 2 is necessary. Thus,

$$\begin{aligned} Y &= \alpha X \\ X &= \frac{1}{\alpha} Y \\ P(Y = y) &= \frac{1}{\alpha} P(X = \frac{y}{\alpha}) \\ P(X = x) &= \alpha P(Y = \alpha x) \end{aligned}$$

More generally,

$$Y = \alpha X$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ ... \\ Y_N \end{pmatrix} = \begin{pmatrix} \alpha_1^T \\ \alpha_2^T \\ ... \\ \alpha_N^T \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ ... \\ X_N \end{pmatrix}$$

$$P(X_1 = x_1, X_2 = x_2, ..., X_N = x_N) = \begin{vmatrix} \alpha_1^T \\ \alpha_2^T \\ ... \\ \alpha_N^T \end{vmatrix} P(Y_1 = \alpha_1^T X, Y_2 = \alpha_2^T X, ..., Y_N = \alpha_N^T X)$$

$$P(X_1 = x_1, X_2 = x_2, ..., X_N = x_N) = |\alpha| P(Y_1 = \alpha_1^T X, Y_2 = \alpha_2^T X, ..., Y_N = \alpha_N^T X)$$

8 Relation to PCA

Given a centered random vector X, PCA defines a linear transform $Y = PX$ such that the resulting vector has uncorrelated components. P is a matrix of orthonormal eigenvectors of the covariance matrix of X.

$$E(XX^T) = P^T DP$$

The covariance matrix of the transformed random vector is diagonalized.

$$\begin{aligned} E(YY^T) &= E(PXX^T P^T) \\ &= PE(XX^T)P^T \\ &= PP^T DPP^T \\ &= D \end{aligned}$$

One can go a step further to define $Y = D^{-1/2}PX$. Now the covariance matrix of the transformed random vector becomes identity.

$$\begin{aligned} E(YY^T) &= D^{-1/2}PE(XX^T)P^T D^{-1/2T} \\ &= D^{-1/2}PP^T DPP^T D^{-1/2T} \\ &= I \end{aligned}$$

Such a transformation of X is called whitening. Whitening is often a preprocessing step of ICA. Intuitively, whitening makes the transformed components uncorrelated and then ICA will make them independent. So whitening solves half the problem of ICA. More formally, whitening restricts the space of mixing matrix to orthogonal matrices thus making the problem easier. To understand why, consider $Y = AS$ such that Y is white.

$$\begin{aligned} E(YY^T) &= I \\ AE(SS^T)A^T &= I \end{aligned}$$

Since signals are independent with unit variances as per an assumption above,

$$\begin{aligned} E(SS^T) &= I \\ AA^T &= I \end{aligned}$$

proving that A must be orthogonal. Hence, in case of a random vector of size N, instead of estimating N^2 parameters of the matrix A, ICA only has to estimate $N(N-1)/2$ parameters since A is orthogonal and an orthogonal matrix has $N(N-1)/2$ degrees of freedom. This is equal to estimating almost half the parameters. Thus, literally too, whitening solves half the problem of ICA.

9 Why Gaussian signals can't be recovered?

Consider a Gaussian random variable,

$$P(X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Consider another random variable $Y = \alpha X$. What does the PDF of Y look like?

$$\begin{aligned} P(Y = y) &= \frac{1}{\alpha} P(X = \frac{y}{\alpha}) \\ P(Y = y) &= \frac{1}{\alpha\sqrt{2\pi\sigma^2}} e^{\frac{-(\frac{y}{\alpha}-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi\alpha^2\sigma^2}} e^{\frac{-(y-\alpha\mu)^2}{2\alpha^2\sigma^2}} \end{aligned}$$

It is clear that Y is also Gaussian with mean $\alpha\mu$ and variance $\alpha^2\sigma^2$. Another well known result with respect to Gaussians is that sum of two

Gaussians is also a Gaussian with mean as sum of individual means and variance as sum of individual variances.

$$\mathcal{N}(\mu_1, \sigma_1^2) + \mathcal{N}(\mu_2, \sigma_2^2) \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

These properties of Gaussian signals make them unrecoverable. Consider two Gaussian signals of zero mean and unit variance as per ICA assumptions and an orthogonal mixing matrix. An orthogonal matrix is considered because it is assumed that whitening has already been done and thus the search space of ICA is only orthogonal matrices.

$$\begin{aligned}
&\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \mathcal{N}(0,1) \\ \mathcal{N}(0,1) \end{pmatrix} \\
&\qquad = \begin{pmatrix} \mathcal{N}(0,a^2) + \mathcal{N}(0,b^2) \\ \mathcal{N}(0,c^2) + \mathcal{N}(0,d^2) \end{pmatrix} \\
&\qquad = \begin{pmatrix} \mathcal{N}(0,a^2 + b^2) \\ \mathcal{N}(0,c^2 + d^2) \end{pmatrix}
\end{aligned}$$

Due to orthogonality of the mixing matrix $a^2 + b^2 = c^2 + d^2 = 1$.

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \mathcal{N}(0,1) \\ \mathcal{N}(0,1) \end{pmatrix} = \begin{pmatrix} \mathcal{N}(0,1) \\ \mathcal{N}(0,1) \end{pmatrix}$$

So, upon the transformation due to any orthogonal matrix the signals remain unchanged. Hence, in case of Gaussian signals only whitening is possible.

10 Algorithm

T training samples are given of the form,

$$\begin{pmatrix} x_1^1 & x_1^2 & ... & x_1^T \\ x_2^1 & x_2^2 & ... & x_2^T \\ ... \\ x_N^1 & x_N^2 & ... & x_N^T \end{pmatrix}$$

The log-likelihood can be written as,

$$L = log \left(\prod_i P(X_1 = x_1^i, X_2 = x_2^i, ..., X_N = x_N^i) \right)$$

Rewriting ICA equation,

$$S = WX$$
$$\begin{pmatrix} S_1 \\ S_2 \\ ... \\ S_N \end{pmatrix} = \begin{pmatrix} w_1^T \\ w_2^T \\ ... \\ w_N^T \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ ... \\ X_N \end{pmatrix}$$

Using the density transformation rule,

$$L(W) = log\left(\prod_i \left(P(S_1 = w_1^T X^i, S_2 = w_2^T X^i, ..., S_N = w_N^T X^i)|W|\right)\right)$$

Since the signals are independent,

$$L(W) = log\left(\prod_i \left(\prod_j P(S_j = w_j^T X^i)|W|\right)\right)$$
$$L(W) = \sum_i \left(\sum_j log(P(S_j = w_j^T X^i)) + log(|W|)\right)$$

$P(S_j = w_j^T X^i)$ is estimated by the sigmoid function.

$$P(S_j = w_j^T X^i) = \frac{e^{-w_j^T X^i}}{1 + e^{-w_j^T X^i}}$$

The log-likelihood equation becomes,

$$L(W) = \sum_i \left(\sum_j log\left(\frac{e^{-w_j^T X^i}}{1 + e^{-w_j^T X^i}}\right) + log(|W|)\right)$$

To maximize $L(W)$ gradient descent can be applied. At convergence, the optimal W can be multiplied to X to recover sought S.