Principal Component Analysis & Independent Component Analysis

Overview

Principal Component Analysis
- Purpose
- Derivation
- Examples

Independent Component Analysis
- Generative Model
- Purpose
- Restrictions
- Computation
- Example

Literature

Notation & Basics

$\mathbf{u}$	Vector
$\mathbf{U}$	Matrix
$\mathbf{u}^T\mathbf{v}, \quad \mathbf{u}, \mathbf{v} \in \mathbf{R}^N$	Dot product written as matrix product
$\mathbf{u}a$	Product of a row vector with scalar as matrix product, and not $a\mathbf{u}$
$\mathbf{u}^2 = \lVert\mathbf{u}\rVert^2 = \mathbf{u}^T\mathbf{u}$	squared norm

Rules for matrix multiplication:

$$\mathbf{UVW} = (\mathbf{UV})\mathbf{W} = \mathbf{U}(\mathbf{VW})$$

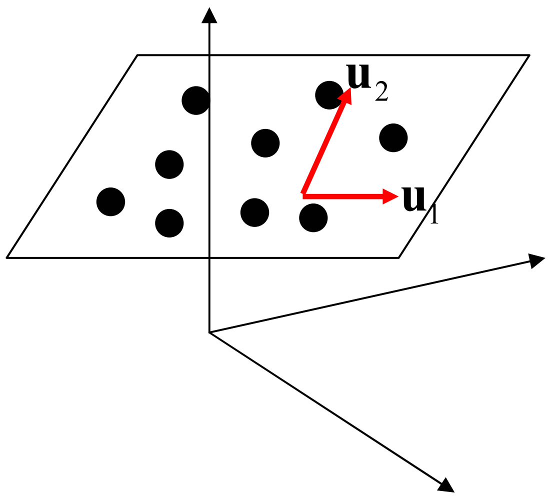
$$(\mathbf{U} + \mathbf{V})\mathbf{W} = \mathbf{UW} + \mathbf{VW}$$

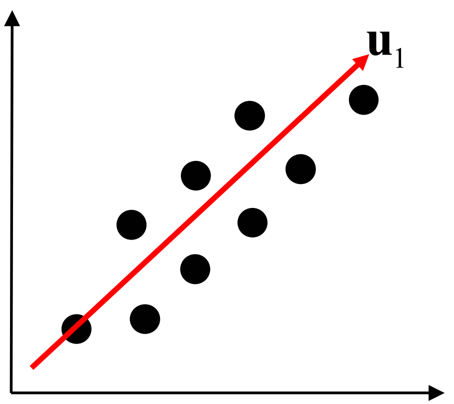
$$(\mathbf{UV})^T = \mathbf{V}^T\mathbf{U}^T$$

Principal Component Analysis (PCA)

Purpose

For a set of samples of a random vector $\mathbf{x} \in \mathbf{R}^N$, discover or reduce the dimensionality and identify meaningful variables.





$$\mathbf{y} = \mathbf{U}\mathbf{x},\ \mathbf{U} : p \times N,\ p < N$$

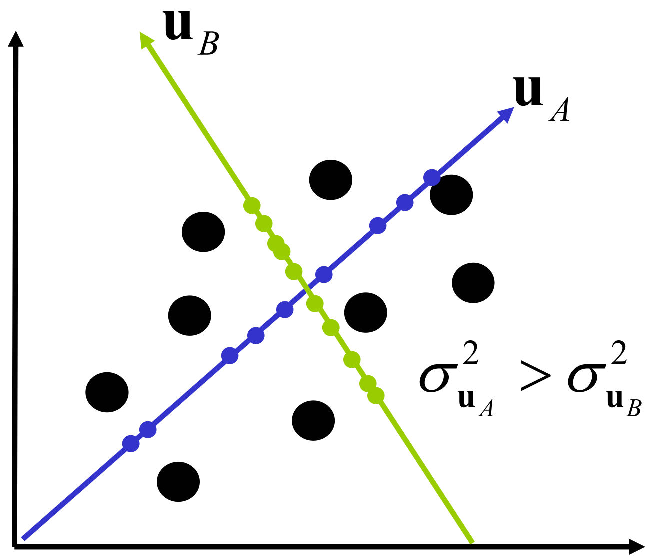
PCA by Variance Maximization

Find the vector $\mathbf{u}_1$, such that the variance of the data along this direction is maximized:

$$\sigma_{\mathbf{u}_1}^2 = E\left\{(\mathbf{u}_1^T\mathbf{x})^2\right\}, E\left\{\mathbf{x}\right\} = 0, \|\mathbf{u}_1\| = 1$$

$$\sigma_{\mathbf{u}_1}^2 = E\left\{(\mathbf{u}_1^T\mathbf{x})(\mathbf{x}^T\mathbf{u}_1)\right\} = \mathbf{u}_1^T E\left\{\mathbf{x}\mathbf{x}^T\right\}\mathbf{u}_1,$$

$$\sigma_{\mathbf{u}_1}^2 = \mathbf{u}_1^T\mathbf{C}\mathbf{u}_1, \quad \mathbf{C} = E\left\{\mathbf{x}\mathbf{x}^T\right\}$$



The solution is the eigenvector $\mathbf{e}_1$ of $\mathbf{C}$ with the largest eigenvalue λ_1.

$$\mathbf{C}\mathbf{e}_1 = \mathbf{e}_1\lambda_1, \lambda_1 = \mathbf{e}_1^T\mathbf{C}\mathbf{e}_1 \Leftrightarrow \sigma_{\mathbf{u}_1}^2 = \lambda_1$$

PCA by Variance Maximization

For a given $p < N$, find p orthonormal basis vectors $\mathbf{u}_i$ such that the variance of the data along these vectors is maximally large, under the constraint of decorrelation:

$$E\left\{(\mathbf{u}_i^T\mathbf{x})(\mathbf{u}_n^T\mathbf{x})\right\} = 0, \quad \mathbf{u}_i^T\mathbf{u}_n = 0 \quad n \neq p$$

The solution are the eigenvectors of $\mathbf{C}$ ordered according to decreasing eigenvalues λ:

$$\mathbf{u}_1 = \mathbf{e}_1, \mathbf{u}_2 = \mathbf{e}_2, \ldots, \mathbf{u}_p = \mathbf{e}_p, \lambda_1 > \lambda_2 \ldots > \lambda_p$$

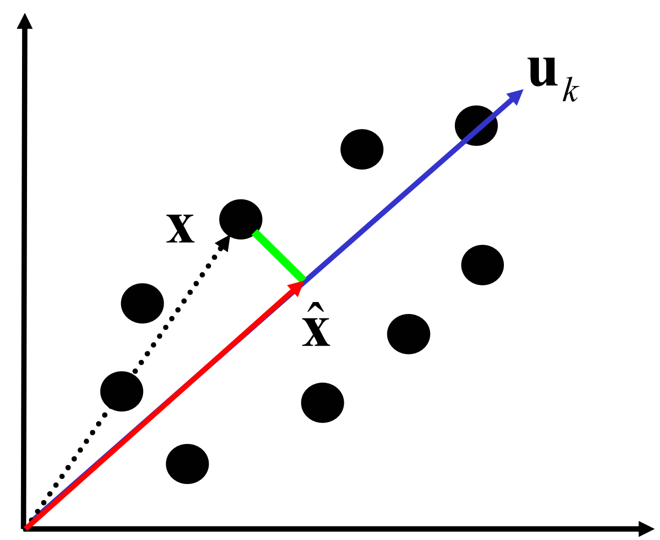
Proof of decorrelation for eigenvectors:

$$E\left\{(\mathbf{e}_i^T\mathbf{x})(\mathbf{e}_n^T\mathbf{x})\right\} = \mathbf{e}_i^T E\left\{\mathbf{x}\mathbf{x}^T\right\}\mathbf{e}_n = \mathbf{e}_i^T\mathbf{C}\mathbf{e}_n = \underbrace{\mathbf{e}_i^T\mathbf{e}_n}_{\text{orthogonal}}\lambda_n = 0$$

PCA by Mean Square Error Compression

For a given $p < N$, find p orthonormal basis vectors such that the *mse* between $\mathbf{x}$ and its projection $\hat{\mathbf{x}}$ into the subspace spanned by the p orthonormal basis vectors is mimimum:

$$mse = E\left\{\left\|\mathbf{x} - \hat{\mathbf{x}}\right\|^2\right\}, \quad \hat{\mathbf{x}}_i = \sum_{k=1}^{p} \mathbf{u}_k\left(\mathbf{x}_k^T \mathbf{u}_k\right), \quad \mathbf{u}_k^T \mathbf{u}_m = \delta_{k,m}$$



Principal Component Analysis (PCA)

$$mse = E\left\{\left\|\mathbf{x}-\hat{\mathbf{x}}\right\|^2\right\} = E\left\{\left\|\mathbf{x}-\sum_{k=1}^{p}\mathbf{u}_k \underbrace{(\mathbf{x}^T\mathbf{u}_k)}_{\text{scalar}}\right\|^2\right\}$$

$$= E\left\{\left\|\mathbf{x}\right\|^2\right\} - 2E\left\{\sum_{k=1}^{p}\mathbf{x}^T\mathbf{u}_k(\mathbf{x}^T\mathbf{u}_k)\right\} + E\left\{\sum_{n=1}^{p}(\mathbf{x}^T\mathbf{u}_n)\mathbf{u}_n^T\sum_{k=1}^{p}\mathbf{u}_k(\mathbf{x}^T\mathbf{u}_k)\right\}$$

$$= E\left\{\left\|\mathbf{x}\right\|^2\right\} - 2E\left\{\sum_{k=1}^{p}(\mathbf{x}^T\mathbf{u}_k)^2\right\} + E\left\{\sum_{k=1}^{p}(\mathbf{x}^T\mathbf{u}_k)^2\right\}$$

$$= E\left\{\left\|\mathbf{x}\right\|^2\right\} - E\left\{\sum_{k=1}^{p}(\mathbf{x}^T\mathbf{u}_k)^2\right\}$$

$$= \mathrm{trace}\left(\mathbf{C}\right) - \sum_{k=1}^{p}\mathbf{u}_k^T\mathbf{C}\mathbf{u}_k, \quad \mathbf{C} = E\left\{\mathbf{x}\mathbf{x}^T\right\}$$

Principal Component Analysis (PCA)

$$mse = \text{trace}\left(\mathbf{C}\right) - \underbrace{\sum_{k=1}^{P} \mathbf{u}_k^T \mathbf{C} \mathbf{u}_k}_{\text{maximize}}, \quad \mathbf{C} = E\left\{\mathbf{x}\mathbf{x}^T\right\}$$

Solution to minimizing *mse* is any (orthonormal) basis of the subspace spanned by the p first eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_p$ of $\mathbf{C}$.

$$mse = \text{trace}\left(\mathbf{C}\right) - \sum_{k=1}^{p} \lambda_k \quad = \sum_{k=p+1}^{N} \lambda_k$$

The *mse* is the sum of the eigenvalues corresponding to the discarded eigenvectors $\mathbf{e}_{P+1}, \ldots, \mathbf{e}_N$ of $\mathbf{C}$: $mse = \sum_{k=p+1}^{N} \lambda_k$

Principal Component Analysis (PCA)

How to determine the number of principal components p?

Linear signal model with unknown number $p < N$ of signals:

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n}, \quad \mathbf{A} = \begin{pmatrix} a_{1,1} & & a_{1,p} \\ & \ddots & \\ a_{N,1} & & a_{N,p} \end{pmatrix} \quad N \times p$$

Signal s_i have 0 mean and are uncorrelated, $\mathbf{n}$ is white noise:

$$E\left\{\mathbf{s}\mathbf{s}^T\right\} = \mathbf{I},\ E\left\{\mathbf{n}\mathbf{n}^T\right\} = \sigma_n^2 \mathbf{I}$$

$$\mathbf{C} = E\left\{\mathbf{x}\mathbf{x}^T\right\} = E\left\{\mathbf{A}\mathbf{s}(\mathbf{A}\mathbf{s})^T\right\} + E\left\{\mathbf{n}\mathbf{n}^T\right\} + \underbrace{E\left\{\mathbf{A}\mathbf{s}\mathbf{n}^T\right\}}_{=0}$$

$$= \mathbf{A}\, E\left\{\mathbf{s}\mathbf{s}^T\right\} \mathbf{A}^T + E\left\{\mathbf{n}\mathbf{n}^T\right\} = \mathbf{A}\mathbf{A}^T + \sigma_n^2 \mathbf{I}$$

$$d_1 > d_2 > \ldots > d_p > d_{p+1} = d_{p+2} = \ldots = d_N = \sigma_n^2$$

$\rightarrow$ cut off when eigenvalues become constants

Computing the PCA

Given a set of samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ of a random vector $\mathbf{x}$ calculate mean and covariance.

$$\tilde{\boldsymbol{\mu}} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{x}_i, \quad \mathbf{x} \to \mathbf{x} - \tilde{\boldsymbol{\mu}}$$

$$\tilde{\mathbf{C}} = \frac{1}{M}\sum_{i=1}^{M}(\mathbf{x}_i - \tilde{\boldsymbol{\mu}})(\mathbf{x}_i - \tilde{\boldsymbol{\mu}})^T$$

Compute eigenvecotors of $\tilde{\mathbf{C}}$ e.g. with QR algorithm

Computing the PCA

If the number of samples M is smaller than the dimensionality N of $\mathbf{x}$:

$$\mathbf{B} = \begin{pmatrix} x_{1,1} & & x_{1,M} \\ & \ddots & \\ x_{N,1} & & x_{N,M} \end{pmatrix}, \tilde{\mathbf{C}} = \mathbf{B}\mathbf{B}^T,\ \mathbf{B} : N \times M,\ \mathbf{B}^T : M \times N$$

$$\mathbf{B}\mathbf{B}^T\mathbf{e} = \mathbf{e}\lambda$$

$$\mathbf{B}^T\mathbf{B}\mathbf{e}' = \mathbf{e}'\lambda'$$

$$\mathbf{B}\mathbf{B}^T(\mathbf{B}\mathbf{e}') = (\mathbf{B}\mathbf{e}')\lambda'$$

$$\mathbf{e} = \mathbf{B}\mathbf{e}',\ \lambda' = \lambda$$

$\rightarrow$ Reducing complexity from $O(N^2)$ to $O(M^2)$

Examples

Eigenfaces for face recognition (Turk&Pentland):

Training:

-Calculate the eigenspace for all faces in the training database

-Project each face into the eigenspace $\rightarrow$ feature reduction

Classification:

-Project new face into eigenspace

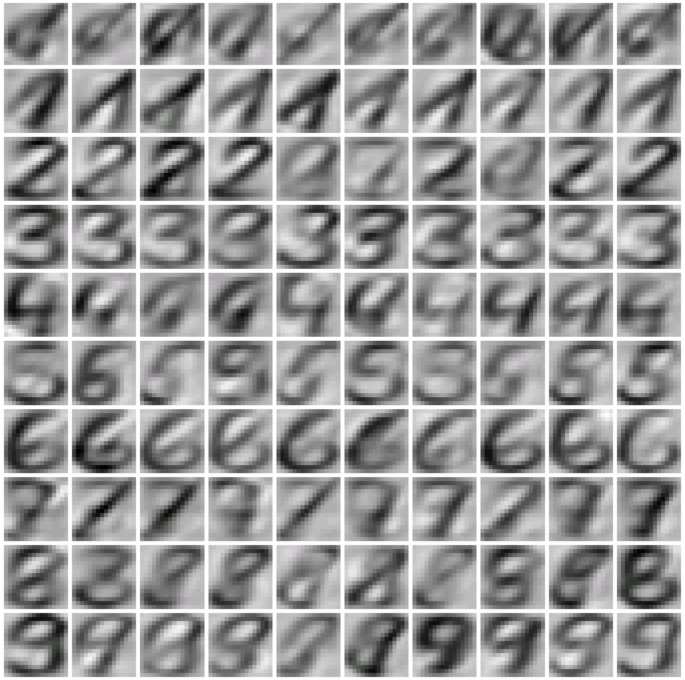
-Nearest neighbor in the eigenspace

Examples cont.

Feature reduction/extraction

Original

Reconstruction with 20 PC



http://www.nist.gov/

Independent Component Analysis (ICA)

Generative model

Noise free, linear signal model:

$$\mathbf{x} = \mathbf{A}\mathbf{s} = \sum_{i=1}^{N} \mathbf{a}_i s_i$$

$\mathbf{x} = \left(x_1, \dots, x_N\right)^T$ Observed variables

$\mathbf{s} = \left(s_1, \dots, s_N\right)^T$ latent signals, independent components

$$\mathbf{A} = \begin{pmatrix} a_{1,1} & & a_{1,N} \\ & \ddots & \\ a_{N,1} & & a_{N,N} \end{pmatrix}$$ unknown mixing matrix, $\mathbf{a}_i = (a_{1,i}, \dots, a_{N,i})^T$

Independent Component Analysis (ICA)

Task

For the linear, noise free signal model, compute **A** and **s** given the measurements **x**.

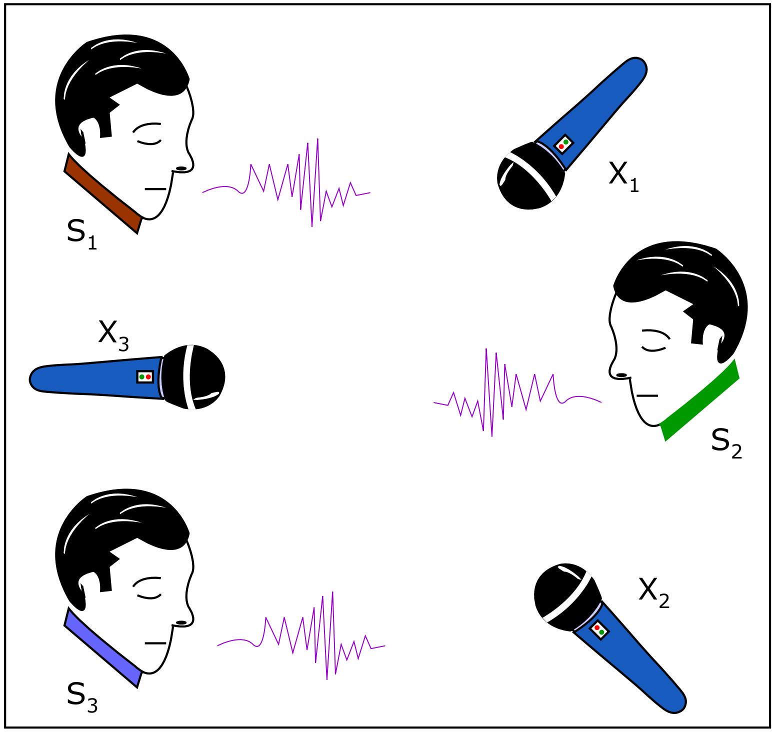


Figure by MIT OCW.

Blind source separation : separate the three original signals s_1, s_2, and s_3 from their mixtures x_1, x_2, and x_3.

Restrictions

1.) Statistical independence

The signals s_i must be statistically independent:

$$p(s_1, s_2, \ldots, s_N) = p_1(s_1)\,p_2(s_2)\ldots p_N(s_N)$$

Independent variables satisfy:

$$E\{g_1(s_1)g_2(s_2)\ldots g_N(s_N)\} = E\{g_1(s_1)\}\,E\{g_2(s_2)\}\ldots E\{g_N(s_N)\}$$

for any $g_i(s) \in L^1$

$$E\{g_1(s_1)g_2(s_2)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_1(s_1)g_2(s_2)\,p(s_1, s_2)\,ds_1 ds_2$$

$$= \int_{-\infty}^{\infty} g_1(s_1)\,p(s_1)\,ds_1 \int_{-\infty}^{\infty} g_2(s_2)\,p(s_2)\,ds_2 = E\{g_1(s_1)\}\,E\{g_2(s_2)\}$$

Restrictions

Statistical independence cont.

$$E\left\{g_1(s_1)g_2(s_2)...g_N(s_N)\right\} = E\left\{g_1(s_1)\right\}E\left\{g_2(s_2)\right\}...E\left\{g_N(s_N)\right\}$$

Independence includes uncorrelatedness:

$$E\left\{(s_i-\mu_i)(s_n-\mu_n)\right\} = E\left\{(s_i-\mu_i)\right\}E\left\{(s_n-\mu_n)\right\} = 0$$

Restrictions

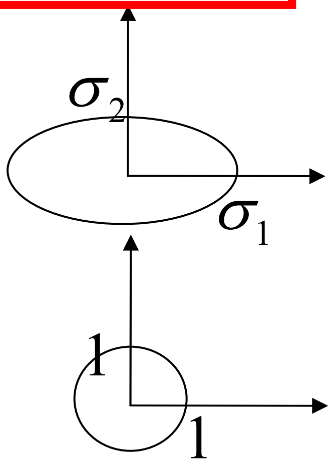
$\boxed{\text{2.) Nongaussian components. The components } s_i \text{ must have a nongaussian distribution otherwise there is no unique solution.}}$

2.) Nongaussian components

The components s_i must have a nongaussian distribution otherwise there is no unique solution.

Example:

given **A** and two gaussian signals:



$$p(s_1, s_2) = \frac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\frac{s_1^2}{2\sigma_1^2} - \frac{s_2^2}{2\sigma_2^2}\right)$$

generate new signals

$$\mathbf{s}' = \underbrace{\begin{pmatrix} 1/\sigma_1 & 0 \\ 0 & 1/\sigma_2 \end{pmatrix}}_{\text{Scaling matrix } \mathbf{S}} \mathbf{s} \Rightarrow p(s_1', s_2') = \frac{1}{2\pi}\exp\left(-\frac{s_1'^2}{2}\right)\exp\left(-\frac{s_2'^2}{2}\right)$$

Restrictions

Nongaussian components cont.

under rotation the components remain independent:

$$\mathbf{s}'' = \underbrace{\begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix}}_{\text{Rotation matrix } \mathbf{R}} \mathbf{s}',\ p(s''_1, s''_2) = \frac{1}{2\pi}\exp(-\frac{s''^2_1}{2})\exp(-\frac{s''^2_2}{2})$$

combine whitening and rotation $\mathbf{B} = \mathbf{RS}$:

$\mathbf{x} = \mathbf{As} = \mathbf{AB}^{-1}\mathbf{s}''$

$\mathbf{AB}^{-1}$ is also a solution to the ICA problem.

Restrictions

3.) Mixing matrix must be invertible
The number of independent components is equal to the number of observerd variables.
Which means that there are no redundant mixtures.

In case mixing matrix is not invertible apply PCA on measurements first to remove redundancy.

Ambiguities

1.) Scale

$$\mathbf{x} = \sum_{i=1}^{N} \mathbf{a}_i s_i = \sum_{i=1}^{N} (\frac{1}{\alpha_i} \mathbf{a}_i)(\alpha_i s_i)$$

Reduce ambiguity by enforcing $E\left\{ s_i{}^2 \right\} = 1$

2.) Order

We cannot determine an order of the independant components

Computing ICA

a) Minimizing mutual information:

$\hat{\mathbf{s}} = \hat{\mathbf{A}}^{-1}\mathbf{x}$

Mutual information: $I(\hat{\mathbf{s}}) = \sum_{i=1}^{N} H(\hat{s}_i) - H(\hat{\mathbf{s}})$

H is the differential etnropy:

$H(\hat{\mathbf{s}}) = -\int p(\hat{\mathbf{s}}) \log_2(p(\hat{\mathbf{s}}))\, d\hat{\mathbf{s}}$

I is always nonnegative and 0 only if the $\hat{s}_i$ are independent.

Iteratively modify $\hat{\mathbf{A}}^{-1}$ such that $I(\hat{\mathbf{s}})$ is minimized.

Computing ICA cont.

b) Maximizing Nongaussianity

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

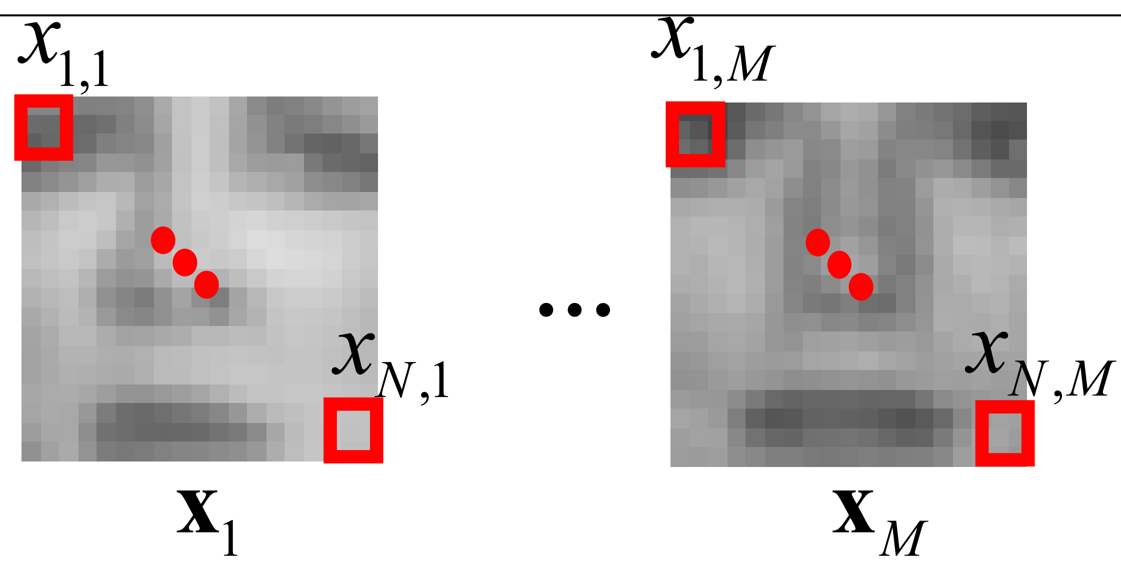
introduce y and $\mathbf{b}$: $y = \mathbf{b}^T\mathbf{x} = \mathbf{b}^T\mathbf{A}\mathbf{s} = \mathbf{q}^T\mathbf{s}$

From central limit theorem:

$y = \mathbf{q}^T\mathbf{s}$ is more gaussian than any of the s_i and becomes least gaussian if $y = s_i$.

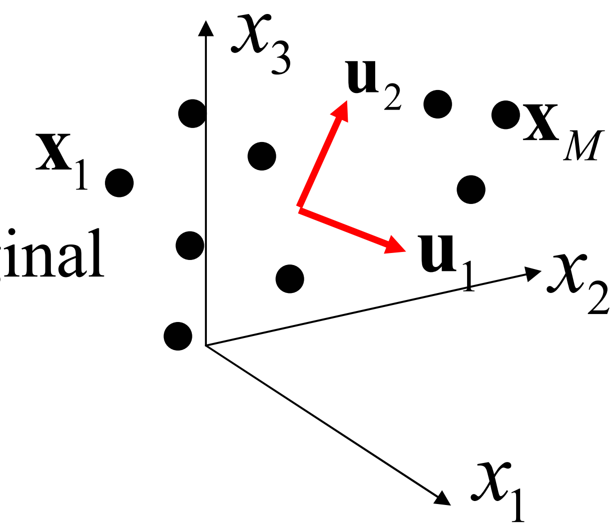
Iteratively modify $\mathbf{b}^T$ such that the "gaussianity" of y is minimized. When a local minimum is reached, $\mathbf{b}^T$ is a row vector of $\mathbf{A}^{-1}$.

PCA Applied to Faces

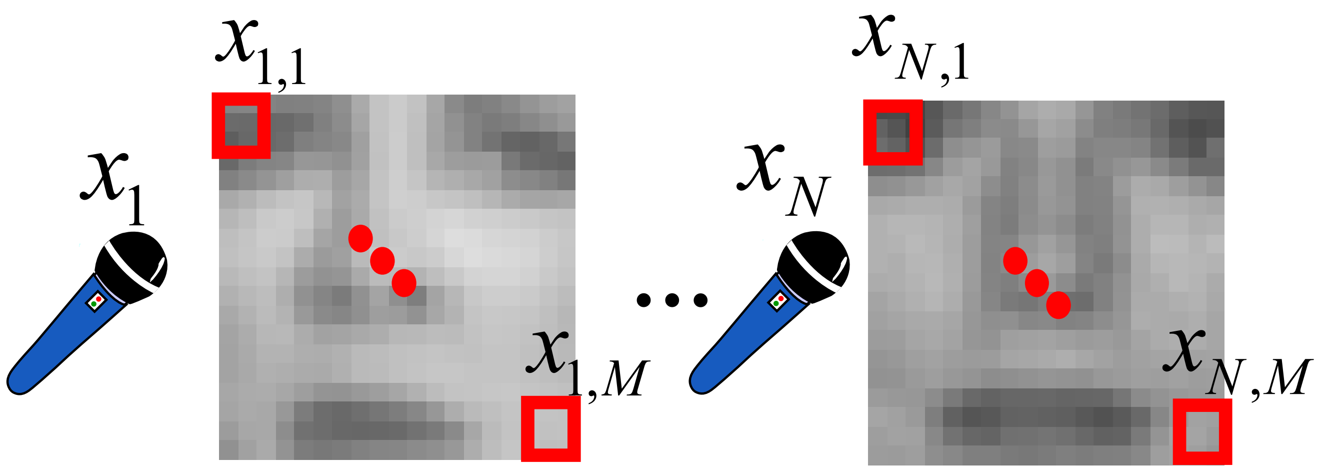


Each pixel is a feature, each face image a point in the feature space. Dimension of feature vector is given by the size of the image.

$\mathbf{u}_i$ are the eigenvectors which can be represented as pixel images in the original cooridinate system $x_1 \ldots x_N$



ICA Applied to Faces



Now each image corresponds to a particular observed variable measured over time (M samples). N is the number of images.

$$\mathbf{x} = \mathbf{A}\mathbf{s} = \sum_{i=1}^{N} \mathbf{a}_i s_i$$

$\mathbf{x} = \left(x_1, \ldots, x_N\right)^T$ Observed variables

$\mathbf{s} = \left(s_1, \ldots, s_N\right)^T$ latent signals, independent components

Features for face recognition

Image removed due to copyright considerations. See Figure 1 in: Baek, Kyungim et. al.
"PCA vs. ICA: A comparison on the FERET data set." International Conference of Computer Vision,
Pattern Recognition, and Image Processing, in conjunction with the 6th JCIS. Durham, NC, March 8-14 2002, June 2001.