Independent component analysis

Theory and applications

Dalila Failli   Lorenzo Pratesi Mariti

Università degli Studi di Firenze

December 17, 2020

Overview

ICA: a statistical perspective

- Independent component analysis is a method for extracting useful information from data.
- It can be seen as a refinement of principal component analysis or factor analysis.
- The independent components can be obtained by maximizing the statistical independence of the estimated components.

Why using ICA

The independent components have often an interpretable physical meaning.

ICA can be used to extract and filter mixed dataset in many real life applications:

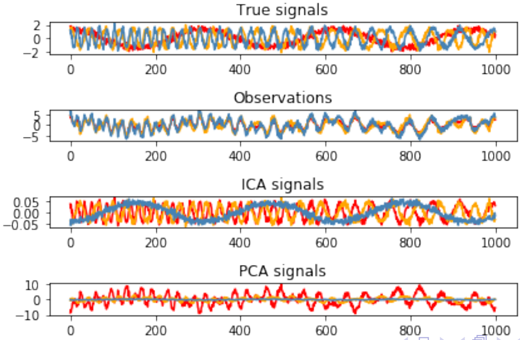
- as a blind source separation (BSS) tool
- in many other applications, such as:
  - biomedical engineering
  - medical imaging
  - speech/face recognition
  - financial time series
- as a previous step for clustering or outlier detection.

Differences from PCA and FA

- Similarities:
  - Feature extraction
  - Dimension reduction
- Differences:
  - PCA and FA are based on the analysis of the covariance matrix (second order moments of the data) to produce **uncorrelated** components.
  - ICA take advantage of higher moments (skewness and kurtosis) to generate components as **independent** as possible.

Microphone outputs

The **PCA** extracts a set of uncorrelated signals from a set of mixtures, which is a new set of voice mixtures.
The **ICA** extracts a set of independent signals, which is a set of single voices.

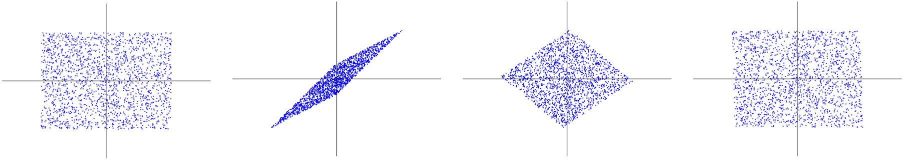


PCA vs ICA

Two mixed independent uniforms

- The independent components result from one further rotation.
- PCA does not find original coordinates, ICA does.

Original composition | Observed mixture | PCA | ICA



IC model

The main assumptions

The observed variables are a linear combinations or mixtures of latent mutually *independent non-Gaussian variables*:

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z}$$

- $\boldsymbol{A}$ (**mixing matrix**) is a *square and invertible matrix* which mixes the latent standardized random vector $\boldsymbol{z}$ (independent components) through the matrix product and reproduces our observed data $\boldsymbol{x}$.
- Both the independent components and the mixing matrix are unknown.
- The main interests are:
  - estimating and testing the unknown mixing matrix
  - inference on the number of the non-Gaussian components

IC model
The main assumptions

- The p-variate random vector $\boldsymbol{x}$ obeys the IC model if:

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z} = \boldsymbol{A}_1\boldsymbol{z}_1 + \boldsymbol{A}_2\boldsymbol{z}_2$$

  - where $\boldsymbol{A} = (\boldsymbol{A}_1, \boldsymbol{A}_2) \in \mathbb{R}^{p \times p}$ is nonsingular
- The p-variate random vector $\boldsymbol{z} = (\boldsymbol{z}_1', \boldsymbol{z}_2')'$ satisfies:
  - $\mathrm{E}[z_i] = \mathbf{0}, \mathrm{Var}(z_i) = 1, i = 1, .., p$
  - $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are independent
  - the components of $\boldsymbol{z}_1$ are independent non-Gaussian and the components of $\boldsymbol{z}_2$ are independent Gaussian.

Central limit theorem

Under very general conditions, the sum of a set of random variables has a distribution that becomes increasingly Gaussian as the number of terms in the sum increases.

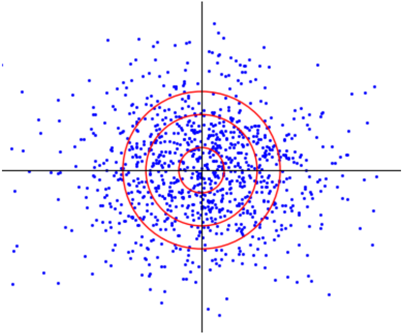
- If we have independent Gaussian components, they become more Gaussian after mixing and it will be impossible to recover the original components.

- To extract the independent components we search for an unmixing matrix that maximizes the non-Gaussianity of the components.

IC model

Issue

- Gaussian models are *not* identifiable:
  - Any orthogonal rotation is equivalent;
  - kurtosis of a Gaussian variable is equal to 0;
  - there is no way to infer the mixing matrix from the mixtures;
  - gaussian variables are forbidden in ICA.



Both $\boldsymbol{z}$ and $\boldsymbol{A}$ are unknown, so:

- We can't determine the **variances** of the independent components.
- Ambiguity of the **sign**.
- We can't determine the **order** of the independent components.

IC model
The goal

ICA tries to "un-mix" the data by estimating an **unmixing matrix** $\boldsymbol{W} \in \mathbb{R}^{q \times p}$, such that:

$$\boldsymbol{Wx} = \boldsymbol{z}_1$$

up to sign changes, location shifts, rescalings and permutations of its rows.

- $\boldsymbol{W}$ has to be found such that sequences $\boldsymbol{z}_1$ are maximally statistically independent.

Methods:

- Non-Gaussianity based ICA
  - Kurtosis based ICA
  - Neg-entropy based ICA
- MLE based ICA
- Mutual information based ICA
- Non-linear ICA
- Tensor ICA

Algorithm:

- Gradient method
- Fast fixed point algorithm

FOBI and JADE

Both methods utilize fourth order cumulants.

FOBI (Fourth-Order Blind Identification)

- One of the first ICA methods (Cardoso, 1989).
- Uses simultaneous diagonalization of the **covariance matrix** $\boldsymbol{S}_1$ and the **matrix based on the fourth moments** $\boldsymbol{S}_2$. Both are scatter matrices with the independence property.
- The unmixing matrix is the simultaneous diagonalizer $\boldsymbol{W}$ satisfying:

$$\boldsymbol{W}\boldsymbol{S}_1\boldsymbol{W}' = \boldsymbol{I}_p \text{ and } \boldsymbol{W}\boldsymbol{S}_2\boldsymbol{W}' = \boldsymbol{D}$$

  where the diagonal elements of $\boldsymbol{D}$ are the eigenvalues of $\boldsymbol{S}_2$.
- To have a unique solution the independent components must have *different kurtosis values*.

FOBI and JADE

JADE (Joint Approximate Diagonalization of Eigenmatrices)

- Developed by Cardoso & Souloumiac, 1993
- The kurtosis values do not have to be different.
- It is increased the number of matrices to be diagonalized: considers all cumulant matrices $\boldsymbol{C}^{ij}(\boldsymbol{x}^{st})$

- The JADE estimate jointly diagonalizes p^2 matrices
  - computationally much heavier than FOBI.

- k-JADE: similar, but faster method. [1]

FastICA

A gaussian variable has the largest entropy among all the random variables of equal variance.

Non-gaussianity is measured using approximations to **neg-entropy** which are:

- more robust than kurtosis-based measures
- fast to compute.

Neg-entropy approximation

$$J(y) = (E[G(y)] - E[G(v)])^2$$

where v is $N(0,1)$ random variable.

Options for G:

- $G(u) = \frac{1}{\alpha} \log \cosh(\alpha u)$
- $G(u) = -\exp(-\frac{u^2}{2})$

FastICA

The algorithm (fixed-point iteration scheme for maximizing neg-entropy)

- Step 1: the data are centered by subtracting the mean of each column of the data matrix $\boldsymbol{X}$.
- Step 2: the data matrix is 'whitened' by projecting the data onto its principal component directions.
- The algorithm estimates a matrix $\boldsymbol{W}$ in order to maximize the neg-entropy approximation under the constraints that $\boldsymbol{W}$ is an orthonormal matrix (so the estimated components are uncorrelated).

FastICA

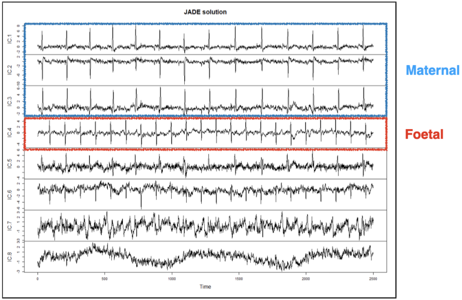
The algorithm (fixed-point iteration scheme for maximizing neg-entropy)

- Step 3: $\boldsymbol{w} \leftarrow E[\boldsymbol{z}g(\boldsymbol{w}'\boldsymbol{z})] - E[g'(\boldsymbol{w}'\boldsymbol{z})]\boldsymbol{w}$
- Step 4: $\boldsymbol{w} \leftarrow \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$
- Step 5: If $\boldsymbol{w}$ has not converged, return to step 3.
- Convergence means that the old and new values point in the same direction (their product is almost equal to 1).

Applications
ECG

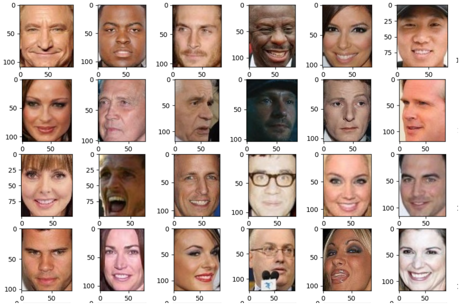
The problem is to separate the signals of maternal and foetal heartbeats. The original data show the maternal heartbeat as a dominant feature.

1. **Dataset**: "foetal_ecg.dat" [2]
2. **Packages**: "JADE" and "BSSasymp" in R
3. **Method**: JADE



We perform a *Dimensionality Reduction/Prediciton* application based on ICA model to aid gender recognition process.

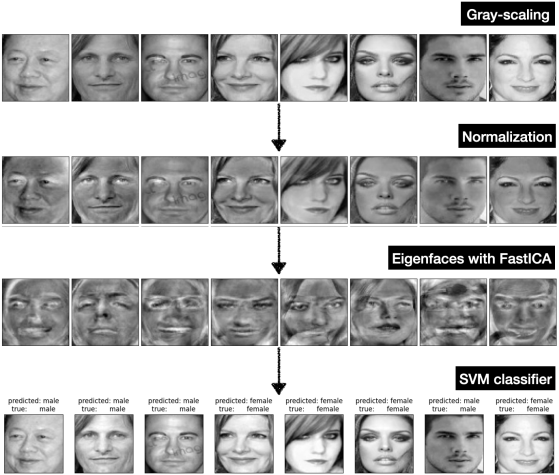
1. **Dataset:** *Gender Classification Dataset* [3]. Amount of train data: female - 23243, male - 23766.
2. **ICA**: Eigenfaces generated captures localized features like nose, eye selectors
3. **SVM Classifier**: Works on the idea of separating members of different class using decision planes.



Example of images in the dataset

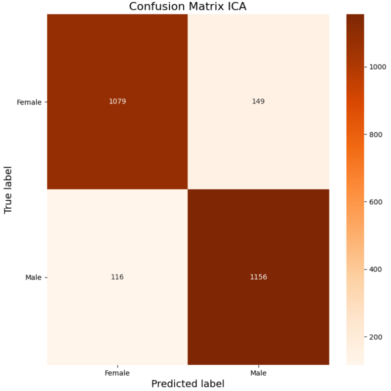
Main workflow (*FastICA* and *SVM* implementation from sklearn package)

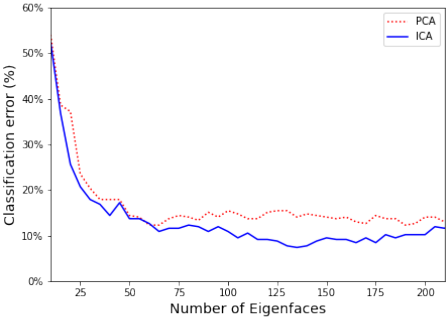
- The training dataset was split into a training and testing set;
- images was rescaled and normalized (local/global centering);
- ICs estimation with FastICA algorithm;
- project the input data on the eigenfaces orthonormal basis;
- fit the SVM classifier to the training set
- perform the final prediciton on test set and validate the results



Applications

Facial Recognition using ICA & SVM Classifier





The prediction accuracy for 150 eigenfaces from 7500 face images is **89.4%**

Bibliography I

Jari Miettinen, Klaus Nordhausen, and Sara Taskinen. "Blind Source Separation Based on Joint Diagonalization in R: The Packages JADE and BSSasymp". In: *Journal of Statistical Software* (2017).

Lieven De Lathauwer. *foetal_ecg*. ftp://ftp.esat.kuleuven.be/pub/SISTA/data/biomedical/foetal_ecg.dat.gz.

Ashutosh Chauhan. *Gender Classification Dataset*. https://www.kaggle.com/cashutosh/gender-classification-dataset.

Klaus Nordhausen and Hannu Oja. "Independent component analysis: A statistical perspective". In: *WIREs Computational Statistics* (2018).

Joni Virta et al. "Independent component analysis for multivariate functional data". In: *Journal of Multivariate Analysis* (2020).

Bibliography II

J. Miettinen and S. Taskinen. "Fourth Moments and Independent Component Analysis". In: *Statistical Science 2015, Vol. 30, No. 3, 372-390* (2015).

Aapo Hyvtirinen, Juha Karhunen, and Erkki Oja. *Independent Component Analsis*. JOHN WILEY SONS, INC., 2001.

James V. Stone. *Independent Component Analsis, A Tutorial Introduction*. Bradford Books, 2004.

The End