Chapter 7
Subspace Methods
Independent Component Analysis

Statistical Methods and Learning in Computer Vision
WS 2012

Claudia Nieuwenhuis
Lehrstuhl für Computer Vision and Pattern Recognition
Fakultät für Informatik
Technische Universität München

Overview

1. **Independent Component Analysis**

What is a 'good' subspace?

- Principal Component Analysis
  - Unsupervised, i.e. data is not associated with classes for classification
  - We find an orthogonal coordinate system, i.e. a basis of the subspace
  - The redundancy in the dataset is minimized
- Linear Discriminant Analysis
  - Supervised, i.e. the data points are associated with classes
  - We find an orthogonal coordinate system, i.e. a basis of the subspace
  - The redundancy in the dataset is minimized
- Independent Component Analysis
  - Unupervised, i.e. the data points are not associated with classes
  - We find linear basis vectors, which are not necessarily orthogonal
  - The independence of different datapoints is maximized

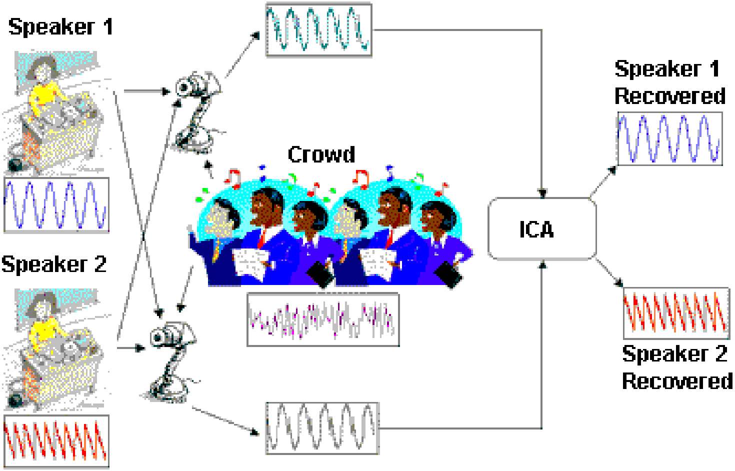
Independent Component Analysis

Left: Principal Components (dimensions of strongest variance)

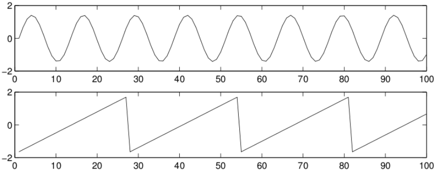
Right: Independent Components (face as superposition of face prototypes (sources))

Independent Component Analysis

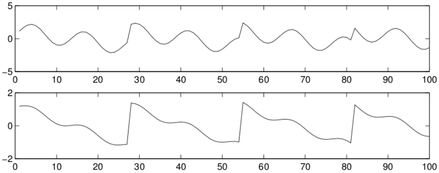
ICA assumes that the given input vector is a linear combination of unknown signals. The goal is to recover the basic signals from the mixtures.



Sources (s_1, s_2)



Mixtures (x_1, x_2)



$$\begin{aligned} x_1(t) &= a_{11}s_1 + a_{12}s_2 \\ x_2(t) &= a_{21}s_1 + a_{22}s_2 \end{aligned}$$

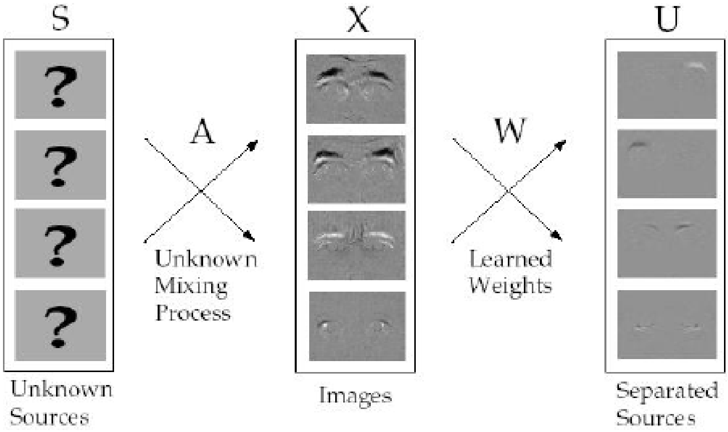
The coefficients a_{ij} depend on the distance to the microphones and are unknown.

The goal of ICA is to estimate the sources s_i and the mixing coefficients a_{ij} given only the mixtures x_i. Since

$$x = As$$

we are searching for the inverse of A to obtain the signals s

$$s = A^{-1}x = Wx$$



Ambiguities of ICA

- We cannot determine the amplitudes / variances of the independent components, assume variance 1.
- We cannot determine the order of the independent components

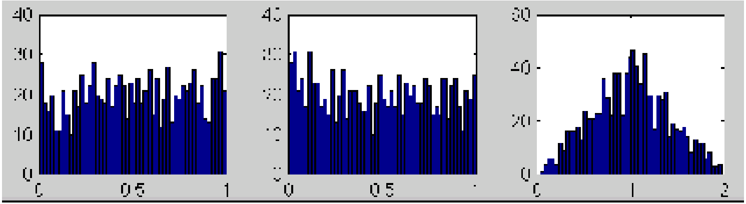
We want to estimate the mixing matrix A and the signals s at the same time, only from the given mixtures x. This is an underconstrained problem. To obtain a solution we have to make additional assumptions.

Assumptions of ICA

- The number of sources s_i is equal to the number of mixtures x_i
- The sources s_i are statistically independent
- None of the sources follows a Gaussian distribution

Central Limit Theorem

The distribution of a sum of **independent** random variables tends towards a Gaussian distribution.



s1 s2 s1 + s2

Idea:
Maximize the non-Gaussianity of $W^T x$ to find the independent components.

Preprocessing

- Centering
  Centering is done to simplify the ICA computation. It means that we subtract the mean from the mixtures x. This ensures that the estimated signals also have mean 0.
- Whitening
  Whitening ensures that the original signals have unit variance and are uncorrelated. It means that we apply PCA to the mixtures x before ICA computation. This strongly reduces the computation time. Dimensionality reduction by removing the eigenvectors of low variance can also be helpful.

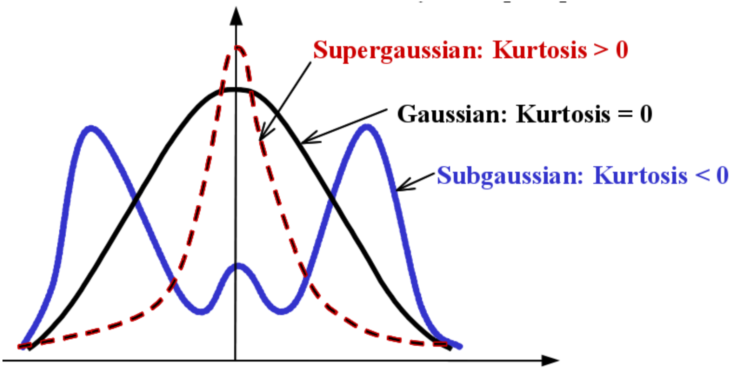
How do we measure non-Gaussianity?

- Kurtosis
- Negentropy
- Mutual Information

Kurtosis

Kurtosis can be considered as a measure of 'peakedness'.

$$kurt(y) = E(y^4) - 3(E(y^2))^2$$



We want to maximize the absolute kurtosis of $W^T x$.

Kurtosis

- For a Gaussian random variable the kurtosis is 0. For most (but not all) non-Gaussian random variables the kurtosis is nonzero.
- $kurt(\alpha x) = \alpha^4 kurt(x)$
- $kurt(x_1 + x_2) = kurt(x_1) + kurt(x_2)$

Based on the independence and unit variance assumption of the signals, the kurtosis is maximal if $w^T x$ equals exactly one of the original signals $\pm s_i$.

Problem:
Kurtosis is very sensitive to outliers, a non-robust measure.

Negentropy

The entropy of a random variable can be interpreted as the degree of information that the observation of the variables gives. The more 'random', i.e. unpredictable or unstructured the variable is, the larger its entropy.

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

A Gaussian random variable has the largest entropy among all random variables of equal variance. Hence the negentropy

$$J(y) = H(y_{gauss}) - H(y)$$

could be used as a measure of non-Gaussianity.

Negentropy

Advantage:
In a statistical sense, negentropy is the optimal measure of non-Gaussianity
Disadvantage:
The computation is very difficult, since it requires an estimate of the pdf of the random variable

Hence, the negentropy is often approximated. Some approximations contain the kurtosis, which leads again to outlier sensitivity. Others give a good compromise between the properties of the classical measures kurtosis and negentropy.

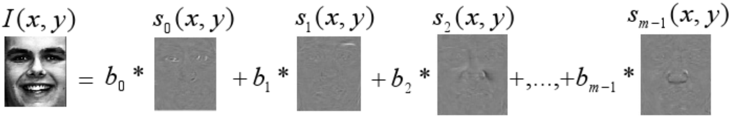
Mutual Information

The mutual information between n scalar random variables $(y_1, ..., y_n)$ with densities $(f_1, ..., f_n)$ can be expressed by the Kullback-Leibler divergence between the joint density $f(y)$ and the product of its marginal densities. It is a measure for independence, which is zero if and only if the variables are statistically independent.

$$MI(y_1, ..., y_n) = \int_{-\infty}^{\infty} f(y) log \frac{f(y)}{\prod_{i=1}^{n} f_i(y)} \, \mathrm{d}y$$

Mutual Information minimization is equivalent to negentropy maximization.

Projection of the original image into the subspace found by ICA



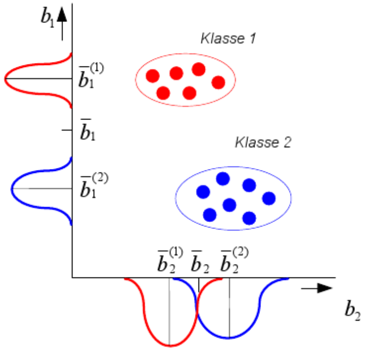
The basis images

- show local features
- show independent features (i.e. one feature cannot be predicted by the other)

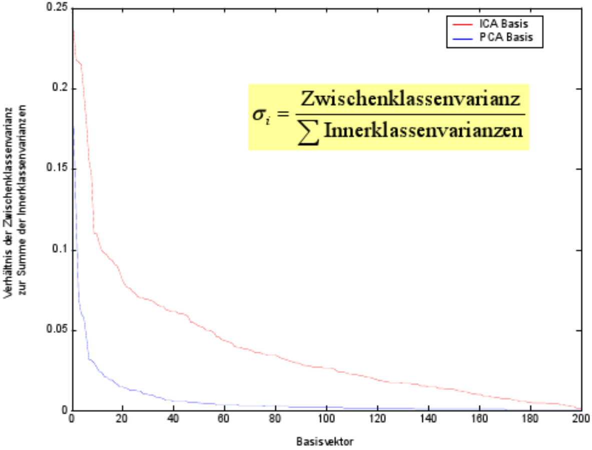
Subspace Dimension

In contrast to PCA, there is no ranking information available to determine the most important independent commponents.

Solution:
If sample classes are known, the class discriminality measurement $\frac{\text{between class variance}}{\sum \text{inner class variance}}$ (very similar to LDA) can be used to rank the independent components.



Subspace Dimension

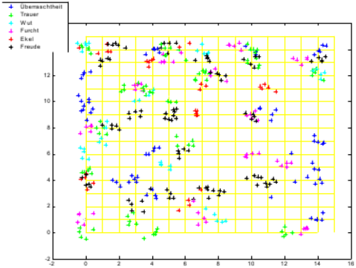


$$\sigma_i = \frac{\text{Zwischenklassenvarianz}}{\sum \text{Innerklassenvarianzen}}$$

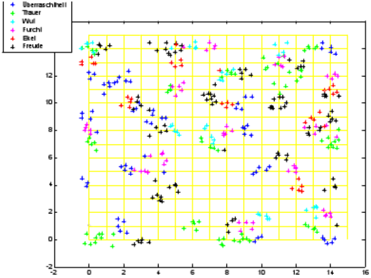
Comparison of class discriminality of a PCA and an ICA basis. The higher the values the better the separation of the classes for the respective component.

Facial Expression Classification

Facial Expression Classification



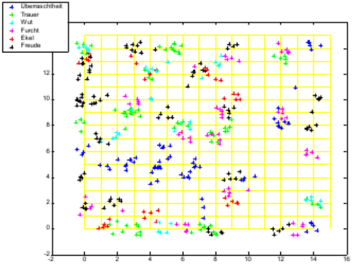
ordered by Eigenvalues



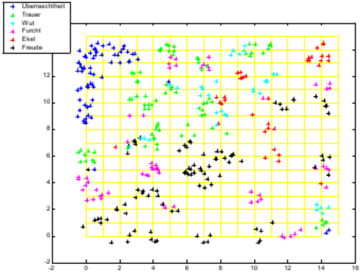
ordered by class discriminality

Facial expression classification based on PCA for components sorted according to a) eigenvalue, b) class discriminality.

Subspace Dimension



no order criterion,
200 dimensional subspace



ordered by class discriminality,
75 selected

Facial expression classification based on ICA.