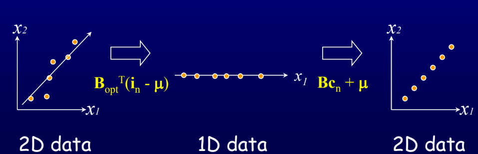
Statistical Linear Models: ICA & FLD

Last Time: PCA

- PCA identifies an m dimensional explanation of n dimensional data where m < n.
- Originated as a statistical analysis technique.
- PCA attempts to minimize the reconstruction error under the following restrictions
  - Linear Reconstruction
  - Orthogonal Factors
- Equivalently, PCA attempts to maximize variance.

Data Loss

- Sample points can be projected via the new m×d projection matrix $\mathbf{B}_{opt}$ and can still be reconstructed, but some information will be lost.



SVD of a Matrix

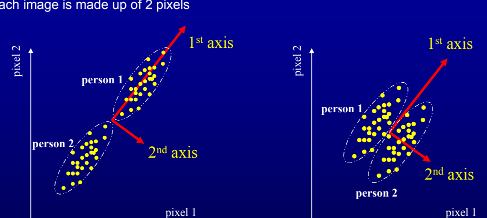
Scatter of matrix: $\mathbf{S}_T = \frac{1}{N-1}(\mathbf{D}-\mathbf{M})(\mathbf{D}-\mathbf{M})^T$

$(\mathbf{D}-\mathbf{M}) = \mathbf{U}\Sigma\mathbf{V}^T$ by svd of $(\mathbf{D}-\mathbf{M})$ ➡ set $\mathbf{B} = \mathbf{U}$

$(\mathbf{D}-\mathbf{M})(\mathbf{D}-\mathbf{M})^T = \mathbf{U}\Sigma^2\mathbf{U}^T$ (svd of $\mathbf{S}_T$) ➡ set $\mathbf{B} = \mathbf{U}$

PCA for Recognition - EigenImages

- Consider a set of images of 2 people under fixed viewpoint & N lighting condition
- Each image is made up of 2 pixels



- Reduce dimensionality by throwing away the axis along which the data varies the least
- The coefficient vector associated with the 1st basis vector is used for classifiction
- Possible classifier: Mahalanobis distance
- Each image is represented by one coefficient vector
- Each person is displayed in N images and therefore has N coefficient vectors

PART I: 2D Vision

- Appearance-Based Methods
  - **Statistical Linear Models:**
    - Principal Component Analysis
    - ICA, FLD ← **Today ➔**
    - Non-negative Matrix Factorization, Sparse Matrix Factorization
  - **Statistical Tensor Models:**
    - Multilinear PCA,
    - Multilinear ICA
  - Person and Activity Recognition

Statistical Linear Models

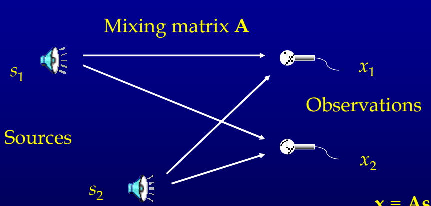
- Generative Models:
  - Second-order methods
    - faithful/accurate data representation - minimal reconstruction (mean-square) error
      - covariance
    - PCA – Principal Component Analysis
    - Factor Analysis
  - Higher Order Methods
    - meaningful representation
      - higher order statistics
    - ICA – Independent Component Analysis
- Descriminant Models:
  - FLD – Fisher Linear Descriminant Analysis

ICA

Blind Signal Separation (BSS) or Independent Component Analysis (ICA) is the identification & separation of mixtures of sources with little prior information.

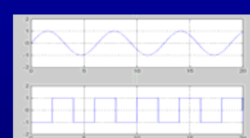
- Applications include:
  - Audio Processing
  - Medical data
  - Finance
  - Array processing (beamforming)
  - Coding
- … and most applications where Factor Analysis and PCA is currently used.
- While PCA seeks directions that represents data best in a $\Sigma|x_0 - x|^2$ sense, ICA seeks such directions that are most independent from each other.

The simple "Cocktail Party" Problem



n sources, m=*n* observations

Motivation



Two Independent Sources

Mixture at two Mics

$$x_1(t) = a_{11}s_1 + a_{12}s_2$$
$$x_2(t) = a_{21}s_1 + a_{22}s_2$$

a_{IJ} ... Depend on the distances of the microphones from the speakers

Independent Component Analysis

Given m signals of length n, construct the data matrix

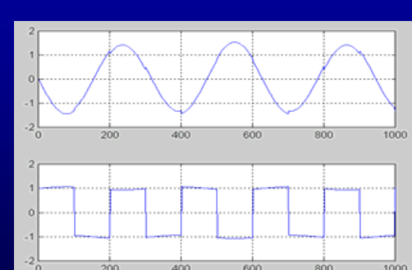
$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_m^T \end{bmatrix}$$

We assume that X consists of m sources such that

X = AS

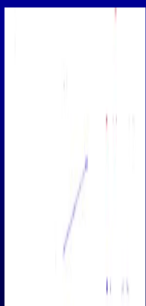
where A is an unknown m by m mixing matrix and S is m independent sources.

Motivation



Get the Independent Signals out of the Mixture

Independent Component Analysis



- *PCA finds the directions that uncorellated*
- *ICA / Blind Source Separation:*
  - Observed data is modeled as a linear combination of independent sources
    - Cocktail Problem: A sound recording at a party is the result of multiple individuals speaking (independent sources)
  - *Finds the directions of maximum independence*
  - ***Statistically independent:***
    - Two variables **x, y** are *statistically independent* ***iff***

$$P(\mathbf{x} \,\&\, \mathbf{y}) = P(\mathbf{x})P(\mathbf{y}).$$

- Equivalently, $E\{g(\mathbf{x})h(\mathbf{y})\} - E\{g(\mathbf{x})\}E\{h(\mathbf{y})\} = 0$

where *g* and *h* are any functions.

Computing Independent Components

- By maximization of nongaussianity: kurtosis
- By maximum likelihood estimation
- By minimization of mutual information
- By tensorial methods
- By nonlinear decorrelation and nonlinear PCA
- By methods using time structure

• Hyvärinen A, Karhunen J, Oja E. Independent component analysis, John Wiley & Sons, Inc., New York, 2001, p. 481

• Material in: http://www.cis.hut.fi/projects/ica/fastica/

Measures of Non-Gaussianity

We need to have a quantitative measure of non-gaussianity for ICA Estimation.

- Kurtosis : gauss=0 (sensitive to outliers)

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

- Entropy : gauss=largest

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

- Neg-entropy :
  - gauss = 0 (difficult to estimate)

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

- Approximations

$$J(y) = \tfrac{1}{12}E\{y^3\}^2 + \tfrac{1}{48}kurt(y)^2$$

$$J(y) \approx [E\{G(y)\} - E\{G(v)\}]^2$$

where *v* is a gaussian rand. variable :

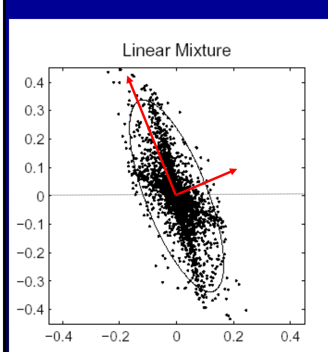
$$G(y) = \tfrac{1}{a}\log\cosh(a.y)$$

$$G(y) = -\exp(-a.u^2/2)$$

Computing IC's using Non-Gausianity

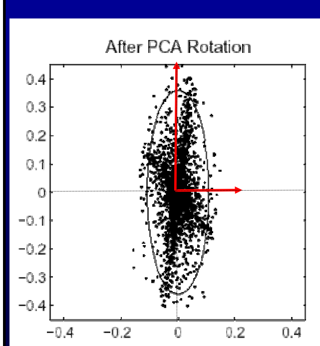
- Kurtosis
  - kurt(y) = $E\{y^4\} - 3(E\{y^2\})^2 = E\{y^4\} - 3$
- for unit-variance data
  - kurt(y) = 0 for gaussian data
  - kurt(y) < 0 for subgaussian data
  - kurt(y) > 0 for supergaussian data
- kurtosis is measured along each possible projection direction over the data
  - a maximum corresponds to one of the IC's
  - other IC's are found from the orthogonal directions with an iterative algorithm
  - rotation matrix R has now been solved

Geometric View of ICA



$$\mathbf{D} = \mathbf{USV}^T$$

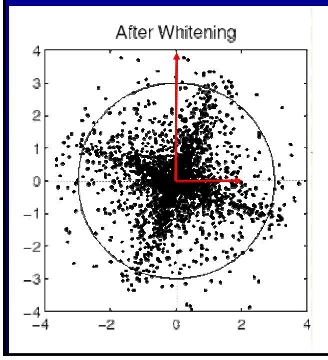
Geometric View of ICA



$$\mathbf{D} = \mathbf{USV}^T$$

$$\mathbf{D}' = \mathbf{U}^T\mathbf{D}$$

Geometric View of ICA

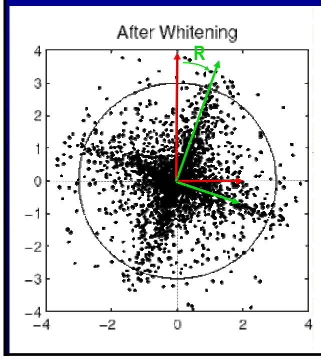


$$D = USV^T$$

$$D' = U^T D$$

$$D'' = S^{-\frac{1}{2}} U^T D$$

Geometric View of ICA



$$D = USV^T$$

$$D' = U^T D$$

$$D'' = S^{-\frac{1}{2}} U^T D$$

$$D''' = RS^{-\frac{1}{2}} U^T D$$

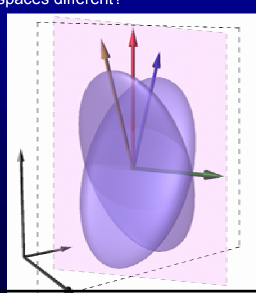
$$D = \underbrace{US^{\frac{1}{2}}R^T}\underbrace{RS^{-\frac{1}{2}}}SV^T$$

$$D = \underbrace{UW^{-1}}WSV^T$$

Independent Components

PCA vs. ICA subspace

- Is ICA and PCA subspace theoretically guaranteed to be the same or different?
- When are the two subspaces different?

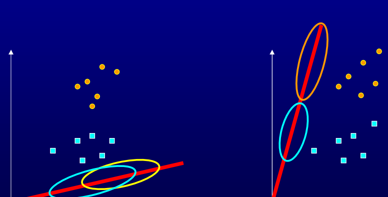


Fisher Linear Discriminant:

FisherFaces

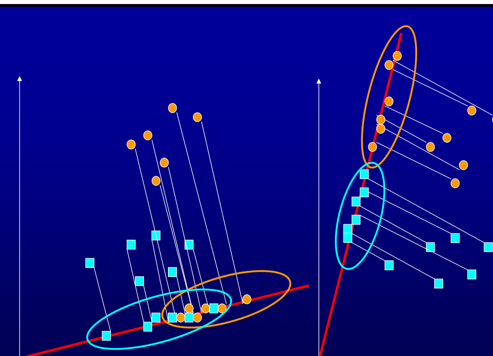
Fisher's Linear Discriminant

- Objective: Find a projection which separates data clusters



Poor separation

Good separation



Poor separation

Good separation

FLD: Theory

- Find a projection that maximize the between-class scatter while minimizing the within-class scatter

FLD: Problem formulation

- N Sample images: $\{\mathbf{i}_1, \cdots, \mathbf{i}_N\}$
- C classes: $\{\mathbf{D}_1, \ \cdots \ \mathbf{D}_c, \ \cdots \ \mathbf{D}_C\}$
- Average of each class: $\boldsymbol{\mu}_i = \frac{1}{N_i} \sum_{\mathbf{i}_k \in \mathcal{D}_c} \mathbf{i}_k$
- Total average: $\boldsymbol{\mu} = \frac{1}{N} \sum_{k=1}^{N} \mathbf{i}_k$

FLD: Data Scatter

- Within-class scatter matrix

$$\mathbf{S}_W = \sum_{c=1}^{C} \sum_{\mathbf{i}_n \in D_c} (\mathbf{i}_n - \mu_c)(\mathbf{i}_n - \mu_c)^T$$

- Between-class scatter matrix

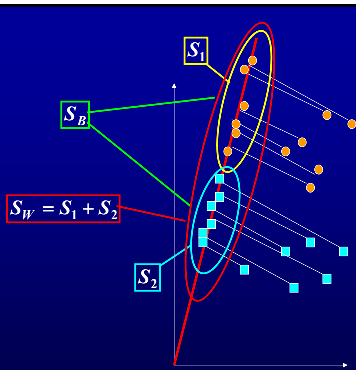
$$\mathbf{S}_B = \sum_{c=1}^{C} |D_c| (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^T$$

- Total scatter matrix

$$\mathbf{S}_T = \mathbf{S}_W + \mathbf{S}_B$$

FLD: Practice *(Cont.)*

- After projection: $\mathbf{y}_k = \mathbf{B}^T \mathbf{x}_k$
- And…
- Between class scatter (of y's): $\widetilde{\mathbf{S}}_B = \mathbf{B}^T \mathbf{S}_B \mathbf{B}$
- Within class scatter (of y's): $\widetilde{\mathbf{S}}_W = \mathbf{B}^T S_W \mathbf{B}$



Good separation

Fisher Linear Discriminant

- The basis matrix B is chosen in order to maximize ratio of the determinant between class scatter matrix of the projected samples to the determinant within class scatter matrix of the projected samples

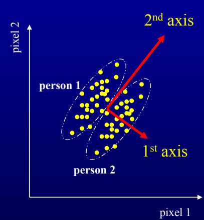
$$\mathbf{B} = \arg\max_{\mathbf{B}} \frac{|\widetilde{\mathbf{S}}_{btw}|}{|\widetilde{\mathbf{S}}_{in}|} = \arg\max_{\mathbf{B}} \frac{|\mathbf{B}^T \mathbf{S}_{btw} \mathbf{B}|}{|\mathbf{B}^T \mathbf{S}_{in} \mathbf{B}|}$$

- B is the set of generalized eigenvectors of S_{Btw} and S_{Win} corresponding with a set of decreasing eigenvalues

$$\mathbf{S}_{btw} \mathbf{B} = \mathbf{S}_{within} \Lambda \mathbf{B}$$

Fisher Linear Discriminant

- Consider a set of images of 2 people under fixed viewpoint & N lighting condition



- Each image is represented by one coefficient vector
- Each person is displayed in N images and therefore has N coefficient vectors