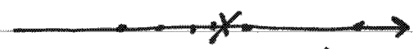
.8.1 PCA

- consider 1-D data, n points
- Imagine we want to characterize the data by a single value



→ could take the mean & minimize distance of data set to representative point
(would be better to take the median)

Now imagine d dimension → we can take the mean vector as the representative $\frac{1}{n}\sum_{k=1}^{n} x_k$

note: the mean vector minimizes the sum of squared distances between a single point and the points in the data set.

note: mean vector is a 0-dimension representation of the data set

Q: What would a 1-D representation be?

A: a line projected through the mean vector

Let e be a unit vector in the direction of our line

$x = m + ae$ is the equation of the line

distance of x from m

$x_k = m + a_k e$

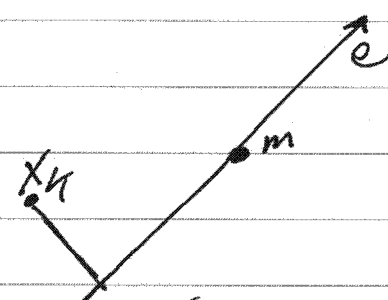
note: optimal coefficients for each data set can be derived by minimizing the squared error criteria function

$$J_1(a_1, \ldots, a_n, e) = \sum_{k=1}^{n} \|(m + a_k e) - x_k\|^2$$

In eqs 82 & 83 text shows $a_k = e^t(x_k - m)$

i.e. the ~~best~~ least squares soln is a projection of x_k onto the line in the direction of e that passes thru m.

ex:



→ the interesting problem is to find the best direction e for the data set.

Ex: consider 3-D data, a cloud of points in the shape of a football.

Q: What's the best direction e for ~~and~~ our football cloud of points

A: a line through the 2 pointy ends

obs: this is the eigenvector of the largest eigenvalue of the scatter matrix

$$S = \sum_{k=1}^{n} (x_k - m)(x_k - m)^t$$

We can extend the idea of a linear projection to d' dimension projection

<u>recall</u>: for linear projection we have $x = m + ae$

$\Longrightarrow$ for d' projection $x = m + \sum_{i=1}^{d'} a_i e_i \quad d' \leq d$

<u>obs</u>: criteria fcn $J_{d'} = \sum_{k=1}^{n} \left\| \left(m + \sum_{i=1}^{d'} a_{ki} e_i \right) - x_k \right\|^2$

is minimized when e_i are the d' largest eigenvectors of S

<u>obs</u>: these eigenvectors are orthogonal

$\rightarrow$ these form a natural basis set for the points x

Q: So what are the "principal components"?

A: the coefficients a_i in $x = m + \sum_{i=1}^{d'} a_i e_i$

<u>Practical issue</u>: PCA focusses on directions in which the scatter cloud S is greatest

$\rightarrow$ it is accounting for variance

!! noise features can be a major problem

3.8.2 Fisher Linear Discriminant

PCA allows ~~us~~ to project onto fewer dimensions
→ these dimensions best represent the data

Q: What if we wanted to project onto dimensions that best discriminate between classes?

→ ~~the~~ In the extreme case project onto 1 dimension

goal: find the best line so that classes are separate

$\mathcal{D} = \{ x_1, .., x_n \}$ n_1 in D_1 (labelled ω_1) + n_2 in D_2 (labelled ω_2)

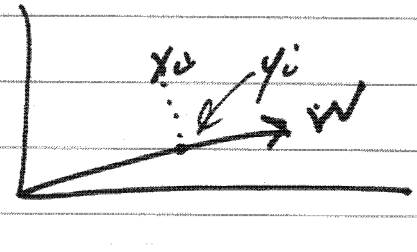
we get $y = w^t x$

y ← scalar; $w^t x$ ← dot product

resulting in $y_1, .., y_n$ in subsets $\mathcal{Y}_1$ + $\mathcal{Y}_2$

if $||w|| = 1$, i.e, a unit vector

→ y_i is the projection of x_i onto a line in the direction w



goal: we want a w s.t. x_i in ω_1 cluster separately from x_i in ω_2

note: such a w may not exist if the distributions are multimodal overlap

Q: How do we find the best w?

~~A~~ this is the Fisher linear discriminant problem

First, find the sample mean $m_i = \frac{1}{n_i} \sum_{x \in D_i} x$

find the ~~projection of that mean vector~~ sample mean of projected points $\tilde{m}_i = \frac{1}{n_i} \sum_{y \in Y_i} y$

note $\tilde{m}_i = \frac{1}{n_i} \sum_{x \in D_i} w^t x = w^t m_i$

i.e. the "sample mean of projected points" is the same as the "projection of the mean vector"

Q: What is the distance between the projected means?

A: $|\tilde{m}_1 - \tilde{m}_2| = |w^t (m_1 - m_2)|$

note: we want to make this large relative to the scatter of each class

Scatter: define scatter of projected points $\tilde{s}_i^2 = \sum_{y \in Y_i} (y - \tilde{m}_i)^2$

def $\tilde{s}_1^2 + \tilde{s}_2^2$ is the within-class scatter

goal: find w that maximizes $J(w) = \dfrac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$

First def. scatter matrices

$$S_i = \sum_{x \in D_i} (x - m_i)(x - m_i)^t$$

$$S_W = S_1 + S_2$$

called "within-class" scatter matrix

$\tilde{s}_i^2$ can be expressed in terms of S_i

recall $y = w^t x$, $\tilde{m}_i = w^t m_i$, $\tilde{s}_i^2 = \sum_{y \in \mathcal{Y}_i} (y - \tilde{m}_i)^2$

$$\Rightarrow \tilde{s}_i^2 = \sum_{x \in D_i} (w^t x - w^t m_i)^2$$

$$= \sum w^t (x - m_i)(x - m_i)^t w$$

$$= w^t S_i w$$

hence $\tilde{s}_1^2 + \tilde{s}_2^2 = w^t S_W w$

called "between-class" scatter matrix

$S_B = (m_1 - m_2)(m_1 - m_2)^t$

similarly $(\tilde{m}_1 - \tilde{m}_2)^2 = w^t S_B w$ (from eq 101)

$$\Rightarrow J(w) = \frac{w^t S_B w}{w^t S_W w}$$

S_W – symmetric + positive semidefinite, usu nonsingular

S_B – " " "

always has rank at most 1 (prod of vectors)

goal: solve $J(w)$ for w.   soln: $w = S_W^{-1}(m_1 - m_2)$

obviously S_W must be nonsingular

maximizes ratio of between-class scatter to within-class scatter

So now we know the line we want to project onto

Q: What point on that line is the classification threshold?

A: choose point w_0 where posteriors are equal

i.e. $p(\omega_1 | \underset{w_0}{\cancel{x}}) = p(\omega_2 | \underset{w_0}{\cancel{x}})$