HW 5

1. (PCA using MSE and population covariance matrix[1]) Assume that $\mathbf{x}$ is a zero-mean p dimensional random vector ($\mathbb{E}[\mathbf{x}] = \mathbf{0}$) with covariance matrix: (10 pts)

$$\mathbf{R} = \mathbb{E}[\mathbf{x}\mathbf{x}^T]$$

We wish to estimate $\mathbf{x}$ with $M \leq p$ *principal directions* as:

$$\hat{\mathbf{x}} = \sum_{i=1}^{M} \alpha_i \mathbf{e}_i$$

where $\mathbf{e}_i$'s are the orthonormal eigenvectors of the covariance matrix $\mathbf{R}$ and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_p]^T$. Show that the minimization of the squared error:

$$J = \|\mathbf{x} - \hat{\mathbf{x}}\|^2$$

with respect to $\alpha_1, \ldots, \alpha_m$ yields:

$$\alpha_i = \mathbf{e}_i^T \mathbf{x}, \; i = 1, 2, \ldots, M$$

as the *principal component*, that is, the projection of the data vector $\mathbf{x}$ onto the eigenvector $\mathbf{e}_i$.

proof: $\because \hat{x} = \sum_{i=1}^{M} \alpha_i e_i$

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

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

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

$\because \sum_{i=1}^{M} \alpha_i x^T e_i = \sum_{i=1}^{M} \alpha_i e_i^T x$

and $e_i^T e_j = \begin{cases} 1, & i=j \\ 0, & i \neq j \end{cases}$ ($\because$ orthonormal)

$\therefore J = x^T x - 2\sum_{i=1}^{M} \alpha_i e_i^T x + \sum_{i=1}^{M} \alpha_i^2$

$\therefore \frac{\partial J}{\partial \alpha_i} = -2 e_i^T x + 2\alpha_i$

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

$\therefore \alpha_i = e_i^T x$, $i = 1, 2, \cdots M$

QED

2. Let $p(\mathbf{x}|\omega_i)$ be arbitrary densities with means $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i$ — not necessarily normal — for $i = 1, 2$. Let y= $\mathbf{w}^T\mathbf{x}$ be a projection, and let the induced one-dimensional densities $p(y|\omega_i)$ have means μ_i and variances σ_i^2. (15 pts)

(a) Show that the criterion function

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

is maximized by

$$\mathbf{w} = (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

(b) If $P(\omega_i)$ is the prior probability for ω_i, show that the criterion function

$$J_2(\mathbf{w}) = \frac{(\mu_1 - \mu_2)^2}{P(\omega_1)\sigma_1^2 + P(\omega_2)\sigma_2^2}$$

is maximized by

$$\mathbf{w} = [P(\omega_1)\boldsymbol{\Sigma}_1 + P(\omega_2)\boldsymbol{\Sigma}_2]^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

(c) Explain which of $J(\mathbf{w}_1)$ and $J(\mathbf{w}_2)$ is "closer" to the criterion that is used by Fisher's LDA.

(a) proof: $\because$ after projection

$$\mu_i = E[y|w_i] = E[w^T x|w_i] = w^T E[x|w_i] = w^T \mu_i$$

$$\sigma_i^2 = Var(y|w_i) = Var(w^T x|w_i) = w^T \Sigma_i w$$

$$\therefore J_1(w) = \frac{(w^T\mu_1 - w^T\mu_2)^2}{w^T\Sigma_1 w + w^T\Sigma_2 w} = \frac{(w^T(\mu_1-\mu_2))^2}{w^T(\Sigma_1+\Sigma_2)w}$$ is generalized Rayleigh quotient

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

$\therefore w = (\Sigma_1+\Sigma_2)^{-1}(\mu_1-\mu_2)$ maximizes $J_1(w)$

QED

(b) proof: similarly $$J_2(w) = \frac{(w^T\mu_1 - w^T\mu_2)^2}{w^T P(w_1)\Sigma_1 w + w^T P(w_2)\Sigma_2 w}$$

$$= \frac{(w^T(\mu_1-\mu_2))^2}{w^T(P(w_1)\Sigma_1 + P(w_2)\Sigma_2)w}$$

$$\therefore w \propto (P(w_1)\Sigma_1 + P(w_2)\Sigma_2)^{-1}(\mu_1-\mu_2)$$

$\therefore w = (P(w_1)\Sigma_1 + P(w_2)\Sigma_2)^{-1}(\mu_1-\mu_2)$ maximizes $J_2(w)$

QED

(c) $\because J_{Fisher}(W) = \frac{(W^T(\mu_1-\mu_2))^2}{W^T S_W W}$, S_W is within-class scatter matrix

$\therefore$ for classical Fisher's LDA, it doesn't explicity incorporate prior probabilities

$\therefore J_1(w)$ is closer to the original Fisher's LDA concept

$\therefore J_2(w)$ considers class prior probability, it can be view as a Bayesian extension of Fisher's LDA