Data Mining & Machine Learning

CS57300
Purdue University

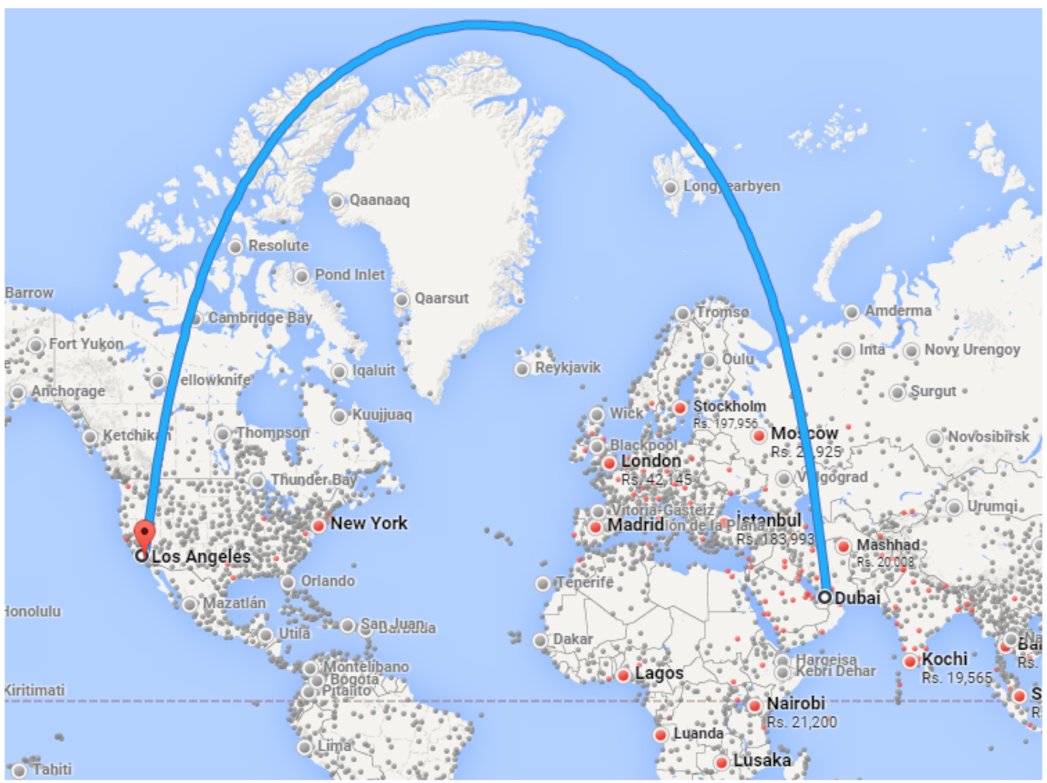
April 10, 2018

Goal

- Visualize high dimensional data (and understand its Geometry)
- Project the data into lower dimensional spaces
- Understand how to model data

A Geometric Embedding Problem

- Ever wondered why flights from U.S. to Europe, India, China, and Middle East seem to take the longest route?
  - E.g. Figure shows Dubai – Los Angeles



This **is** the shortest path.
We were looking at the wrong geometric embedding

PURDUE UNIVERSITY

Embedding of High Dimensional Data

- Example: Bank Loans
  1. Amount Requested
  2. Interest Rate Percentage
  3. Loan Length in Months
  4. Loan Title
  5. Loan Purpose
  6. Monthly Payment
  7. Total Amount Funded
  8. Debt-To-Income Ratio Percentage
  9. FICO Range (https://en.wikipedia.org/wiki/FICO)
  10. Status (1 = Paid or 0 = Not Paid)

- What is the best low dimensional embedding?

Today

- Principal Component Analysis (PCA) is a linear projection that maximizes the variance of the projected data
  - The output of PCA is a set of k orthogonal vectors in the original p-dimensional feature space, the k *principal components*, $k \le p$
  - PCA gives us uncorrelated components, which are generally not independent components; for that you need independent component analysis
- Independent Component Analysis (ICA)
  - A weaker form of independence is uncorrelatedness. Two random variables x and y are said to be uncorrelated if their covariance is zero
  - But uncorrelatedness does not imply independence (nonlinear dependencies)
  - We would like to learn hidden independence in the data

PCA Formulation

- Goal Find a linear projection that maximizes the variance of the projected data
  - Given a p-dimensional observed r.v., $\mathbf{x}$, find $\mathbf{U}$ and $\mathbf{z}$ such that

$$\mathbf{x} = \mathbf{U}\mathbf{z}$$

  and $\mathbf{z}$ has independent normally distributed components z_i

- Due to non-uniqueness, $\mathbf{U}$ is assumed unitary

- Best practices (Standardization = Centering + Scaling):
  - Centered PCA: The variables x are first centered should they have 0 mean
  - Scaled PCA: Each variable is scaled to have unit variance (divide column by standard deviation)

PCA Properties

- Principal Component Analysis (PCA) is a linear projection that maximizes the variance of the projected data
  - The output of PCA is a set of k orthogonal vectors in the original p-dimensional feature space, the k *principal components*, $k \leq p$
  - Principal components are weighted combinations of the original features
  - The projections onto the principal components are uncorrelated
  - No other projection onto k dimensions captures more of the variance
  - The first principal component is the direction in the feature space along which the data has the most variance
  - The second principal component is the direction orthogonal to the first component with the most variance

PCA Algorithm

- Let X be a N by p matrix with N measurements of dimension p
- $A = X X^T$
- Let $A = P \Lambda P^T$, where L is the eigenvalue diagonal matrix and P is the eigenvector matrix

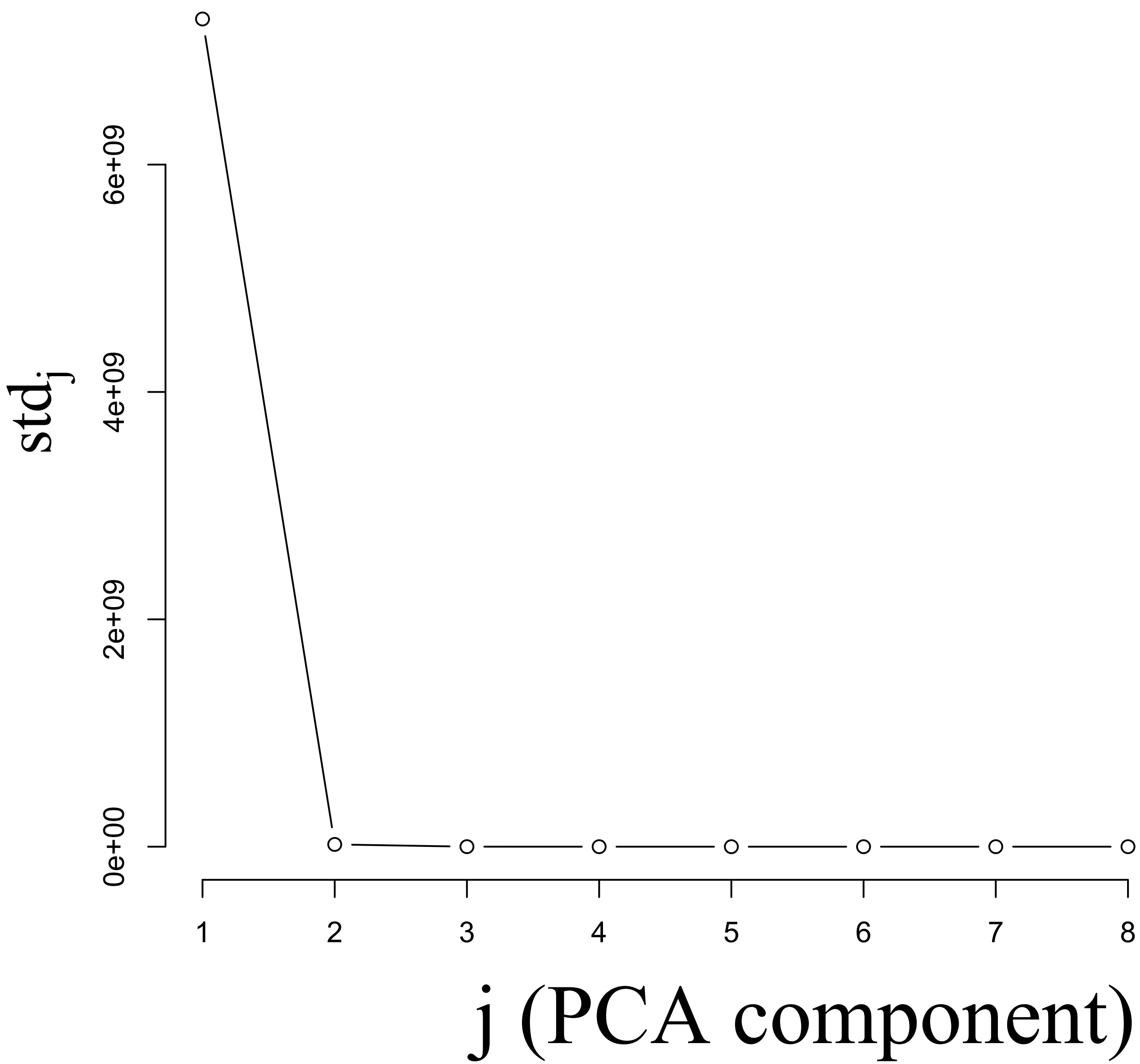
- PCA projection = P X

- Standard deviation of component j

$$\mathrm{std}_j = \sqrt{\frac{1}{p}(PAP^T)_j}$$

How Many Components to Use?

- Find steep drop in standard deviation (std_j)



Independent Component Analysis (ICA) Formulation

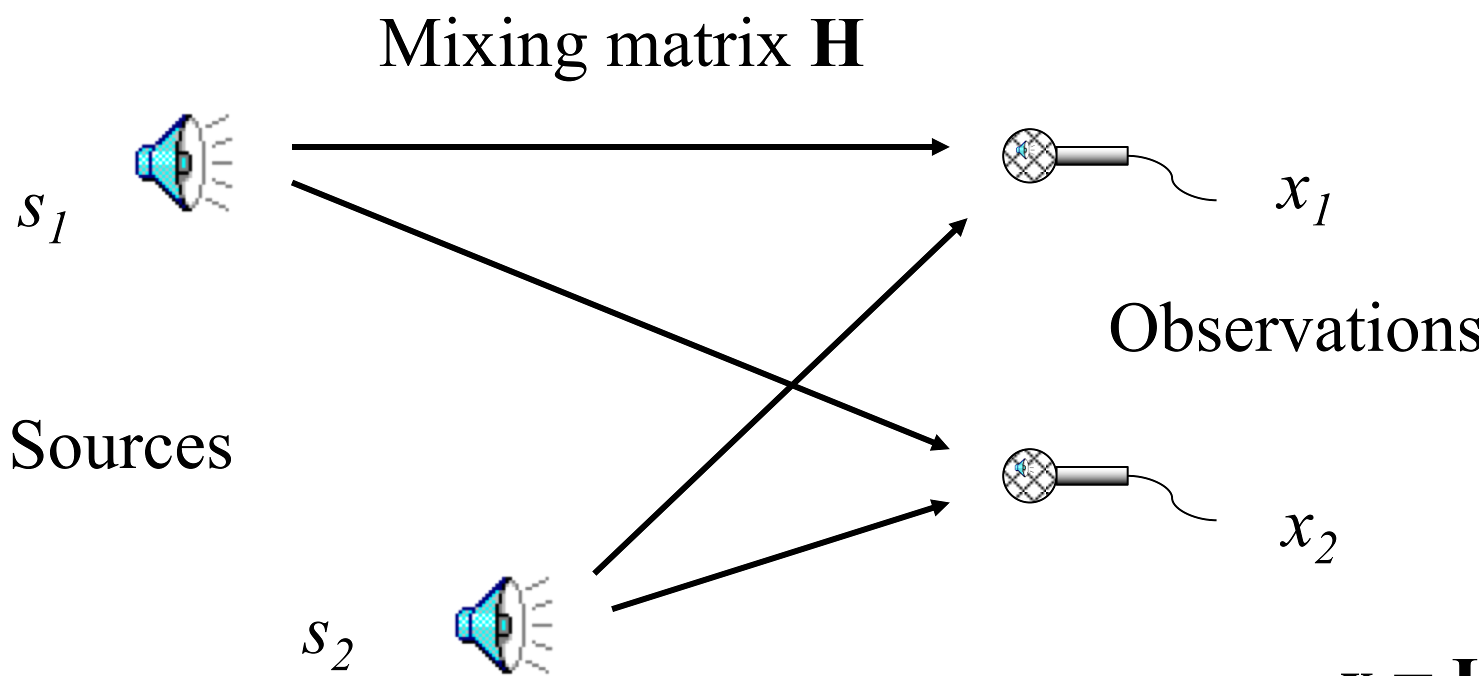
- Goal:
  - Given a p-dimensional observed r.v., $\mathbf{x}$, find $\mathbf{H}$ and $\mathbf{s}$ such that

$$\mathbf{x} = \mathbf{H}\mathbf{s}$$

    $\mathbf{s}$ has mutually statistically independent components s_i

  - Known as *"blind source separation"* problem

Blind Source Separation: The "Cocktail Party" Problem

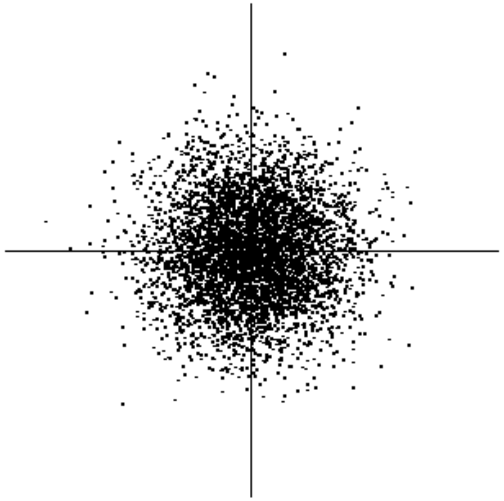


$\mathbf{x} = \mathbf{Hs}$

k sources, N=k observations

Restrictions

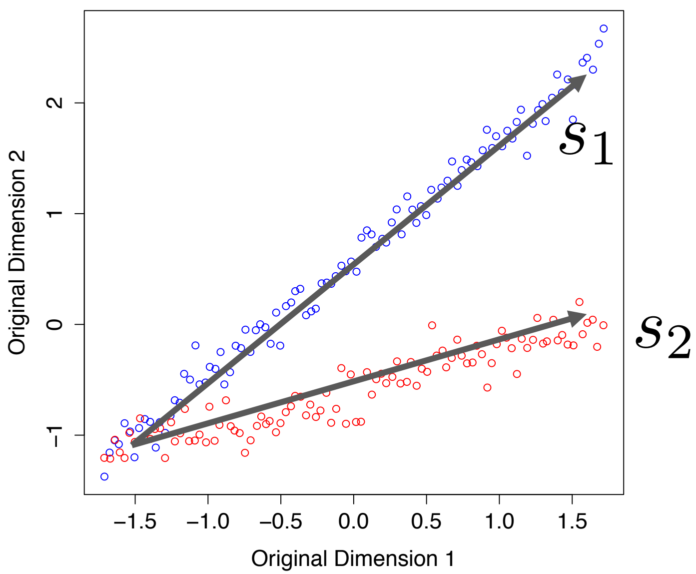
- s_i must be statistically independent
  - $p(s_1,s_2) = p(s_1)p(s_2)$
- ICA not for Gaussian distributions
  - The joint density of unit variance s_1 & s_2 is symmetric.
    So no information about the directions (col vectors) in mixing matrix H.
    Thus, H can't be estimated.
  - If only one IC is gaussian, the estimation is still possible.

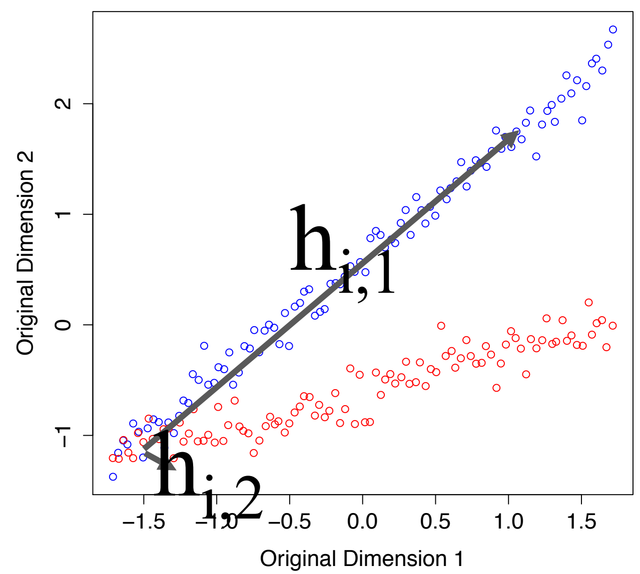


$$p(x_1, x_2) = \frac{1}{2\pi}\exp\left(-\frac{x_1^2 + x_2^2}{2}\right)$$

ICA Linear representation

$$x_i = h_{i,1}s_1 + h_{i,2}s_2$$





- Find vectors that describe the data set the best.
- Each point: linear combination of

$$x_i = h_{i,1}s_1 + h_{i,2}s_2$$

ICA Uniqueness

- If $\mathbf{s}$ has independent components s_i, so has $\mathbf{\Lambda P s}$ where $\mathbf{\Lambda}$ is invertible diagonal and $\mathbf{P}$ is a permutation

- If $(\mathbf{H}, \mathbf{s})$ is a solution, then $\mathbf{H\Lambda P}$ **and** $\mathbf{P}^{\mathrm{T}}\mathbf{\Lambda}^{-1}\mathbf{s}$ are also solutions.
  - *Essential uniqueness*: unique up to a trivial transformations, i.e. a scale-permutation
  - Whole equivalence class of solutions $\Rightarrow$ Just find one representative solution

An ICA Algorithm (Example of noise-free ICA)

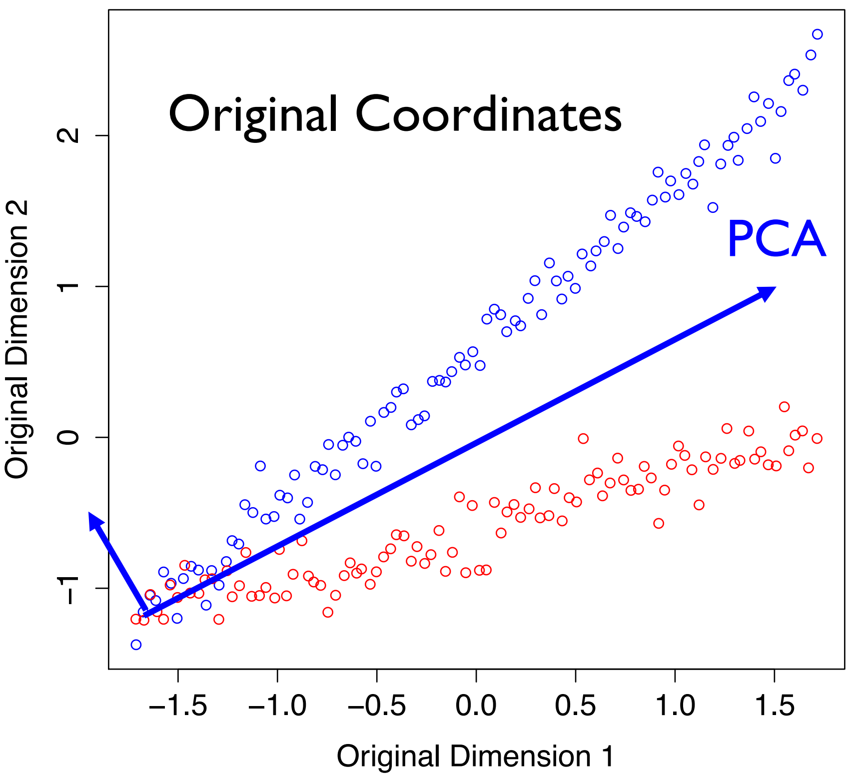
- Observed values $\mathbf{x}$ such that $\mathbf{x} = \mathbf{H}\,\mathbf{s}$
- One way is to minimize mutual information
  - Equivalent to the well known Kullback-Leibler divergence between the joint density of $\mathbf{s}$ and the product of its marginal densities
- Define distribution family of $\mathrm{s_i} \sim g$ **(assumed known, often tanh)**
- Let $W = H^{-1}$
- Recall sources are independent $p(x) = \prod_{i=1}^{N} p_s(w_i^T x) \cdot |\det W|^N$
- Given a training set $\{\mathrm{x}^{(\mathrm{i})};\mathrm{i} = 1,\ldots,\mathrm{N}\}$, the log likelihood of $\mathbf{W}$ is given by

$$\log P[X|W] = \sum_{i=1}^{N} \left(\sum_{j=1}^{p} \log g'(w_j^T x^{(i)}) + N \log|\det W| \right)$$

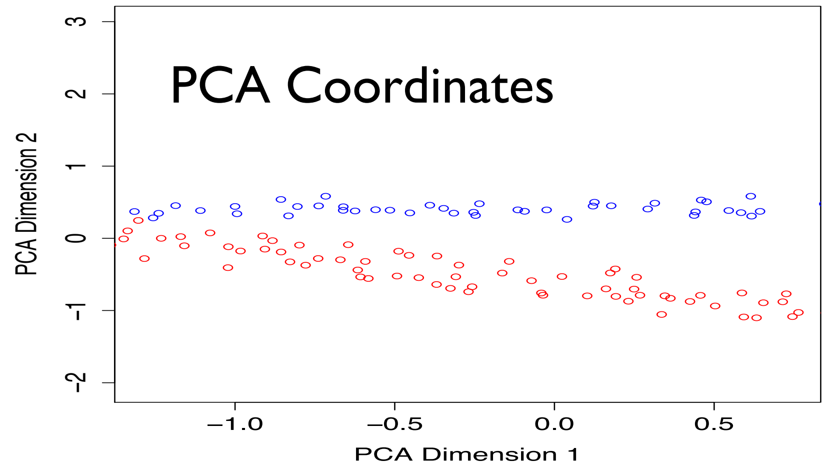
(Pham et al. 1992) D.T. Pham, P. Garrat and C. Jutten, Separation of a mixture of independent sources through a maximum likelihood approach, EUSIPC, 1992

PCA v.s. ICA

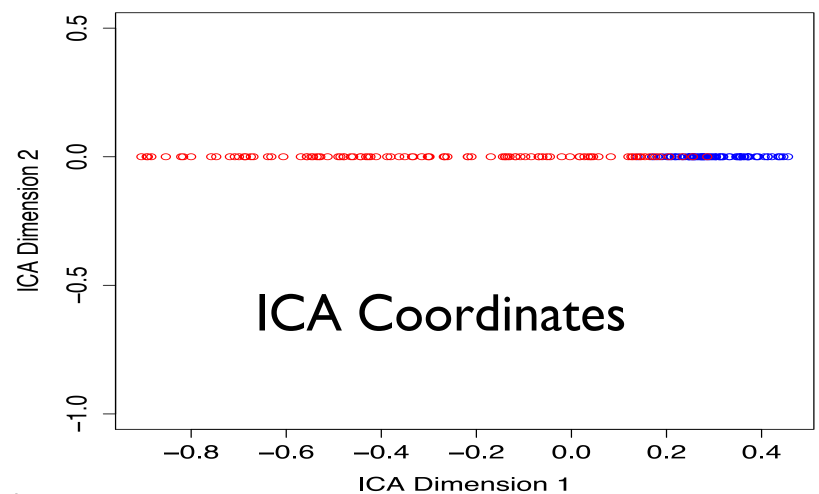
- Which method (PCA or ICA) provides best projection?



PCA orthogonality bad for non-orthogonal data



ICA: forcing independence but not orthogonality shows true dimension of data



R Code: https://www.cs.purdue.edu/homes/ribeirob/courses/Spring2016/extra/PCA_vs_ICA.R

Correlation vs Independence

- Example 1: Mixture of 2 identically distributed sources
- Consider the mixture of two independent sources

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} s_1 \\ s_2 \end{pmatrix}$$

- where $E[s^2_i] = 1$ and $E[s_i] = 0$.
- Then x_i are uncorrelated as

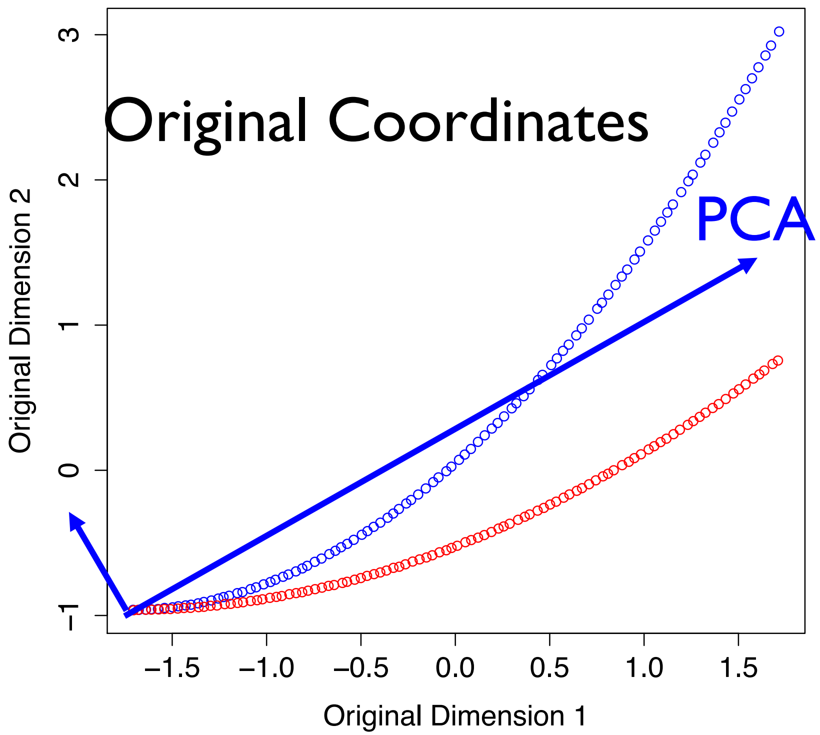
$$E[x_1 x_2] - E[x_1]E[x_2] = E[s_1^2] - E[s_2^2] = 0$$

But x_i are not independent since, say

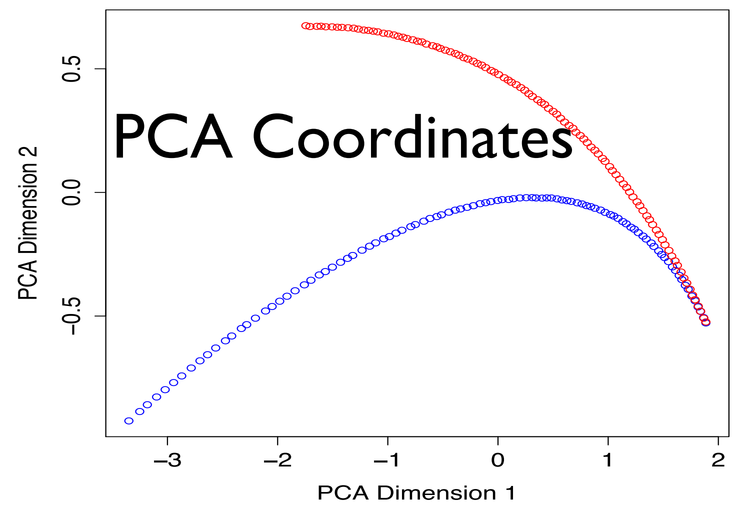
$$E[x_1^2 x_2^2] - E[x_1^2]E[x_2^2] = E[s_1^4] + E[s_2^4] - 6 \neq 0$$

PCA v.s. ICA (Round 2)

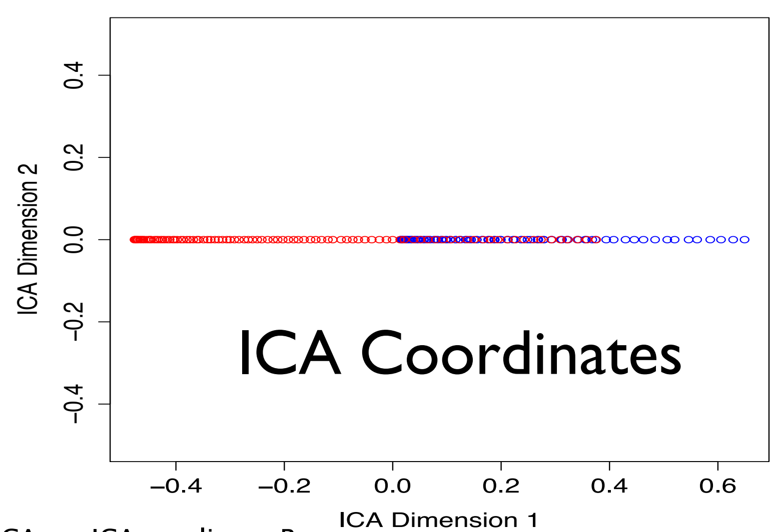
- Which method (PCA or ICA) provides best projection?



PCA



ICA (works better even though data is non-linear)



R Code: https://www.cs.purdue.edu/homes/ribeirob/courses/Spring2016/extra/PCA_vs_ICA_nonlinear.R