EE-559

HOMWORK-5.

1. $X \rightarrow$ Zero mean p-dimensional random vector.

i.e $E[X] = 0$

Covariance matrix : $R = E[X X^T]$

estimation: $\hat{X} = \sum_{i=1}^{M} \alpha_i e_i$

e_i = orthonormal eigenvectors of covariance matrix R

$\alpha = [\alpha_1 \cdots \alpha_p]^T$

To show: $J = \|x - \hat{x}\|^2 \xrightarrow{\min} \alpha_i = e_i^T x \quad i = 1, 2 \cdots M$

as principal component

Proof:

$$\|x - \hat{x}\|^2 = (x - \hat{x})^T (x - \hat{x})$$

Substituting $\hat{x} = \sum_{i=1}^{M} \alpha_i e_i$

$$J = \|x - \sum_{i=1}^{M} \alpha_i e_i\|^2 = (x - \sum_{i=1}^{M} \alpha_i e_i)^T (x - \sum_{i=1}^{M} \alpha_i e_i)$$

$$= x^T x - x^T \sum_{i=1}^{M} \alpha_i e_i - \left(\sum_{i=1}^{M} \alpha_i e_i\right)^T x + \left(\sum_{i=1}^{M} \alpha_i e_i\right)^T \left(\sum_{i=1}^{M} \alpha_i e_i\right)$$

$$= x^T x - \sum_{i=1}^{M} \alpha_i e_i x^T - \sum_{i=1}^{M} \alpha_i e_i^T x + \sum_{i=1}^{M} \sum_{j=1}^{M} \alpha_i \alpha_j e_i^T e_j$$

Let's use $\sum_{j=1}$ to separate & distinguish

$\because e_i$ = orthonormal,

$e_i^T e_j = \delta_{ij}$ where, $\delta_{ij} = 1$ if $i = j$

$\delta_{ij} = 0$ if $i \neq j$

$\therefore \sum_{i=1}^{M} \sum_{j=1}^{M} \alpha_i \alpha_j e_i^T e_j = \sum_{i=1}^{M} \alpha_i^2$

minimising J,

take derivative wrt α_i, & = 0

$$\frac{\partial J}{\partial \alpha_i} = \frac{\partial}{\partial \alpha_i}\left[x^T x - 2\sum_{i=1}^{M} \alpha_i x^T e_i + \sum_{i=1}^{M} \alpha_i^2 \right]$$

$$= -2x^T e_i + 2\alpha_i = 0$$

$$\Rightarrow \alpha_i = x^T e_i$$

i.e $\alpha_i = e_i^T x$ {dot product is commutative}

⇒ Principal components α_i are given by projection of the data vector x onto eigen vectors e_i.

2. $P(x|\omega_i) \rightarrow$ means $= \mu_i$

Co-variance matrix : Σ_i $i = 1, 2$

$y = w^T x$: projection.

$P(y|\omega_i) \rightarrow$ mean μ_i

var σ_i^2

(a) TO show:

$$J_1(w) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$ is max by

$$w = (\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2)$$

{reference : class notes}.

$$\mu_i = \frac{1}{\#D_i} \sum_{x \in D_i} x \quad 0 \le \mu_i \le 1$$

$$= \frac{1}{\#D_i} w^T \sum_{x \in D_i} x = w^T \mu_i$$

$\therefore$ $Var(y) = Var(w^T x)$

$= w^T Cov(x) w$

$\Rightarrow \sigma_i^2 = w^T \Sigma_i w$

Let $V = \mu_1 - \mu_2$

$\Sigma = \Sigma_1 + \Sigma_2$

$\therefore J_1(w) = \dfrac{(w^T(\mu_1 - \mu_2))^2}{w^T(\Sigma_1 + \Sigma_2)w}$

$= \dfrac{(w^T V)^2}{w^T \Sigma w}$

to maximise, we use Lagrange multipliers:-

$L(w, \lambda) = \dfrac{(w^T V)^2}{w^T \Sigma w} - \lambda(w^T w - 1)$

Considering the Rayleigh quotient:

$\dfrac{(w^T V)^2}{w^T \Sigma w}$. the Rayleigh quotient is maximised when w is $\propto \Sigma^{-1} V$

i.e $w \propto (\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2)$

(b) Let projection,

$y = w^T x$

with $\mu_i' = w^T \mu_i$ → mean & covariance matrix Σ_i

$\sigma_y^{(i)} = w^T \Sigma_i w$

Considering contributions from each class ω_1 & ω_2, weighted by their probabilities $P(\omega_1)$ and $P(\omega_2)$

Variance of y is

$\sigma_y^2 = P(\omega_1)\sigma_y^{(1)} + P(\omega_2)\sigma_y^{(2)}$

Where, $\sigma_y^{(1)} = w^T \Sigma_1 w$ $\quad \sigma_y^{(2)} = w^T \Sigma_2 w$

$$\therefore \sigma_y^2 = P(\omega_1) \underbrace{w^T \Sigma_1 w}_{\text{scalar}} + P(\omega_2) \underbrace{w^T \Sigma_2 w}$$

$$\therefore \sigma_y^2 = w^T (P(\omega_1)\Sigma_1 w) + w^T (P(\omega_2)\Sigma_2 w)$$

$$\sigma_y^2 = w^T (P(\omega_1)\Sigma_1 + P(\omega_2)\Sigma_2) w$$

$$\therefore \sigma_y^2 = w^T (P(\omega_1)\Sigma_1 + P(\omega_2)\Sigma_2) w$$

$$J_2(w) = \frac{(w^T(\mu_1 - \mu_2))^2}{w^T (P(\omega_1)\Sigma_1 + P(\omega_2)\Sigma_2) w}$$

Considering Rayleigh quotient:

$$\frac{(w^T a)^2}{w^T B w} \quad \text{where } a = \mu_1 - \mu_2$$

$$B = P(\omega_1)\Sigma_1 + P(\omega_2)\Sigma_2$$

it is max when $w \propto B^{-1} a$

i.e $w \propto (P(\omega_1)\Sigma_1 + P(\omega_2)\Sigma_2)^{-1} (\mu_1 - \mu_2)$

$\therefore J(w)$ is maximised when

$$w = [P(\omega_1)\Sigma_1 + P(\omega_2)\Sigma_2]^{-1} (\mu_1 - \mu_2)$$

(c) The objective function for Fisher's LDA can be written as:

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

$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$ (between-class scatter matrix).

$S_W = \Sigma_1 + \Sigma_2$ (within class scatter matrix)

$J_1(w)$ is closer to the criterion used by Fisher's LDA because it compares between class mean difference to the within-class variance without involving prior probabilities. Similar to original Fisher's LDA of max ratio of between-class scatter to within-class scatter.