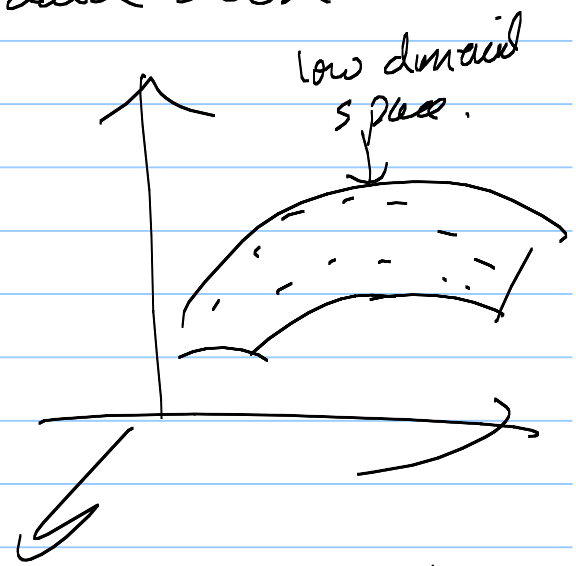
(1)

Spring 2014

Note Title 10/15/2006

Principal Component Analysis (PCA)

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 → e.g. multidimensional scaling. Too many to deal with in this course.

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

(2) <u>CONVENTION</u>: $\underline{\mu}^T \underline{\mu}$ is a scalar $\mu_1^2 + \mu_2^2 + \dots + \mu_D^2$

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

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

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

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

<u>Compute</u> the covariance:

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

<u>Next</u> 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.

$\lambda_1 > \lambda_2 > \dots > \lambda_N$

<u>PCA</u> reduces the dimension by

by keeping the eigenvectors $\underline{e}_i$ with $\lambda_i < T$
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)

Formally:

Project. $\underline{x} - \underline{\mu} = \sum_{\nu=1}^{D} a_\nu \underline{e}_\nu$

where the coefficients are given by

$a_\nu = (\underline{x} - \underline{\mu}) \cdot \underline{e}_\nu$ (orthogonality, means $\underline{e}_\nu \cdot \underline{e}_\mu = \delta_{\nu\mu}$ Kronecker delta)

Hence $\underline{x} = \underline{\mu} + \sum_{\nu=1}^{D} \{ (\underline{x} - \underline{\mu}) \cdot \underline{e}_\nu \} \underline{e}_\nu$

no dimension reduction (no compression)

Then, approximate

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

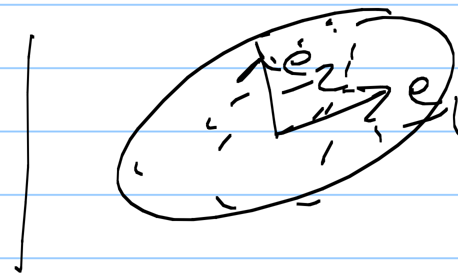
Projects the data onto the M-dim subspace.

$\underline{\mu} + \sum_{\nu=1}^{M} b_\nu \underline{e}_\nu$. //

(4)

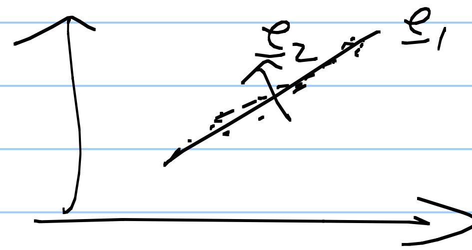
In 2-dimensions

Visually



The eigenvectors of $\underline{\underline{K}}$ correspond to the second order moments of the data.

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



PCA and Gaussian Distribution

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

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

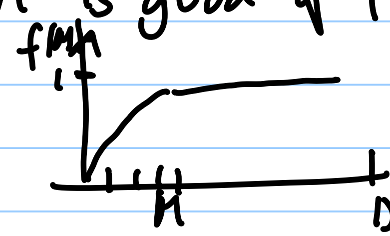
to get $\hat{\underline{\mu}}, \hat{\underline{\underline{\Sigma}}}$. And then throw away the directions where the variance is small.

(5) When is PCA good?

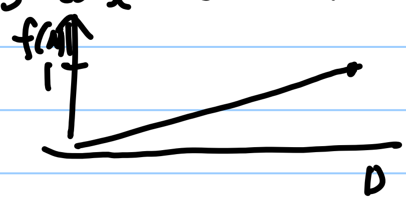
It is almost always a good technique to try, because it is so simple. Obtain the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_D$ and plot $f(M) = \dfrac{\sum_{i=1}^{M} \lambda_i}{\sum_{i=1}^{D} \lambda_i}$. , $f(M)$ increases with M and take max value 1, at $M=D$.

PCA is good if $f(M)$ asymptotes rapidly to 1. This happens if the first few eigenvalues are big and the remainder are small.

E.g.



PCA is bad if all the eigenvalues are roughly equal. The worst case is when all eigenvalues are the same. Then we have



This can happen. For example if the data is a set of strings.

$(1, 0, 0, 0 \ldots) = \underline{x}_1$

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

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

(6) What is PCA doing?

There are two equivalent ways to interpret PCA:

(1) Minimize the projection error.

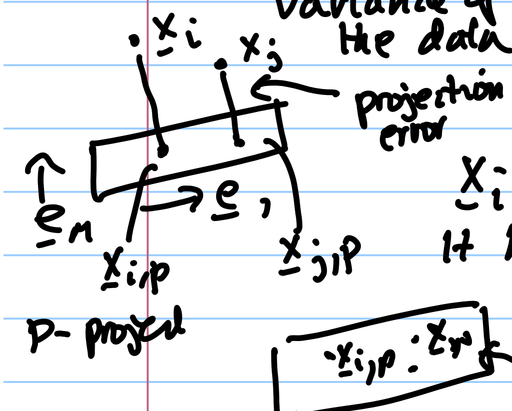
(2) Maximize the variance of the projection.

Consider $\frac{1}{N}\sum_{i=1}^{N} (\underline{x}_i-\underline{\mu})^2$. This is the variance of the data. It is independent of the projection.

We can express $(\underline{x}_i-\underline{\mu})^2 = \sum_{\nu=1}^{D} \{(\underline{x}_i-\underline{\mu})\cdot\underline{e}_\nu\}^2$

$\boxed{\underline{e}_\nu\text{'s are the eigenvectors of the correlation}}$

Hence $\underbrace{\frac{1}{N}\sum_{i=1}^{N}(\underline{x}_i-\underline{\mu})^2}_{\text{variance of the data}} = \underbrace{\frac{1}{N}\sum_{i=1}^{N}\sum_{\nu=1}^{M}\{(\underline{x}_i-\underline{\mu})\cdot\underline{e}_\nu\}^2}_{\text{variance of data within plane } \underline{e}_1, \dots \underline{e}_M} + \underbrace{\frac{1}{N}\sum_{i=1}^{N}\sum_{\nu=M+1}^{D}\{(\underline{x}_i-\underline{\mu})\cdot\underline{e}_\nu\}^2}_{\text{projection error}}$



$\underline{x}_i$ projects to $\underline{x}_{i,p} = \underline{\mu} + \sum_{\nu=1}^{M}\{(\underline{x}_i-\underline{\mu})\cdot\underline{e}_\nu\}\underline{e}_\nu$

It has projection error $\sum_{\nu=M+1}^{D}\{(\underline{x}_i-\underline{\mu})\cdot\underline{e}_\nu\}^2$.

The projected points have variance.

The sum of the projection error and the variance of projection are constant. So maximizing one is equivalent to minimizing the other.

Also this relationship can be expressed in terms of the eigenvalues.

It reduces to $\sum_{\nu=1}^{D}\lambda_\nu = \sum_{\nu=1}^{M}\lambda_\nu + \sum_{\nu=M+1}^{D}\lambda_\nu$

To see this, $\frac{1}{N}\sum_{i=1}^{N}(\underline{x}_i-\underline{\mu})\cdot(\underline{x}_i-\underline{\mu}) = \mathrm{Trace}(\underline{C}) = \sum_{\nu=1}^{D}\lambda_\nu$.

The variance of the projection is $\sum_{\nu=1}^{M}\lambda_\nu$, by similar reasoning.

Hence projection error is $\sum_{\nu=M+1}^{D}\lambda_\nu$. //

(7) To understand why PCA minimizes, or maximizes these terms we must express the criterion slightly differently. Then we use Singular Value Decomposition (SVD), which is advanced material

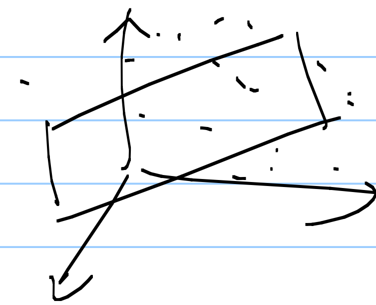
$$J(\underline{\mu}, \{a\}, \{e\}) = \sum_{k=1}^{N} \left\| \left(\underline{\mu} + \sum_{i=1}^{M} a_{ki} \underline{e}_i\right) - \underline{x}_k \right\|^2$$

<u>Minimize</u> J w.r.t. $\underline{\mu}, \{a\}, \{e\}$.

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

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

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



<u>Minimizing J</u>, gives the $\{\underline{\hat{e}}_i\}$'s to be the eigenvectors of the covariance matrix $\underline{\underline{K}} = \frac{1}{N} \sum_{k=1}^{N} (\underline{x}_k - \underline{\mu})(\underline{x}_k - \underline{\mu})^T$

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

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

(8)

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

We can re-express the criteria as

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

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.

(9)

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 $\underline{\underline{X}}$ $\quad X_{ak}$

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 ($d_{\mu\nu}=0, \mu\neq\nu$)
$$\underline{\underline{D}} = \begin{pmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_M} \end{pmatrix}$$
, 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 $\tilde{\underline{\underline{X}}}$ defined on previous page, we get that
$$(\tilde{\underline{\underline{X}}}\,\tilde{\underline{\underline{X}}}^T) = \sum_{k=1}^{N} (\underline{x}_k - \underline{\mu})(\underline{x}_k - \underline{\mu})^T$$

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 $\underline{\underline{X}}\,\underline{\underline{X}}^T$.
(Calculate the eigenvectors for the smallest matrix, then deduce those of the bigger matrix - $D < N$)

(10)

Minimize:

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

we set.

$$\begin{cases} \psi_{a\nu} = \sqrt{d_{\nu\nu}}\, e_a^{\nu} \\ \phi_{\nu k} = \sqrt{d_{\nu\nu}}\, f_k^{\nu} \end{cases}$$

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

<u>But</u> there is an ambiguity.

$$\sum_{\nu=1}^{M} \psi_{a\nu} \phi_{\nu k} = \left(\underline{\underline{\Psi}}\, \underline{\underline{\Phi}}\right)_{ak}$$

matrix multiplied.

$$= \left(\underline{\underline{\Psi}}\, \underline{\underline{A}}\, \underline{\underline{A}}^{-1} \underline{\underline{\Phi}}\right)_{ak}$$

for any $M \times M$ 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

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.

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

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 $\underline{\underline{X}}\,\underline{\underline{X}}^T \underline{f}^\nu = \tau^\nu \underline{f}^\nu$

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

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

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

<u>Hence:</u> $\underline{\underline{X}}^T\underline{e}^\mu \propto \underline{f}^\mu$, $\underline{\underline{X}}\,\underline{f}^\mu \propto \underline{e}^\mu$ $\quad \lambda^\mu = \tau^\mu$

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

(12)

Claim: we can express

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

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

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 $\sum_\mu x_\mu \underline{e}^\mu \underline{e}^{\mu T}$ //

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

In coordinates:

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

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

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

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

(13)

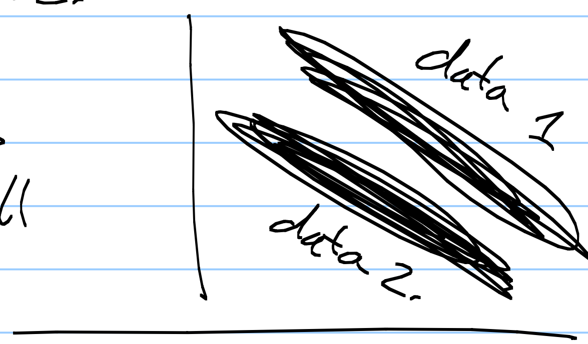
Fisher's Linear Discriminant

Note Title 10/15/2006

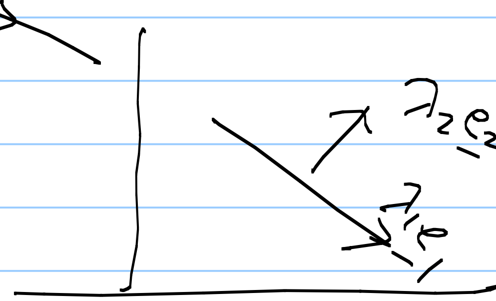
PCA may not be the best way to reduce the dimension if the goal is discrimination.

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

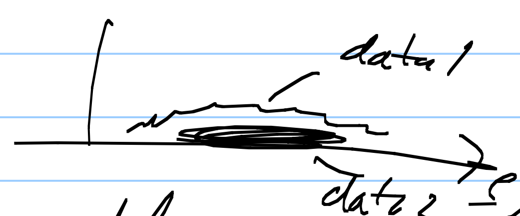
If you put both sets of data into PCA, you will get this



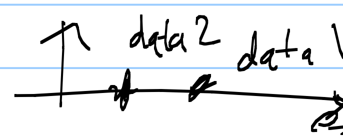
The best axis, according to PCA is in the worst direction for segmentation.



Projecting datasets onto $\underline{e}$ gives.



The second direction $(\underline{e}_2)$ would be far better. How to get this.



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

n_1 samples $\underline{x}_i$ from class $\mathcal{X}_1$
n_2 samples $\underline{x}_i$ from class $\mathcal{X}_2$

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

Define the sample means

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

Define scatter matrices

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

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

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

(15)

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

$$\hat{m}_i = \frac{1}{N_i} \sum_{\underline{x} \in \mathcal{X}_i} \underline{\omega} \cdot \underline{x} = \underline{\omega} \cdot \underline{m}_i$$

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

The scatter of the projected points is

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

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

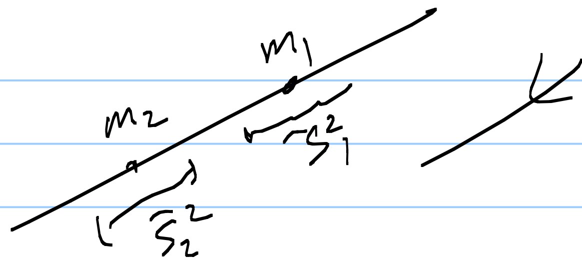
Fisher's Criterion:

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

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

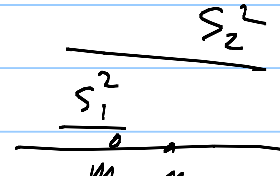
maximizes the ratio of the between-class distance to the within-class scatter.

(16)



This is a good projection direction.

BAD 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 $\underline{w}^T \underline{\underline{S}}_B \underline{w} - \underbrace{\lambda}_{\text{Lagrange multiplier}} (\underline{w}^T \underline{\underline{S}}_w \underline{w} - \underbrace{\tau}_{\text{constant}})$

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

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

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

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

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

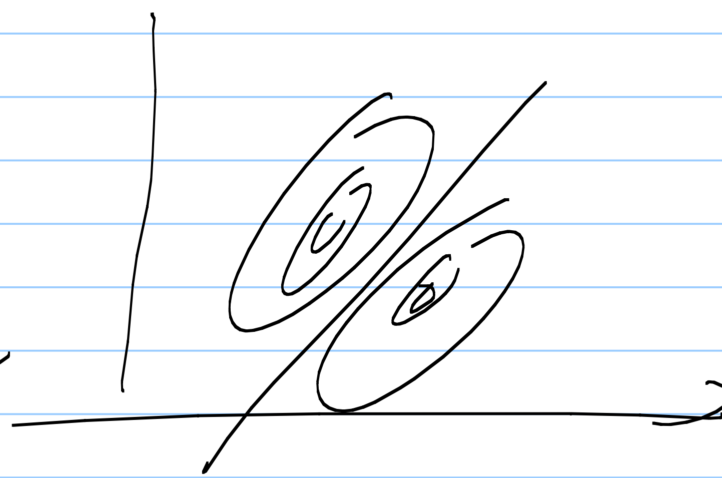
Hence $\underline{\underline{S}}_B \hat{\underline{w}} \propto (\underline{m}_1 - \underline{m}_2)$.

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.

(18)

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

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

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

Multiple Discriminant Analysis:

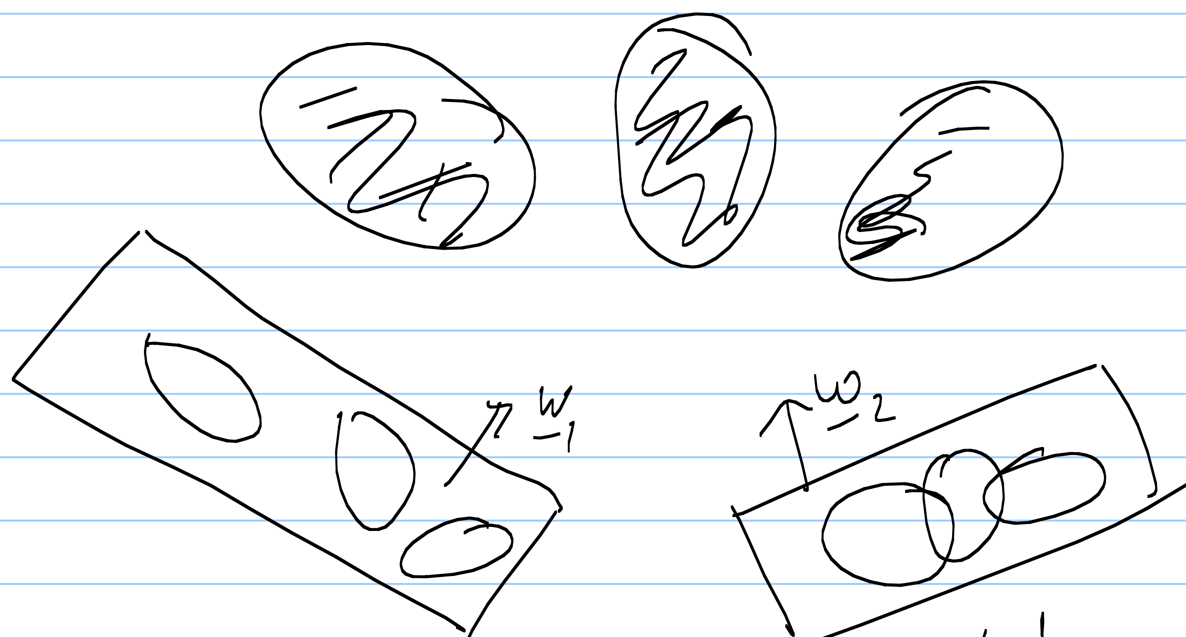
Seek vectors $\omega_i : i = 1, \ldots c-1$,
Project samples to c-1 dim space:

$$(\omega_i \cdot x, \ldots, \omega_{c-1} \cdot x) = \underline{\underline{\omega}}^T \underline{x}.$$

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

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.