Independent component analysis

A new concept? Comon (1994)

Geelon So, agso@eng.ucsd.edu
Unsupervised learning reading group — February 9, 2022

Blind identification: statistical background

Cocktail party problem

Setting: imagine a party

- there are d tables
  - there are independent conversations going on at each table
- there are $n \geq d$ microphones
  - each microphone picks up a mixture of the conversations

Question: can we unmix the recording to recover the independent conversations?

- This is called the *source separation problem.*

Formal problem setting

Consider the **linear statistical model**:

$$y = Mx$$

- $x \in \mathbb{R}^d$ is drawn from p_x, which has statistically independent components
- $M \in \mathbb{R}^{n \times d}$ is a full column rank matrix and only $y \in \mathbb{R}^n$ is observed

Statistically independent components

We say that a density p_x on $\mathbb{R}^d$ has **statistically independent components** if:

$$p_x(x) = \prod_{i=1}^{d} p_{x_i}(x_i),$$

where $x = (x_1, \ldots, x_d)$ and p_{x_i} is a density on $\mathbb{R}$.

Full column rank

$$y = Mx$$

Our assumption that $M \in \mathbb{R}^{n \times d}$ has full column rank means that $n \geq d$.

- In the cocktail party problem, this means that unmixing the audio is possible.

Blind identification

Question: can we recover M from seeing independent realizations of y from our model:

$$y = Mx$$

where $x \sim p_x$ has independent components and M is full column rank?

Inherent indeterminations

Given $x \sim p_x$ with independent components, define the following:

- $P \in \mathbb{R}^{d \times d}$ a permutation
- $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ is an invertible axis-aligned scaling

Non-identifiability: from $y \sim p_y$ alone, we cannot distinguish between:

$$y = Mx \qquad \text{and} \qquad y = Nz$$

- $N = M\Lambda^{-1}P^\top$ has full column rank
- $z = P\Lambda x$ has independent components
  - Cocktail party problem: z is a re-numbering of the tables through P, and an adjustment of the individual volumes of each conversation through Λ.
  - However, for non-Gaussian randomness, these are the only issues with identifiability!

Statistical independence and orthogonal transformations

Informal theorem: a rotation of the space cannot preserve non-Gaussian independent randomness unless it is a permutation or reflection.

Theorem

Let $x \in \mathbb{R}^d$ have independent components that are not Gaussian nor deterministic (i.e. point masses). Let $U \in \mathbb{R}^{d \times d}$ be orthogonal and let z be:

$$z = Ux.$$

The following are equivalent:

(i) *The components of z_i are pairwise independent.*

(ii) *The components of z_i are mutually independent.*

(iii) $U = P\Lambda$, *P permutation, Λ diagonal.*

Implication: identifiability

Let $\mathcal{P}$ be a family of distributions p_x over $\mathbb{R}^d$ satisfying:

- p_x has independent components
- p_x has covariance $I_{d \times d}$

Corollary (Identifiability)

Let $M \in \mathbb{R}^{n \times d}$ be full column rank. Let $p_x \in \mathcal{P}$, with components that are not Gaussian nor deterministic. If there exists $N \in \mathbb{R}^{n \times d}$ such that:

$$Mx = Nz$$

where $z \sim p_z$ and $p_z \in \mathcal{P}$. Then for P permutation and Λ diagonal:

$$M = NP\Lambda.$$

Proof of corollary

Proof.

- M has full column rank, so it has inverse A.
- Since $Mx = Nz$, we must have $x = ANz$.
- Both x and z have covariance $I_{d\times d}$, so $U = (AN)^\top$ is orthogonal and:

$$z = Ux.$$

- Both x and z have independent components, so by Theorem, $U = P\Lambda$, implying:

$$Mx = Nz = NUx = NP\Lambda x.$$

- $M = NP\Lambda$.

□

Darmois' theorem

Theorem (Darmois' theorem)

Let $x_1, \ldots, x_d$ be independent random variables. Let:

$$X_1 = \sum a_i x_i \qquad \text{and} \qquad X_2 = \sum b_i x_i.$$

Then if X_1 and X_2 are independent, then whenever $a_i b_i \neq 0$, then x_i is Gaussian.

Operationalizing blind identification

Simplifying the problem by whitening the data

Remark: without loss of generality, we may assume that the transformation M is orthogonal,

$$y = Mx$$

for otherwise, we could simply whiten the observed data using SVD and work with the preprocessed data.

- As a result, the problem becomes one of finding an orthogonal matrix U such that Uy has independent components.

Independent component analysis

Let p_y be a distribution over $y \in \mathbb{R}^n$ with covariance Σ_y.

Definition (ICA)

The ***independent component analysis*** *(ICA) of p_y is a factorization of Σ_y:*

$$\Sigma_y = A \Sigma_x A^\top$$

where (A, Σ_x) satisfies:

(a) *A has full column rank d and Σ_x is diagonal real positive*

(b) *when $M = A\Sigma_x^{1/2}$, an observation $y \sim p_y$ can be written as:*

$$y = Mx$$

where $x \sim p_x$ for some distribution p_x over $\mathbb{R}^d$ with covariance $I_{d \times d}$

(c) *the components of x are 'the most independent possible'.*

Constructing an optimization problem

Question: can we specify part (c) of ICA as an optimization problem?

- Can we measure/maximize how independent components of a random vector are?

Idea: given a distance function δ on distributions, we can check whether a distribution p on $\mathbb{R}^d$ has independent components:

$$\delta\left(p, \prod_{i=1}^{d} p_i\right).$$

- We can aim to maximize a **contrast function** $\Psi(p) := -\delta\left(p, \prod_{i=1}^{d} p_i\right)$.
  - The contrast function Ψ is maximized at zero when p is independent.

Desiderata for a contrast function

Let $\mathcal{Q} \subset \{p_x : p_x \text{ a density on } \mathbb{R}^d \text{ for random variable } x\}$.

Definition (Contrast function)

*A **contrast** is a mapping $\Psi : \mathcal{Q} \to \mathbb{R}$ if it satisfies:*

(i) *Ψ does not change if the components x_i are permuted:*

$$\Psi(p_x) = \Psi(p_{Px}), \qquad \forall P \text{ permutation}$$

(ii) *Ψ is invariant by 'scale' change,*

$$\Psi(p_x) = \Psi(p_{\Lambda x}), \qquad \forall \Lambda \text{ invertible diagonal}$$

(iii) *if x has independent components, then:*

$$\Psi(p_{Ax}) \leq \Psi(p_x), \qquad \forall A \text{ invertible}$$

Desiderata for a contrast function

Definition (Discriminating)

A contrast is said to be ***discriminating*** *over* $\mathcal{Q}$ *if equality holds:*

$$\Psi(p_{Ax}) = \Psi(p_x)$$

only when A *is of the form* $P\Lambda$ *for some permutation* P *and diagonal* Λ*, when* x *has independent components with* $p_x \in \mathcal{Q}$*.*

Identifying independent components through optimization

Recall setting: let $\mathcal{P}$ be a family of distributions over $\mathbb{R}^d$ with independent components and covariance $I_{d\times d}$. Let $M \in \mathbb{R}^{d\times d}$ be orthogonal and $p_x \in \mathcal{P}$, with non-Gaussian and non-deterministic components.

Idea: define a discriminating contrast function Ψ of the form $\Psi(p) := -\delta\left(p, \prod_{i=1}^{d} p_i\right)$

- Let $z \in \mathrm{ICA}_{a,b}$ means that z satisfies properties (a) and (b) of ICA. Then:

$$\Psi(p_x) = \max_{z\in\mathrm{ICA}_{a,b}} \Psi(p_z).$$

- If z does not have independent components, then:

$$\Psi(p_z) \leq \Psi(p_x).$$

Measuring independence

Definition (KL divergence)

*The **KL-divergence** between two distributions p and q on $\mathbb{R}^d$ is:*

$$\mathrm{KL}(p \| q) = \int p(x) \log \frac{p(x)}{q(x)} \, dx.$$

Definition (Average mutual information)

*The **average mutual information** $I(p)$ of a distribution p on $\mathbb{R}^d$ is:*

$$I(p) = \mathrm{KL}\left(p \, \middle\| \, \prod_{i=1}^{d} p_i \right).$$

Choosing the contrast function

Theorem

Let $\mathcal{Q}$ be the set of mean zero densities on $\mathbb{R}^d$ with covariance $I_{d\times d}$. Then the map $\Psi = -I$ is a contrast function over $\mathcal{Q}$. Moreover, it is discriminating over random vectors whose components are non-Gaussian.

Proof of contrast function

Proof.

Note that we can consider only orthogonal maps $U \in \mathbb{R}^{d\times d}$.

- If U is a permutation, then $\Psi(p_{Ux}) = \Psi(p_x)$ since permuting the components does not change the KL divergence.
- If U is diagonal, then $\Psi(p_{Ux}) = \Psi(p_x)$ is a reflection and does not change the KL divergence.
- If U is an arbitrary orthogonal map and x has independent components, then:

$$\Psi(p_{Ux}) \leq \Psi(p_x) = 0,$$

since KL divergence is non-negative.

□

Proof of discrimination

Proof.

Let x have independent non-Gaussian components. Recall that $\Psi(p_x) = 0$.

($\Longrightarrow$) $\Psi(p_{Ax}) = \Psi(p_x)$ implies $A = P\Lambda$ for permutation P and scaling Λ.

- If $\Psi(p_{Ax}) = \Psi(p_x) = 0$, then Ax must have independent components (KL property).
- By earlier theorem, since x has non-Gaussian components, $A = P\Lambda$.

($\Longleftarrow$) If $A = P\Lambda$, then $\Psi(p_{Ax}) = \Psi(p_x)$.

- Since Ax still has independent components, $\Psi(p_{Ax}) = 0$ (KL property).

$\square$

Estimation problem

Sample access

- We do not have access directly to the densities themselves, but to draws of data.
- We can approximate ICA by estimating the average mutual information.

Analyzing the mutual information

Definition (Differential entropy)

*The **differential entropy** of p is:*

$$S(p) = \int p(x) \log \frac{1}{p(x)} \, dx.$$

- If z has covariance I, then:

$$S(p_{Az}) = S(p_z) - \frac{1}{2} \log \det AA^\top.$$

- The density with largest entropy with matching covariance is Gaussian:

$$S(p_{Az}) \leq S(\mathcal{N}(0, AA^\top)).$$

Analyzing the mutual information

Definition (Negentropy)

Let p be a density with covariance $I_{d\times d}$. The ***negentropy*** *of p is defined:*

$$\mathcal{J}(p) = S(\mathcal{N}(0, I)) - S(p).$$

- While the differential entropy may be negative, the negentropy is always positive, is invariant by linear invertible changes of coordinates, and vanishes if and only if p is Gaussian. Thus, negentropy is a measure of distance from normality.

Analyzing the average mutual information

Fact (Expansion of average mutual information)

The mutual information may be written:

$$I(p_x) = \mathcal{J}(p_x) - \sum_{i=1}^{d} \mathcal{J}(p_{x_i}) + \frac{1}{2} \log \frac{\prod_{i=1}^{d} \mathrm{Var}(x_i)}{\det(\mathrm{Cov}(x))}$$

Approximation to the negentropy

Recall that the nth cumulant is defined by:

$$\kappa_n = K^{(n)}(0)$$

the nth derivative of the cumulant-generating function $K(t) = \log \mathbb{E}[e^{tx}]$.

Fact

Let z be mean zero with standard covariance and is a sum of m independent random variables. Then:

$$\mathcal{J}(p_z) = \frac{1}{12}\kappa_3^2 + \frac{1}{48}\kappa_4^2 + \frac{7}{48}\kappa_3^4 - \frac{1}{8}\kappa_3^2\kappa_4 + o(m^{-2}).$$

References

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.