Independent Component Analysis

Kaiwen Zhou

Contents

Readings

In addition to the lecture notes, the following are required readings:

- Chapter 2.7, 5.3-5.5, 7, 8.3, 8.4 hyvarinenetal2001ica

1 Contextualization: Blind Source Separation Problem

Suppose we have a set of p independent signals $s_{i,t}, i = 1 \dots p$ which we wish to infer, from noise-free observation of p - unknown set of mixtures $x_{i,t}, i = 1 \dots p$ of all the signals:

$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t, \quad \mathbf{A} \in \mathbb{R}^{p \times p}$$

Is it possible to infer $\mathbf{A}$ and separate each signal $s_{i,t}$ from the rest?

We call this the blind source separation problem, because it arises in a situation where we try to disentangle the sound of each of p-musicians in an orchestra by recording their symphony with $p-$ distinct microphones by blindly positioning the microphones in different parts of the symphony hall.

Does PCA Solve BSS?

Because the sources $s_{i,t}$ are independent, they are therefore uncorrelated. Without loss of generality we can also assume that each signal is demeaned and has unit variance:

$$\mathbb{E}[s_i] = 0, \quad \mathbb{E}[s_i s_j] = \delta_{ij}$$

In this case, it is tempting to say that (1) is simply a linear factor model such as the one analyzed in the Probabilistic PCA framework so solution is

$$\mathbf{A}^* = \mathbf{W}^* = \mathbf{V}\mathbf{\Sigma} \text{ - the PCA components of } \mathbf{X}$$
$$\mathbf{S}^* = \mathbf{Z}^* = \mathbf{U} \text{ - the left SVD components of } \mathbf{X}$$

PCA is not unique I

However, as we discussed in class that the PCA solution (3) to the latent factor model is not unique!

Indeed, suppose $\mathbf{R}$ is an arbitrary $p \times p$ unitary (rotation) matrix, $\mathbf{R}'\mathbf{R} = \mathbf{I}_p$. Then the following would also be a solution to the latent factor problem:

$$\tilde{\mathbf{W}}' = \mathbf{R}\mathbf{W}^{*\prime} = \mathbf{R}\mathbf{\Sigma}\mathbf{V}'$$
$$\tilde{\mathbf{Z}} = \mathbf{Z}^*\mathbf{R}' = \mathbf{U}\mathbf{R}'$$

as we clearly have $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}' = \mathbf{Z}^*\mathbf{W}^{*\prime} = \tilde{\mathbf{Z}}\tilde{\mathbf{W}}'$.

PCA is not unique II

Since PCA only looks at solutions to the latent factor problem which have zero mean and unit covariance, any rotation of such solution would also have zero mean and covariance.

We need a further constraint in order to solve the problem uniquely. Is there an objective way to do it in the BSS problem?

PCA is not unique III

What are the implications of this lack of uniqueness to the use case of risk modeling and/or portfolio allocation via PCA.

Is prior knowledge the only way one can separate independent signals such as is the case with paying attention to only one instrument in a symphony?

Independence and covariance I

Independence is a much stronger condition than zero cross-covariance:

Definition

Two random variables s_1 and s_2 are independent if their joint cdf (and equivalently pdf) density factorizes into the marginals of each random variable, $F(s_1, s_2) = F(s_1) F(s_2)$ (and equivalently $p(s_1, s_2) = p(s_1) p(s_2)$)

Independence and covariance II

It is clear that due to the factorization of the joint density that for two independent variables $\mathbb{E}[s_1 s_2] = \mathbb{E}[s_1] \mathbb{E}[s_2]$ and therefore independence implies zero cross-covariance $\operatorname{cov}(s_1, s_2) = 0$.

Furthermore, it is also clear that $\mathbb{E}[f(s_1) g(s_2)] = \mathbb{E}[f(s_1)] \mathbb{E}[g(s_2)]$ for any two functions f and g. If we take $f(x') = g(x') = I_{x'>x}$, we see this is equivalent to the definition of independence so we have the following trivial

Proposition

Two random variables s_1 and s_2 are independent if and only if for any transformations $f(x), g(x)$, the random variables $f(s_1)$ and $g(s_2)$ have zero covariance.

Gaussian variables and independence

Proposition

If s_1 and s_2 are Gussian then s_1 and s_2 are independent iff $\operatorname{cov}(s_1, s_2) = 0$.

As a consequence, if the underlying signal s in the BSS problem is a multivariate Gaussian, then any of the infinitely many PCA solutions is a valid BSS solution.

Whitening

Staring with any observed signal $\mathbf{X}$ it is convenient to "whiten" it by mapping it to its PCA latent factors:

$$\mathbf{X} \rightarrow_{\text{whiten}} (\mathbf{X} - \boldsymbol{\mu}) \mathbf{W}_{PCA} \left(\mathbf{W}'_{PCA} \mathbf{W}_{PCA} \right)^{-1}$$

The transformed signal will therefore have the same first two moments as if it were a mean-zero independent multivariate Gaussian signal i.e.

$$\mathbb{E}[\mathbf{X}_{\text{whitened}}] = 0, \quad \operatorname{cov}(\mathbf{X}_{\text{whitened}}) = I_p$$

BSS on whitened data

Let's assume that $\mathbf{X}$ is whitened as in (7) from now on and drop the subscripts for simplicity.

Then, the BSS problem reduces to finding the p-dimensional rotation matrix $\mathbf{R}, \mathbf{R}'\mathbf{R} = \mathbf{I}_p$ in the linear model:

$$\mathbf{x}_t = \mathbf{R}\mathbf{s}_t \Longleftrightarrow \mathbf{X} = \mathbf{S}\mathbf{R}$$

such that the latent $\mathrm{s}_t = \{s_{i,t}\}$ are "maximally" independent.

What do we mean by "maximally" independent? We follow hyvarinenetal2001ica and present several computationally feasible quantifications of independence.

Cumulants and independence I

Given that independence and zero covariance are equivalent for Gaussian signals, in which case the the BSS solutions are the non-unique PCA solutions, our only hope to obtain a unique BSS solution is to consider the case where the undelying signals $s_{i,t}$ are non-Gaussian.

Since Gaussian distribution is the only distribution with zero moments beyond the second, we will look for independence criteria linked to higher moments.

Cumulants and independence II

Definition

The moments m_i of a univariate random variable X are the Taylor coefficients of the Moment generating function:

$$M_X(t) = \mathbb{E}\left[e^{itX}\right] = \sum_{k=0}^{\infty} \mathbb{E}\left[X^k\right] \frac{(it)^k}{k!} = \sum_{k=0}^{\infty} m_k \frac{(it)^k}{k!}$$

Definition

The cumulants κ_i of X are the Taylor coefficients of the logarithm of the Moment generating function:

$$K_X(t) = \ln M_X(t) = \sum_{k=0}^{\infty} \kappa_k \frac{(it)^k}{k!}$$

Cumulants and independence III

Proposition

For a zero-mean r.v.:

$$\kappa_1 = m_1 = \mathbb{E}[X], \kappa_2 = m_2 = \mathbb{E}\left[X^2\right], \kappa_3 = m_3 = E\left[X^3\right]$$
$$\kappa_4 = m_4 - 3m_2 = E\left[X^4\right] - 3E\left[X^2\right]$$

Note that for a Gaussian r.v. $\kappa_n = 0, n \geq 3$

Cumulants and independence IV

Proposition (Cumulant additivity under independence)

If X and Y are independent r.v.s then

$$K_{X+Y}(t) = \ln \mathbb{E}\left[e^{it(X+Y)}\right] = \ln \mathbb{E}\left[e^{itX} e^{itY}\right] =$$
$$\ln \mathbb{E}\left[e^{itX}\right] + \ln \mathbb{E}\left[e^{itY}\right] = K_X(t) + K_Y(t)$$

Proposition (Homogeneity under scaling)

The ℓ-th cumulant is homogeneous of order ℓ :

$$\kappa_\ell(cX) = c^\ell \kappa_\ell(X) \text{ and therefore}$$
$$\kappa_\ell(aX + bY) = a^\ell \kappa_\ell(X) + b^\ell \kappa_\ell(Y)$$

Cumulants and independence V

Corollary

$$|\kappa_\ell(aX + bY)| \leq \left|a^\ell\right| \cdot |\kappa_\ell(X)| + \left|b^\ell\right| \cdot |\kappa_\ell(Y)|$$

where if $a^2 + b^2 = 1$ and $\ell > 2$ equality is only achieved if either $a^2 = 0$ or $b^2 = 0$.

Therefore a mixture of two independent r.v.'s with finite cumulants has smaller than or equal abs-cumulant than the abs-cumulant of each variable.

The above corollary easily generalizes to higher dimensions

Connection to Central Limit Theorem

The cumulant inequality above is somewhat related to the Central Limit Theorem as it shows that the abs-cumulant of a sum of r.v.'s would be smaller than the abs-cumulants of each of the individual r.v.'s so the sum would look more Gaussian.

This statement is true even for r.v.'s unlike the statement of CLT which is valid for $N \to \infty$ assets. However an implicit assumption in the result above is that the cumulants of the given order $\ell \geq 3$ are finite, which is a stronger than the finite second moment assumption of CLT.

ICA by Cumulant Optimization I

Basic idea

1. Whiten the data, i.e. assume $\mathbf{X}$ is demeaned and $\mathbf{X}'\mathbf{X} = \mathbf{I}$. As a result the mixing matrix $\mathbf{A}$ which we seek to recover is restricted to be an orthogonal rotation matrix with $\mathbf{A}'\mathbf{A} = \mathbf{I}$ and the separated signal is $\mathbf{S} = \mathbf{X}\mathbf{A}$
2. Find $\mathbf{b}_1^* = \min_{\mathbf{b},|\mathbf{b}|^2=1} |\kappa_\ell(\mathbf{x} \cdot \mathbf{b})|$ for any fixed $\ell \geq 3$.

Typically the method uses $\ell = 3, 4$ corresponding to skewnessand kurtosis-ICA.

- by the cumulant inequality property, $\mathbf{b}_1^* = \pm\delta_{i_1,i}$ where i_1 is the index of the signal with maximal abs-cumulant
- therefore $\mathbf{x} \cdot \mathbf{b}_1^* = \pm s_{i_1}$ is the independent signal with the maximum kurtosis up to a sign!

3. Let $\mathbf{X} \to \mathbf{X}\left(\mathbf{I} - \mathbf{b}_1^*\mathbf{b}_1^{*\prime}\right)$, i.e. project out the contribution of s_{i_1}. Repeat steps 2-3 $p-1$ times

ICA by Cumulant Optimization II

Note that this algorithm is similar to the Gram-Schmidt procedure of finding the principal components of the data (i.e. eigenvectors of the second cumulant which is the covariance matrix of the data).

Note also that the algorithm implicitly assumes that the underlying signal is indeed a linear combination of independent non-Gaussian signals. This is not always the case (in fact for most signals it isn't). Therefore it is not guaranteed that the solution of this optimization procedure will produce independent factors.

Gaussian distribution and maximal entropy

Definition (Differential Entropy)

The differential entropy of a r.v. X with $pdf p(x)$ is defined as $H[X] = -\mathbb{E}[\ln p(X)]$.

Proposition

Of all possible distributions of a given r.v. X, with a fixed mean and covariance, the Gaussian distribution has the largest entropy.

Intuitively, this result is due to the fact that the Gaussian distribution is only determined by its mean and covariance and therefore is the least informative of all distributions of fixed mean and covariance.

Negentropy and maximal non-gaussianity I

Corollary (Positivity of Negentropy)

Given a r.v. x with distribution $p(\mathrm{x})$ of fixed mean and covariance, the following quantity is always positive or zero:

$$J[\mathrm{x}] = H\left[\mathrm{x}_{\text{gauss}}\right] - H[\mathrm{x}] \geq 0$$

Here $\mathrm{x}_{\text{gauss}}$ is a r.v. with the same mean and covariance as x

Therefore maximizing non-gaussianity of $J\left[\mathbf{A}'\mathrm{x}\right]$ is another way to reveal the de-mixing transformation which makes the data maximally non-gaussian.

Negentropy and maximal non-gaussianity II

Negentropy is in general not known, because it requires knowledge of the distribution $p(\mathrm{x})$ so as to compute the entropy $H[\mathrm{x}]$. However, if we assume that this distribution is close to a Gaussian, we can use the Gram-Challier expansion and write it as (see hyvarinenetal2001ica Ch 5.52):

$$J(x) \sim \frac{1}{12}\kappa_3(x)^2 + \frac{1}{48}\kappa_4(x)^2$$

so maximization would boil down again to maximizing abs-cumulants as before.

References

[1] Rasmussen, et al. (2006). Gaussian Processes In Machine Learning.

[2] Sheffield, S. (2007). "Gaussian Free Fields for Mathematicians". In: 139 (3-4), pp. 521-541.