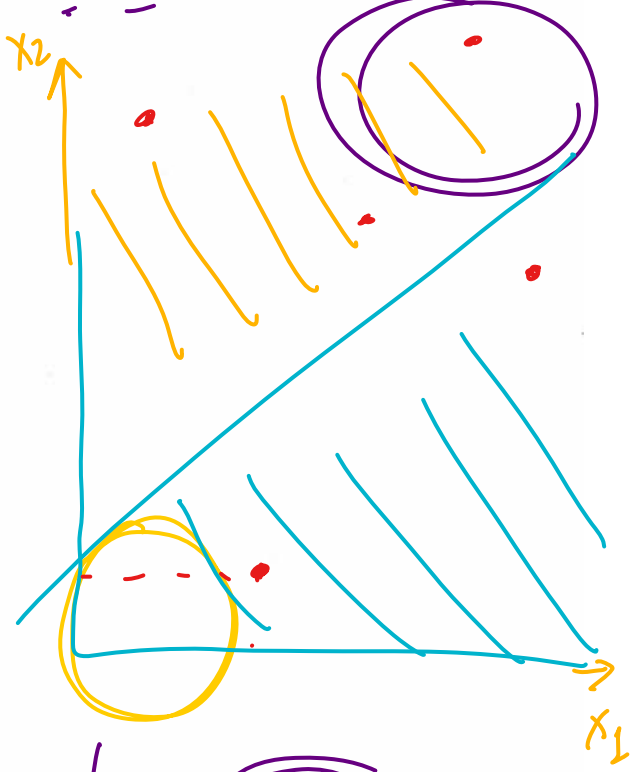
Discriminant Analysis

$X = (X_1, X_2, \ldots, X_p)'$, $p \geq 2$

$\Pi_1 \equiv N_p(\underset{\sim}{\mu_1}, \Sigma)$

$\Pi_2 \equiv N_p(\underset{\sim}{\mu_2}, \Sigma)$

$\mu_1, \mu_2, \Sigma =$ ΑΓΝΩΣΤΑ



$p = 2$

$\underset{\sim}{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \mathrm{Var}X_1 & \mathrm{Cov}X_1X_2 \\ \mathrm{Cov}X_1X_2 & \mathrm{Var}X_2 \end{pmatrix} \right)$

2×1 $\quad 2\times 1$ $\quad 2\times 2$

Πρόβλημα Ταξινόμησης

H_0: Άθως

H_1: Γλώσσα

$H_0: X \in \Pi_1$ (κέρινα) vs. $H_1: X \in \Pi_2$ (μηλέ)

Έστω $f(\,|\mu_i, \Sigma)$ η σ.π.π. του Π_i $\Longrightarrow$

X ταξινομείται στον Π_1 αν

$$\lambda = \frac{f(X|\mu_1, \Sigma)}{f(X|\mu_2, \Sigma)} > C$$

Λόγω Πιθανοφανειών

Likelihood Ratio Test

$f(x|\mu_1, \Sigma)$ για να H_0

$f(x|\mu_2, \Sigma)$ για να H_1

$X \in \Pi_1$ αν $\frac{f(X|\mu_1,\Sigma)}{f(X|\mu_2,\Sigma)} > 1$

Αν $f \sim N_p(\mu_i, \Sigma) \Longrightarrow$

Θέλω Τύπου I : λάθους → βάζω χαμηλά
Θέλω Τύπου II : ευκολία → λάθους
Ελαχιστοποίηση = $\alpha = 0.01, 0.05, \ldots$

$$(X-\mu_1)'\Sigma^{-1}(X-\mu_1) < (X-\mu_2)'\Sigma^{-1}(X-\mu_2) + C \Rightarrow$$

$$-X'\Sigma^{-1}\mu_1 - \mu_1'\Sigma^{-1}X + \mu_1'\Sigma^{-1}\mu_1 < -X'\Sigma^{-1}\mu_2 - \mu_2'\Sigma^{-1}X + \mu_2'\Sigma^{-1}\mu_2 + C$$

$$-2\mu_1'\Sigma^{-1}X + \mu_1'\Sigma^{-1}\mu_1 < -2\mu_2'\Sigma^{-1}X + \mu_2'\Sigma^{-1}\mu_2 + C \Rightarrow$$

$$(\mu_1-\mu_2)'\Sigma^{-1}X > \mu_1'\Sigma^{-1}\mu_1 - \mu_2'\Sigma^{-1}\mu_2 + \underbrace{\mu_1'\Sigma^{-1}\mu_2 - \mu_2'\Sigma^{-1}\mu_1}_{0} + C$$

$$(\mu_1-\mu_2)'\Sigma^{-1}X > (\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2) + C$$

$$\boxed{(\mu_1-\mu_2)'\Sigma^{-1}X > C}$$

Το C καθορίζεται μέσω των σφαλμάτων:

$$e_{12} = P[\, X \in \Pi_1 \,/\, X \in \Pi_2 \,]$$

$$e_{21} = P[\, X \in \Pi_2 \,/\, X \in \Pi_1 \,]$$

$$e_{12} = P[\, (\mu_1 - \mu_2)' \Sigma^{-1} X \geq C \,/\, X \in \Pi_2 \,]$$

$$e_{21} = P[\, (\mu_1 - \mu_2)' \Sigma^{-1} X \leq C \,/\, X \in \Pi_1 \,]$$

$X \sim N(\mu, \sigma^2)$

$aX \sim N(a\mu, a^2\sigma^2)$

$$X \in \Pi_1 \Rightarrow (\mu_1-\mu_2)'\Sigma^{-1}X \sim N_p\left((\mu_1-\mu_2)'\Sigma^{-1}\mu_1, (\mu_1-\mu_2)'\Sigma^{-1}\Sigma\Sigma^{-1}(\mu_1-\mu_2)\right)$$

$$X \in \Pi_2 \Rightarrow (\mu_1-\mu_2)'\Sigma^{-1}X \sim N_p\left((\mu_1-\mu_2)'\Sigma^{-1}\mu_2, \Delta^2\right)$$

$$\Delta^2 = (\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)$$

$$\|\underset{\sim}{X}\|^2 = \underset{\sim}{X}'\underset{\sim}{X}$$

$$(ZW)' = W'Z'$$

Ara

$$e_{12} = P\left[Z > \frac{C - (\mu_1 - \mu_2)' \Sigma^{-1} \mu_2}{\Delta} \right] = 1 - \Phi\left(\frac{C - (\mu_1 - \mu_2)' \Sigma^{-1} \mu_2}{\Delta} \right)$$

$$e_{21} = P\left[Z < \frac{C - (\mu_1 - \mu_2)' \Sigma^{-1} \mu_1}{\Delta} \right] = \Phi\left(\frac{C - (\mu_1 - \mu_2)' \Sigma^{-1} \mu_1}{\Delta} \right)$$

Απαιτούμε:

$$\boxed{e_{12} = e_{21}} \Rightarrow$$

$$1 - \Phi\left(\frac{c - (\mu_1 - \mu_2)' \Sigma^{-1} \mu_2}{\Delta}\right) = \Phi\left(\frac{c - (\mu_1 - \mu_2)' \Sigma^{-1} \mu_1}{\Delta}\right) \Rightarrow$$

$$\Phi\left(\frac{-c + (\mu_1 - \mu_2)' \Sigma^{-1} \mu_2}{\Delta}\right) \equiv \Phi\left(\frac{c - (\mu_1 - \mu_2)' \Sigma^{-1} \mu_1}{\Delta}\right) \Rightarrow$$

$$-c + (\mu_1 - \mu_2)' \Sigma^{-1} \mu_2 \equiv c - (\mu_1 - \mu_2)' \Sigma^{-1} \mu_1 \Rightarrow$$

$$\boxed{c = \frac{1}{2} (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 + \mu_2)}$$

Άρα $X \in \Pi_1$ όταν

$$(\mu_1-\mu_2)' \Sigma^{-1} X \geq \frac{1}{2}(\mu_1-\mu_2)' \Sigma^{-1}(\mu_1+\mu_2)$$

$\Updownarrow$

$$(X-\mu_1)' \Sigma^{-1}(X-\mu_1) < (X-\mu_2)' \Sigma^{-1}(X-\mu_2) \iff$$

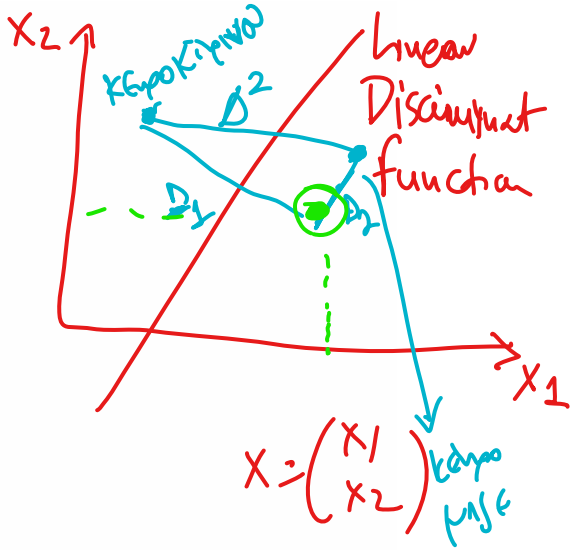
$$\| \Sigma^{-1/2}(X-\mu_1) \|^2 < \| \Sigma^{-1/2}(X-\mu_2) \|^2$$

$$\| \Delta_1^2 \| < \| \Delta_2^2 \|$$

Απόσταση X από Π_1 — Απόσταση X από Π_2

$$\|X\|^2 = X' \cdot X$$

$$\| \Sigma^{-1/2}(X-\mu_1) \|^2 = (X-\mu_1)' \cdot \Sigma^{-1/2} \cdot \Sigma^{-1/2}(X-\mu_1)$$



Άρα η μικρότερη απόσταση οδηγεί στην κατάταξη του X:

$$X \in \Pi_i \quad \text{αν} \quad \Delta_i^2 = \min\left\{ \Delta_1^2, \Delta_2^2 \right\}, \; i=1,2.$$

Γενικεύοντας σε K πληθυσμούς:

$$X \in \Pi_i \quad \text{αν} \quad \Delta_i^2 = \min\left\{ \Delta_1^2, \ldots, \Delta_K^2 \right\}$$

Επίσης

$$\Delta^2 = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) = \| \Sigma^{-1/2} (\mu_1 - \mu_2) \|^2$$

Απόσταση μεταξύ Π_1 & Π_2

= Mahalanobis Distance

Αν μ_1, μ_2, Σ = άγνωστα $\Longrightarrow$

$$\boxed{(\hat{\mu}_1 - \hat{\mu}_2)' \hat{S}^{-1} X > \frac{1}{2} (\hat{\mu}_1 - \hat{\mu}_2)' \hat{S}^{-1} (\hat{\mu}_1 + \hat{\mu}_2)}$$

όπου

$$\hat{\mu}_1 = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad , \quad \hat{\mu}_2 = \bar{Y} = \frac{1}{m} \sum_{i=1}^{m} Y_i$$

$$S_1 = \frac{1}{n-1} \Sigma (X_i - \bar{X})(X_i - \bar{X})' \quad , \quad S_2 = \frac{1}{m-1} \Sigma (X_i - \bar{X})(X_i - \bar{X})'$$

$$\boxed{S = \frac{(n-1) S_1 + (m-1) S_2}{n+m-2}} = \begin{pmatrix} \widehat{Var} X_1 & \widehat{Cov} X_1 X_2 \\ \widehat{Cov} X_1 X_2 & \widehat{Var} X_2 \end{pmatrix}$$

$$\widehat{Cov}(X_1, X_2) = E(X_1 - \mu_1)(X_2 - \mu_2)$$

$$= \frac{1}{n} \sum_{i=1}^{n} (X_{i,1} - \bar{X}_1)(X_{i,2} - \bar{X}_2)$$

$$\hat{E} X_1 = \frac{1}{n} \Sigma X_{i,1} = \bar{X}_1$$

$$\hat{E} X_2 = \frac{1}{n} \Sigma X_{i,2} = \bar{X}_2$$

$$\Delta^2 = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1} (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) \equiv \underset{\sim}{d}' \Sigma^{-1} \cdot \underset{\sim}{d} \equiv \sum_i \sum_j s^{ij} d_j \cdot d_i = \sum_i l_i \cdot d_i$$

$$= \underset{\sim}{d}' \cdot \underset{\sim}{l} \equiv \underset{\sim}{l}' \cdot \underset{\sim}{d} \Rightarrow (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1} = \underset{\sim}{l}' \equiv \sum_j s^{ij} d_j$$

ναι για ευκλήρεις:

$$\hat{\underset{\sim}{l}}' = (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2) \cdot \hat{\Sigma}^{-1} = (\hat{\underset{\sim}{\mu}}_1 - \hat{\underset{\sim}{\mu}}_2)' S^{-1}$$

Γ_1 αν $(\hat{\underset{\sim}{\mu}}_2 - \hat{\underset{\sim}{\mu}}_2)' S^{-1} \cdot \underset{\sim}{x} > \frac{1}{2} (\hat{\underset{\sim}{\mu}}_1 - \hat{\underset{\sim}{\mu}}_2)' S^{-1} (\hat{\underset{\sim}{\mu}}_1 + \hat{\underset{\sim}{\mu}}_2) \Rightarrow$

$$(\hat{\underset{\sim}{\mu}}_1 - \hat{\underset{\sim}{\mu}}_2)' S^{-1} \cdot \underset{\sim}{x} > \frac{(\hat{\underset{\sim}{\mu}}_1 - \hat{\underset{\sim}{\mu}}_2)' S^{-1} \hat{\underset{\sim}{\mu}}_1 + (\hat{\underset{\sim}{\mu}}_1 - \hat{\underset{\sim}{\mu}}_2)' S^{-1} \hat{\underset{\sim}{\mu}}_2}{2} \Rightarrow$$

$$\hat{\underset{\sim}{l}}' \cdot \underset{\sim}{x} > \frac{\hat{\underset{\sim}{l}}' \hat{\underset{\sim}{\mu}}_1 + \hat{\underset{\sim}{l}}' \hat{\underset{\sim}{\mu}}_2}{2}$$

$$\hat{\underset{\sim}{l}}' \cdot \underset{\sim}{x} > \frac{1}{2} \left(\hat{\underset{\sim}{l}} \bar{\underset{\sim}{x}}_1 + \hat{\underset{\sim}{l}} \bar{\underset{\sim}{x}}_2 \right)$$

$$\underset{\sim}{l} = (\hat{\underset{\sim}{\mu}}_1 - \hat{\underset{\sim}{\mu}}_2) \cdot S^{-1}$$

$\underset{\sim}{X}_{11}, \ldots \underset{\sim}{X}_{1n_1}$ δείγμα από $N_p(\underset{\sim}{\mu}_1, \Sigma)$.

$\underset{\sim}{X}_{21}, \ldots, \underset{\sim}{X}_{2n_2}$ δείγμα από $N_p(\underset{\sim}{\mu}_2, \Sigma)$.

H_0: $\underset{\sim}{X}, \underset{\sim}{X}_{11}, \ldots \underset{\sim}{X}_{1n_1} \in \Pi_1$ & $\underset{\sim}{X}_{21}, \ldots, \underset{\sim}{X}_{2n_2} \in \Pi_2$

vs.

H_0: $\underset{\sim}{X}_{11}, \ldots, \underset{\sim}{X}_{1n_1} \in \Pi_1$ & $\underset{\sim}{X}, \underset{\sim}{X}_{21}, \ldots \underset{\sim}{X}_{2n_2} \in \Pi_2$

2023

Παράδειγμα (Johnson and Wichern ~~2007~~, Example 11.3): $p = 2$, $n_1 = 30$, $n_2 = 22$, $X_1 = \log_{10}(\text{AHF activity})$, $X_2 = \log_{10}(\text{AHF} - \text{like antigen})$. Μας ενδιαφέρει η ανίχνευση φορέων hemophilia A στον γυναικείο πληθυσμό. Από τα δεδομένα προέκυψε ότι τα διανύσματα των δειγματικών μέσων τιμών για τους υπό-πληθυσμούς υγιών και μη-υγιών γυναικών είναι, αντίστοιχα,

$$\bar{\mathbf{x}}_1 = (-0.0065, -0.0390)', \bar{\mathbf{x}}_2 = (-0.2483, -0.0262)'.$$

Άσκηση 11.32?

Επίσης, ο $\mathbf{S}_{\text{pooled}}^{-1}$ είναι ίσος με

$$\mathbf{S}_{\text{pooled}}^{-1} \equiv \mathbf{S}^{-1} = \begin{pmatrix} 131 & -90 \\ -90 & 108 \end{pmatrix},$$

και άρα

$$\boldsymbol{l}' = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1} \equiv \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = (37.5438, -28.8036)',$$

δηλ. η LDF είναι

$$\boldsymbol{l}'\mathbf{x} = 37.61x_1 - 28.92x_2,$$

με

$$\bar{y}_1 = \boldsymbol{l}'\bar{\mathbf{x}}_1 = 0.8793,\ \bar{y}_2 = \boldsymbol{l}'\bar{\mathbf{x}}_2 = -10.0768 \text{ και } \bar{y} = -4.5987.$$

- Αν έρθει μια νέα παρατήρηση με διάνυσμα $\mathbf{x}_0 = (-0.210, -0.044)'$, αυτή θα τοποθετηθεί στον Π2 (δηλ. η γυναίκα ταξινομείται ως φορέας της hemophilia A) αφού

$$\boldsymbol{l}'\mathbf{x}_0 = 37.61 \cdot (-0.210) - 28.92 \cdot (-0.044) = -6.6168 \not\geq -4.5987.$$

- Αξίζει να σημειωθεί ότι η παραπάνω εφαρμογή της LDF του Fisher, υποθέτει ίσα κόστη εσφαλμένης ταξινόμησης στους δύο πληθυσμούς αλλά και ίσες εκ των προτέρων πιθανότητες ταξινόμησης σε αυτούς.

- Έστω τώρα ότι οι πιθανότητες ταξινόμησης είναι γνωστές και ίσες με $p_1 = 0.75$ και $p_2 = 0.25$, αντίστοιχα. Έστω επίσης ότι τα κόστη εσφαλμένης ταξινόμησης είναι επίσης ίσα. Τότε, η παρατήρηση με διάνυσμα $\mathbf{x}_0$ ταξινομείται στον Π1 αν

$$\mathbf{x}_0' \mathbf{S}_{\text{pooled}}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \frac{1}{2} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)' \mathbf{S}_{\text{pooled}}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) > \log(p_2/p_1).$$

Άρα, με άμεση αντικατάσταση, διαπιστώνουμε ότι

$$\mathbf{x}_0' \mathbf{S}_{\text{pooled}}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = -6.6168,$$

$$\frac{1}{2} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)' \mathbf{S}_{\text{pooled}}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = -4.5987, \quad \log(p_2/p_1) = -1.0986$$

και τελικά, αφού

$$-6.6168 - (-4.5987) = -2.0181 \not\geq -1.0986,$$

η παρατήρηση αυτή θα ταξινομηθεί στον Π2.