Principal Component Analysis and Linear Discriminant Analysis for Feature Reduction

Jieping Ye

Department of Computer Science and Engineering

Arizona State University

http://www.public.asu.edu/~jye02

Outline of lecture

- What is feature reduction?

- Why feature reduction?

- Feature reduction algorithms
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)

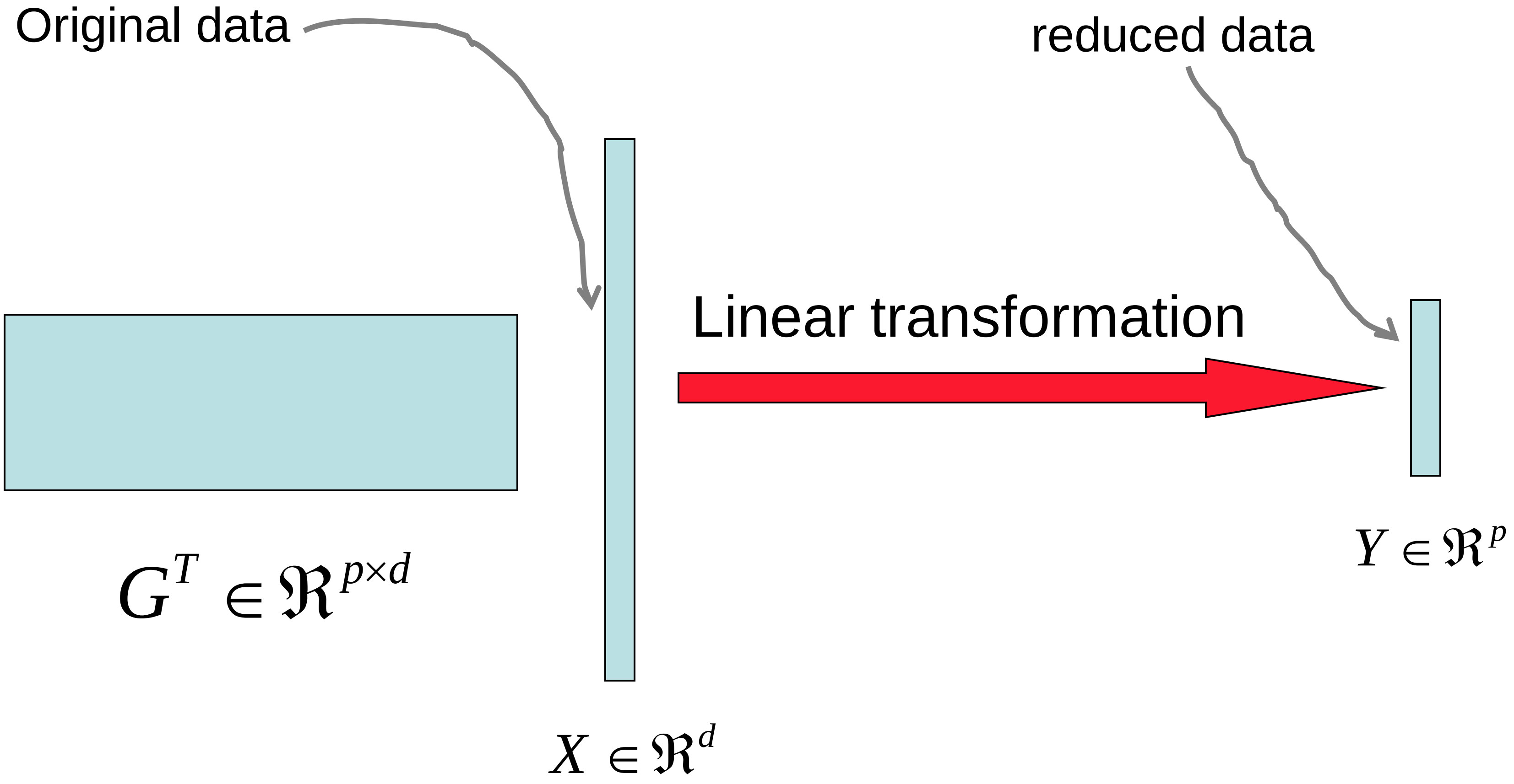
What is feature reduction?

- Feature reduction refers to the mapping of the original high-dimensional data onto a lower-dimensional space.
  - Criterion for feature reduction can be different based on different problem settings.
    - Unsupervised setting: minimize the information loss
    - Supervised setting: maximize the class discrimination

- Given a set of data points of p variables $\{x_1, x_2, \cdots, x_n\}$
  Compute the linear transformation (projection)

$$G \in \Re^{d \times p} : x \in \Re^d \rightarrow y = G^T x \in \Re^p \ (p << d)$$

What is feature reduction?



Original data

reduced data

Linear transformation

$G^T \in \Re^{p \times d}$

$X \in \Re^d$

$Y \in \Re^p$

$$G \in \Re^{d \times p} : X \to Y = G^T X \in \Re^p$$

Feature reduction versus feature selection

- Feature reduction
  - All original features are used
  - The transformed features are linear combinations of the original features.
- Feature selection
  - Only a subset of the original features are used.
- Continuous versus discrete

Outline of lecture

- What is feature reduction?
- Why feature reduction?
- Feature reduction algorithms
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)

Why feature reduction?

- Most machine learning and data mining techniques may not be effective for high-dimensional data
  - Curse of Dimensionality
  - Query accuracy and efficiency degrade rapidly as the dimension increases.

- The intrinsic dimension may be small.
  - For example, the number of genes responsible for a certain type of disease may be small.

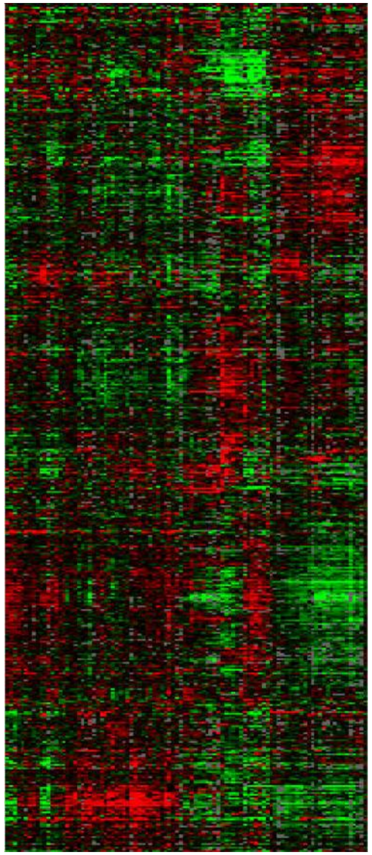
Why feature reduction?

- Visualization: projection of high-dimensional data onto 2D or 3D.
- Data compression: efficient storage and retrieval.
- Noise removal: positive effect on query accuracy.

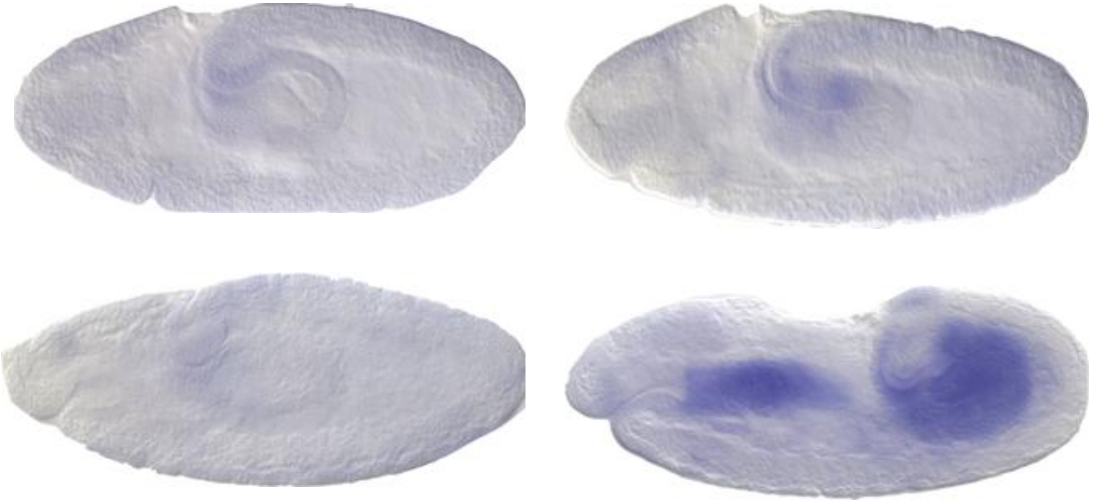
Applications of feature reduction

- Face recognition
- Handwritten digit recognition
- Text mining
- Image retrieval
- Microarray data analysis
- Protein classification

High-dimensional data in bioinformatics



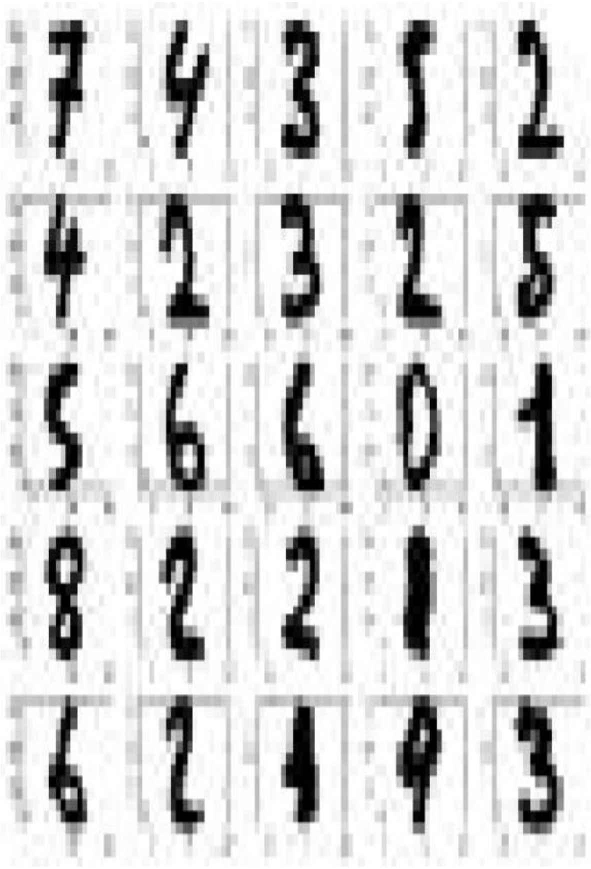
Gene expression



Gene expression pattern images

High-dimensional data in computer vision

Face images



Handwritten digits

Outline of lecture

- What is feature reduction?
- Why feature reduction?
- Feature reduction algorithms
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)

Feature reduction algorithms

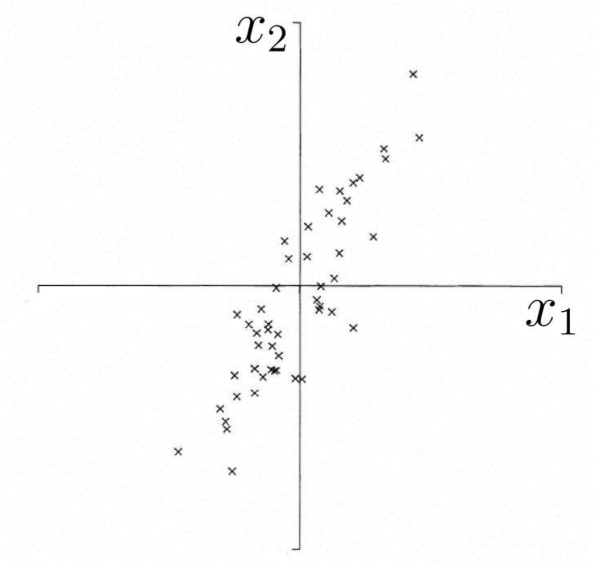
- Unsupervised
  - Latent Semantic Indexing (LSI): truncated SVD
  - Independent Component Analysis (ICA)
  - Principal Component Analysis (PCA)
  - Canonical Correlation Analysis (CCA)
- Supervised
  - Linear Discriminant Analysis (LDA)
- Semi-supervised
  - Research topic

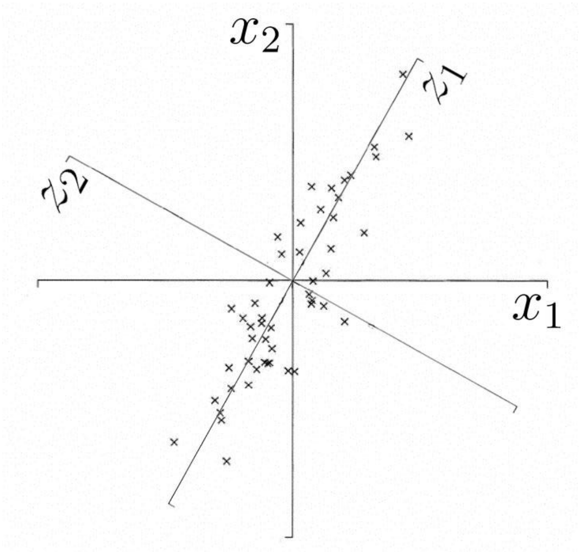
What is Principal Component Analysis?

- Principal component analysis (PCA)
  - Reduce the dimensionality of a data set by finding a new set of variables, smaller than the original set of variables
  - Retains most of the sample's information.
  - Useful for the compression and classification of data.

- By information we mean the variation present in the sample, given by the correlations between the original variables.
  - The new variables, called principal components (PCs), are uncorrelated, and are ordered by the fraction of the total information each retains.

Geometric picture of principal components (PCs)



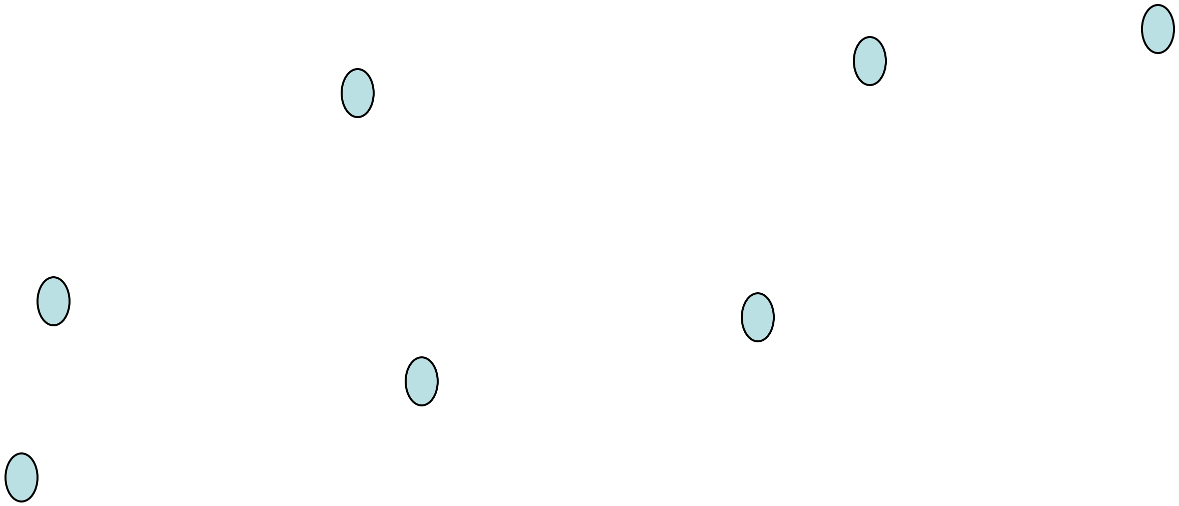


- the 1st PC $\boldsymbol{z}_1$ is a minimum distance fit to a line in X space

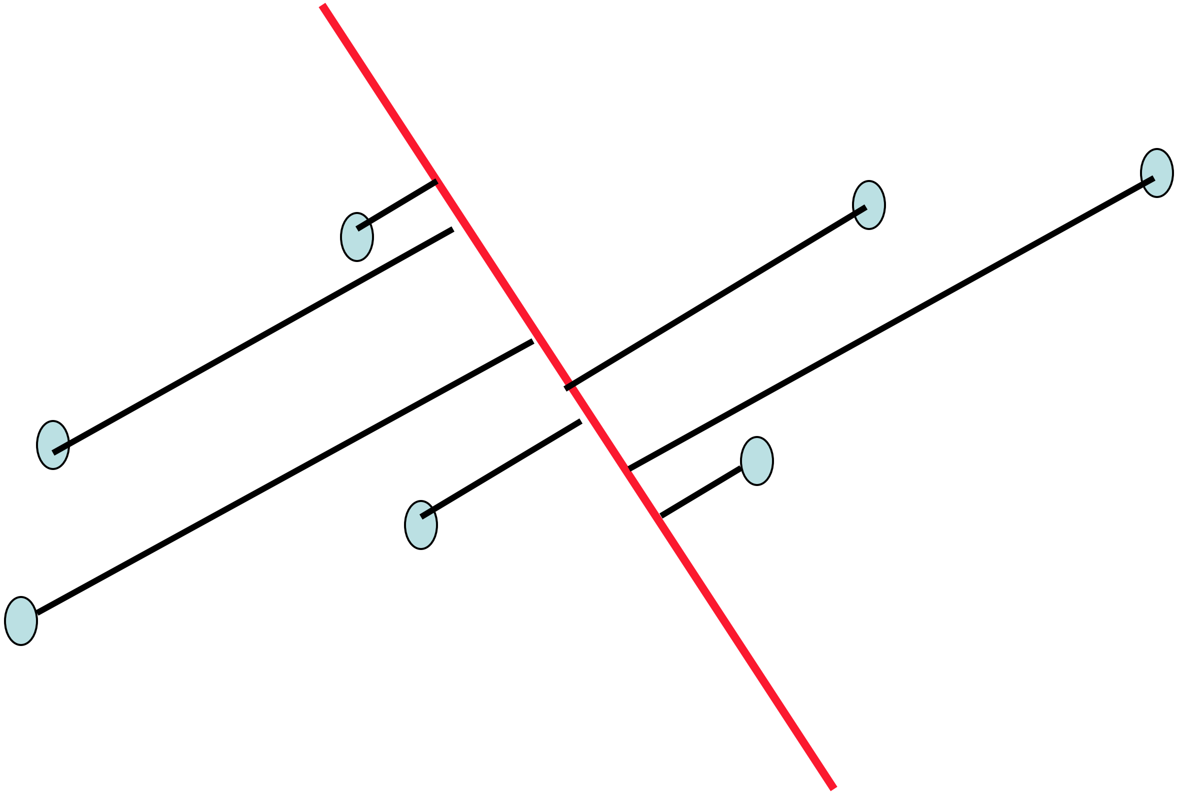
- the 2nd PC $\boldsymbol{z}_2$ is a minimum distance fit to a line in the plane perpendicular to the 1st PC

PCs are a series of linear least squares fits to a sample, each orthogonal to all the previous.

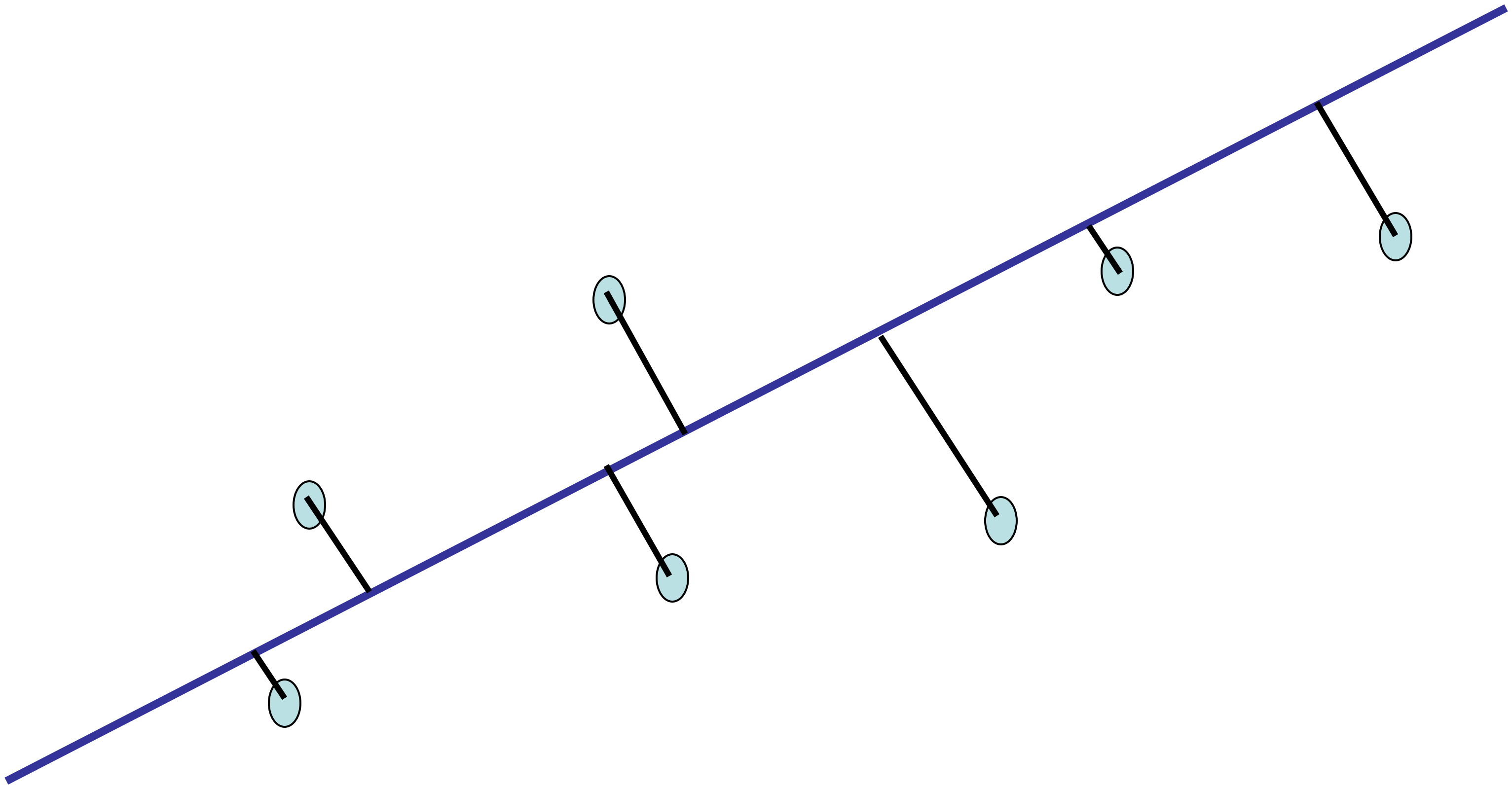
Geometric picture of principal components (PCs)



Geometric picture of principal components (PCs)



Geometric picture of principal components (PCs)



Algebraic definition of PCs

Given a sample of *n* observations on a vector of *p* variables

$$\{x_1, x_2, \cdots, x_n\} \in \Re^d$$

define the first principal component of the sample by the linear transformation

$$z_1 = a_1^T x_j = \sum_{i=1}^{d} a_{i1} x_{ij}, \quad j = 1,2,\cdots,n.$$

where the vector

$$a_1 = (a_{11}, a_{21}, \cdots, a_{d1})$$

$$x_j = (x_{1j}, x_{2j}, \cdots, x_{dj})$$

is chosen such that $\mathrm{var}[z_1]$ is maximum.

Algebraic derivation of PCs

To find a_1 first note that

$$\mathrm{var}[z_1] = E((z_1 - \overline{z_1})^2) = \frac{1}{n}\sum_{i=1}^{n}\left(a_1^T x_i - a_1^T \overline{x}\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} a_1^T \left(x_i - \overline{x}\right)\left(x_i - \overline{x}\right)^T a_1 = a_1^T S a_1$$

where $$S = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(x_i - \overline{x}\right)^T$$

is the covariance matrix.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ is the mean.}$$

Algebraic derivation of PCs

To find a_1 that maximizes $\mathrm{var}[z_1]$ subject to $a_1^T a_1 = 1$

Let λ be a Lagrange multiplier

$$L = a_1^T S a_1 - \lambda(a_1^T a_1 - 1)$$

$$\frac{\partial}{\partial a_1} L = S a_1 - \lambda a_1 = 0$$

$$\Rightarrow (S - \lambda I_p) a_1 = 0$$

therefore a_1 is an eigenvector of S

corresponding to the largest eigenvalue $\lambda = \lambda_1$.

Algebraic derivation of PCs

We find that a_2 is also an eigenvector of S

whose eigenvalue $\lambda = \lambda_2$ is the second largest.

In general

$$\mathrm{var}[z_k] = a_k^T S a_k = \lambda_k$$

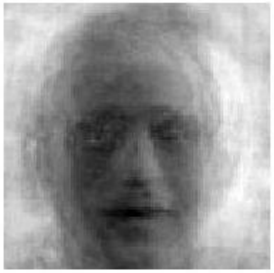
- The *k*th largest eigenvalue of S is the variance of the *k*th PC.
- The *k*th PC z_k retains the *k*th greatest fraction of the variation in the sample.

Algebraic derivation of PCs

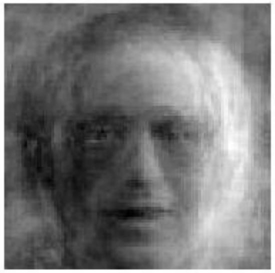
- Main steps for computing PCs
  - Form the covariance matrix S.
  - Compute its eigenvectors: $\left\{a_i\right\}_{i=1}^{d}$
  - The first p eigenvectors $\left\{a_i\right\}_{i=1}^{p}$ form the p PCs.
  - The transformation G consists of the p PCs:

$$G \leftarrow [a_1, a_2, \cdots, a_p]$$

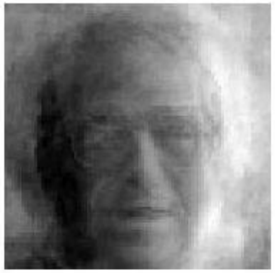
PCA for image compression



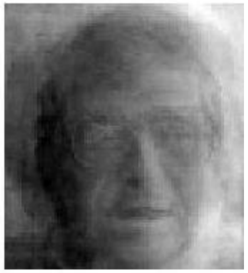
p=1



p=2

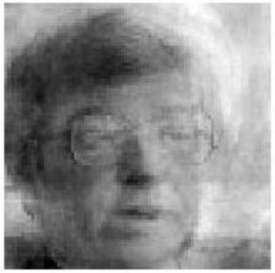


p=4

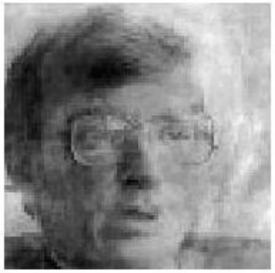


p=8

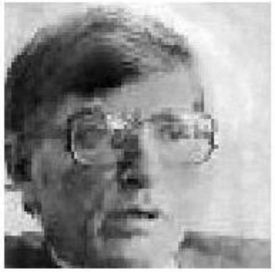
p=16



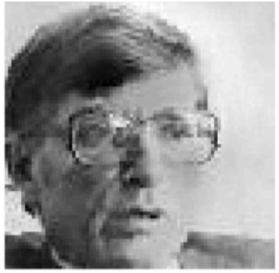
p=32



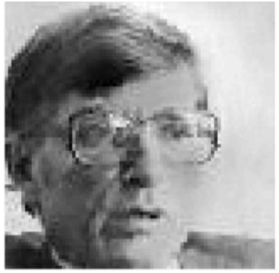
p=64



p=100



Original Image



Outline of lecture

- What is feature reduction?
- Why feature reduction?
- Feature reduction algorithms
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis

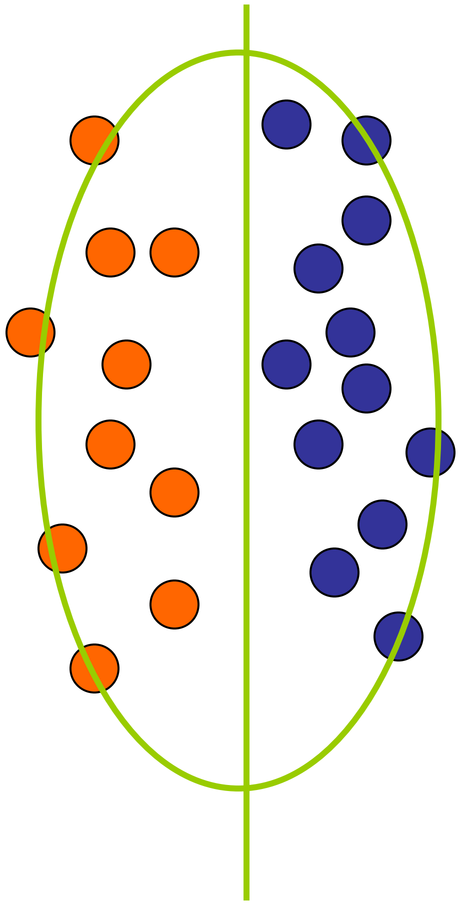
- First applied by M. Barnard at the suggestion of R. A. Fisher (1936), Fisher linear discriminant analysis (FLDA):

  - Dimension reduction
    - Finds linear combinations of the features $\mathbf{X}$=X_1,...,X_d with large ratios of between-groups to within-groups sums of squares - discriminant variables;

  - Classification
    - Predicts the class of an observation **X** by the class whose mean vector is closest to **X** in terms of the discriminant variables

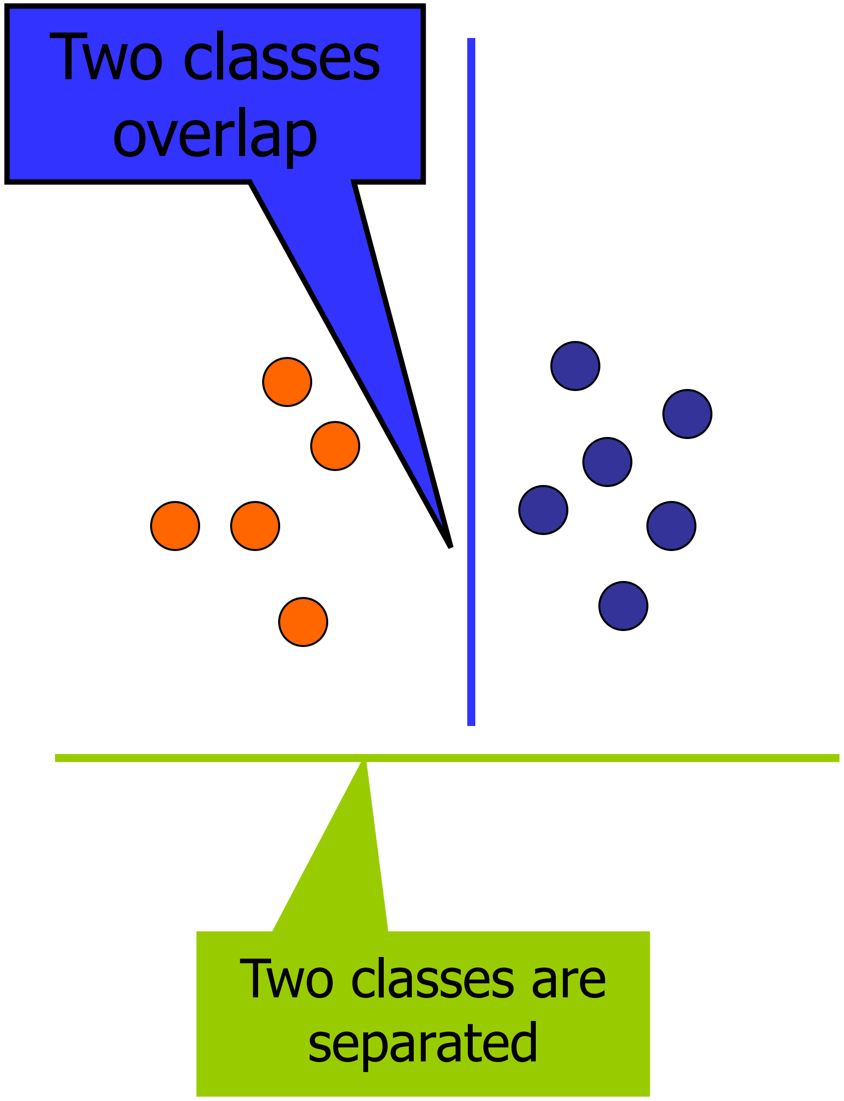
Is PCA a good criterion for classification?

- Data variation determines the projection direction
- What’s missing?
  - Class information



What is a good projection?

- Similarly, what is a good criterion?
  - Separating different classes

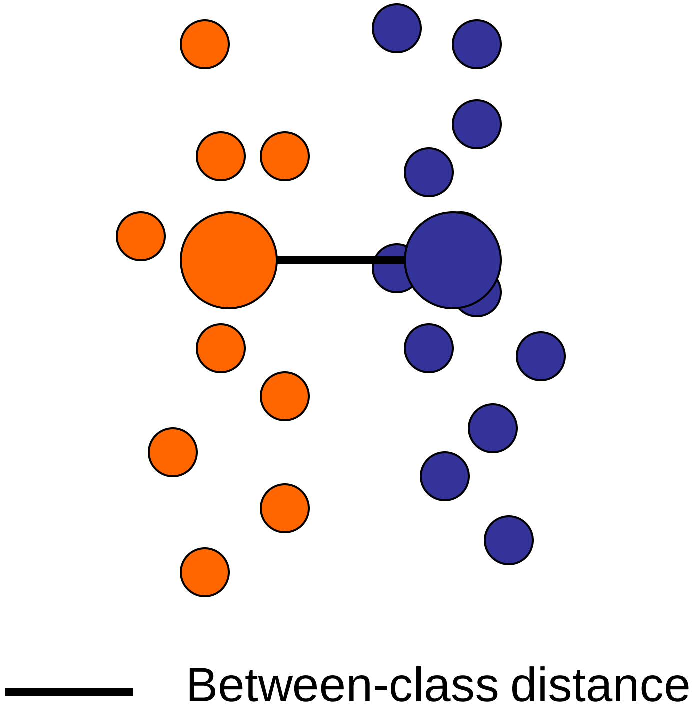


Two classes overlap

Two classes are separated

What class information may be useful?

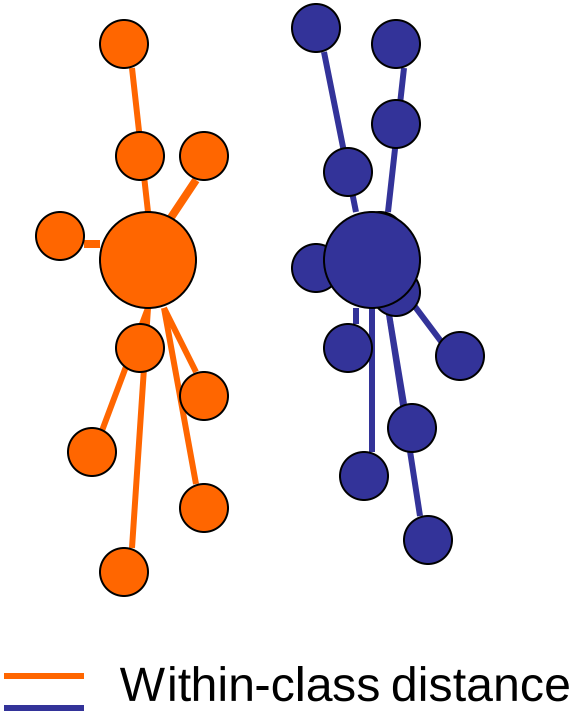
- Between-class distance
  - Distance between the centroids of different classes



What class information may be useful?

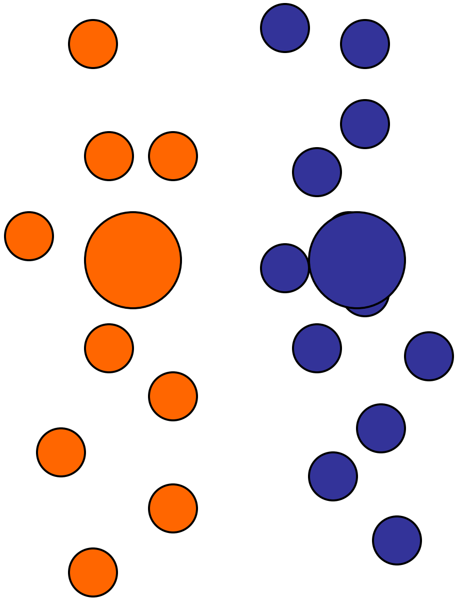
- Between-class distance
  - Distance between the centroids of different classes

- Within-class distance
  - Accumulated distance of an instance to the centroid of its class



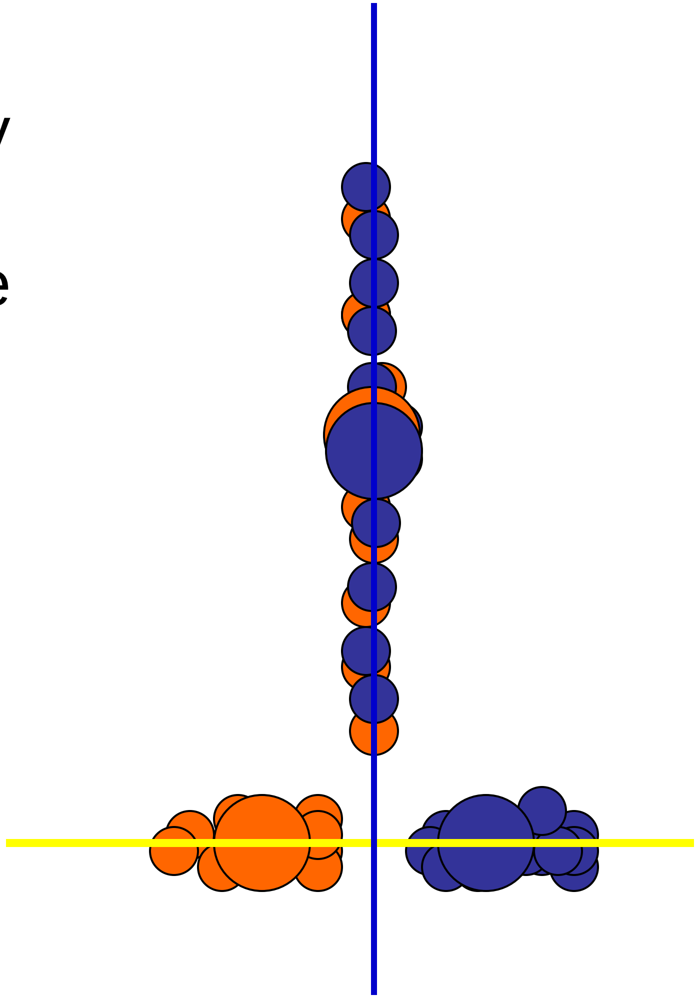
Linear discriminant analysis

- Linear discriminant analysis (LDA) finds most discriminant projection by maximizing between-class distance and minimizing within-class distance



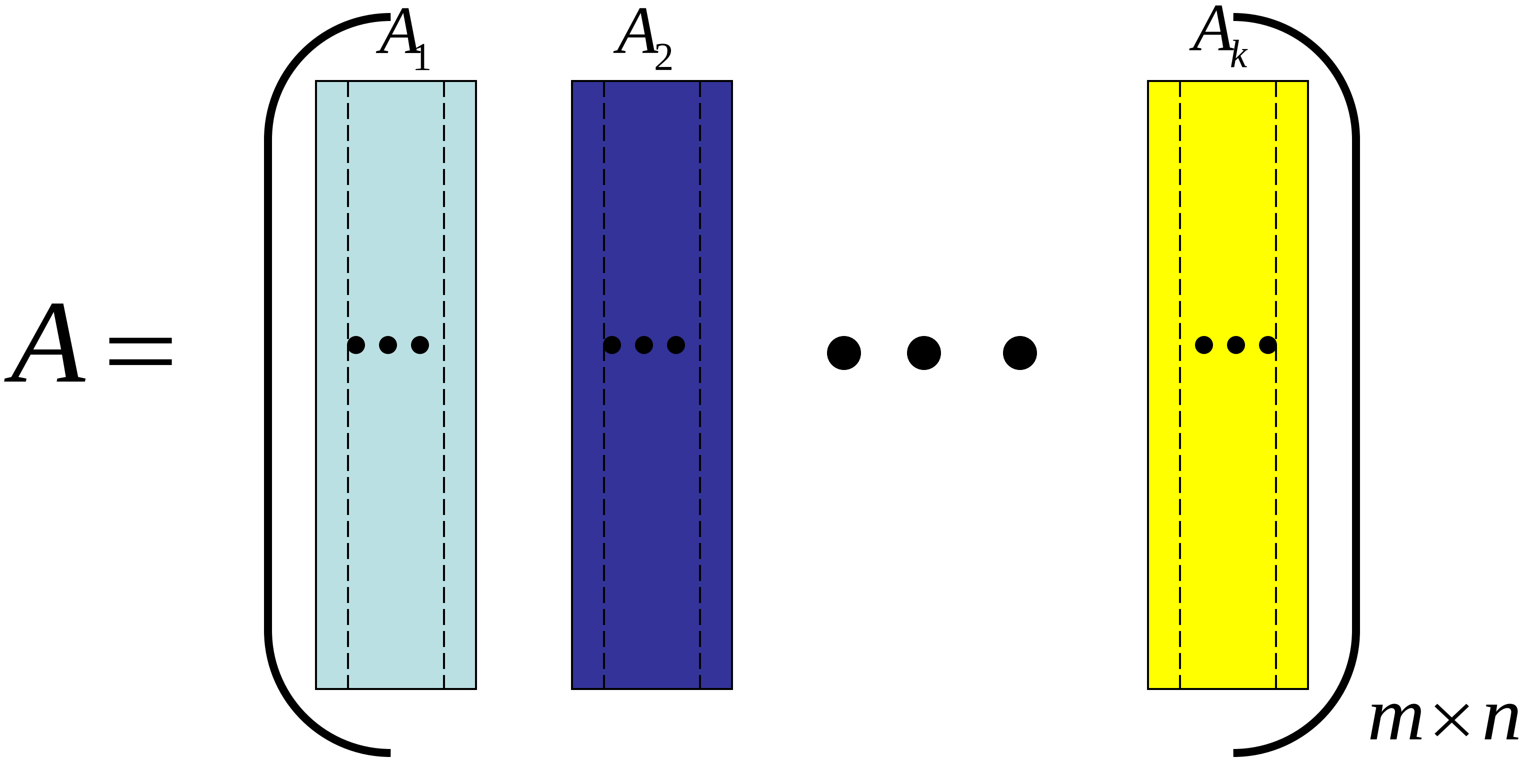
Linear discriminant analysis

- Linear discriminant analysis (LDA) finds most discriminant projection by maximizing between-class distance and minimizing within-class distance



Notations

Training data from different from 1, 2, …, k



Data matrix

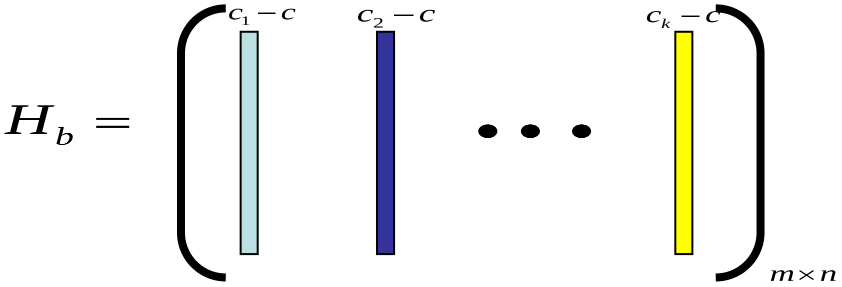
Notations

- Between-class scatter

$$S_b = H_b H_b^T$$

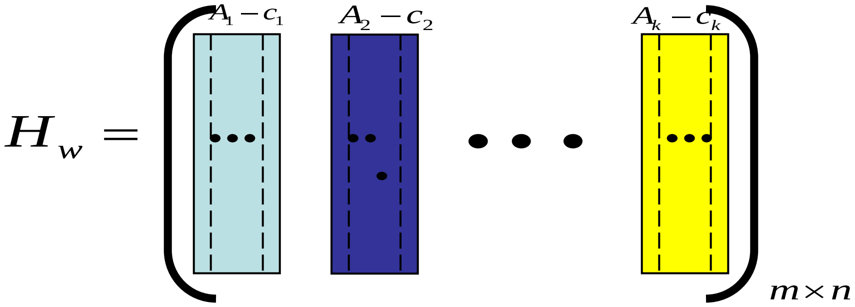
c_i is the centroid of ith class

c is the centroid of all classes



- Within-class scatter

$$S_w = H_w H_w^T$$



- Properties:
  - Between-class distance = trace of between-class scatter (I.e., the summation of diagonal elements of the scatter)
  - Within-class distance = trace of within-class scatter

Discriminant criterion

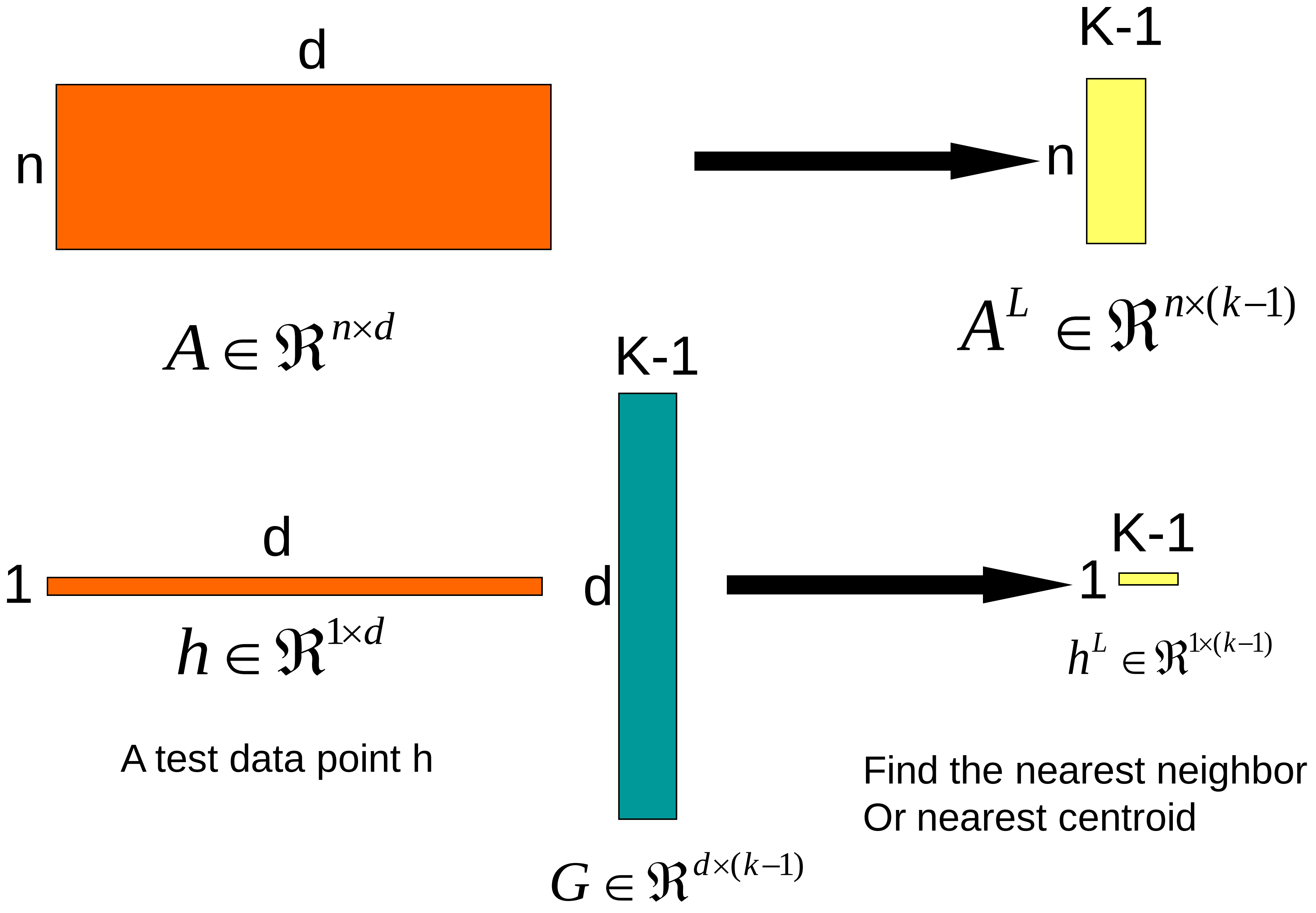
- Discriminant criterion in mathematical formulation

$$\arg\max_{G} \frac{\mathrm{trace}(G^T S_b G)}{\mathrm{trace}(G^T S_w G)}$$

  - Between-class scatter matrix S_b
  - Within-class scatter matrix S_w

- The optimal transformation is given by solving a generalized eigenvalue problem $S_w^{-1} S_b$

Graphical view of classification



$A \in \Re^{n \times d}$

$A^L \in \Re^{n \times (k-1)}$

$h \in \Re^{1 \times d}$

$h^L \in \Re^{1 \times (k-1)}$

$G \in \Re^{d \times (k-1)}$

A test data point h

Find the nearest neighbor
Or nearest centroid

Applications

- Face recognition
  - Belhumeour *et al.*, PAMI'97
- Image retrieval
  - Swets and Weng, PAMI'96
- Gene expression data analysis
  - Dudoit *et al.,* JASA'02; Ye *et al.,* TCBB'04
- Protein expression data analysis
- Lilien *et al.,* Comp. Bio.'03
- Text mining
  - Park *et al.,* SIMAX'03; Ye *et al.*, PAMI'04
- Medical image analysis
  - Dundar, SDM'05

Issues in LDA

- S_w is required to be nonsingular.
  - Singularity or undersampled problem (when n<d)
  - Example: gene expression data (d is around few thousands and n is around few hundreds), images, text documents

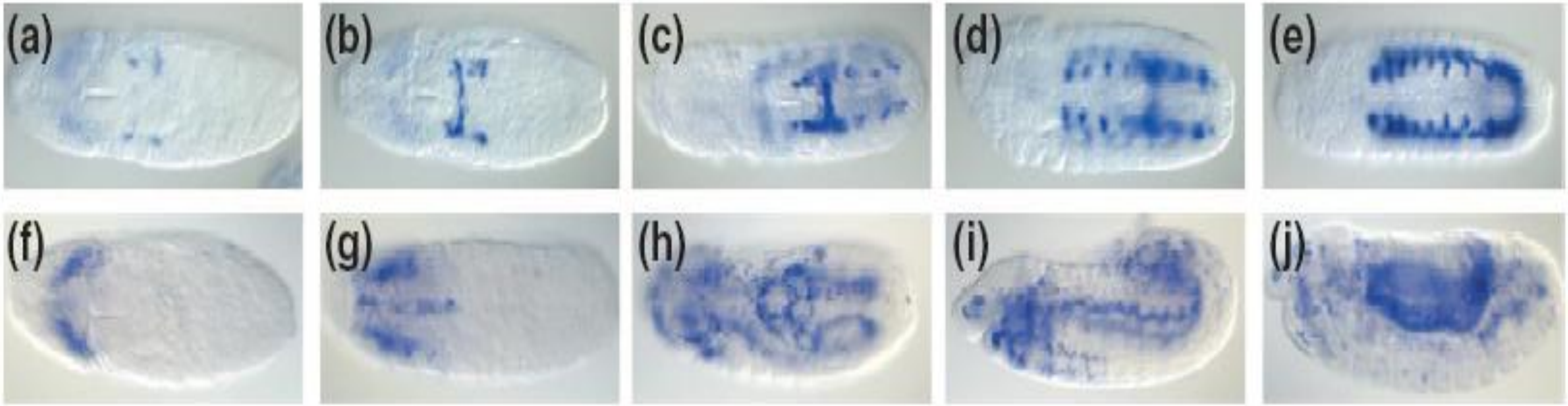
- Approaches
  - PCA+LDA (PCA: Principal Component Analysis)
  - Regularized LDA:
  - Uncorrelated LDA
  - Orthogonal LDA

Summary

- Feature reduction is an important pre-processing step in many applications.

- Unsupervised versus supervised
  - PCA and LDA

- Research problems:
  - Semi-supervised feature reduction
  - Nonlinear feature reduction
  - Determination of the reduced dimension in PCA



(a-e) Series of five embryos stained with a probe (*bgm*)
(f-j) Series of five embryos stained with a probe (*CG4829*)

- Computational and theoretical issues in machine learning and data mining
  - Dimensionality reduction
  - Clustering and classification
  - Semi-supervised learning
  - Kernel methods

– Their applications to bioinformatics
  - Expression pattern images
  - Microarray gene expression data
  - Protien sequences and structures

- *Are there any other expression patterns that are similar to the pattern I have observed*?
- *Which genes show extensive overlap in expression patterns*?
- *What is the extent and location of the overlap between gene expression patterns*?
- *Is there a change in the expression pattern of a gene when another gene's expression is altered*?

Project:
Machine learning approaches for biological image informatics

To answer the above questions, investigators generally rely on their own, a collaborator's, or senior mentor's knowledge, which has been gained by following the published literature over many years or even decades. It does not scale to enormous data.

We propose to develop computational approaches for answering these questions automatically.