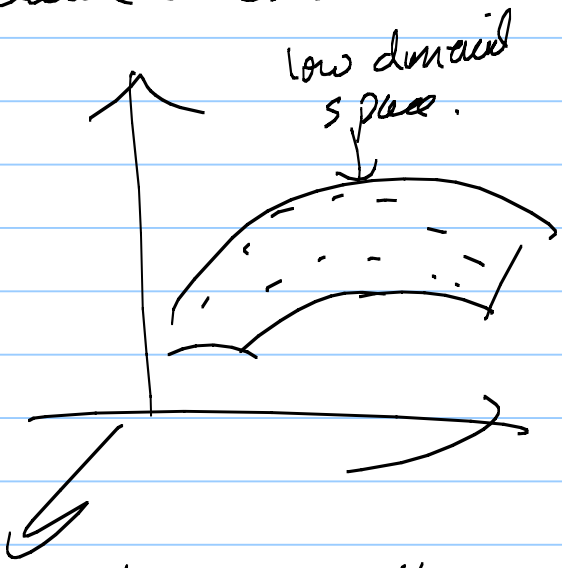
(1)

Spring 2013.

Principal Component Analysis (PCA)

Note Title 10/15/2006

One way to deal with the curse of dimensionality is to project data down onto a space of low dimensions.



There are a number of different techniques for doing this → eg. multidimensional scaling. Too many to deal with in this course.

Now we discuss the most basic method - Principal Component Analysis. (PCA)

(2)

CONVENTION: $\underline{\mu}^T\underline{\mu}$ is a scalar $\mu_1^2+\mu_2^2+\cdots+\mu_D^2$

$\underline{\mu}\,\underline{\mu}^T$ is a matrix $\begin{pmatrix} \mu_1^2 & \mu_1\mu_2 & \mu_1\mu_3 & \cdots \\ & \mu_2^2 & & \\ & & \ddots & \end{pmatrix}$

Spring 2013

(N.B. different convention than blackboard notes - but same as in book)

Data samples $\underline{x}_1, \ldots, \underline{x}_N$ in D-dim space.

Compute the mean $\underline{\mu} = \frac{1}{N}\sum_{i=1}^{N} \underline{x}_i$.

Compute the covariance:

$$\underline{\underline{K}} = \frac{1}{N}\sum_{i=1}^{N} (\underline{x}_i - \underline{\mu})(\underline{x}_i - \underline{\mu})^T.$$

Next compute the eigenvalues and eigenvectors of $\underline{\underline{K}}$

Solve $\underline{\underline{K}}\,\underline{e} = \lambda\,\underline{e}$

Note: $\underline{\underline{K}}$ is symmetric - so eigenvalues are real, eigenvectors are orthogonal.

$\lambda_1 > \lambda_2 > \ldots > \lambda_N$

PCA reduces the dimension by by keeping the eigenvectors $\underline{e}_i$ with $\lambda_i < T$ T threshold

Let M eigenvectors be kept.

Then project data $\underline{x}$ onto the subspace spanned by the first M eigenvectors. (After substracting out the mean!

(3)

Spring 2013

Formally:

Project. $\underline{x} - \underline{\mu} = \sum_{v=1}^{D} a_v \underline{e}_v$

where the coefficients are given by

$a_v = (\underline{x} - \underline{\mu}) \cdot \underline{e}_v$

(orthogonality, means $\underline{e}_v \cdot \underline{e}_\mu = \delta_{v\mu}$ ✗ Kronecker delta)

Here $\underline{x} = \underline{\mu} + \sum_{v=1}^{D} \{ (\underline{x} - \underline{\mu}) \cdot \underline{e}_v \} \underline{e}_v$

no dimension reduction (no compression)

Then, approximate

$\underline{x} \approx \underline{\mu} + \sum_{v=1}^{M} \{ (\underline{x} - \underline{\mu}) \cdot \underline{e}_v \} \underline{e}_v$.

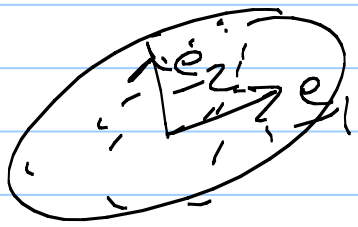
Projects the data into the M-dim subspace.

$\underline{\mu} + \sum_{v=1}^{M} b_v \underline{e}_v$. //

Spring 2013

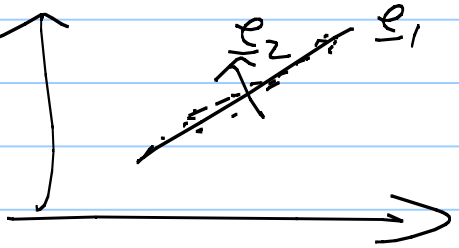
In 2-dimensions

Visually



The eigenvectors of Σ correspond to the second order moments of the data.

If the data lies (almost) on a straight line, then $\lambda_1 \gg 0, \lambda_2 \approx 0$



PCA and Gaussian Distributions

PCA is equivalent to performing ML estimation of the parameters of a Gaussian

$$p(x \mid \mu, \Sigma) = \frac{1}{\sqrt{2\pi}\sqrt{\det \Sigma}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}.$$

to get $\hat{\mu}, \hat{\Sigma}$. And then throw away the directions where the variance is small.

Spring 2013

Cost Function for PCA.

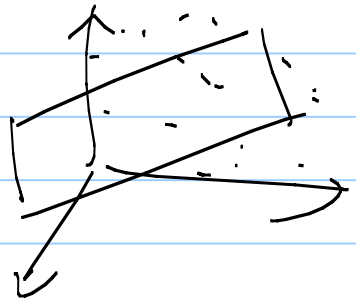
$$J(\underline{\mu}, \{a_{ki}\}, \{\underline{e}_i\}) = \sum_{k=1}^{N} \left\| \left(\underline{\mu} + \sum_{i=1}^{M} a_{ki}\,\underline{e}_i\right) - \underline{x}_k \right\|^2$$

Minimize J w.r.t. $\underline{\mu}, \{a_{ki}\}, \{\underline{e}_i\}$.

Data $\{\underline{x}_k : k = 1 \text{ to } N\}$.

The $\{a_{ki}\}$ are projection coefficients.

Intuition: find the M-dimensional subspace s.t. the projections of the data onto this subspace have minimal error.



Minimizing J, gives the $\{\hat{\underline{e}}_i\}$'s to be the eigenvectors of the covariance matrix

$$\underline{\underline{K}} = \frac{1}{N} \sum_{k=1}^{N} (\underline{x}_k - \underline{\mu})(\underline{x}_k - \underline{\mu})^T$$

$$\underline{\mu} = \frac{1}{N} \sum_{k=1}^{N} \underline{x}_k$$

$$\hat{a}_{ki} = (\underline{x}_k - \hat{\underline{\mu}}) \cdot \hat{\underline{e}}_i$$ the projection coefficients.

Spring 2013

To understand this fully, you must understand Singular Value Decomposition (SVD)

We can re-express the criteria as

$$J[\underline{\mu}, \{a_i\}, \{e_i\}] = \sum_{k=1}^{N} \sum_{b=1}^{D} \left((\mu_b - x_{bk}) + \sum_{i=1}^{m} a_{ki} e_{ib} \right)^2$$

where b denotes the vector components.

This is an example of a general class of problem.

Let $$E[\psi, e] = \sum_{a=1,k=1}^{a=D,k=N} \left(\tilde{x}_{ak} - \sum_{v=1}^{m} \psi_{av} \phi_{vk} \right)^2$$

Goal: minimize $E[\psi, e]$ w.r.t. ψ, e.

This is a bilinear problem, that can be solved by SVD.

Note: $\tilde{x}_{ak} = x_{ak} - \mu_a$

the position of the point, relative to the mean.

(7)

Spring 2013

Note: $\underline{\underline{X}}$ is not a square matrix. So it has no eigenvalues or eigenvectors.

SVD

We can express any $N \times D$ matrix $\underline{\underline{X}}$ X_{ak} in form

$$\underline{\underline{X}} = \underline{\underline{E}}\ \underline{\underline{D}}\ \underline{\underline{F}}$$

$$X_{ak} = \sum_{\mu,\nu=1}^{M} e_{a\mu} d_{\mu\nu} f_{\nu k}$$

where $\underline{\underline{D}} = \{d_{\mu\nu}\}$ is a diagonal matrix ($d_{\mu\nu}=0, \mu \neq \nu$)

$$\underline{\underline{D}} = \begin{pmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_M} \end{pmatrix}$$

, where the $\{\lambda_i\}$ are eigenvalues of $\underline{\underline{X}}\,\underline{\underline{X}}^T$ (equivalently of $\underline{\underline{X}}^T\underline{\underline{X}}$).

μ & ν label the eigenvectors.

$\underline{\underline{E}} = \{e_{a\mu}\}$ are eigenvectors of $(\underline{\underline{X}}\,\underline{\underline{X}}^T)_{ab}$

$\underline{\underline{F}} = \{f_{\nu k}\}$ are eigenvectors of $(\underline{\underline{X}}^T\underline{\underline{X}})_{kl}$

Note: For $\bar{\underline{\underline{X}}}$ defined on previous page, we get that

$$(\bar{\underline{\underline{X}}}\,\bar{\underline{\underline{X}}}^T) = \sum_{k=1}^{N} (\underline{x}_k - \underline{\mu})(\underline{x}_k - \underline{\mu})^T$$

Note: If $(\underline{\underline{X}}\,\underline{\underline{X}}^T)\underline{e} = \lambda \underline{e}$

then $(\underline{\underline{X}}^T\underline{\underline{X}})(\underline{\underline{X}}^T\underline{e}) = \lambda(\underline{\underline{X}}^T\underline{e})$

This relates the eigenvectors of $\underline{\underline{X}}\,\underline{\underline{X}}^T$ and of $\underline{\underline{X}}^T\underline{\underline{X}}$.

(Calculate the eigenvectors for the smallest matrix, then deduce those of the bigger matrix - $D < N$)

(8)

Spring 2013

Minimize:

$$E[\Psi, e] = \sum_{a=1,k=1}^{a=D,k=N} \left(\tilde{X}_{ak} - \sum_{v=1}^{M} \Psi_{av} \phi_{vk} \right)^2$$

we set.

$$\begin{cases} \Psi_{av} = \sqrt{d_{vv}}\, e_a^v \\ \phi_{vk} = \sqrt{d_{vv}}\, f_k^v \end{cases}$$

Take M biggest terms in the SVD expansion of $\underline{\underline{X}}$.

But there is an ambiguity.

$$\sum_{v=1}^{M} \Psi_{av} \phi_{vk} = \underline{\underline{\Psi}}\, \underline{\underline{\phi}}_{ak}$$ matrix multiplied.

$$= \underline{\underline{\Psi}}\, \underline{\underline{A}}\, \underline{\underline{A}}^{-1} \underline{\underline{\phi}}_{ak}$$

for any MxM invertible matrix $\underline{\underline{A}}$.

$$\underline{\underline{\Psi}} \to \underline{\underline{\Psi}}\, \underline{\underline{A}}$$
$$\underline{\underline{\phi}} \to \underline{\underline{A}}^{-1} \underline{\underline{\phi}}$$

This gets rid of the ambiguity.

For the PCA problem — we have constraints that the projection directions are orthogonal unit eigenvectors

(9)

Spring 2013

Relate SVD to PCA (Linear Algebra)

Start with an $n \times m$ matrix $\underline{\underline{X}}$.

$\underline{\underline{X}}\,\underline{\underline{X}}^T$ is a symmetric $n \times n$ matrix

$\underline{\underline{X}}^T\underline{\underline{X}}$ is a symmetric $m \times m$ matrix.

$\left(\underline{\underline{X}}\,\underline{\underline{X}}^T\right)^T = \underline{\underline{X}}\,\underline{\underline{X}}^T$

By standard linear algebra.

$\underline{\underline{X}}\,\underline{\underline{X}}^T \underline{e}^\mu = \lambda^\mu \underline{e}^\mu$ n eigenvalues λ^μ eigenvectors $\underline{e}^\mu$

eigenvectors are orthogonal $\underline{e}^\mu \cdot \underline{e}^\nu = \delta^{\mu\nu}$.

Similarly $\underline{\underline{X}}^T\underline{\underline{X}}\,\underline{f}^\nu = \tau^\nu \underline{f}^\nu$ m eigenvalues τ^ν eigenvectors $\underline{f}^\nu$ $\underline{f}^\mu \cdot \underline{f}^\nu = \delta^{\mu\nu}$.

The $\{\underline{e}^\mu\}$ and $\{\underline{f}^\mu\}$ are related because

$$\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)\left(\underline{\underline{X}}^T\underline{e}^\mu\right) = \lambda^\mu\left(\underline{\underline{X}}^T\underline{e}^\mu\right)$$

$$\left(\underline{\underline{X}}\,\underline{\underline{X}}^T\right)\left(\underline{\underline{X}}\,\underline{f}^\mu\right) = \tau^\mu\left(\underline{\underline{X}}\,\underline{f}^\mu\right)$$

Hence: $\underline{\underline{X}}^T\underline{e}^\mu \propto \underline{f}^\mu$, $\underline{\underline{X}}\,\underline{f}^\mu \propto \underline{e}^\mu$ $\lambda^\mu = \tau^\mu$

If $n > m$, then there are n eigenvectors $\{\underline{e}_\mu\}$ and m eigenvecton $\{\underline{f}_\mu\}$. So some $\underline{f}_\mu$ relate to several $\{\underline{e}_\mu\}$.

Spring 2013

Claim: we can express

$$\underline{\underline{X}} = \sum_\mu \alpha^\mu \underline{e}^\mu \underline{f}^{\mu T} \quad \text{for some } \alpha^\mu$$

$$\underline{\underline{X}}^T = \sum_\mu \alpha^\mu \underline{f}^\mu \underline{e}^{\mu T} \quad \left(\text{we will solve for } \alpha^\mu \text{ later.}\right)$$

Verify the claim $\underline{\underline{X}}\,\underline{f}^\nu = \sum_\mu \alpha^\mu \underline{e}^\mu \underline{f}^{\mu T} \underline{f}^\nu$

$= \sum_\mu \alpha^\mu \delta_{\mu\nu} \underline{e}^\mu = \alpha^\nu \underline{e}^\nu$. //

$$\underline{\underline{X}}\,\underline{\underline{X}}^T = \sum_{\mu,\nu} \alpha^\nu \underline{e}^\nu \underline{f}^{\nu T} \alpha^\mu \underline{f}^\mu \underline{e}^{\mu T}$$

$$= \sum_{\mu\nu} \alpha^\nu \alpha^\mu \underline{e}^\nu \delta_{\mu\nu} \underline{e}^{\mu T} = \sum_\mu (\alpha^\mu)^2 \underline{e}^\mu \underline{e}^{\mu T}.$$

Similarly $\underline{\underline{X}}^T\underline{\underline{X}} = \sum_\mu (\alpha^\mu)^2 \underline{f}^\mu \underline{f}^{\mu T}$, so $(\alpha^\mu)^2 = \lambda^\mu$ //

(Because we can express a symmetric matrix in form $\sum_\mu \chi_\mu \underline{e}^\mu \underline{e}^{\mu T}$ //

$$\underline{\underline{X}} = \sum_\mu \alpha^\mu \underline{e}^\mu \underline{f}^{\mu T} \quad \text{is the SVD of } \underline{\underline{X}}$$

In coordinates: $X_{ai} = \sum_\mu \alpha^\mu e^\mu_a f^\mu_i$

$$X_{ai} = \sum_{\mu,\nu} e^\mu_a \alpha^\mu \delta_{\mu\nu} f^\nu_i$$

$$\underline{\underline{X}} = \underline{\underline{E}}\;\underline{\underline{D}}\;\underline{\underline{F}} \qquad E_{a\mu} = e^\mu_a,\ \underline{\underline{D}}_{\mu\nu} = \alpha^\mu \delta_{\mu\nu}$$

$$F_{\nu i} = f^\nu_i$$

Spring 2013

Effectiveness of PCA.

In practice, PCA is often effective at minimum reduction of data dimension

But it will not be effective for some problems -

For example, if the data is a set of strings $(1,0,0,0\ldots1) = \underline{x}_1$

$(0,1,0,0\ldots1) = \underline{x}_2$

$(0,0,0,0\ldots,0,1) = \underline{x}_n$

then the eigenvalues do not fall off as PCA requires.

(12)

Fisher's Linear Discriminant

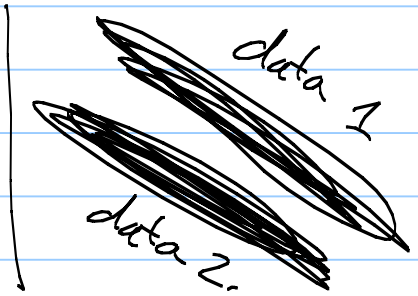
Spring 2013

Note Title 10/15/2006

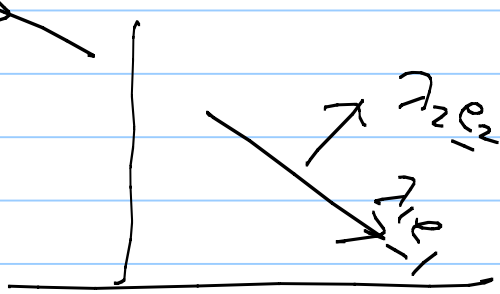
PCA may not be the best way to reduce the dimensions if the goal is discrimination.

Suppose you want to discriminate between two classes of data 1 & 2.

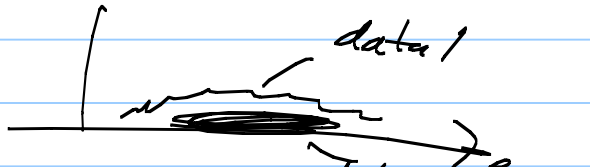
If you put both sets of data into PCA, you will get this



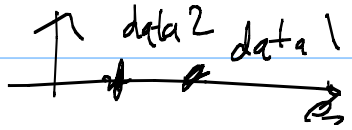
The best axis, according to PCA is in the worst direction for segmentation.



Projecting datasets onto e_1 gives



The second direction (e_2) would be far better. How to get this



(13)

Spring 2013

Fisher's Linear Discriminant gives a way to find a better projection direction.

n_1 samples $\underline{x}_i$ from class χ_1
n_2 samples $\underline{x}_i$ from class χ_2

Goal: find a vector $\underline{w}$, project data onto this axis (i.e. $\underline{x}_i \cdot \underline{w}$) so that the data is well separated.

Define the sample means

$$\underline{m}_i = \frac{1}{N_i} \sum_{\underline{x} \in \chi_i} \underline{x}$$

Define scatter matrices

$$\underline{\underline{S}}_i = \sum_{\underline{x} \in \chi_i} (\underline{x} - \underline{m}_i)(\underline{x} - \underline{m}_i)^T$$

Define the between-class scatter.

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

within-class scatter

$$\underline{\underline{S}}_W = \underline{\underline{S}}_1 + \underline{\underline{S}}_2 .$$

(14)

Spring 2013

Now project onto an (unknown) direction $\underline{w}$.

$$\hat{m}_i = \frac{1}{N_i} \sum_{\underline{x} \in \chi_i} \underline{w} \cdot \underline{x} = \underline{w} \cdot \underline{m}_i$$

// The means of the projections are the projections of the means.

The scatter of the projected points is

$$\hat{s}_i^2 = \sum_{x \in \chi_i} (\underline{w} \cdot \underline{x} - \underline{w} \cdot \underline{m}_i)^2.$$

$$= \underline{w}^T \underline{\underline{S}}_i \underline{w}.$$

Fisher's Criterion:

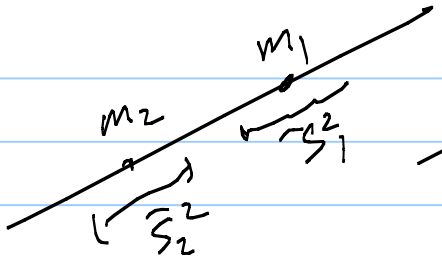
Choose the projection direction $\underline{w}$ to maximize:

$$J(\underline{w}) = \frac{|\hat{\underline{m}}_1 - \hat{\underline{m}}_2|^2}{\hat{s}_1^2 + \hat{s}_2^2}$$

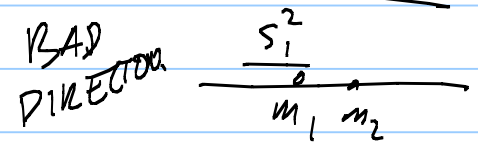
maximizes the ratio of the between-class distance to the within-class scatter.

15)

Spring 2013



This is a good projection direction.



Result: The projection direction that maximizes $J(\underline{w})$ is

$$\underline{w} = \underline{\underline{S}}_w^{-1}(\underline{m}_1 - \underline{m}_2).$$

Proof.

maximize $\underline{w}^T \underline{\underline{S}}_B \underline{w} - \lambda(\underline{w}^T \underline{\underline{S}}_w \underline{w} - \tau)$

λ: Lagrange multiplier; τ: constant.

$$\frac{\partial}{\partial \underline{w}}$$

$$\rightarrow \quad \underline{\underline{S}}_B \underline{w} - \lambda \underline{\underline{S}}_w \underline{w} = 0.$$

Hence $\quad \underline{\underline{S}}_w^{-1} \underline{\underline{S}}_B \underline{w} = \lambda \underline{w}.$

But $\quad \underline{\underline{S}}_B = (\underline{m}_1 - \underline{m}_2)^T (\underline{m}_1 - \underline{m}_2)$

so $\quad \underline{\underline{S}}_B \cdot \underline{w} = \rho(\underline{m}_1 - \underline{m}_2)$ for some ρ.

Hence $\quad \underline{\underline{S}}_B \hat{\underline{w}} \propto (\underline{m}_1 - \underline{m}_2).$

(16)

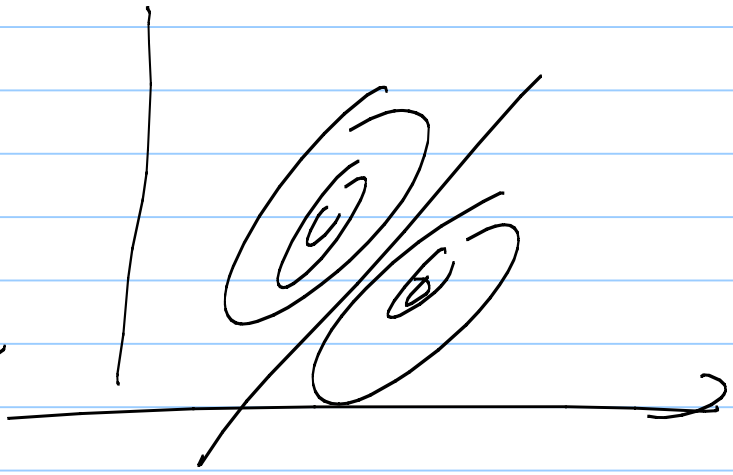
Spring 2017

Fisher's Linear Discriminant

If the data comes from two Gaussians with same covariance $\underline{\underline{\Sigma}}$ and means $\underline{\mu}_1, \underline{\mu}_2$.

Then the Bayes classifier is a straight line whose normal is the direction $\underline{\omega}$

$$\underline{\omega} \cdot \underline{x} + \omega_0 = 0, \qquad \underline{\omega} = \underline{\underline{\Sigma}}^{-1}(\underline{\mu}_1 - \underline{\mu}_2).$$



But. if the data comes from two Gaussian with different covariances, then Bayes classifier is a quadratic curve, so it differs from Fisher's linear discriminant.

Spring 2013

Multiple Classes:

For C classes, compute C-1 discriminants
Project D-dimensional features into C-1 space.

Within class

$$\underline{\underline{S}}_W = \underline{\underline{S}}_1 + \ldots + \underline{\underline{S}}_{C-1}$$

Between-class

$$\underline{\underline{S}}_B = \underline{\underline{S}}_{total} - \underline{\underline{S}}_W$$

$\underline{\underline{S}}_{total}$ is the scatter matrix for all the classes.

$$= \sum_{i=1}^{C} n_i (\underline{m}_i - \underline{m})(\underline{m}_i - \underline{m})^T$$

Multiple Discriminant Analysis:

Seek vectors $\omega_i : i = 1, \ldots C-1$,
project samples to C-1 dim space:

$$(\omega_1 \cdot x, \ldots, \omega_{C-1} \cdot x) = \underline{\underline{W}}^T \underline{x}.$$

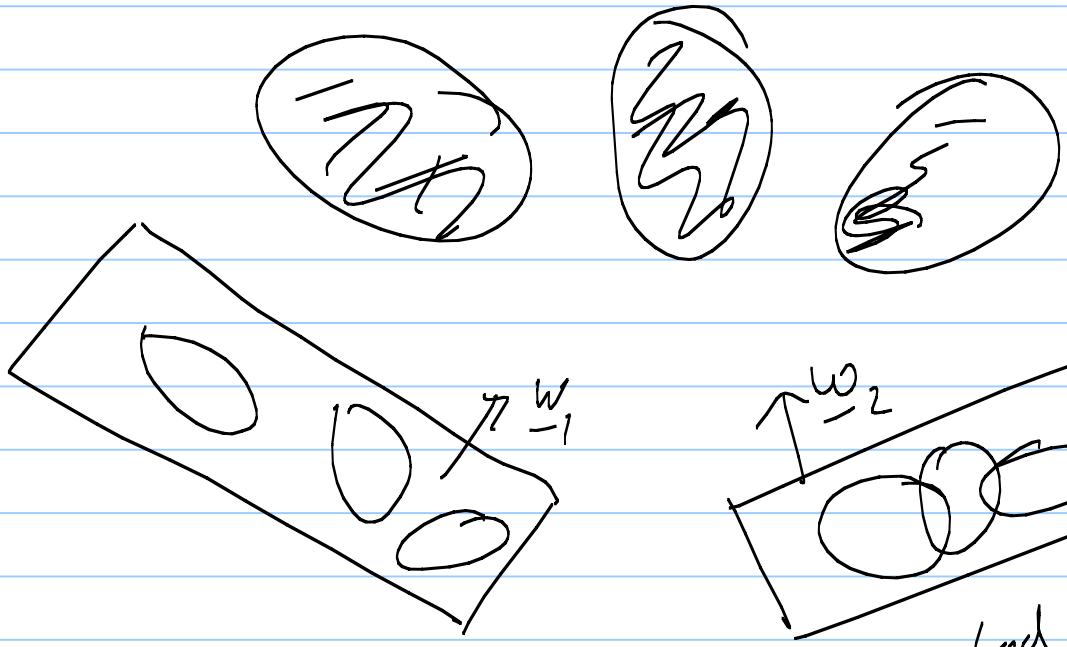
Criterion is $J(W) = \dfrac{|\underline{\underline{W}}^T \underline{\underline{S}}_B \underline{\underline{W}}|}{|\underline{\underline{W}}^T \underline{\underline{S}}_W \underline{\underline{W}}|}$

$|\cdot|$ is determinant

(12)

Spring 2013

The Solution is given by the eigenvectors, whose eigenvalues are the c-1 largest in $\underline{\underline{S}}_B \underline{w} = \lambda \underline{\underline{S}}_w \underline{w}$.



$\underline{w}_1$ is a good projection of the data.

$\underline{w}_2$ is a bad projection.