Independent Components Analysis

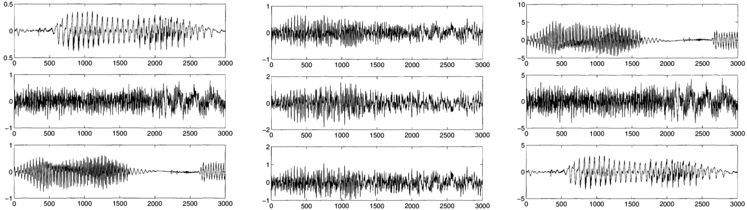
Ramesh Srinivasan

October 31, 2024

High Dimensional Data

- We introduce Principal Components Analysis as a method of dimensionality reduction.
- The core of the idea was to learn a new basis for the data points, as a rotation of the original data.
- Our motivation is that some form of dimensionality reduction such that predictors (p) < observations (n) will lead to better predictive models.
- We find it somewhat challenging to interepret more than one PCA component for a number of reasons.

Cocktail Party Problems



Left: Original Signals Middle: Mixture recorded with 3 microphones at different positions in the room. Right: ICA reconstructed signals

The ICA model

$$x_1(t) = a_{11}s_1(t) + a_{12}s_2(t) + a_{13}s_3(t)$$
$$x_2(t) = a_{21}s_1(t) + a_{22}s_2(t) + a_{23}s_3(t)$$
$$x_3(t) = a_{31}s_1(t) + a_{32}s_2(t) + a_{33}s_3(t)$$

$$\boldsymbol{x} = \boldsymbol{As}$$

- in the ICA model, we assume that each mixture x_j as well as each independent component s_k is a **random variable**, instead of a proper time signal.
- The above point is an important one to understand that ICA (and for that matter PCA) **are not time series analysis methods**
- The above model seems no different than how we think about PCA, with s being the scores and $\boldsymbol{A} = \boldsymbol{V}^T$ where V are the components (eigenvectors).
- In PCA, s are uncorrelated, while ICA makes a stronger claim that s are statistically independent.
- Its worth noting that ICA is only one of many Blind Source Separation (BSS) method. There are approaches particularly tuned to capture oscillatory time series (SOBI) that may produce better data models in specific applications.

What is independence

- Consider the case of two scalar random variables with joint probability distribution $p(y_1, y_2)$
- The marginal pdf of y_1 is

$$p_1(y_1) = \int p(y_1, y_2) dy_2$$

- Then, y_1 and y_2 are statistically independent if and only if

$$p(y_1, y_2) = p_1(y_1) p_2(y_2)$$

- For n variables, this would be the product of n terms.
- Uncorrelated does not imply independent. But independent can imply uncorrelated.
- Many ICA estimation methods will take advantage of this by first transforming the data to be uncorrelated.

Whitening

- Random variables are uncorrelated if their covariance matrix $\boldsymbol{\Sigma}$ is diagonal
- Whiteness is a stricter definition, requiring $\boldsymbol{\Sigma} = \boldsymbol{I}$.
- A whitening transformation of the data x can be calculated given the eigenvectors $\boldsymbol{V}$ and a diagonal matrix of eigenvalues $\boldsymbol{D}$ as

$$\boldsymbol{z} = \boldsymbol{V}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{V}^T\boldsymbol{x}$$

- This matrix representation is useful, because we will usually estimate ICA on $\boldsymbol{z}$. In principle you could also get this by standardizing your variables x and applying PCA.
- If you believe your underlying signals are also Gaussian, whitening is as far as you can go, as the higher order moments vanish.

Mixing and Unmixing

- The intuition of ICA is that nongaussianity is a proxy for independence.
- Loosely speaking, the sum of independent random variables $\boldsymbol{x}$ usually has a distribution that is closer to Gaussian than the original variables $\boldsymbol{s}$.

$$\boldsymbol{s} = \boldsymbol{A}\boldsymbol{s}$$

- The problem of ICA is to estimate

$$\boldsymbol{s} = \boldsymbol{A}^{-1}\boldsymbol{x}$$

- In ICA jargon $\boldsymbol{A}$ is called the **mixing** matrix and its inverse is called the **unmixing** matrix. (This is true for PCA as well, except the inverse is simply the transpose).
- Our objective is to estimate the mixing matrix $\boldsymbol{A}$ and indepedent components $\boldsymbol{s}$ given the data $\boldsymbol{x}$

Nongaussianity is independence

- Assume all the $s_i \in \boldsymbol{s}$ are not Gaussian. If they were Gaussian, we only need PCA. For the sake of this argument, assume they are identically distributed.
- Any one component s_i must be a linear combination of the $\boldsymbol{x}$. Lets estimate it with the proxy variable $\hat{y}$ which is a mixture of $\boldsymbol{x}$ with weights $\boldsymbol{b}$

$$\hat{y} = \boldsymbol{b}^T \boldsymbol{x} = \sum_i b_i x_i$$

- We can substitute the ICA model into the above to observe

$$\hat{y} = \boldsymbol{b}^T \boldsymbol{A} \boldsymbol{s} = \boldsymbol{q}^T \boldsymbol{s} = \sum_i q_i s_i$$

- Ideally $\boldsymbol{b}^T$ is one of the rows of the inverse of $\boldsymbol{A}$. In this case, we have exactly estimated the independent component and only 1 of the q_i is 1 and the rest are zero.
- The intuition here is that since $\boldsymbol{q}^T \boldsymbol{s}$ is a **mixture** of $\boldsymbol{s}$ it will be more Gaussian, and will be least Gaussian if only 1 of the q_i is 1 and the rest are zero.
- Thus a (potential) method to estimate the independent components is to maximize the nongaussianity of $\boldsymbol{b}^T \boldsymbol{x}$

Kurtosis

- A direct way to think about nongaussianity in kurtosis (or the 4th moment). If we assume, unit variance

$$K(x) = E(x^4) - 3$$

- Conveniently, kurtosis is linear over independent variables. i.e., if x_1 and x_2 are independent

$$K(x_1 + x_2) = K(x_1) + K(x_2)$$

$$K(\alpha x_1) = \alpha^4 K(x_1)$$

- In the case of two variables

$$\hat{y} = \boldsymbol{b}^T \boldsymbol{As} = \boldsymbol{q}^T \boldsymbol{s} = q_1 s_1 + q_2 s_2$$

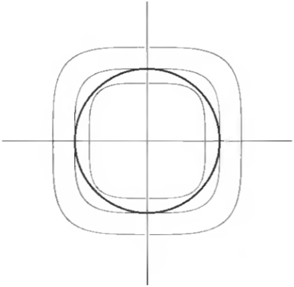
$$K(\hat{y}) = K(q_1 s_1 + q_2 s_2) = q_1^4 K(s_1) + q_2^4 K(s_2)$$

Since the variance of each $\boldsymbol{s}$ is 1,

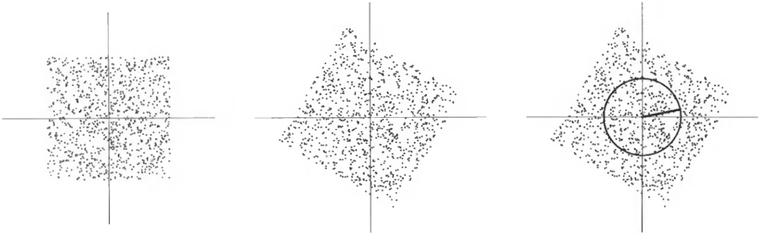
$$E[y^2] = q_1^2 + q_2^2 = 1$$

Optimization Problem for Kurtosis

- maximize
$$|q_1^4 K(s_1) + q_2^4 K(s_2)|$$
with
$$q_1^2 + q_2^2 = 1$$
- if the data is whitened, we can write this in matrix form as find $\boldsymbol{w}$ that maximizes $K(\boldsymbol{w}^T \boldsymbol{z})$ with $\|w\| = 1$



ICA from Kurtosis



Left: Samples from Two Independent Uniform Distributions Middle: Mixture that has been whitened. Right: Optimal ***w*** points to the corners.

Negentropy as nongaussianity

- The entropy of a random variable is related to the information that the observation of the variable gives. The more unpredictable or unstructured a variable is, the larger its entropy.
- For discrete random variables, Entropy H is

$$H(Y) = -\sum_i P(Y = a)i)log(P(Y = a_i)$$

- for a continuous random variable

$$H(y) = -\int f(y)log(f(y))dy$$

- A Gaussian variable has the largest entropy of all random variables of equal variance.
- Entropy is smaller for any other distributions; the more concentrated a distribution the smaller the entropy
- Negentropy $J(y)$ is defined as

$$J(y) = H(y_gauss) - H(y)$$

where $H(y_gauss)$ is the entropy of a Gaussian random variable with the same variance as the observed data.

- This is the optimal measure of nongaussianity but difficult to estimate.

Frame Title

-
- The classical way to approximate negentropy is higher order cumulants like the kurtosis.

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

- for symmetric distributions the first term vanishes and we are left with kurtosis.
- We can generalize the higher-order cumulant approximation by using any non-quadratic functions.

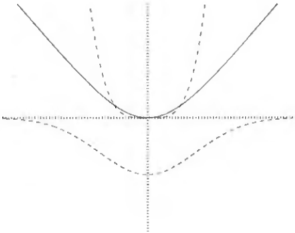
$$J(y) \propto (E[G(y)] - E[G[\nu]])$$

where ν is a Gaussian random variable with unit variance.

Widely used functions include

$$G(y) = \frac{1}{a_1} log(cosh(a_1 y)$$

$$G(y) == exp(-\frac{y^2}{2})$$



FastICA

- Fast ICA uses a Gradient algorithm to find a fast solution. Most of the time it works well. Other times, you might resort to the other methods discussed in the paper.
- Center the data
- Whiten the data
- Choose an initial (random) vector $\boldsymbol{w}$ of unit norm
- update. Here G' and G'' are the first and second derivatives of G

$$\boldsymbol{w} \leftarrow E[\boldsymbol{z}G'(\boldsymbol{w}^T\boldsymbol{z})] - E(G''(\boldsymbol{w}^T\boldsymbol{z})]\boldsymbol{w}$$

- normalize

$$w \leftarrow \frac{\boldsymbol{w}}{||\boldsymbol{w}||}$$

- iterate to convergence

Tough questions

- which G to use.
- How to estimate more than one $\boldsymbol{w}$. Most common approach is to deflate $\boldsymbol{z}$ by removing the component w_i