Principal Component Analysis

♣ *Motivation*

Principal **C**omponent **A**nalysis (PCA) is a multivariate statistical technique that is often useful in reducing dimensionality of a collection of unstructured random variables for analysis and interpretation.

♣ *Problem Statement*

Let $\mathbf{X}$ be a m-dimensional random vector with covariance matrix C. The problem is to *consecutively* find the unit vectors $\mathbf{a}_1,\ \mathbf{a}_2,\ \ldots,\ \mathbf{a}_m$ such that $y_i = \mathbf{x}^t\mathbf{a}_i$ with $Y_i = \mathbf{X}^t\mathbf{a}_i$ satisfies

1. $\text{var}(Y_1)$ is the maximum.
2. $\text{var}(Y_2)$ is the maximum subject to $\text{cov}(Y_2, Y_1)$=0.
3. $\text{var}(Y_k)$ is the maximum subject to $\text{cov}(Y_k, Y_i)$=0, where $k = 3, 4, \cdots, m$ and $k > i$.

• Y_i is called the *i-th* principal component • Feature extraction by PCA is called PCP

♡ *The Solution*

Let $(\lambda_i, \mathbf{u}_i)$ be the pairs of eigenvalues and eigenvectors of C such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$ and $\|\mathbf{u}_i\|_2 = 1,\ \forall\ 1 \leq i \leq m$. Then $\mathbf{a}_i = \mathbf{u}_i$ and $\text{var}(Y_i)$=λ_i for $1 \leq i \leq m$.

Methodology of Practical PCA

Given observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in R^m$.

1. Compute the *mean vector* $\mathbf{m} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$
2. Compute the *covariance matrix* $C = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^t$ by MLE
3. Compute the eigenvalue/eigenvector pairs $(\lambda_j, \mathbf{u}_j)$ of C, $1 \leq j \leq m$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$.
4. Compute the first d principal components $y_i^{(j)} = \mathbf{x}_i^t \mathbf{u}_j$, for each observation $\mathbf{x}_i$, $1 \leq i \leq n$, along the direction $\mathbf{u}_j$, $j = 1, 2, \cdots, d$.

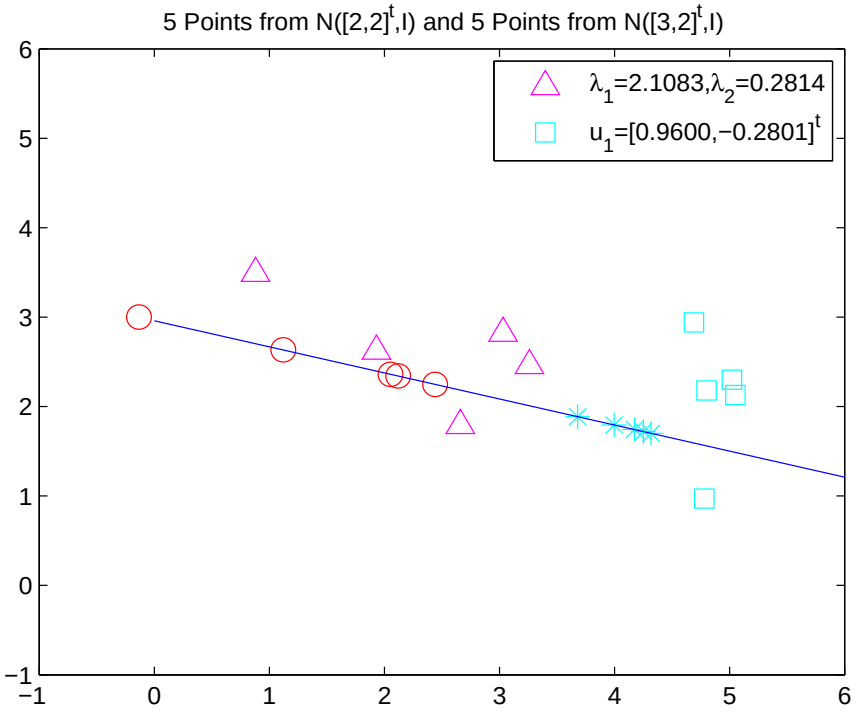
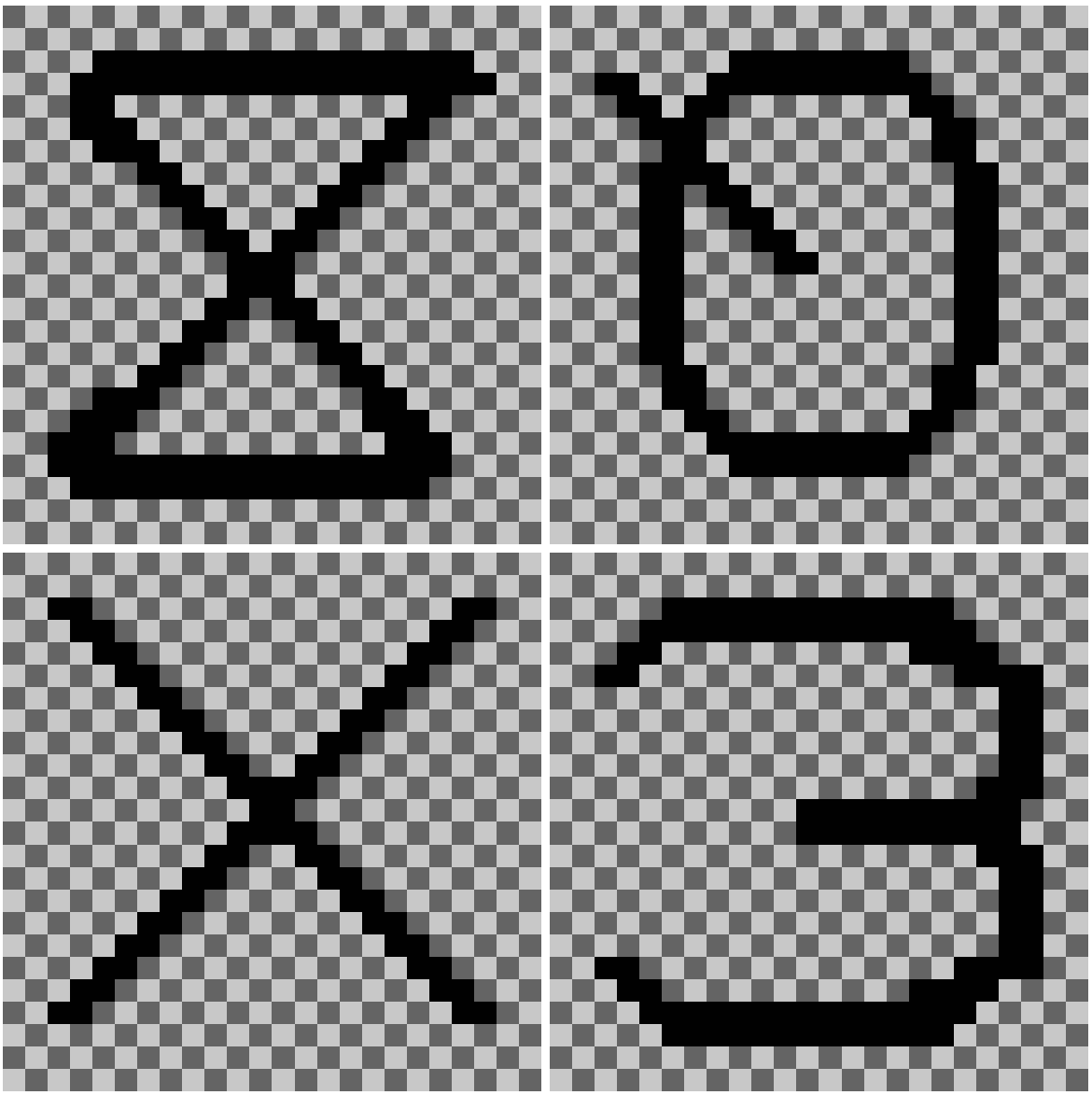


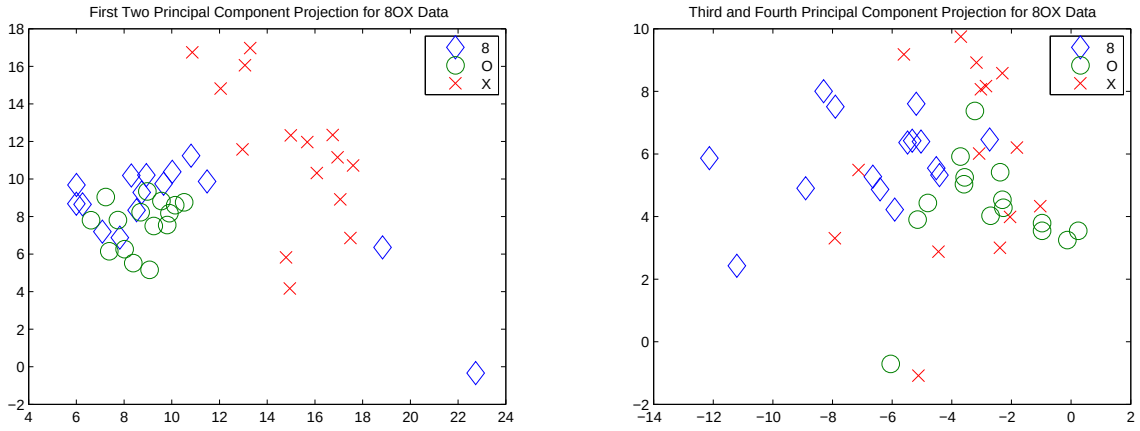
Figure 1: An illustration of PCP on 10 2d Points.

Images of Characters 8,O,X,3



Principal Component Projection of 8OX Data

□ The 8OX data set is derived from Munson's hand printed Fortran character set. Included are 15 patterns from each of the characters '8', 'O', 'X'. Each pattern consists of 8 feature measuremnts.



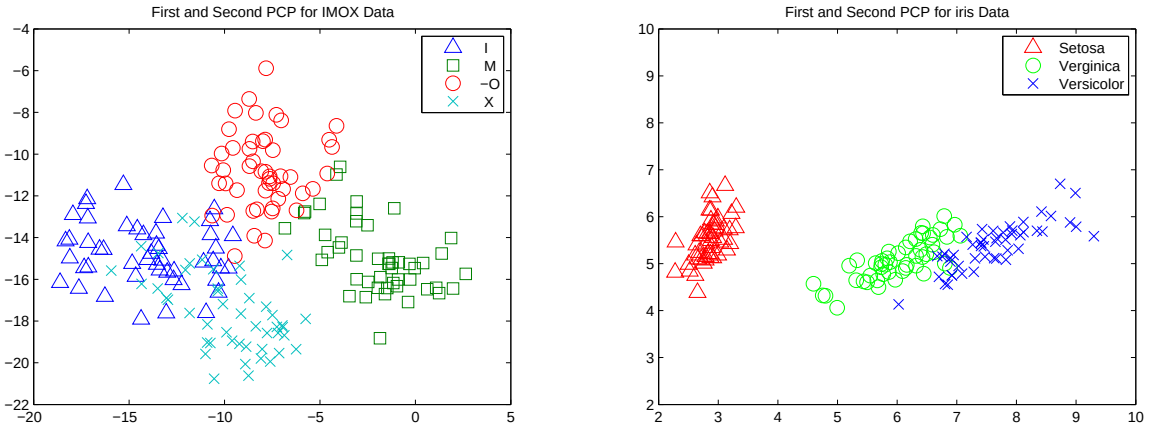
(a) (b)

Figure 2: The (1st,2nd) and (3rd,4th) PCP of 8OX Data

Principal Component Projection of IMOX and iris Data

□ The IMOX data set contains 8 feature measurements on each character of 'I', 'M', 'O', 'X'. It contains 192 patterns, 48 in each character. This data set is also derived from Munson's database.

□ The iris data set contains four feature measurements of three species of iris flowers: *setosa, virginica, versicolor*. It contains 50 patterns from each species on each of four features: sepal length, sepal width, petal length, petal width. This data set has been frequently used for the study of *clustering and classification*.



(a) (b)

Figure 3: PCP of (a) IMOX and (b) iris Data Sets

.

Generate Two Sets of Points in Elongated Regions

```
% Script File: pcaNo.m
% Generate a set of long shaped data in two categories and
% show that PCA does ot pick up the desired direction
%
n=20; d=2; r=5;
X1=random('Uniform',-r,r,n,1);
Y1=random('Uniform',0,1,n,1);
X2=random('Uniform',-r,r,n,1);
Y2=random('Uniform',-1,0,n,1);
for i=1:21
  Xh(i)=-5.5+0.5*i;
  Yh(i)=0;
end
for i=1:n
  X(i,1)=X2(i,1);
  X(i,2)=Y2(i,1);
  X(i+n,1)=X1(i,1);
  X(i+n,2)=Y1(i,1);
end
[n,d]=size(X);
C=cov(X);
[U D]=eig(C);
L=diag(D); L',  U  % principal component directions
plot(X1,Y1,'b^',X2,Y2,'ro',Xh,Yh,'g-'); axis([-r, r, -2 2]); grid;
legend('The 1st principal direction is [-1.0, 0.0]');
title('An Example of PCA fails')
```

Script File: Compute the First K Principal Components

```
% Script file: PCA.m
% Find the first K Principal Components of data X (n rows, d columns)
% X contains n pattern vectors with d features
%
function Y=PCA(X,K)
[n,d]=size(X);
C=cov(X);
[U D]=eig(C);
L=diag(D);
[sorted index]=sort(L,'descend');
Xproj=zeros(d,K);        % initiate a projection matrix
for j=1:K
   Xproj(:,j)=U(:,index(j));
end
Y=X*Xproj;               % first K principal components
```

An Example that PCA Fails

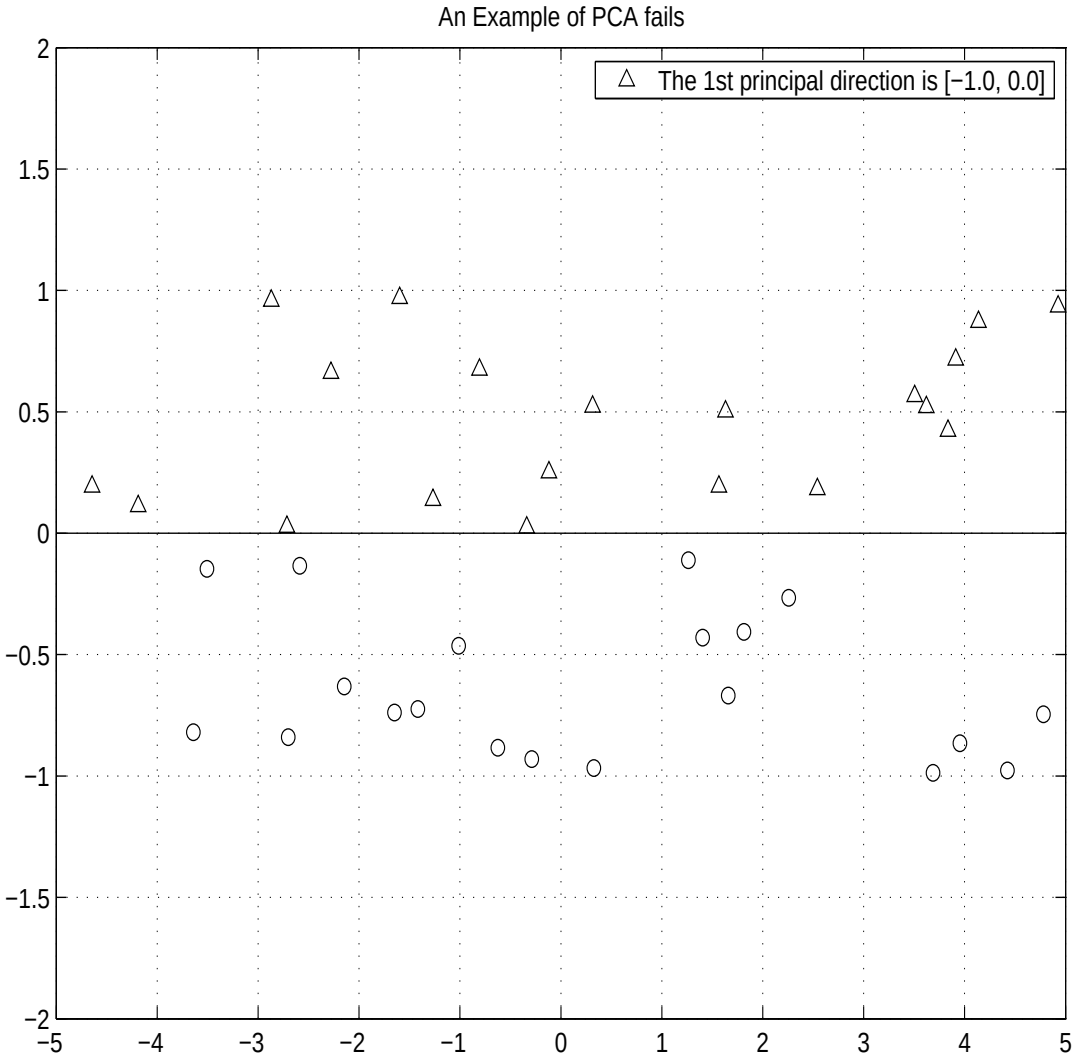


Figure 4: An Example that PCA Fails

.

Generate Two Sets of Points from Gaussian Distributions

```
% Script File: pcaYes.m
% Generate a set of elliptical-shaped data in two categories and
% show that PCA really picks up the desired direction
%
n=20; d=2;
X1=random('Normal',2.0,1,n,1);
Y1=random('Normal',2.0,1,n,1);
X2=random('Normal',-2.0,1,n,1);
Y2=random('Normal',-2.0,1,n,1);
Xh=-4:0.5:4;
Yh=-4:0.5:4;
for i=1:n
  X(i,1)=X2(i,1);
  X(i,2)=Y2(i,1);
  X(i+n,1)=X1(i,1);
  X(i+n,2)=Y1(i,1);
end
[n,d]=size(X);
C=cov(X);
[U D]=eig(C);
L=diag(D);  L', U  % principal component directions
plot(X1,Y1,'b^',X2,Y2,'ro',Xh,Yh,'g-'); axis([-4,4, -4,4]); grid;
legend('The 1st principal direction is [1.0, 1.0]');
title('An Example of PCA works')
```

An Example that PCA Works

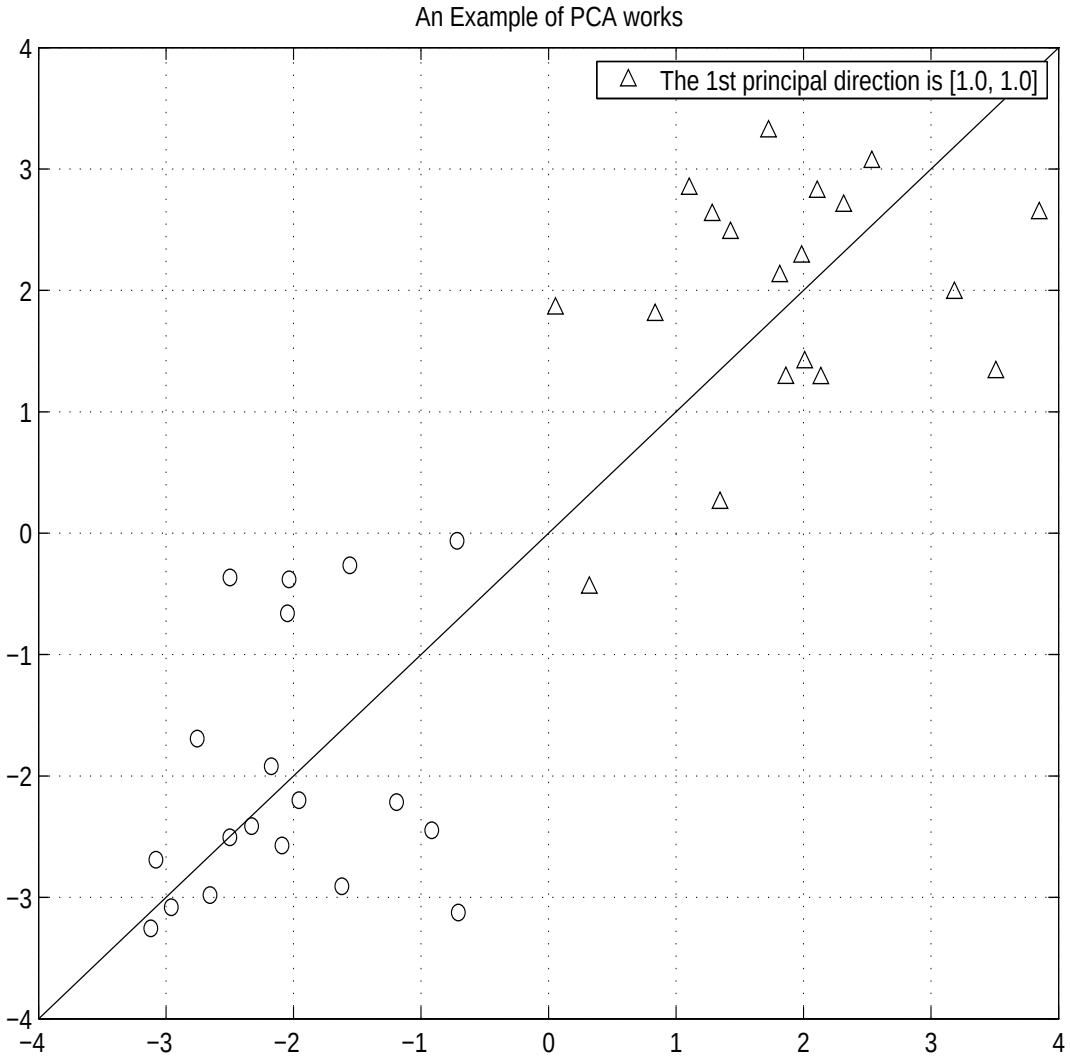


Figure 5: An Example that PCA Works

Fundamentals of Linear Discriminant Analysis

Given the training patterns $\mathbf{x}_1,\ \mathbf{x}_2, \ldots, \mathbf{x}_n$ from K categories, where $n_1+n_2+\ldots+n_K = n$. Let the between-class scatter matrix B, the within-class scatter matrix W, and the total scatter matrix T be defined below.

$$B \ = \ \sum_{i=1}^{K} n_i(\mathbf{u}_i - \mathbf{u})(\mathbf{u}_i - \mathbf{u})^t, \ \textit{where } \mathbf{u}_i \textit{ is the mean of ith category}, \mathbf{u} \ = \ \frac{1}{n}\sum_{i=1}^{n} \mathbf{x_i},$$

$$W \ = \ \sum_{i=1}^{K} \sum_{\mathbf{x}\in\omega_i} (\mathbf{x} - \mathbf{u}_i)(\mathbf{x} - \mathbf{u}_i)^t.$$

$$T \ = \ \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{u})(\mathbf{x}_i - \mathbf{u})^t.$$

Show that $B \ + \ W \ = \ T$.

Linear discriminant analysis for a *dichotomous* problem attempts to find an optimal direction $\mathbf{w}$ for projection which maximizes a *Fisher's discriminant ratio*

$$J(\mathbf{w}) = \frac{(m_1 - m_2)^2}{n_1 s_1^2 + n_2 s_2^2} = \frac{n}{n_1 n_2} \times \frac{\mathbf{w}^t B \mathbf{w}}{\mathbf{w}^t W \mathbf{w}} = \frac{n}{n_1 n_2} \times J_2(\mathbf{w}) \tag{1}$$

where

$$y_i = \mathbf{w}^t \mathbf{x}_i, \quad 1 \le i \le n_1, \quad y_j = \mathbf{w}^t \mathbf{x}_j, \quad n_1 < j \le n, \quad n_1 + n_2 = n$$

$$m_k = \mathbf{w}^t \mathbf{u}_k, \quad k = 1, 2$$

$$s_1^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} (y_i - m_1)^2$$

$$s_2^2 = \frac{1}{n_2} \sum_{j=1+n_1}^{n} (y_j - m_2)^2$$

Let $n \ = \ n_1 n_2$, the problem could be reduced to solving the generalized eigenvalue problem of

$$B\mathbf{w} = \lambda W \mathbf{w}, \quad \textit{where } \lambda = J_2(\mathbf{w}).$$

Discriminant Analysis

The objective of this method is to find the *optimal* set of discriminant vectors in order to separate the predefined classes of objects or events. The material is based on the paper [Duchene and Leclercq, pp.978∼983, IEEE Trans. PAMI 1988]

Let the between-class scatter matrix B, the within-class scatter matrix W, and the total scatter matrix T be defined below.

$$B = \sum_{i=1}^{K} n_i(\mathbf{u}_i - \mathbf{u})(\mathbf{u}_i - \mathbf{u})^t,$$

where

$$\mathbf{u}_i \textit{ is the mean of ith category}, \mathbf{u} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x_i} \textit{ is the sample mean}$$

$$W = \sum_{i=1}^{K} \sum_{\mathbf{x}\in\omega_i} (\mathbf{x} - \mathbf{u}_i)(\mathbf{x} - \mathbf{u}_i)^t$$

$$T = \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{u})(\mathbf{x}_i - \mathbf{u})^t, \quad \textit{where} \quad T = B + W$$

Define the criterion

$$C_1 = \frac{\mathbf{v}^t B \mathbf{v}}{\mathbf{v}^t T \mathbf{v}}, \qquad C_2 = \frac{\mathbf{v}^t B \mathbf{v}}{\mathbf{v}^t W \mathbf{v}}$$

The classical discriminant analysis finds an optimal set of discriminant vectors by the following steps.

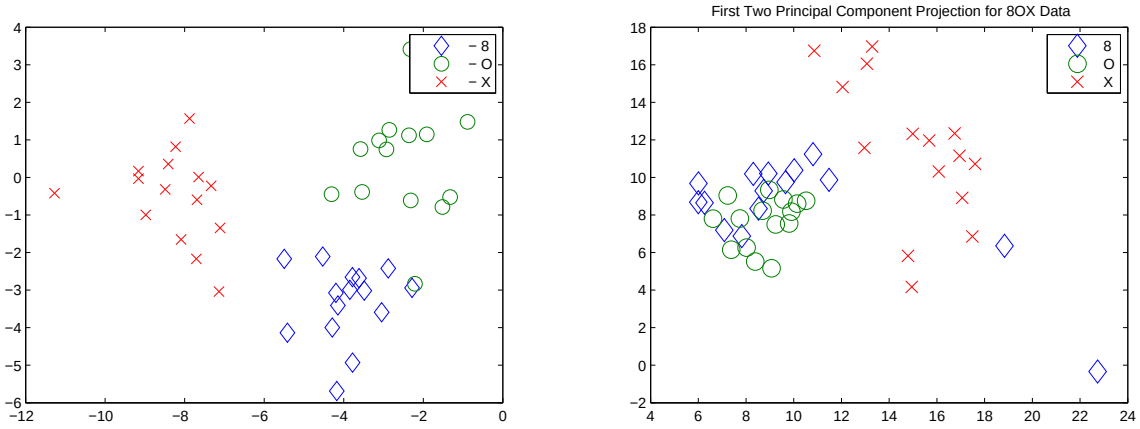
(1) Look for a unit vector $\mathbf{u}_1$ which maximizes C_2, where $\mathbf{u}_1$ could be the eigenvector corresponding to the largest eigenvalue of $W^{-1}B$.

(2) Look for a unit vector $\mathbf{u}_2$ which maximizes C_2 subject to $\mathbf{u}_2^t W \mathbf{u}_1 = 0$.

(3) Look for a unit vector $\mathbf{u}_k$ which maximizes C_2 subject to $\mathbf{u}_k^t W \mathbf{u}_j = 0$ for $k \geq 3$, $1 \leq j < k$.

$\{\mathbf{u}_j\}$ is an optimal set of vectors which best discriminates the patterns. Note that $\mathbf{u}_j$ may not be orthogonal vectors. Duchene and Leclercq suggest that $\mathbf{u}_k^t \mathbf{u}_j = 0$ and $\mathbf{u}_k^t W \mathbf{u}_k = 1$ be used for step (3) and showed by experiments that their proposed method improved the traditional one.

A Comparison of LDA and PCA on 8OX Data



(a) (b)

Figure 6: A Comparison of LDA and PCA on 8OX Data

.

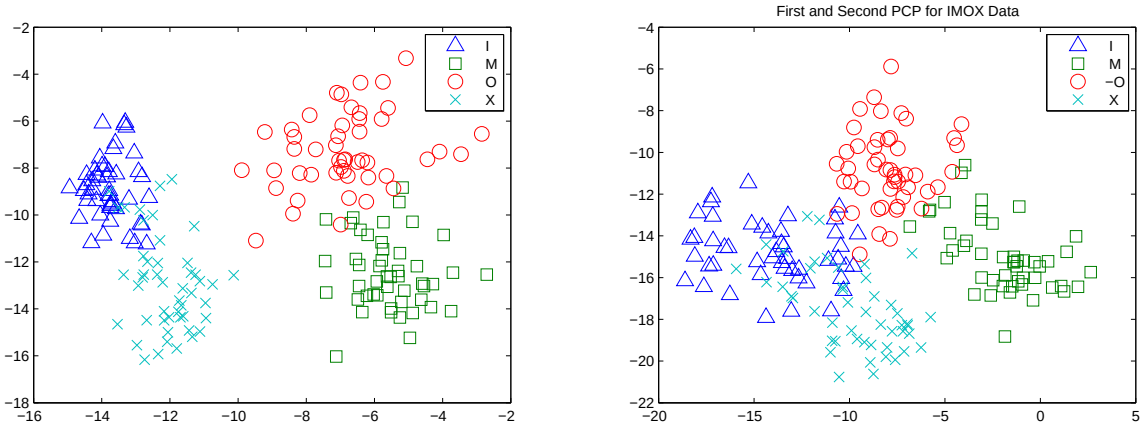
Matlab Codes for Projection Based on LDA

```
% lda8OX.m - Linear Discriminant Projection for data8OX.txt
%
fin=fopen('data8OX.txt');
nf=8;     n=45;                        % nf features, n patterns
L(1)=15;  L(2)=30;  L(3)=45;           % L(3)=n
fgetl(fin); fgetl(fin); fgetl(fin);    % skip 3 header lines
A=fscanf(fin,'%f',[1+nf n]); A=A';     % read input data
d=8;  nk=15;  X=A(:,1:d);
%
% (a) - Covariance Matrix T,  [n d]=size(X);  n=45,  d=8
%
X1=X(1:L(1),:);  X2=X(1+L(1):L(2),:);  X3=X(1+L(2):L(3),:);
m1=mean(X1);    m2=mean(X2);    m3=mean(X3);
mu=mean(X);     T=cov(X);
W1=cov(X1);     W2=cov(X2);     W3=cov(X3);
W=(nk-1)*(W1+W2+W3);
B=nk*((m1-mu)'*(m1-mu)+(m2-mu)'*(m2-mu)+(m3-mu)'*(m3-mu));
s=0.0001;
C=(inv(W+s*eye(d)))*(B+eps);
%
% (b) - Compute Eigenvalues of W^{-1}B
%
[U D]=eig(C);
Lambda=diag(D);
[Cat index]=sort(Lambda,'descend');
%
% (c) - Compute Percentage of Variance Retained
%
R(1)=Cat(1);
for i=2:d
  R(i)=R(i-1)+Cat(i);
end
S=R(d);
for i=1:d
  R(i)=R(i)/S*100;
end
format short;
```

```
L', R

%
% (d) - LDA for 8OX data set
%
K=2;
Xproj=zeros(K,d);                       % initiate a projection matrix
for i=1:K
   Xproj(i,:)=U(:,index(i))';
end
Y=(Xproj*X')';                          % first K discriminant components
X1=Y(1:L(1),1);      Y1=Y(1:L(1),2);
X2=Y(1+L(1):L(2),1); Y2=Y(1+L(1):L(2),2);
X3=Y(1+L(2):L(3),1); Y3=Y(1+L(2):L(3),2);
plot(X1,Y1,'d',X2,Y2,'O',X3,Y3,'X','markersize',10);
legend(' 8',' O',' X')
axis([-16, -2,  -18, 2]); grid;
title('First Two Linear Discriminant Projection for data8OX')
```

A Comparison of LDA and PCA on IMOX Data

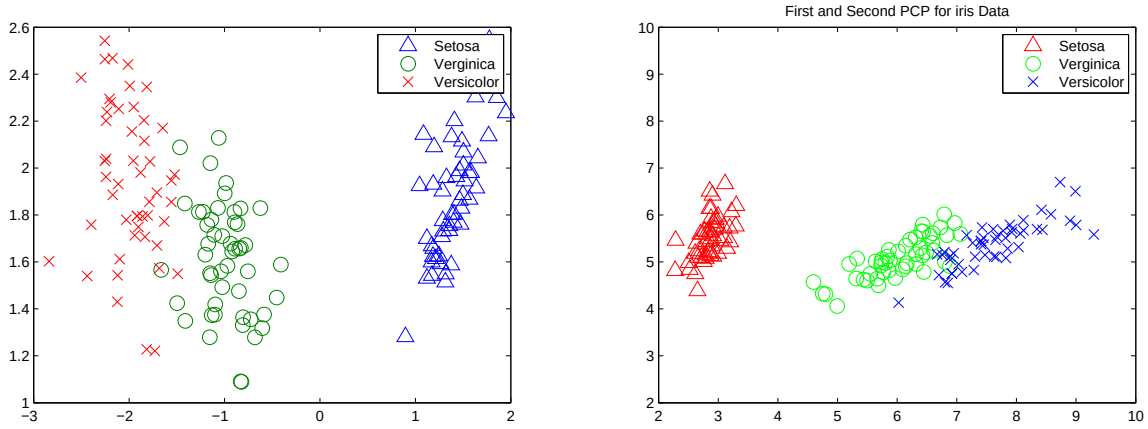


(a) (b)

Figure 7: A Comparison of LDA and PCA on IMOX Data

.

A Comparison of LDA and PCA on iris Data



(a) (b)

Figure 8: A Comparison of LDA and PCA on iris Data

.

Some Exercises for Linear Discriminant Analysis

(1) Let $p(\mathbf{x}|\omega_i)$ be arbitrary densities with mean vectors $\mathbf{u}_i$ and covariance matrices C_i (not necessarily normal) for $i = 1, 2$. Let $y = \mathbf{w}^t\mathbf{x}$ be a projection, and let the induced densities $p(y|\omega_i)$ have means and variances, μ_i and σ_i^2, respectively.

(a) Show that the criterion function

$$J_1(\mathbf{w}) = (\mu_1 - \mu_2)^2/(\sigma_1^2 + \sigma_2^2)$$

is maximized by $\mathbf{w} = (C_1 + C_2)^{-1}(\mathbf{u}_1 - \mathbf{u}_2)$

Hint: $E(\mathbf{w}^t\mathbf{X}|\omega_i) = \mu_i$ and $Var(\mathbf{w}^t\mathbf{X}|\omega_i) = \sigma_i^2$ for $i = 1, 2$.

(b) If the prior probability for ω_i is denoted by $p_i = P(\omega_i)$, show that

$$J_2(\mathbf{w}) = (\mu_1 - \mu_2)^2/(p_1\sigma_1^2 + p_2\sigma_2^2)$$

is maximized by $\mathbf{w} = [p_1C_1 + p_2C_2]^{-1}(\mathbf{u}_1 - \mathbf{u}_2)$

(c) The Fisher linear discriminant function employs that a linear function $\mathbf{w}^t\mathbf{x}$ for which the criterion function J is maximized, where

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

where $y_j = \mathbf{x}_j^t\mathbf{w}$ if $\mathbf{x}_j \in \omega_1$ and $z_k = \mathbf{x}_k^t\mathbf{w}$ if $\mathbf{x}_k \in \omega_2$, and

$$m_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} y_j, \quad m_2 = \frac{1}{n_2}\sum_{k=1}^{n_2} z_k, \quad s_1^2 = \frac{1}{n_1}\sum_{j=1}^{n_1}(y_j - m_1)^2, \quad s_2^2 = \frac{1}{n_2}\sum_{k=1}^{n_2}(z_k - m_2)^2,$$

(d) To which of these criterion functions J_1 or J_2 is the $J(\mathbf{w})$ more closely related?

(2) Let $A \in R^{n\times n}$ be a positive definite matrix and $\mathbf{x}, \mathbf{b} \in R^n$, $\beta \in R$. Find the criterion such that $g(\mathbf{x}) = \frac{1}{2}\mathbf{x}^tA\mathbf{x} - \mathbf{b}^t\mathbf{x} - \beta$ is minimized. What is the minimum value of $g(\mathbf{x})$, $\mathbf{x} \in R^n$?