Linear Dimensionality Reduction

Outline

- Introduction
- Principal Component Analysis
- Factor Analysis
- Multidimensional Scaling
- Linear Discriminant Analysis

Introduction

Why Dimensionality Reduction?

- Scientific: understand structure of data (visualization)
- Statistical: fewer dimensions allows better generalization
- Computational: compress data for efficiency (both time/space)
- Direct: use as a model for anomaly detection

Feature Selection vs. Extraction

- Feature selection:
  - Choosing $K < D$ important features and discarding the remaining D-K features.
  - Subset selection algorithms
- Feature extraction:
  - Projecting the original D dimensions to $K(< D)$ new dimensions.
  - Unsupervised methods (without using output information):
    - Principal component analysis (PCA)
    - Factor analysis (FA)
    - Multidimensional scaling (MDS)
  - Supervised methods (using output information):
    - Linear discriminant analysis (LDA)
  - The linear methods above also have nonlinear extensions.

预备知识——特征向量和特征矩阵

- 设A有n个特征值及特征向量，则

$$
\begin{aligned}
A * x_1 &= \lambda_1 * x_1 \\
A * x_2 &= \lambda_2 * x_2 \\
&\vdots \\
A * x_n &= \lambda_n * x_n
\end{aligned}
$$

- 将上面的写到一起成矩阵形式：

$$
A * \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix} * \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots\dots\dots & & & \\ 0 & 0 & \dots & \lambda_n \end{pmatrix}
$$

- 若(x1, x2, . . . , xn)可逆，则左右两边都求逆，则方阵A可直接通过特征值和特征向量进行唯一的表示，令

$$Q=(x_{1,}, x_2, \dots, x_{n,})$$

$$\Sigma = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$$

上述公式表示为：$A = Q\Sigma Q^{-1}$

协方差矩阵

- 样本X和样本Y的协方差(Covariance):

$$Cov(X,Y)=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{(n-1)}$$

- 当样本是n维数据时，它们的协方差实际上是协方差矩阵（对称方阵）
  - 比如对于3维数据(x, y, z)，计算它的协方差就是:

$$C=\begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix}$$

Principal Component Analysis

Principal Component Analysis

- PCA finds a linear mapping from the D-dimensional input space to a K-dimensional space (K < D) with minimum information loss according to some criterion.
- Projection of x on the direction of w: $z = \mathbf{w}^T\mathbf{x}$
- Finding the first principal component w1 s.t. var(z1) is maximized:

$$\begin{aligned}\operatorname{var}(z_1) = \operatorname{var}(\mathbf{w}_1^T\mathbf{x}) &= E[(\mathbf{w}_1^T\mathbf{x} - \mathbf{w}_1^T\mu)^2] \\ &= E[\mathbf{w}_1^T(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T\mathbf{w}_1] \\ &= \mathbf{w}_1^T E[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T]\mathbf{w}_1 = \mathbf{w}_1^T\mathbf{\Sigma}\mathbf{w}_1\end{aligned}$$

where

$$\operatorname{cov}(\mathbf{x}) = E[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T] = \mathbf{\Sigma}$$

- Maximization of var(z1) subject to ||w1|| = 1 can be solved as a constrained optimization problem using a Lagrange multiplier.
- Maximization of Lagrangian:

$$\mathbf{w}_1^T \Sigma \mathbf{w}_1 - \alpha(\mathbf{w}_1^T \mathbf{w}_1 - 1)$$

- Taking the derivative of the Lagrangian w.r.t. w1 and setting it to 0, we get an eigenvalue equation for the first principal component w1:

$$\Sigma \mathbf{w}_1 = \alpha \mathbf{w}_1$$

- Because we have

$$\mathbf{w}_1^T \Sigma \mathbf{w}_1 = \alpha \mathbf{w}_1^T \mathbf{w}_1 = \alpha$$

- we choose the eigenvector with the largest eigenvalue for the variance to be maximum.

- The second principal component w2 should also maximize the variance var(z2), subject to the constraints that ||w2||= 1 and that w2 is orthogonal to w1.
- Maximization of Lagrangian:

$$\mathbf{w}_2^T \mathbf{\Sigma} \mathbf{w}_2 - \alpha(\mathbf{w}_2^T \mathbf{w}_2 - 1) - \beta(\mathbf{w}_2^T \mathbf{w}_1 - 0)$$

- Taking the derivative of the Lagrangian w.r.t. w2 and setting it to 0, we get the following equation:

$$2\mathbf{\Sigma}\mathbf{w}_2 - 2\alpha\mathbf{w}_2 - \beta\mathbf{w}_1 = 0$$

- We can show that β = 0 and hence have this eigenvalue equation:

$$\mathbf{\Sigma}\mathbf{w}_2 = \alpha\mathbf{w}_2$$

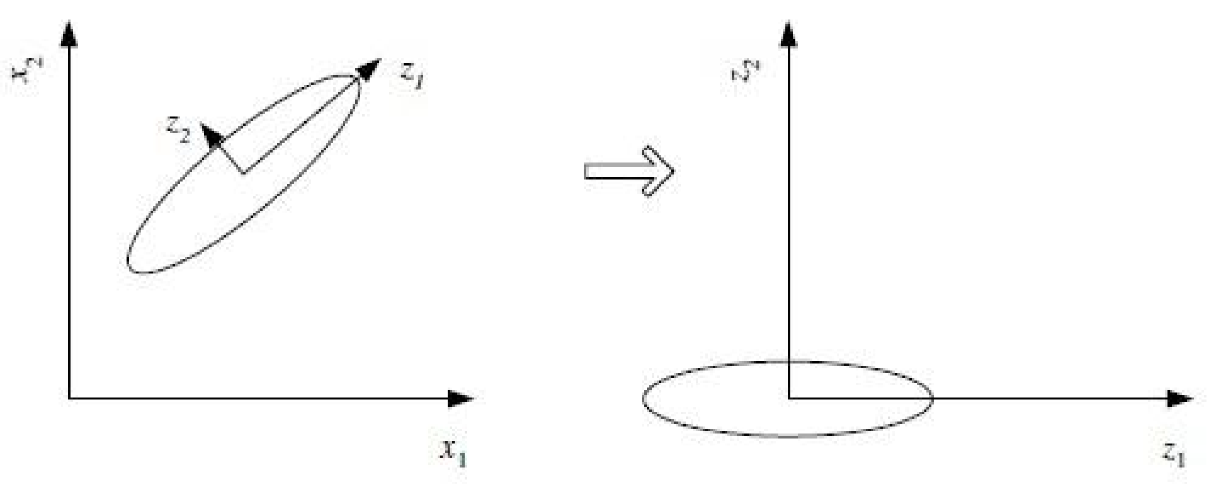
- implying that w2 is the eigenvector of ∑ with the second largest eigenvalue.

What PCA Does

- Transformation of data:

  $z = W^T (x - m)$

- where the columns of $W = [w1; w2; : : :]$ are the eigenvectors of $\sum$ and m is the sample mean.
- Centering the data at the origin and rotating the axes:



If the variance on z2 is too small, it can be ignored to reduce the dimensionality from 2 to 1.

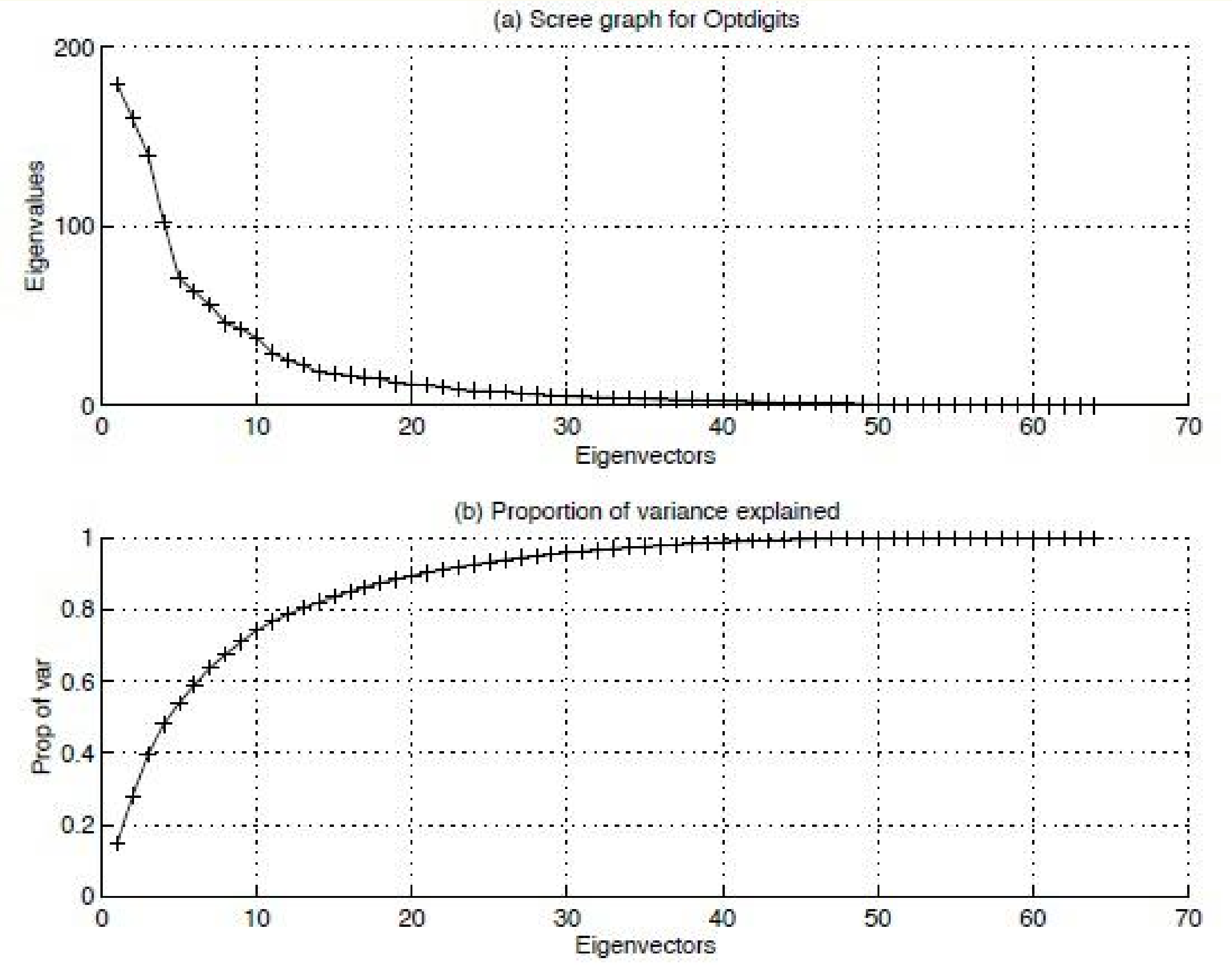
How to Choose K

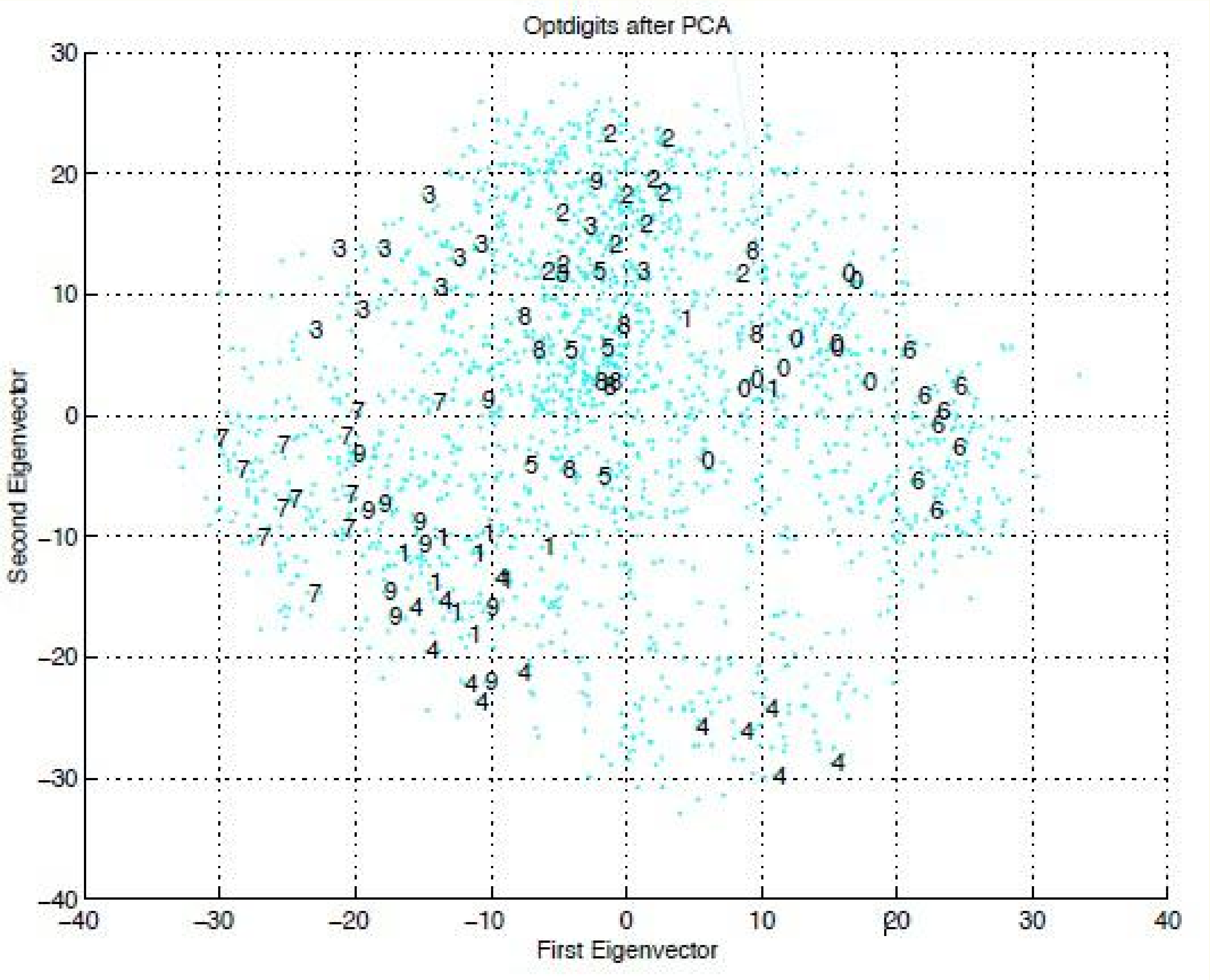
- Proportion of variance (PoV) explained:

$$\frac{\lambda_1 + \lambda_2 + \ldots + \lambda_K}{\lambda_1 + \lambda_2 + \ldots + \lambda_D}$$

where λ_i are sorted in descending order.

- Typically, stop at PoV > 0:9
- Scree graph plotting PoV against K; stop at "elbow".


(a) Scree graph for Optdigits
Eigenvalues
Eigenvectors
200
100
0
0
10
20
30
40
50
60
70
(b) Proportion of variance explained
Prop of var
Eigenvectors
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70


Optdigits after PCA
Second Eigenvector
First Eigenvector

Factor Analysis

Definition

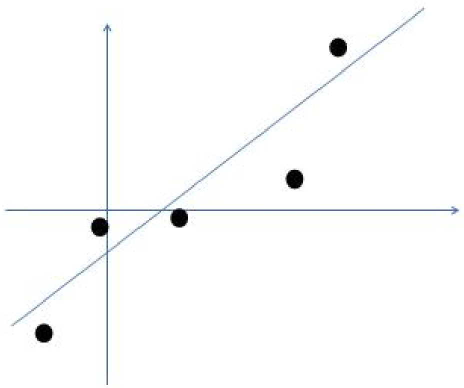
- FA assumes that there is a set of latent factors z_j which when acting in combination to generate the observed variables **x**.
- The goal of FA is to characterize the dependency among the observed variables by means of a smaller number of factors.
- $X = \mu + AF + \varepsilon$
- What exactly that mean?

A simple Example

- $X = \mu + AF + \varepsilon$
- It is easy to say whether a lady is beautiful or not, but the definition of beautiful can be very different. But we can attribute it to outer and inter beauty, which relates to the “F” above
- besides, the ε relates to some special taste of the individuals, which we don’t care much in FA.

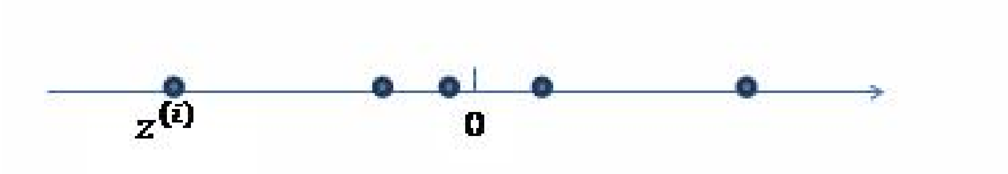
Another example

- Suppose we have 5 samples in a 2-D plane
- Lets prove it could be present with points from 1-D plane

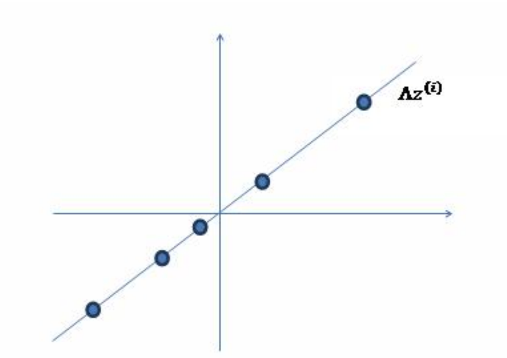


Another example

- Now we have 5 points with Gaussian distribution in the 1-D plane

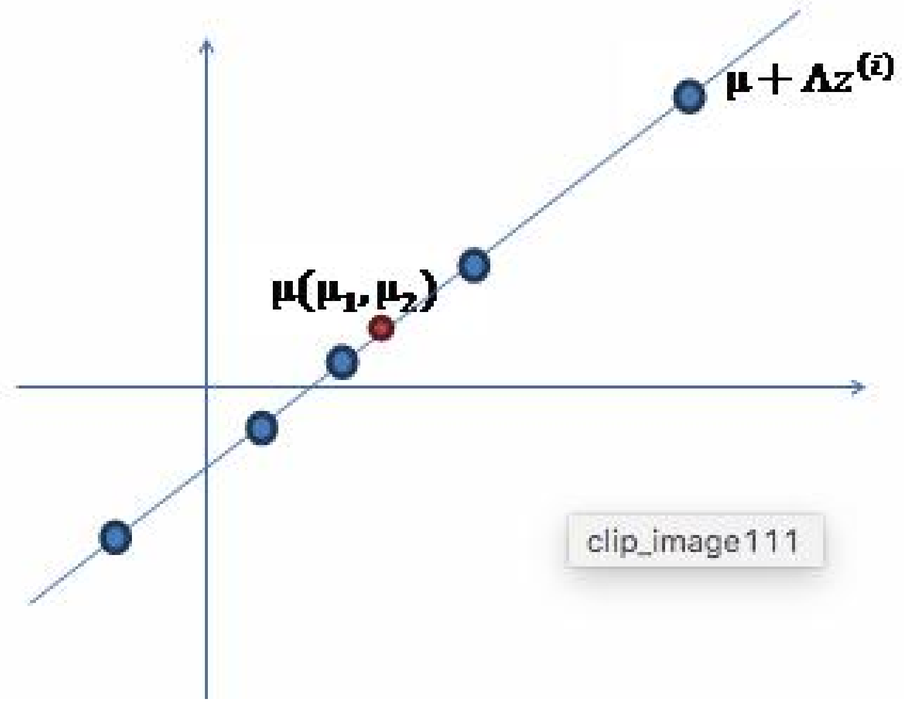


- The reason to choose Gaussian distribution is that its μ = 0 and var = 1
- Now we project it to the 2-D plane with Covariance Matrix



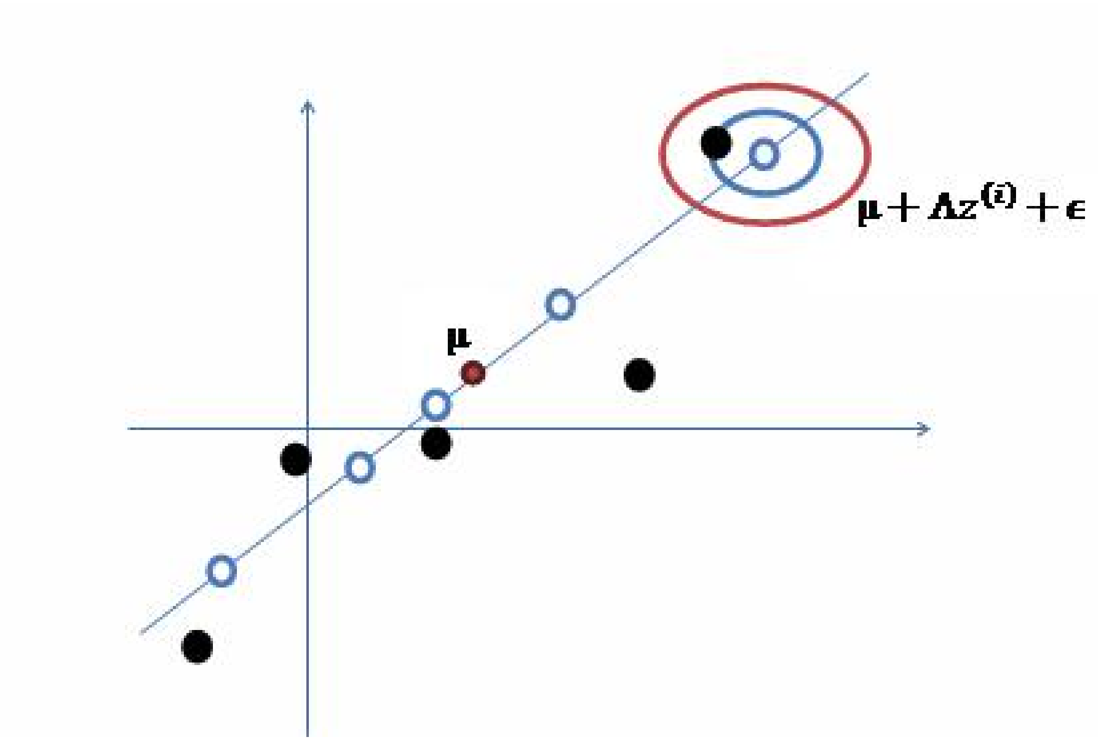
Another example

- Plus the μ



Another example

- Plus the ε



Summary

- From the example above, we know that the samples of high dimensional planes can be reduced to lower dimensional.
- We have already known μ， and we don’t care much of the special variation εj, so our main task in FA is to find A to project Z to X, that is, find the covariance Matrix “A”
- Unfortunately, usually Z could not be found easily, which means we cannot use a function to present Z
- How to solve it?

The solution of covariance Matrix

$$\Lambda = \left(\sum_{i=1}^{m} (x^{(i)} - \mu)\mu_{z^{(i)}|x^{(i)}}^T \right) \left(\sum_{i=1}^{m} \mu_{z^{(i)}|x^{(i)}} \mu_{z^{(i)}|x^{(i)}}^T + \Sigma_{z^{(i)}|x^{(i)}} \right)^{-1}.$$

$$\mu = \frac{1}{m} \sum_{i=1}^{m} x^{(i)}.$$

$$\Phi = \frac{1}{m} \sum_{i=1}^{m} x^{(i)} x^{(i)T} - x^{(i)} \mu_{z^{(i)}|x^{(i)}}^T \Lambda^T - \Lambda \mu_{z^{(i)}|x^{(i)}} x^{(i)T} + \Lambda (\mu_{z^{(i)}|x^{(i)}} \mu_{z^{(i)}|x^{(i)}}^T + \Sigma_{z^{(i)}|x^{(i)}}) \Lambda^T$$

Detailed solution

- http://www.cnblogs.com/jerrylead/archive/2011/05/11/2043317.html

The difference with PCA

- PCA meanly to reduce the dimension, it involves extracting linear composites of observed variables.
- FA is based on a formal model predicting observed variables from theoretical latent factors
- The direction of FA is opposite to that of PCA (from courseware PPT):
  - PCA (from $\mathbf{x}$ to $\mathbf{z}$): $\mathbf{z} = \mathbf{W}T\ (\mathbf{x} - \mu)$
  - FA (from $\mathbf{z}$ to $\mathbf{x}$): $\mathbf{x} - \mu = \mathbf{V}\mathbf{z} + \varepsilon$

Multidimensional Scaling

definition

- Problem formulation:
  - Given the pairwise distances between pairs of points in some space (but the exact coordinates of the points and their dimensionality are unknown).
  - We want to embed the points in a lower-dimensional space such that the pairwise distances in this space are as close as possible to those in the original space.
- The projection to the lower-dimensional space is not unique because the pairwise distances are invariant to such operations as translation, rotation and reflection.

task

- The I objects could have I^2 distances

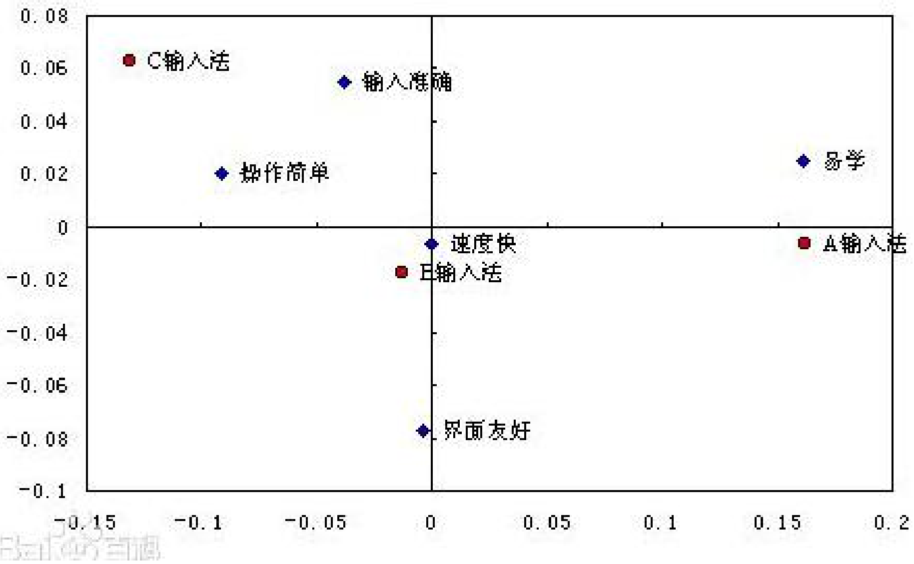
$$\Delta := \begin{pmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,I} \\ \delta_{2,1} & \delta_{2,2} & \cdots & \delta_{2,I} \\ \vdots & \vdots & & \vdots \\ \delta_{I,1} & \delta_{I,2} & \cdots & \delta_{I,I} \end{pmatrix}.$$

- The task of MDS is to find $\|x_i - x_j\| \approx \delta_{i,j}$
- And we can take it as an optimize problem:

$$\min_{x_1,\ldots,x_I} \sum_{i<j} (\|x_i - x_j\| - \delta_{i,j})^2.$$

example

- It located multi variances into 2-D or 3-D planes and calculate their distances to reflect their similarities and difference.
- Perceptual Mapping (知觉图)



Fomula

$$\mathbf{B} = \mathbf{X}\mathbf{X}^T$$

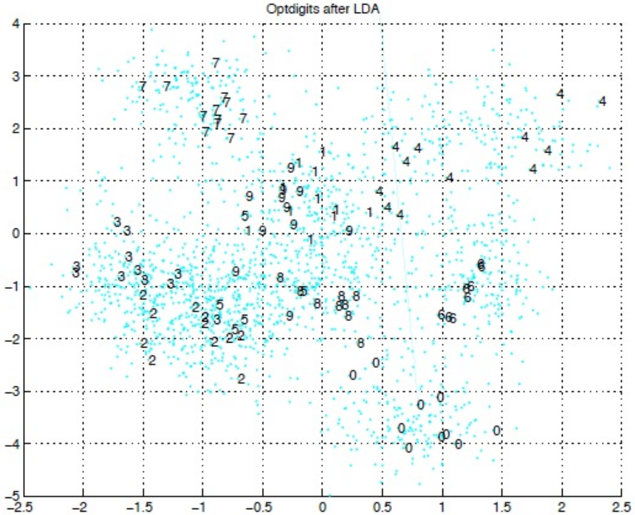
$$\mathbf{B} = \mathbf{C}\mathbf{D}\mathbf{C}^T = \mathbf{C}\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{C}^T = (\mathbf{C}\mathbf{D}^{1/2})(\mathbf{C}\mathbf{D}^{1/2})^T$$

where C is the matrix whose columns are the eigenvectors of B and D1/2 is the diagonal matrix whose diagonal elements are the square roots of the eigenvalues.

Linear Discriminant Analysis

- Unlike PCA, FA and MDS, LDA is a supervised dimensionality reduction method.
- LDA is typically used with a classifier for classification problems.
- Goal: the classes are well-separated after projecting to a low-dimensional space by utilizing the label information (output information).

Example



2-Class Case

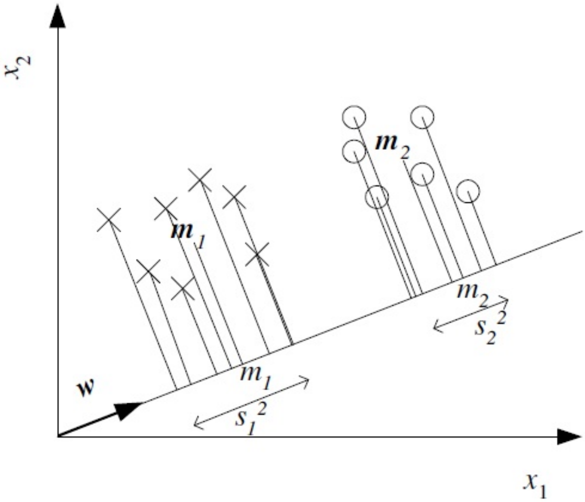
- Given sample $\mathcal{X} = \{(\mathbf{x}^{(i)}, y^{(i)})\}$, where $y^{(i)} = 1$ if $\mathbf{x}^{(i)} \in C_1$ and $y^{(i)} = 0$ if $\mathbf{x}^{(i)} \in C_2$.
- Find vector $\mathbf{w}$ on which the data are projected such that the examples from C_1 and C_2 are as well separated as possible.
- Projection of $\mathbf{x}$ onto $\mathbf{w}$ (dimensionality reduced from D to 1):

$$z = \mathbf{w}^T \mathbf{x}$$

- $\mathbf{m}_j \in \mathbb{R}^D$ and $m_j \in \mathbb{R}$ are sample means of C_j before and after projection:

$$\begin{aligned} m_1 &= \frac{\sum_i \mathbf{w}^T \mathbf{x}^{(i)} y^{(i)}}{\sum_i y^{(i)}} = \mathbf{w}^T \mathbf{m}_1 \\ m_2 &= \frac{\sum_i \mathbf{w}^T \mathbf{x}^{(i)} (1 - y^{(i)})}{\sum_i (1 - y^{(i)})} = \mathbf{w}^T \mathbf{m}_2 \end{aligned}$$

Projection



Between-Class Scatter

- Between-class scatter:

$$
\begin{aligned}
(m_1 - m_2)^2 &= (\mathbf{w}^T\mathbf{m}_1 - \mathbf{w}^T\mathbf{m}_2)^2 \\
&= \mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{w} \\
&= \mathbf{w}^T\mathbf{S}_B\mathbf{w}
\end{aligned}
$$

where

$$
\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T
$$

Within-Class Scatter

- Within-class scatter:

$$
\begin{aligned}
s_1^2 &= \sum_i (\mathbf{w}^T\mathbf{x}^{(i)} - m_1)^2 y^{(i)} \\
&= \sum_i \mathbf{w}^T(\mathbf{x}^{(i)} - \mathbf{m}_1)(\mathbf{x}^{(i)} - \mathbf{m}_1)^T \mathbf{w} y^{(i)} \\
&= \mathbf{w}^T\mathbf{S}_1\mathbf{w}
\end{aligned}
$$

where $\mathbf{S}_1 = \sum_i (\mathbf{x}^{(i)} - \mathbf{m}_1)(\mathbf{x}^{(i)} - \mathbf{m}_1)^T y^{(i)}$. Similarly, $s_2^2 = \mathbf{w}^T\mathbf{S}_2\mathbf{w}$ with $\mathbf{S}_2 = \sum_i (\mathbf{x}^{(i)} - \mathbf{m}_2)(\mathbf{x}^{(i)} - \mathbf{m}_2)^T(1 - y^{(i)})$.
So

$$s_1^2 + s_2^2 = \mathbf{w}^T\mathbf{S}_W\mathbf{w}$$

where $\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$.

Fisher's Linear Discriminant

- Fisher's linear discriminant refers to the vector $\mathbf{w}$ that maximizes the Fisher criterion (a.k.a. generalized Rayleigh quotient):

$$J(\mathbf{w}) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2} = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

- Taking the derivative of J w.r.t. $\mathbf{w}$ and setting it to 0, we obtain the following generalized eigenvalue problem:

$$\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w}$$

or, if $\mathbf{S}_W$ is nonsingular, an equivalent eigenvalue problem:

$$\mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{w} = \lambda \mathbf{w}$$

Fisher's Linear Discriminant (2)

- Alternatively, for the 2-class case, we note that

$$\mathbf{S}_B\mathbf{w} = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{w} = c(\mathbf{m}_1 - \mathbf{m}_2)$$

  for some constant c and hence $\mathbf{S}_B\mathbf{w}$ is in the same direction of $\mathbf{m}_1 - \mathbf{m}_2$.
- So we get

$$\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2) = (\mathbf{S}_1 + \mathbf{S}_2)^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

  The constant factor is irrelevant and hence is discarded.

$K > 2$ Classes

- Find the matrix $\mathbf{W} \in \mathbb{R}^{D \times K}$ such that

$$\mathbf{z} = \mathbf{W}^T \mathbf{x} \in \mathbb{R}^K$$

- Within-class scatter matrix for class C_k:

$$\mathbf{S}_k = \sum_i y_k^{(i)} (\mathbf{x}^{(i)} - \mathbf{m}_k)(\mathbf{x}^{(i)} - \mathbf{m}_k)^T$$

  where $y_k^{(i)} = 1$ if $\mathbf{x}^{(i)} \in C_k$ and 0 otherwise.
- Total within-class scatter matrix:

$$\mathbf{S}_W = \sum_{k=1}^{K} \mathbf{S}_k$$

$K > 2$ Classes (2)

- Between-class scatter matrix:

$$\mathbf{S}_B = \sum_{k=1}^{K} N_k(\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^T$$

where $\mathbf{m}$ is the overall mean and $N_k = \sum_i y_k^{(i)}$.

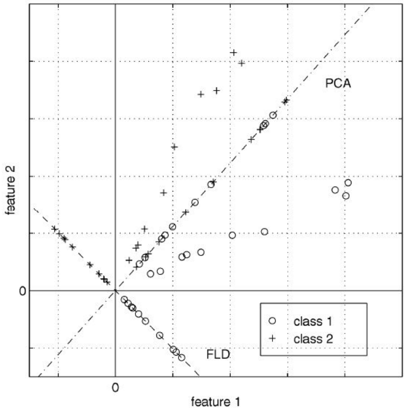
- The optimal solution is the matrix $\mathbf{W}$ that maximizes

$$J(\mathbf{W}) = \frac{\mathrm{Tr}(\mathbf{W}^T\mathbf{S}_B\mathbf{W})}{\mathrm{Tr}(\mathbf{W}^T\mathbf{S}_W\mathbf{W})}$$

which corresponds to the eigenvectors of $\mathbf{S}_W^{-1}\mathbf{S}_B$ with the largest eigenvalues.

- Take new dimensionality $d \leq K - 1$: since $\mathbf{S}_W$ is the sum of K rank-1 matrices and only $K - 1$ of them are independent, $\mathbf{S}_B$ has a maximum rank of $K - 1$.

Application in Face Recognition: PCA vs. LDA



- PCA (Eigenface) maps features to a subspace that contains most energy.
- FLD (Fisherface) maps features to a subspace that most separate the classes.

Application in Face Recognition: PCA vs. LDA (2)

- PCA is an unsupervised dimension reduction algorithm, while LDA is supervised.
- PCA is good at outlier cleaning, and LDA could learn the within-class deviation.
- These two methods only extract 1st and 2nd statistical moments.
- The combination of PCA and LDA could enhance the performance.
- PCA serves as the first-step processing of several kinds of face recognition technique.
- Techniques of dimension reduction are frequently used in face recognition.