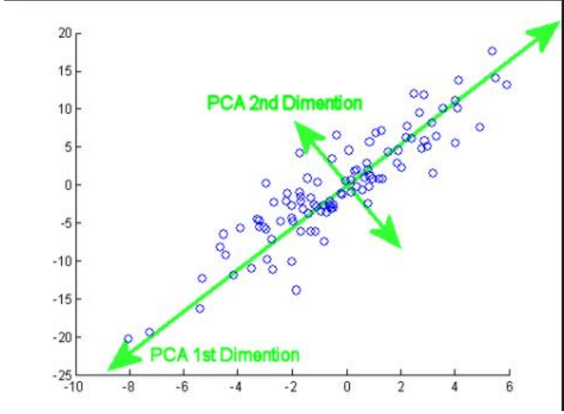
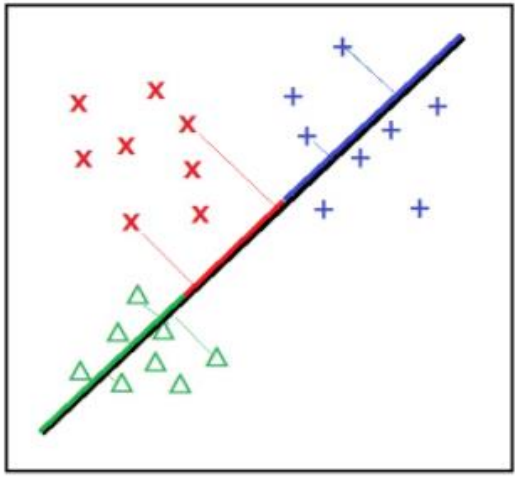


Principal Component Analysis, and Fisher Linear Discriminant

Dimensionality reduction / feature extraction



Principal Component Analysis

- Originated from the work by Pearson(1901).
- Its purpose is to derive new features (variables) in the decreasing order of importance.
- Dimensionality can be reduced without loosing much information and structure present in the data.

PRINCIPAL COMPONENT ANALYSIS [PCA]

* A method to reduce the dimensionality of the data.
* PCA seeks a projection that best represents the data in a least squares sense.
* In the new space data is so represented that the features become uncorrelated.
* In the new space distributions might become a simpler one.

PRINCIPAL COMPONENT ANALYSIS [PCA]

* A method to reduce the dimensionality of the data.
* PCA seeks a projection that best represents the data in a least squares sense.
* In the new space data is so represented that the features become uncorrelated.
* In the new space distributions might become a simpler one.

Short comings

* Needs to find covariance matrix and its eigen vectors
* Discriminating components between classes might be lost

Objective :-

Let $\mathcal{D} = \{x_1, \ldots, x_n\}$ be the set of patterns of dimensionality d

We want to find $\mathcal{D}' = \{x_1', x_2', \ldots, x_n'\}$ where each x_i' is of dim. d' such that $d' < d$, and

$$J = \sum_{i=1}^{n} \| x_i - x_i' \|^2$$ should be minimum possible one.

What is zero dim. projection for the data?

i.e., we want to represent the data set by just one pattern (x_0).

$$J = \sum_i \| x_i - x_0 \|^2$$

What is zero dim. projection for the data?

i.e., we want to represent the data-set by just one pattern (X_0).

$$J = \sum_i \| X_i - X_0 \|^2$$

It is easy to see that J is minimized when X_0 is sample mean (centroid).

i.e., $X_0 = \frac{1}{n} \sum_i X_i$

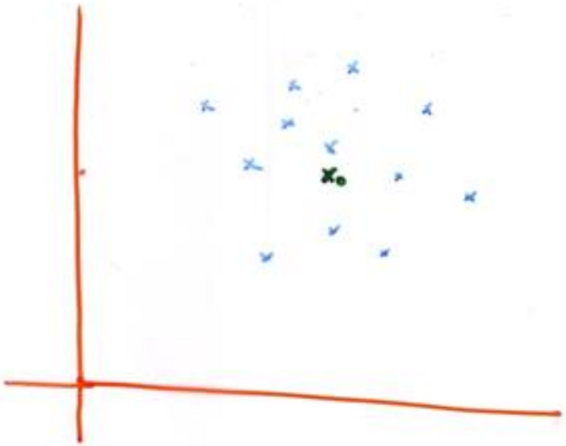
What is zero dim. projection for the data?

i.e., we want to represent the data set by just one pattern (x_0).

$$J = \sum_i \| x_i - x_0 \|^2$$

It is easy to see that J is minimized when x_0 is sample mean (centroid).

i.e., $x_0 = \frac{1}{n} \sum_i x_i$



What is good 1-Dim Representation

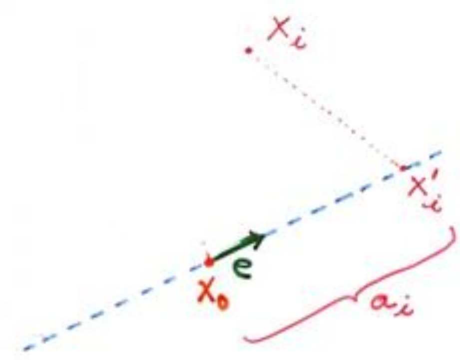
Let the data is projected onto a line passing through the centroid (X_0)

What is good 1-Dim Representation

Let the data is projected onto a line passing through the centroid (X_0)

Let $X_i' = \text{Proj}(X_i)$

$\vec{e}$ is unit vector in the direction of the line.



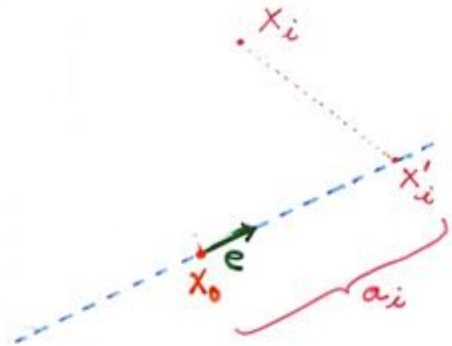
Then

$$X_i' = X_0 + a_i e$$

a_i → scalar dist. between X_0 & X_i'

What is good 1-Dim Representation

Let the data is projected onto a line passing through the centroid (x_0)



Let $x_i' = \text{Proj}(x_i)$

$\vec{e}$ is unit vector in the direction of the line.

Then

$$x_i' = x_0 + a_i e$$

→ scalar dist. between x_0 & x_i'

$$J = \sum_i \left\| (x_0 + a_i e) - x_i \right\|^2$$

We need to find a_i & e

To find $a_1, a_2, \dots, a_n$

$$J = \sum_i \left\| (x_0 + a_i e) - x_i \right\|^2$$

$$= \sum_i \left\| a_i e - (x_i - x_0) \right\|^2$$

$$= \sum a_i^2 \|e\|^2 - 2 \sum a_i e^T (x_i - x_0) + \sum \|x_i - x_0\|^2$$

$$= \sum_i a_i^2 - 2 \sum_i a_i e^T (x_i - x_0) + \sum_i \|x_i - x_0\|^2$$

Now, To find a_j

$$\frac{\partial J}{\partial a_j} = 2 a_j - 2 e^T (x_j - x_0) = 0$$

$$a_j = e^T (x_j - x_0)$$

To find $a_1, a_2, \ldots, a_n$

$$J = \sum_i \left\| (x_0 + a_i e) - x_i \right\|^2$$

$$= \sum_i \left\| a_i e - (x_i - x_0) \right\|^2$$

$$= \sum a_i^2 \|e\|^2 - 2 \sum a_i e^T (x_i - x_0) + \sum \|x_i - x_0\|^2$$

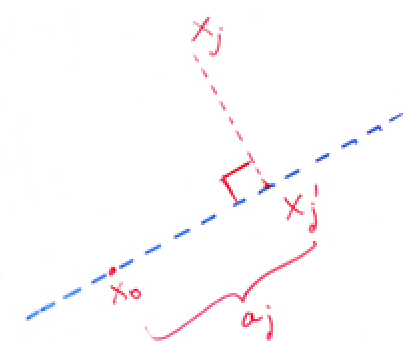
$$= \sum_i a_i^2 - 2 \sum_i a_i e^T (x_i - x_0) + \sum_i \| x_i - x_0 \|^2$$

Now, To find a_j

$$\frac{\partial J}{\partial a_j} = 2 a_j - 2 e^T (x_j - x_0) = 0$$

$$a_j = e^T (x_j - x_0)$$

i.e., Perpendicularly Project onto the line



To Find $\vec{e}$

$$J = \sum a_i^2 - 2\sum a_i e^T(x_i - x_0) + \sum \|x_i - x_0\|^2$$

$$\because \quad a_i = e^T(x_i - x_0),$$

$$J = \sum a_i^2 - 2\sum a_i^2 + \sum \|x_i - x_0\|^2$$

$$= -\sum a_i^2 \qquad + \sum \|x_i - x_0\|^2$$

$$= -\sum e^T(x_i - x_0)(x_i - x_0)^T e \quad + \sum \|x_i - x_0\|^2$$

$$= -\left[e^T\left(\sum (x_i - x_0)(x_i - x_0)^T \right) e \right] + \sum \|x_i - x_0\|^2$$

$$= -e^T S e \quad + \sum \|x_i - x_0\|^2$$

To Find $\vec{e}$

$$J = \sum a_i^2 - 2\sum a_i e^T(x_i - x_0) + \sum \|x_i - x_0\|^2$$

$$\because \quad a_i = e^T(x_i - x_0),$$

$$J = \sum a_i^2 - 2\sum a_i^2 + \sum \|x_i - x_0\|^2$$

$$= -\sum a_i^2 \qquad + \sum \|x_i - x_0\|^2$$

$$= -\sum e^T(x_i - x_0)(x_i - x_0)^T e \quad + \sum \|x_i - x_0\|^2$$

$$= -\left[e^T \left(\sum (x_i - x_0)(x_i - x_0)^T \right) e \right] + \sum \|x_i - x_0\|^2$$

$$= -e^T S e \quad + \sum \|x_i - x_0\|^2$$

where S = Scatter matrix $= \sum_i (x_i - x_0)(x_i - x_0)^T$

$S = n \cdot$ (sample covariance Matrix)

$$J = -e^T S e + \sum \| x_i - x_0 \|^2$$

To minimize J, $-e^T S e$ should be minimized subject to the constraint

$\|e\| = 1$ or $e^T e - 1 = 0$

This is "Constrained Optimization" problem.

We can use the method of "Lagrange multipliers

Constrained Optimization

Minimize $f(v)$

subject to $g_j(v) \leq 0$, for $1 \leq j \leq n$.

Lagrangian, $\mathcal{L} = f(v) + \sum_{j=1}^{n} \underbrace{\alpha_j}_{\text{Lagrange Multiplier}} g_j(v)$

Necessary cond. at optimal v are :

(i) $\nabla_v \mathcal{L} = 0$

$$\left.\begin{array}{ll} \text{(ii)} & \alpha_j \geq 0 \\ \text{(iii)} & \alpha_j \, g_j(v) = 0 \end{array}\right\} \text{for all } j = 1 \text{ to } n$$

But, we are with equality constraint.

- So, gradient w.r.t. primal variables and gradient w.r.t. dual variables can be equated to zero.
  - Ofcourse, the Lagrange multipliers should be nonnegative.

Minimize $\quad -e^T s e$

such that $\quad e^T e - 1 = 0$

$$\mathcal{L} = (-e^T s e) + \alpha (e^T e - 1)$$

$$\nabla_e \mathcal{L} = \frac{\partial \mathcal{L}}{\partial e} = -2se + 2\alpha e = 0$$

Minimize $\quad -e^T s e$

such that $\quad e^T e - 1 = 0$

$$\mathcal{L} = (-e^T s e) + \alpha (e^T e - 1)$$

$$\nabla_e \mathcal{L} = \frac{\partial \mathcal{L}}{\partial e} = -2se + 2\alpha e = 0$$

$$se = \alpha e$$

scatter matrix

scalar

vector

Minimize $\quad -e^T S e$

such that $\quad e^T e - 1 = 0$

$$\mathcal{L} = (-e^T S e) + \alpha(e^T e - 1)$$

$$\nabla_e \mathcal{L} = \frac{\partial \mathcal{L}}{\partial e} = -2Se + 2\alpha e = 0$$

$$\underbrace{S}_{\text{scatter matrix}} e = \underbrace{\alpha}_{\text{scalar}} \underbrace{e}_{\text{vector}}$$

This is eigen value (vector) problem

$\left.\begin{array}{l} \alpha \text{ is eigen value} \\ e \text{ is eigen vector} \end{array}\right\}$ for S

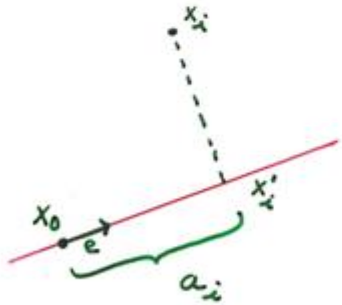
$\because$ $-e^T S e$ minimized $\Rightarrow$ $-e^T \alpha e$ minimized

$\Rightarrow$ α should be maximum.

i.e., e is the eigen vector for S for which eigen value is maximum.

e gives maximum Variance Direction

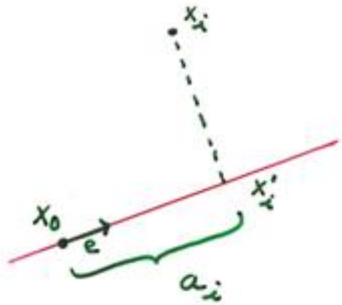
$$a_i = e^T (x_i - x_0)$$



$\{x_1, x_2, \dots, x_n\}$ is represented as $\{a_1, a_2, \dots, a_n\}$

e gives maximum Variance Direction

$$a_i = e^T (x_i - x_0)$$



$\{x_1, x_2, \dots, x_n\}$ is represented as $\{a_1, a_2, \dots, a_n\}$

Variance of $\{a_1, \dots, a_n\}$

$$= \frac{1}{n} \sum (a_i - a_0)^2 = \frac{1}{n} \sum a_i^2$$

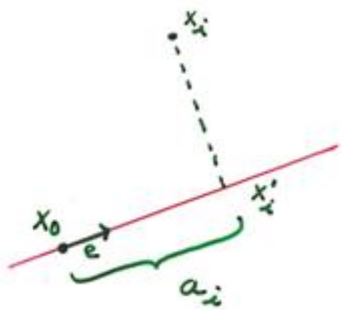
$$[\because a_0 = 0]$$

$$= \frac{1}{n} e^T \left[\sum (x_i - x_0)(x_i - x_0)^T \right] e$$

$$= \frac{1}{n} e^T S e$$

e gives maximum Variance Direction

$$a_i = e^T (x_i - x_0)$$



$\{x_1, x_2, \dots, x_n\}$ is represented as $\{a_1, a_2, \dots, a_n\}$

Variance of $\{a_1, \dots, a_n\}$

$$= \frac{1}{n} \sum (a_i - a_0)^2 = \frac{1}{n} \sum a_i^2$$

$$[\because a_0 = 0]$$

$$= \frac{1}{n} e^T \left[\sum_i (x_i - x_0)(x_i - x_0)^T \right] e$$

$$= \frac{1}{n} e^T S e$$

* The new space is so found that $e^T S e$ is maximum possible one.

i.e., The data is Projected onto that line over which the variance is large.

Generalization : To find d' dimensions

$$x_i' = x_0 + [a_{i1} e_1 + \cdots + a_{id'} e_{d'}]$$

We get, $J = \sum \left\| x_0 + \sum_{j=1}^{d'} a_{ij} e_j - x_i \right\|^2$

Generalization : To find d' dimensions

$$x_i' = x_0 + [a_{i1} e_1 + \cdots + a_{id'} e_{d'}]$$

We get, $J = \sum \left\| x_0 + \sum_{j=1}^{d'} a_{ij} e_j - x_i \right\|^2$

It is easy to show that

$$\left.\begin{matrix} e_1 \\ e_2 \\ \vdots \\ e_{d'} \end{matrix}\right\}$$ are eigen vectors of S for which eigen values are maximum possible.

$$S e_1 = \alpha_1 e_1$$
$$S e_2 = \alpha_2 e_2$$
$$\vdots$$
$$S e_{d'} = \alpha_{d'} e_{d'}$$

$$\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_{d'}$$

Generalization : To find d' dimensions

$$x_i' = x_0 + [a_{i1} e_1 + \cdots + a_{id'} e_{d'}]$$

We get, $J = \sum \left\| x_0 + \sum_{j=1}^{d'} a_{ij} e_j - x_i \right\|^2$

It is easy to show that

$\left.\begin{matrix} e_1 \\ e_2 \\ \vdots \\ e_{d'} \end{matrix}\right\}$ are eigen vectors of S for which eigen values are maximum possible.

$$S e_1 = \alpha_1 e_1$$
$$S e_2 = \alpha_2 e_2$$
$$\vdots$$
$$S e_{d'} = \alpha_{d'} e_{d'}$$

$$\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_{d'}$$

$\boxed{\text{Further, } e_1, e_2, \ldots, e_{d'} \text{ are orthonormal}}$

Generalization : To find d' dimensions

$$x_i' = x_0 + [a_{i1} e_1 + \cdots + a_{id'} e_{d'}]$$

We get, $J = \sum \left\| x_0 + \sum_{j=1}^{d'} a_{ij} e_j - x_i \right\|^2$

It is easy to show that

$\left.\begin{matrix} e_1 \\ e_2 \\ \vdots \\ e_{d'} \end{matrix}\right\}$ are eigen vectors of S for which eigen values are maximum possible.

$$S e_1 = \alpha_1 e_1$$
$$S e_2 = \alpha_2 e_2$$
$$\vdots$$
$$S e_{d'} = \alpha_{d'} e_{d'}$$

$\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_{d'}$

Because the scatter matrix S is real and symmetric, Eigen values are real and nonnegative.

Further, $e_1, e_2, \ldots, e_{d'}$ are orthonormal

Representation in New Space

$$X_i' = X_0 + [a_{i1} e_1 + \cdots + a_{id'} e_{d'}]$$

$\because$ x_0, $e_1, \ldots, e_{d'}$ are fixed, So

X_i' can be represented as $\begin{bmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{id'} \end{bmatrix} = Y_i$

Representation in New Space

$$X_i' = X_0 + [a_{i1} e_1 + \cdots + a_{id'} e_{d'}]$$

$\because$ x_0, $e_1, \ldots, e_{d'}$ are fixed, So

X_i' can be represented as $\begin{bmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{id'} \end{bmatrix} = Y_i$

Since

$$\begin{aligned} a_{i1} &= e_1^T(X_i - X_0) \\ a_{i2} &= e_2^T(X_i - X_0) \\ &\vdots \\ a_{id'} &= e_{d'}^T(X_i - X_0) \end{aligned}$$

$$Y_i = \underbrace{\begin{bmatrix} e_1^T \\ e_2^T \\ \vdots \\ e_{d'}^T \end{bmatrix}}_{\text{Transformation matrix}} (X_i - X_0)$$

Let us call this, $P = \begin{bmatrix} - e_1 - \\ - e_2 - \\ \vdots \\ - e_{d'} - \end{bmatrix}_{d' \times d}$

The projection matrix P

- $P^t P = I$, But P need not be square.

  If P is square, i.e., use all eigen vectors, then P is orthogonal.

  Not only P is orthogonal, P is a rotation matrix.

The projection matrix P

- $P^t P = I$, But P need not be square.

  If P is square, i.e., use all eigen vectors, then P is orthogonal.

  Not only P is orthogonal, P is a rotation matrix.

- PCA basically, translates (so that origin becomes centroid) and does rotation (so that features are uncorrelated).

Data

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Ref: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

Data

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Mean subtracted Data

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

Ref: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

Data

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

$\Rightarrow$

Mean subtracted Data

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

$\Rightarrow$

$$cov = \begin{pmatrix} .61 & .61 \\ .61 & .71 \end{pmatrix}$$

Ref: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

Data

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

$\Rightarrow$

Mean subtracted Data

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

$\Rightarrow$

$$cov = \begin{pmatrix} .61 & .61 \\ .61 & .71 \end{pmatrix}$$

Transformed Data (Single eigenvector)

x
-.827970186
1.77758033
-.992197494
-.274210416
-1.67580142
-.912949103
.0991094375
1.14457216
.438046137
1.22382056

The data after transforming using only the most significant eigenvector

Ref: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

Data

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Mean subtracted Data

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

$$cov = \begin{pmatrix} .61 & .61 \\ .61 & .71 \end{pmatrix}$$

Eigenvalues are 1.28, 0.049

Corresponding Eigenvectors are $\begin{pmatrix} -.67 \\ -.73 \end{pmatrix}$ and $\begin{pmatrix} -.73 \\ .67 \end{pmatrix}$

Transformed Data (Single eigenvector)

x
-.827970186
1.77758033
-.992197494
-.274210416
-1.67580142
-.912949103
.0991094375
1.14457216
.438046137
1.22382056

The data after transforming using only the most significant eigenvector

Ref: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

Data

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

$\Rightarrow$

Mean subtracted Data

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

$\Rightarrow$

$$cov = \begin{pmatrix} .61 & .61 \\ .61 & .71 \end{pmatrix}$$

Eigenvalues are 1.28, 0.049

Corresponding Eigenvectors are $\begin{pmatrix} -.67 \\ -.73 \end{pmatrix}$ and $\begin{pmatrix} -.73 \\ .67 \end{pmatrix}$

What is the projection matrix?

Transformed Data (Single eigenvector)

x
-.827970186
1.77758033
-.992197494
-.274210416
-1.67580142
-.912949103
.0991094375
1.14457216
.438046137
1.22382056

The data after transforming using only the most significant eigenvector

Ref: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

Data

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

$\Rightarrow$

Mean subtracted Data

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

$\Rightarrow$

$$cov = \begin{pmatrix} .61 & .61 \\ .61 & .71 \end{pmatrix}$$

Eigenvalues are 1.28, 0.049

Corresponding Eigenvectors are $\begin{pmatrix} -.67 \\ -.73 \end{pmatrix}$ and $\begin{pmatrix} -.73 \\ .67 \end{pmatrix}$

What is the projection matrix?

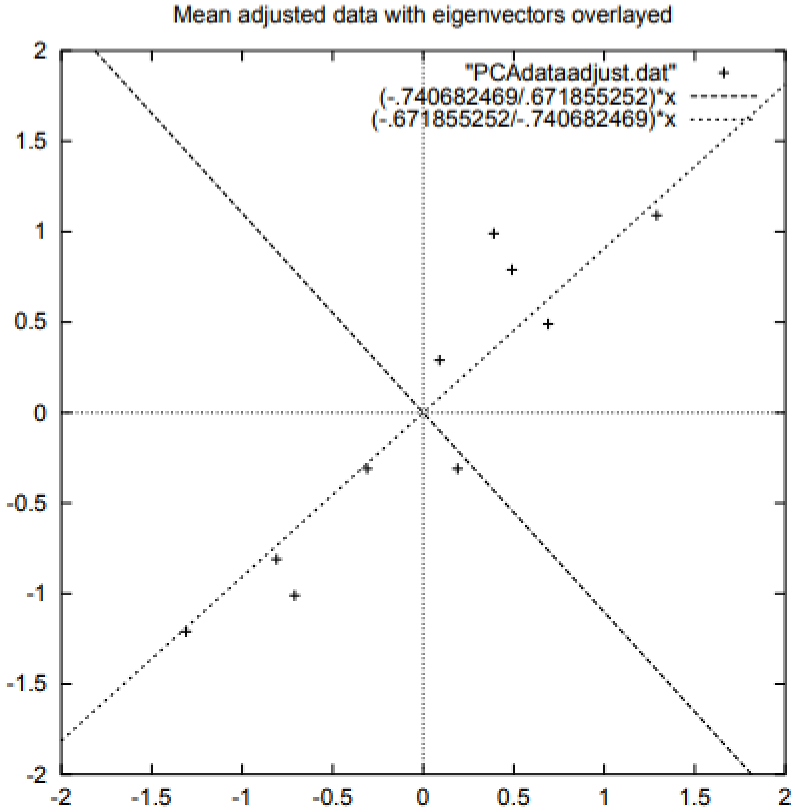
$$\boldsymbol{P} = \begin{pmatrix} -.67 \\ -.73 \end{pmatrix}^t$$

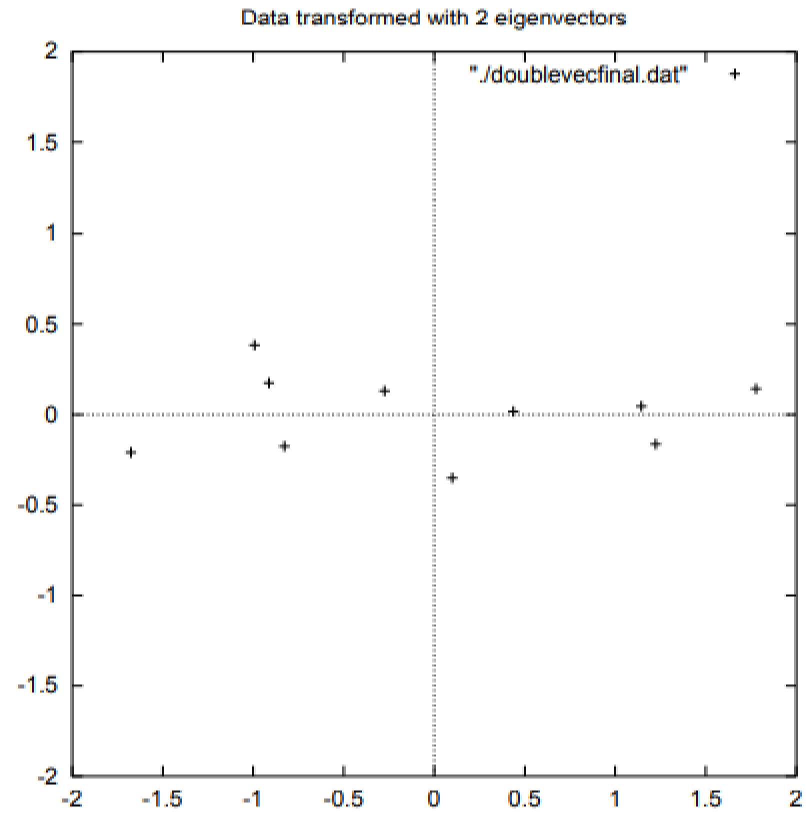
Transformed Data (Single eigenvector)

x
-.827970186
1.77758033
-.992197494
-.274210416
-1.67580142
-.912949103
.0991094375
1.14457216
.438046137
1.22382056

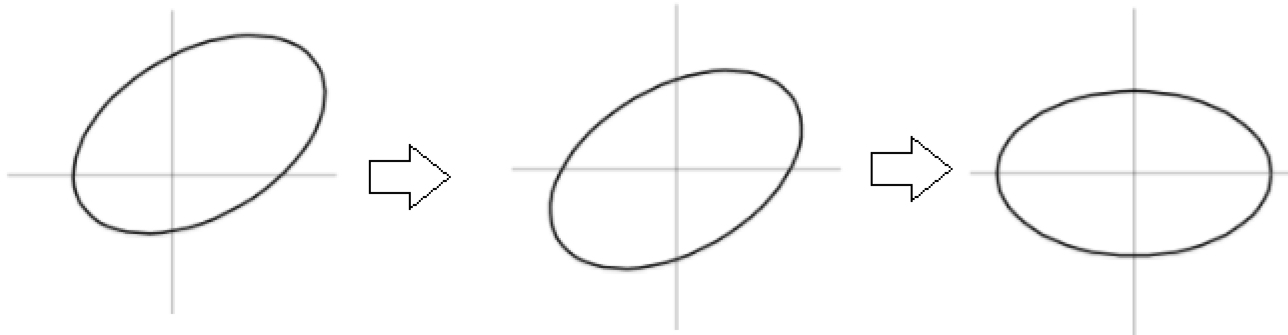
The data after transforming using only the most significant eigenvector

Ref: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf





Ref: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf



HOW IS THAT FEATURES ARE UNCORRELATED IN THE NEW SPACE?

In the new space features are un-correlated !

$$Y_i = P(X_i - X_0) = PX_i - PX_0$$

Mean in the new space, $Y_0 = P(X_0 - X_0) = \vec{0}$

Scatter matrix in new space $= S'$

$$S' = \sum_i (Y_i - Y_0)(Y_i - Y_0)^T = \sum_i Y_i Y_i^T$$

$$= \sum_i (PX_i - PX_0)(PX_i - PX_0)^T$$

$$= \sum_i P(X_i - X_0)(X_i - X_0)^T P^T$$

$$= PSP^T$$

S is the scatter matrix of original space

$$= \begin{bmatrix} - e_1 - \\ - e_2 - \\ \vdots \\ - e_{d'} - \end{bmatrix} S \cdot \begin{bmatrix} | & | & & | \\ e_1 & e_2 & \cdots & e_{d'} \\ | & | & & | \end{bmatrix}$$

$$= \begin{bmatrix} - e_1 - \\ - e_2 - \\ \vdots \\ - e_{d'} - \end{bmatrix} \begin{bmatrix} \alpha_1 e_1 & \cdots & \alpha_{d'} e_{d'} \end{bmatrix}$$

$$= \begin{bmatrix} \alpha_1 & & & 0 \\ & \alpha_2 & & \\ & & \ddots & \\ 0 & & & \alpha_{d'} \end{bmatrix}$$

A Transformation; Covariance Matrix = I

The following transformation matrix can give

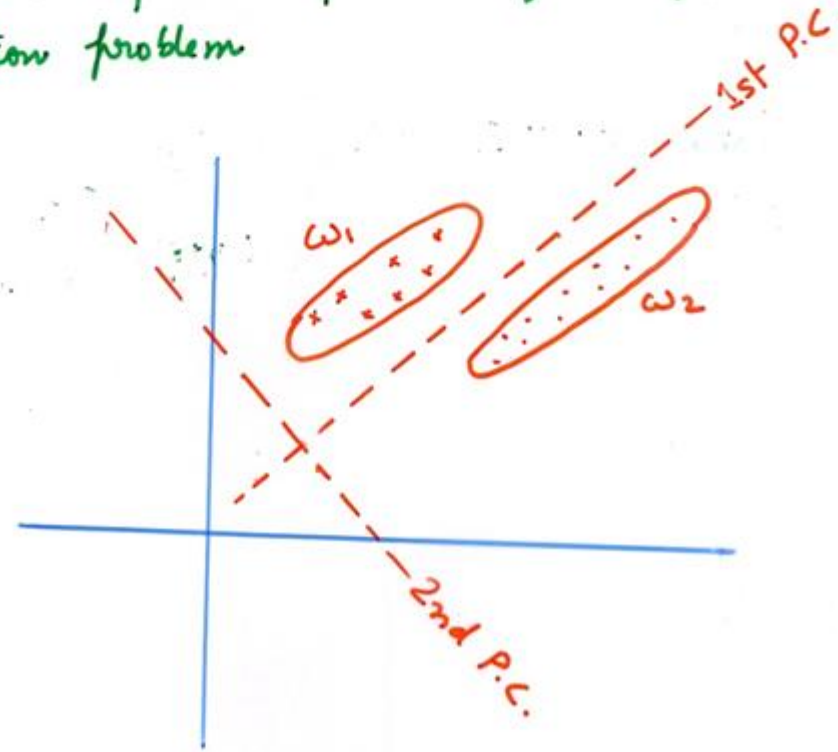
$\Sigma = I$ in the new space

$$\begin{pmatrix} (\alpha_1)^{-1/2} & & & & \\ & (\alpha_2)^{-1/2} & & & \\ & & . & & \\ & & & . & \\ & & & & (\alpha_d)^{-1/2} \end{pmatrix} P$$

Drawback of PCA

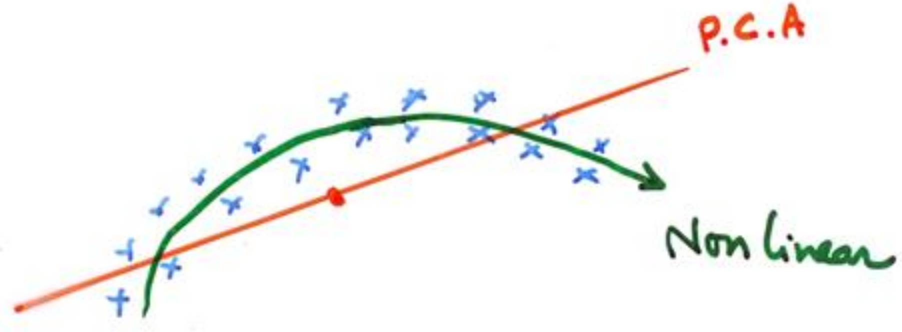
* PCA seeks directions that are efficient for representation
  But doesnot take class-labels into account

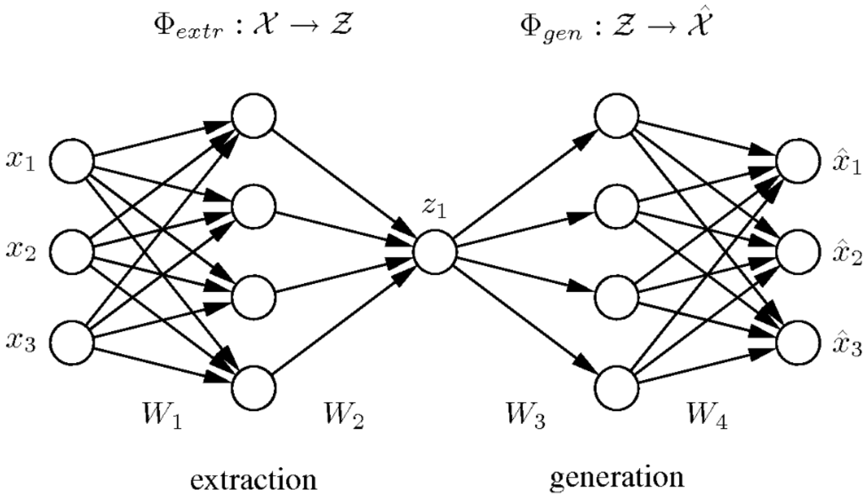
  So, the new space maynot be good for classification problem



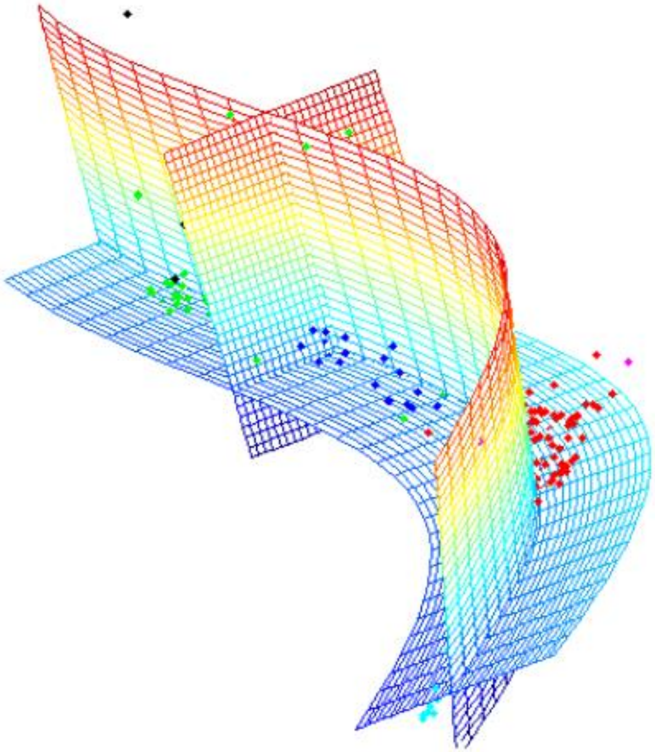
* 2nd P.C. is better than 1st P.C.

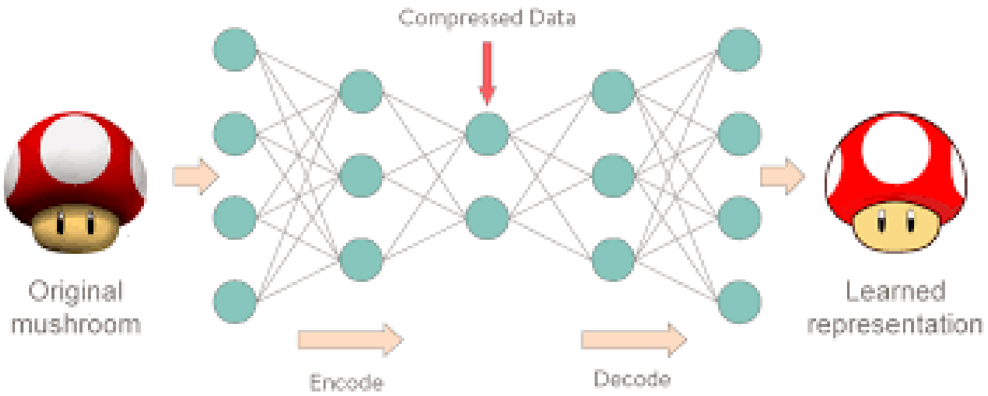
Non-linear Transformation can give better results





Autoencoder networks and Kernel PCA can find these kind of non-linear mappings..





Autoencoder networks and Kernel PCA can find these kind of non-linear mappings..

Have you realized that PCA is unsupervised?

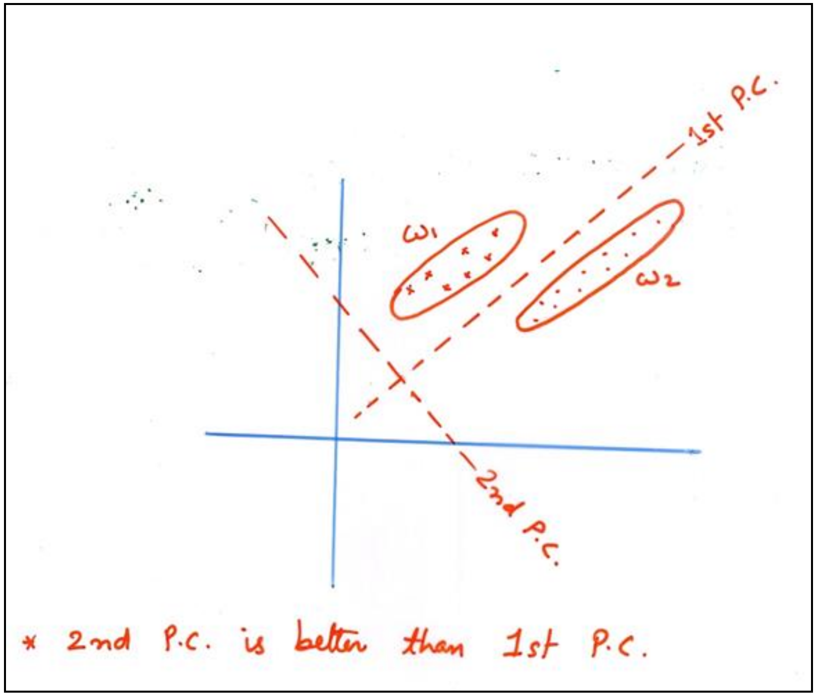
FISHER LINEAR DISCRIMINANT (FLD)

PCA is unsupervised

- With a labeled (training) data set, for the classification problem, principal components based dimensionality reduction may be bad.

PCA is unsupervised

- With a labeled (training) data set, for the classification problem, principal components based dimensionality reduction may be bad.



Fisher Linear Discriminant

- The objective is to find linear projections of the patterns which is good for classification.
- Class-labels are taken into account.

Fisher Linear Discriminant

$\Omega = \{\omega_1, \omega_2\}$

D_1 = Set of patterns belonging to ω_1

D_2 = " " " ω_2

Let $\vec{W}$ denotes direction of a line

$$Y_1 = \{ y_i = W^T x_i \mid x_i \in D_1 \}$$

$$Y_2 = \{ y_j = W^T x_j \mid x_j \in D_2 \}$$

Objective : Find W s.t. Y_1 & Y_2 are well seperated

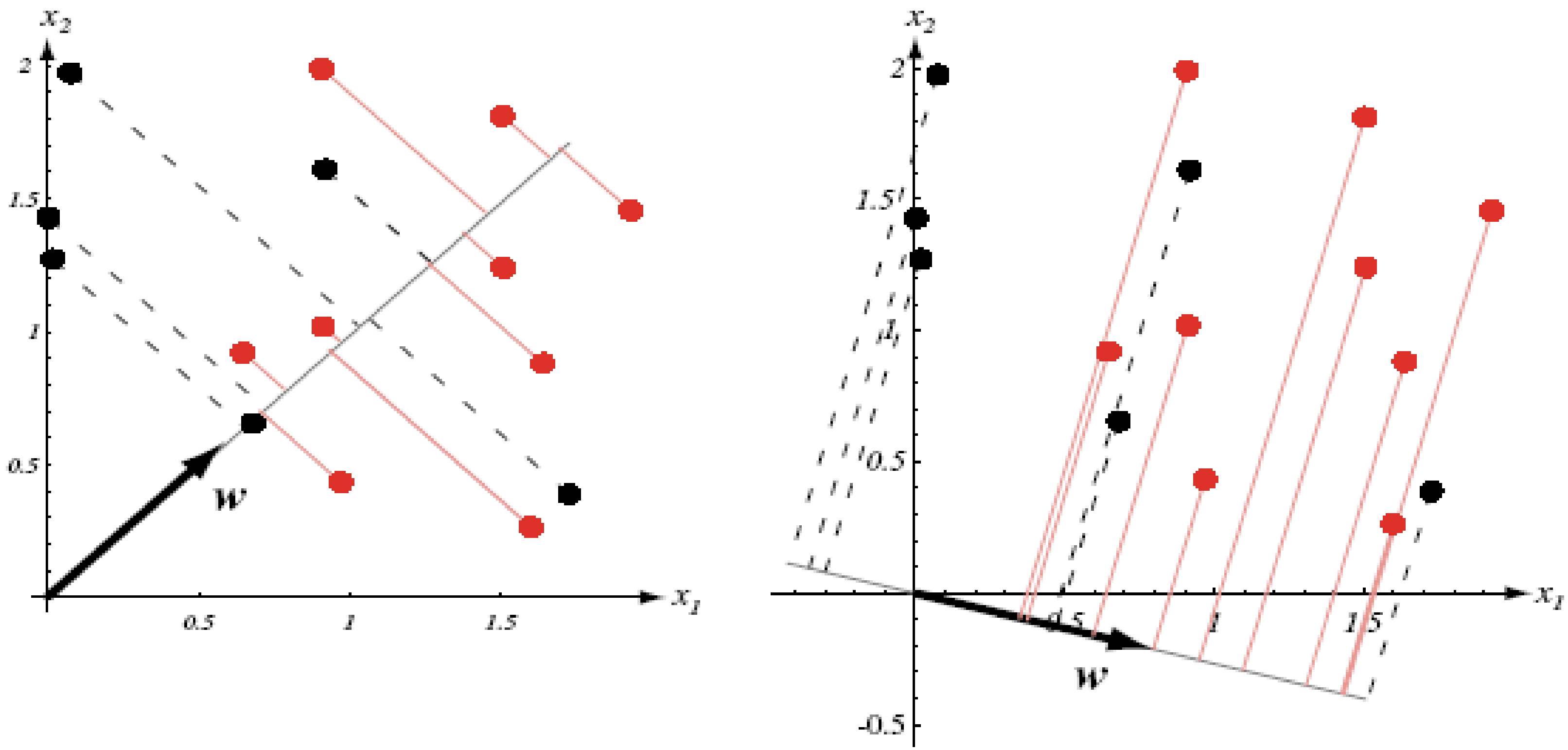
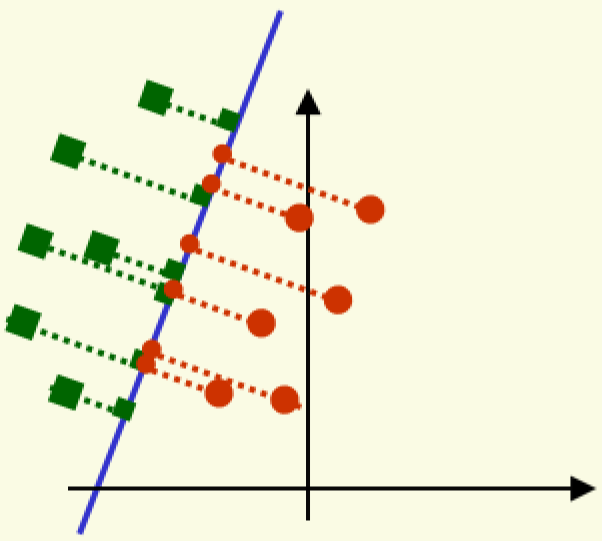
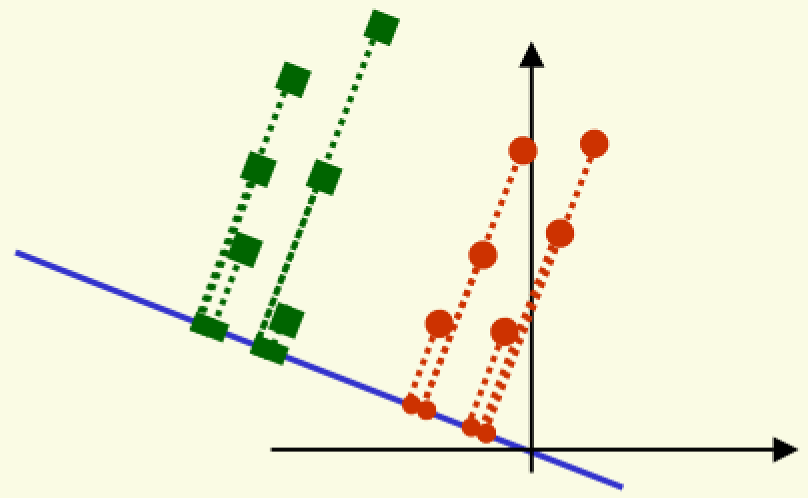


FIGURE 3.5. Projection of the same set of samples onto two different lines in the directions marked **w**. The figure on the right shows greater separation between the red and black projected points. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



bad line to project to, classes are mixed up



good line to project to, classes are well separated

$$J(w) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \quad \text{maximize}$$

where $\tilde{m}_1$ = Mean of samples in Y_1

$\tilde{m}_2$ = " " Y_2

$\tilde{s}_1^2$ = Within class scatter for Y_1

$$= \sum_{y \in Y_1} (y - \tilde{m}_1)^2$$

$\tilde{s}_2^2$ = Within class scatter for Y_2

$$(\tilde{m}_1 - \tilde{m}_2)^2 = \left[W^T (m_1 - m_2) \right]^2$$

m_2 → mean of patterns in D_2

m_1 → " " " D_1

$$= W^T (m_1 - m_2)(m_1 - m_2)^T W$$

$$= W^T S_B W$$

S_B → Between Class Scatter

$$(\tilde{m}_1 - \tilde{m}_2)^2 = \left[W^T (m_1 - m_2) \right]^2$$

m_2 → mean of patterns in D_2

m_1 → " " " D_1

$$= W^T (m_1 - m_2)(m_1 - m_2)^T W$$

$$= W^T S_B W$$

S_B → Between Class Scatter

$$\tilde{s}_i^2 = \sum_{y \in Y_i} (y - \tilde{m}_i)^2$$

$$= \sum_{x \in D_i} (W^T x - W^T m_i)^2$$

$$= \sum W^T (x - m_i)(x - m_i)^T W$$

$$= W^T S_i W$$

S_i → Within class scatter for D_i

$$(\tilde{m}_1 - \tilde{m}_2)^2 = \left[W^T(m_1 - m_2)\right]^2$$

$m_2 \rightarrow$ mean of patterns in D_2

$m_1 \rightarrow$ " " " D_1

$$= W^T(m_1 - m_2)(m_1 - m_2)^T W$$

$$= W^T S_B W$$

$S_B \rightarrow$ Between Class Scatter

$$\tilde{s}_i^2 = \sum_{y \in Y_i} (y - \tilde{m}_i)^2$$

$$= \sum_{x \in D_i} (W^T x - W^T m_i)^2$$

$$= \sum W^T (x - m_i)(x - m_i)^T W$$

$$= W^T S_i W$$

$S_i \rightarrow$ Within class scatter for D_i

$$\tilde{s}_1^2 + \tilde{s}_2^2 = W^T S_1 W + W^T S_2 W$$

$$= W^T (S_1 + S_2) W$$

$$= W^T S_W W$$

$S_W \rightarrow$ Total "within class scatter".

$$J(W) = \frac{W^T S_B W}{W^T S_W W}$$

W that maximizes J must satisfy

$$S_B W = \lambda S_W W \longrightarrow (1)$$

$$S_W^{-1} S_B W = \lambda W$$

W is eigen vector for $S_W^{-1} S_B$

λ is eigen value

$$J(W) = \frac{W^T S_B W}{W^T S_W W}$$

W that maximizes J must satisfy

$$S_B W = \lambda S_W W \quad \text{(1)}$$

$$S_W^{-1} S_B W = \lambda W$$

W is eigen vector for $S_W^{-1} S_B$

λ is eigen value

Reason is given in the supplementary slides towards the end.

$$J(W) = \frac{W^T S_B W}{W^T S_W W}$$

W that maximizes J must satisfy

$$S_B W = \lambda S_W W \longrightarrow (1)$$

$$S_W^{-1} S_B W = \lambda W$$

W is eigen vector for $S_W^{-1} S_B$

λ is eigen value

But there is no need to solve the eigen value problem.

$$S_B W = (m_1 - m_2) \underbrace{(m_1 - m_2)^T W}_{\text{scalar}}$$

$$= K(m_1 - m_2)$$

From (1),

$$K(m_1 - m_2) = \lambda S_W W$$

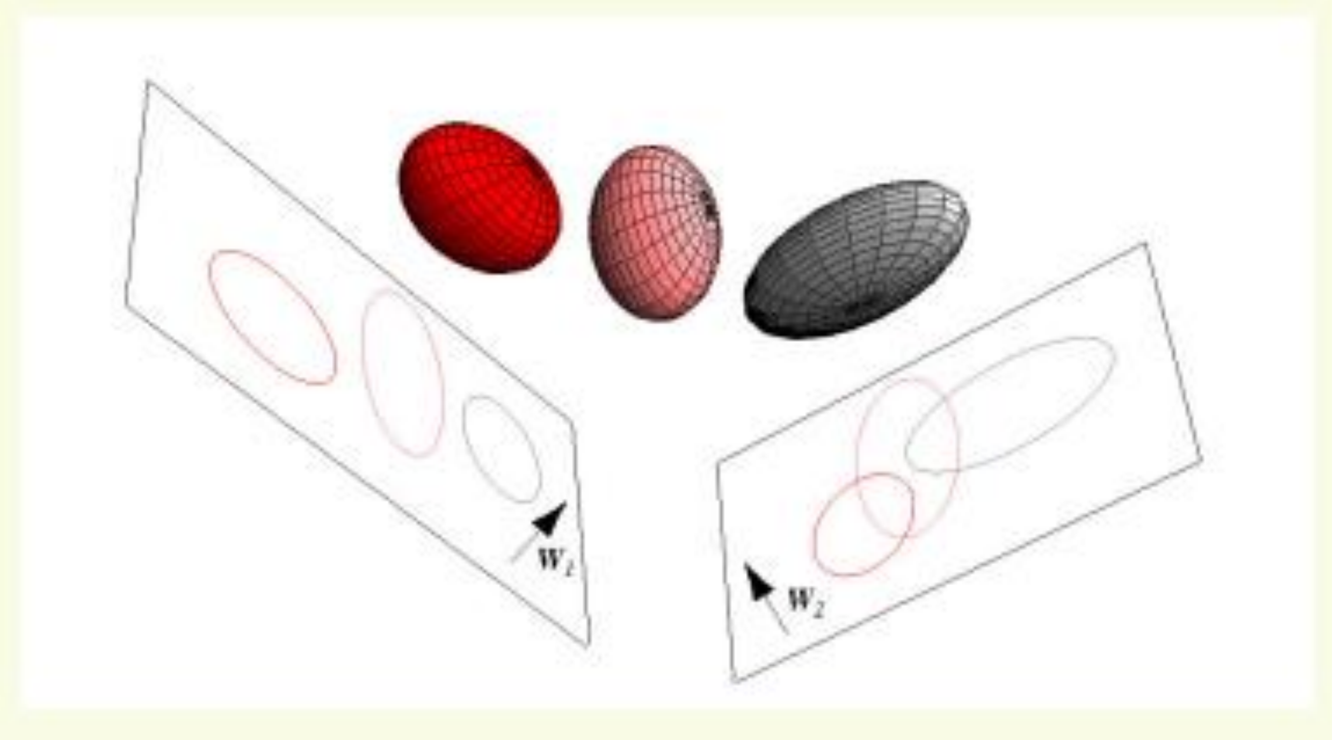
$$W = \frac{K}{\lambda} S_W^{-1} (m_1 - m_2)$$

Since only direction of W is important

$$W = S_W^{-1} (m_1 - m_2)$$

Extension to get more than 1D data

- Multiple discriminant analysis. Popularly known as Linear Discriminant Analysis (LDA).



- Similar to Kernel PCA, kernel Fisher discriminant is there.

Some supplementary material

SUPPLEMENTARY

Fisher Linear Discriminant Derivation

- Thus our objective function can be written:

$$J(v) = \frac{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2} = \frac{v^t S_B v}{v^t S_W v}$$

- Minimize $J(v)$ by taking the derivative w.r.t. v and setting it to 0

$$\frac{d}{dv} J(v) = \frac{\left(\frac{d}{dv} v^t S_B v\right) v^t S_W v - \left(\frac{d}{dv} v^t S_W v\right) v^t S_B v}{\left(v^t S_W v\right)^2}$$

$$= \frac{(2 S_B v) v^t S_W v - (2 S_W v) v^t S_B v}{\left(v^t S_W v\right)^2} = 0$$

Instead of *W*, *V* is used.

Ref: http://www.csd.uwo.ca/~olga/Courses/CS434a_541a/Lecture8.pdf

Fisher Linear Discriminant Derivation

- Need to solve $v^t S_W v (S_B v) - v^t S_B v (S_W v) = 0$

$$\Rightarrow \frac{v^t S_W v (S_B v)}{v^t S_W v} - \frac{v^t S_B v (S_W v)}{v^t S_W v} = 0$$

$$\Rightarrow S_B v - \underbrace{\frac{v^t S_B v (S_W v)}{v^t S_W v}}_{=\lambda} = 0$$

$$\Rightarrow \underbrace{S_B v = \lambda S_W v}_{\textit{generalized eigenvalue problem}}$$

Instead of *W*, *V* is used.

Ref: http://www.csd.uwo.ca/~olga/Courses/CS434a_541a/Lecture8.pdf

- http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf