2) Binary classification problem ; $\theta \in \{0,1\}$ denotes class label

$X|(\theta=0) \sim N_p(\mu_0, \Sigma)$ and $X|(\theta=1) \sim N_p(\mu_1, \Sigma)$

Σ is +DEF , 0-1 Loss, prior dist of θ is $P(\theta=0) = 1/2$ and $P(\theta=1) = 1/2$, so $\theta \sim Ber(1/2)$

(a) Derive the Bayes rule for classifying a new obs. $X \in \mathbb{R}^p$.

→ Bayes rules minimize post. exp. loss

Under 0-1 Loss, Bayes Rule classifies observation to the posterior mode → whichever occurs more, $\theta=1$ or $\theta=0$

So, when $\dfrac{P(\theta=1|X)}{P(\theta=0|X)} > 1$, then X gets assigned to class 1

We know $P(\theta_1|X) \propto p(X|\theta_1)\,p(\theta_1) = \{(2\pi)^p|\Sigma|\}^{-1/2} \exp\{-(X-\mu_1)^T\Sigma^{-1}(X-\mu_1)/2\} \cdot \frac{1}{2}$

and $P(\theta_2|X) \propto p(X|\theta_2)\,p(\theta_2) = \{(2\pi)^p|\Sigma|\}^{-1/2} \exp\{-(X-\mu_0)^T\Sigma^{-1}(X-\mu_0)/2\} \cdot \frac{1}{2}$

So, Bayes rule assigns X to $\theta = 1$ if :

$$\cancel{\{(2\pi)^p|\Sigma|\}^{-1/2}} \exp\{-(X-\mu_1)^T\Sigma^{-1}(X-\mu_1)/2\} \cdot \tfrac{1}{2} > \cancel{\{(2\pi)^p|\Sigma|\}^{-1/2}} \exp\{-(X-\mu_0)^T\Sigma^{-1}(X-\mu_0)/2\} \cdot \tfrac{1}{2}$$

$$\Leftrightarrow \exp\{-(X-\mu_1)^T\Sigma^{-1}(X-\mu_1)/2\} > \exp\{-(X-\mu_0)^T\Sigma^{-1}(X-\mu_0)/2\}$$

$$\Leftrightarrow \exp\{-(X-\mu_1)^T\Sigma^{-1}(X-\mu_1)/2 + (X-\mu_0)^T\Sigma^{-1}(X-\mu_0)/2\} > 1$$

I probably would have stepped here...

↓ Had to look ahead to part (d) to see δ and $\bar{\mu}$ defined

$$\Leftrightarrow -(X-\mu_1)^T\Sigma^{-1}(X-\mu_1) + (X-\mu_0)^T\Sigma^{-1}(X-\mu_0) > 0$$

$$\Leftrightarrow \cancel{-X^T\Sigma^{-1}X} + \mu_1^T\Sigma^{-1}X + X^T\Sigma^{-1}\mu_1 - \mu_1^T\Sigma^{-1}\mu_1 + \cancel{X^T\Sigma^{-1}X} - X^T\Sigma^{-1}\mu_0 - \mu_0^T\Sigma^{-1}X + \mu_0^T\Sigma^{-1}\mu_0 > 0$$

$$\Leftrightarrow X^T\Sigma^{-1}(\mu_1-\mu_0) + \underbrace{(\mu_1^T-\mu_0^T)}_{(\mu_1-\mu_0)^T}\Sigma^{-1}X + \underbrace{(-\mu_1^T+\mu_0^T)}_{-(\mu_1-\mu_0)^T}\Sigma^{-1}\underbrace{(\mu_1+\mu_0)}_{(\mu_1+\mu_0)}$$

Let $\delta = (\mu_1 - \mu_0)$ and $\bar{\mu} = \dfrac{\mu_0+\mu_1}{2}$

$$\Leftrightarrow X^T\Sigma^{-1}\delta - \delta^T\Sigma^{-1}X - \delta^T\Sigma^{-1}2\bar{\mu} > 0$$

$$(\delta^T\Sigma^{-1}X)^T - \delta^T\Sigma^{-1}X - 2\delta^T\Sigma^{-1}\bar{\mu} > 0$$

$$2\delta^T\Sigma^{-1}(X-\bar{\mu}) > 0$$

$$\Leftrightarrow \boxed{\delta^T\Sigma^{-1}(X-\bar{\mu}) > 0} \checkmark$$

(b) Derive the misclassification rate R^* of the Bayes rule.

$R^* = \underline{R(0,d)} + \underline{R(1,d)}$ ← BINARY! So, 2 risks to consider!

$R^* = P(\text{choose } X \text{ in } \theta=1 \mid \theta=0)\cdot P(\theta=0) + P(\text{choose } X \text{ in } \theta=0 \mid \theta=1)\cdot P(\theta=1)$

← since their δ in (d) is $\hat{\mu}_0-\hat{\mu}_1$ and their $\mu = \bar{\mu}$ from our (a)

$$R^* = \underbrace{\tfrac{1}{2} P(-\delta^T\Sigma^{-1}(X-\mu) > 0 \mid \theta=0)} + \tfrac{1}{2} P(-\delta^T\Sigma^{-1}(X-\mu) \le 0 \mid \theta=1)$$

$$\tfrac{1}{2} P(-\delta^T\Sigma^{-1}(X-\mu) > 0 \mid \theta=0)$$

$$= \tfrac{1}{2} P(\delta^T\Sigma^{-1}(X-\mu) \le 0 \mid \theta=0)$$

$$= \tfrac{1}{2} P(\delta^T\Sigma^{-1}(X-\mu_0+\mu_0-\mu) \le 0 \mid \theta=0) \quad ; \quad X\mid\theta=0 \sim N_p(\mu_0, \Sigma)$$

$$\Rightarrow X-\mu_0\mid\theta=0 \sim N_p(0,\Sigma)$$

$$= \tfrac{1}{2} P(\delta^T\Sigma^{-1}(X-\mu_0) \le \delta^T\Sigma^{-1}(\mu-\mu_0) \mid \theta=0) \qquad \Rightarrow \delta^T\Sigma^{-1}(X-\mu_0)\mid\theta=0 \sim N_p(0, \delta^T\Sigma^{-1}\delta)$$

$$= \tfrac{1}{2} P\left(\frac{\delta^T\Sigma^{-1}(X-\mu_0)}{\sqrt{\delta^T\Sigma^{-1}\delta}} \le \frac{\delta^T\Sigma^{-1}(\mu-\mu_0)}{\sqrt{\delta^T\Sigma^{-1}\delta}}\right) \qquad \Rightarrow \frac{\delta^T\Sigma^{-1}(X-\mu_0)}{\sqrt{\delta^T\Sigma^{-1}\delta}}\,\Big|\,\theta=0 \sim N_p(0,1)$$

$$= \tfrac{1}{2}\Phi\left(\frac{\delta^T\Sigma^{-1}(\mu-\mu_0)}{\sqrt{\delta^T\Sigma^{-1}\delta}}\right), \text{ where } \delta=\mu_0-\mu_1 \text{ and } \mu=\frac{\mu_0+\mu_1}{2} \quad ; \quad \mu-\mu_0 = \frac{\mu_1-\mu_0}{2} = -\tfrac{1}{2}\delta$$

Similarly, $\tfrac{1}{2} P(-\delta^T\Sigma^{-1}(X-\mu) \le 0 \mid \theta=1) = \tfrac{1}{2}\Phi\left(-\frac{\delta^T\Sigma^{-1}(\mu-\mu_1)}{\sqrt{\delta^T\Sigma^{-1}\delta}}\right)$

$$= \tfrac{1}{2}\Phi\left(-\frac{\delta^T\Sigma^{-1}(-\frac{1}{2}\delta)}{\sqrt{\delta^T\Sigma^{-1}\delta}}\right) = \tfrac{1}{2}\Phi\left(\tfrac{1}{2}(\delta^T\Sigma^{-1}\delta)^{1/2}\right)$$

So, $R^* = \tfrac{1}{2}\Phi\left(\tfrac{1}{2}(\delta^T\Sigma^{-1}\delta)^{1/2}\right) + \tfrac{1}{2}\Phi\left(\tfrac{1}{2}(\delta^T\Sigma^{-1}\delta)^{1/2}\right)$

$$= \Phi\left(\tfrac{1}{2}(\delta^T\Sigma^{-1}\delta)^{1/2}\right)$$

(C) Let X_{0i} $(i=1,\dots,n_0)$ be iid samples class $\theta=0$ and X_{1i} $(i=1,\dots,n_1)$ be iid samples class $\theta=1$.

$X_{0i} \perp\!\!\!\perp X_{1i}$. Derive MLEs of (μ_0, μ_1, Σ).

The likelihood of $X_{0i} = \prod_{i=1}^{n_0} \{(2\pi)^p |\Sigma|\}^{-1/2} \exp\{-(x_{0i}-\mu_0)^T \Sigma^{-1} (x_{0i}-\mu_0)/2\}$

$\Rightarrow$ Log-likelihood: $l_n(\mu_0, \Sigma) = \sum_{i=1}^{n_0} \left\{ -\frac{1}{2} \log\left[(2\pi)^p |\Sigma|\right] - \frac{(x_{0i}-\mu_0)^T \Sigma^{-1} (x_{0i}-\mu_0)}{2} \right\}$

Then, $\frac{d l_n(\mu_0, \Sigma)}{d\mu_0} = \sum_{i=1}^{n_0} \frac{2(x_{0i}-\mu_0)\Sigma^{-1}(-1)}{2} \overset{set}{=} 0 \rightarrow \sum_{i=1}^{n_0} \hat{\mu}_0 - x_{0i} = 0 \rightarrow n_0 \hat{\mu}_0 = \sum_{i=1}^{n_0} x_{0i}$

$\Rightarrow \underbrace{\boxed{\hat{\mu}_0 = \frac{1}{n_0} \sum_{i=1}^{n_0} x_{0i}}}_{= \bar{X}_0}$, and similarly, $\underbrace{\boxed{\hat{\mu}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{1i}}}_{= \bar{X}_1}$

Now, to find the MLE of Σ, we rewrite the Log-likelihood in terms of trace:

$$l_n(\mu_0, \Sigma) = -\frac{p n_0}{2} \log(2\pi) - \frac{n_0}{2} \log(|\Sigma|) - \sum_{i=1}^{n_0} (x_{0i}-\mu_0)^T \Sigma^{-1} (x_{0i}-\mu_0)$$

$$\propto -\frac{n_0}{2} \log(|\Sigma|) - \frac{1}{2} tr\Big(\overset{\text{b/c scalar}}{\sum_{i=1}^{n_0}} (x_{0i}-\mu_0)^T \Sigma^{-1} (x_{0i}-\mu_0)\Big)$$

$$= -\frac{n_0}{2} \log\left(\frac{1}{|\Sigma^{-1}|}\right) - \frac{1}{2} tr\Big(\Sigma^{-1} \underbrace{\sum_{i=1}^{n_0} (x_{0i}-\mu_0)(x_{0i}-\mu_0)^T}_{\text{symmetric}}\Big)$$

Note, as given: $\frac{d \log(|\Sigma|)}{d\Sigma} = \Sigma^{-1}$ and $\frac{d\, tr(\Sigma^{-1}\{\sum_{i=1}^{n_0}(x_{0i}-\mu_0)(x_{0i}-\mu_0)^T\})}{d\Sigma^{-1}} = \sum_{i=1}^{n_0} (x_{0i}-\hat{\mu}_0)(x_{0i}-\hat{\mu}_0)^T$ plug in $\hat{\mu}_0$

This makes this part much easier, so we take derivative wrt Σ^{-1}

Note $\frac{d \log\left(\frac{1}{|\Sigma^{-1}|}\right)}{d\Sigma^{-1}} = -\frac{d \log(|\Sigma^{-1}|)}{d\Sigma^{-1}} = -(\Sigma^{-1})^{-1} = -\Sigma$

Since $X_{0i} \perp\!\!\!\perp X_{1i}$, then $l_n(\mu_0, \mu_1, \Sigma) = l_n(\mu_0, \Sigma) + l_n(\mu_1, \Sigma)$ *

$$= -\frac{n_0}{2} \log\left(\frac{1}{|\Sigma^{-1}|}\right) - \frac{1}{2} tr\Big(\Sigma^{-1} \sum_{i=1}^{n_0} (x_{0i}-\mu_0)(x_{0i}-\mu_0)^T\Big)$$
$$- \frac{n_1}{2} \log\left(\frac{1}{|\Sigma^{-1}|}\right) - \frac{1}{2} tr\Big(\Sigma^{-1} \sum_{i=1}^{n_1} (x_{1i}-\mu_1)(x_{1i}-\mu_1)^T\Big)$$

* $\Rightarrow \frac{d l_n(\mu_0, \mu_1, \Sigma)}{d\Sigma^{-1}} = -\frac{n_0}{2}(-\Sigma) - \frac{n_1}{2}(-\Sigma) - \frac{1}{2} \sum_{i=1}^{n_0} (x_{0i}-\bar{x}_0)(x_{0i}-\bar{x}_0)^T - \frac{1}{2} \sum_{i=1}^{n_1} (x_{1i}-\bar{x}_1)(x_{1i}-\bar{x}_1)^T \overset{set}{=} 0$

$$\hat{\Sigma} = \left(\frac{1}{n_0+n_1}\right) \left(\sum_{i=1}^{n_0} (x_{0i}-\bar{x}_0)(x_{0i}-\bar{x}_0)^T + \sum_{i=1}^{n_1} (x_{1i}-\bar{x}_1)(x_{1i}-\bar{x}_1)^T \right)$$

(d) If we replace (μ_0, μ_1, Σ) in Bayes rule with $(\hat{\mu}_0, \hat{\mu}_1, \hat{\Sigma})$, prove that the misclassification rate of the resulting rule, i.e. the probability of classifying X to a wrong class given the training data $\{X_{0i}\}_{i=1}^{n_0}$ and $\{X_{1i}\}_{i=1}^{n_1}$, is given by: $\frac{1}{2}\Phi\left(\frac{\hat{\delta}^T\hat{\Sigma}^{-1}(\mu_1-\hat{\mu})}{\sqrt{\hat{\delta}^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\delta}}}\right) + \frac{1}{2}\Phi\left(-\frac{\hat{\delta}^T\hat{\Sigma}^{-1}(\mu_0-\hat{\mu})}{\sqrt{\hat{\delta}^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\delta}}}\right)$

From (b), we have $R^* = \frac{1}{2}P(\delta^T\Sigma^{-1}(X-\bar{\mu})>0 \mid \theta=0) + \frac{1}{2}P(\delta^T\Sigma^{-1}(X-\bar{\mu}) \le 0 \mid \theta=1)$

→ replacing (μ_0, μ_1, Σ) with $(\hat{\mu}_0, \hat{\mu}_1, \hat{\Sigma})$, we have:

$$R^* = \underbrace{\frac{1}{2}P\left((\hat{\mu}_1-\hat{\mu}_0)^T\Sigma^{-1}\left(X-\frac{\hat{\mu}_1+\hat{\mu}_0}{2}\right)>0 \mid \theta=0\right)} + \frac{1}{2}P\left((\hat{\mu}_1-\hat{\mu}_0)^T\Sigma^{-1}\left(X-\frac{\hat{\mu}_1+\hat{\mu}_0}{2}\right)\le 0 \mid \theta=1\right)$$

$$= \frac{1}{2}P\left(\hat{\delta}^T\hat{\Sigma}^{-1}(X-\hat{\mu})>0 \mid \theta=0\right)$$

$$= \frac{1}{2}P\left(\hat{\delta}^T\hat{\Sigma}^{-1}(X-\mu_0+\mu_0-\hat{\mu})>0 \mid \theta=0\right)$$

; $X \mid \theta=0 \sim N_p(\mu_0, \Sigma) \Rightarrow X-\mu_0 \mid \theta=0 \sim N(0,\Sigma)$

$\Rightarrow \hat{\Sigma}^{-1}(X-\mu_0) \mid \theta=0 \sim N(0, \hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1})$

$\Rightarrow \hat{\delta}^T\hat{\Sigma}^{-1}(X-\mu_0) \mid \theta=0 \sim N(0, \hat{\delta}^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\delta})$

$\Rightarrow \frac{\hat{\delta}^T\hat{\Sigma}^{-1}(X-\mu_0)}{\sqrt{\hat{\delta}^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\delta}}} \mid \theta=0 \sim N(0,1)$

$$= \frac{1}{2}P\left(\hat{\delta}^T\hat{\Sigma}^{-1}(X-\mu_0) > \hat{\delta}^T\hat{\Sigma}^{-1}(\hat{\mu}-\mu_0) \mid \theta=0\right)$$

$$= \frac{1}{2}P\left(\frac{\hat{\delta}^T\hat{\Sigma}^{-1}(X-\mu_0)}{\sqrt{\hat{\delta}^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\delta}}} > \frac{\hat{\delta}^T\hat{\Sigma}^{-1}(\hat{\mu}-\mu_0)}{\sqrt{\hat{\delta}^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\delta}}} \mid \theta=0\right)$$

$$= \frac{1}{2}\Phi\left(-\frac{\hat{\delta}^T\hat{\Sigma}^{-1}(\hat{\mu}-\mu_0)}{\sqrt{\hat{\delta}^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\delta}}}\right)$$

* Note: My $\hat{\delta}$ is $-\hat{\delta}$ of what is given

Similarly, $\frac{1}{2}P\left((\hat{\mu}_1-\hat{\mu}_0)^T\Sigma^{-1}(X-\hat{\mu})\le 0 \mid \theta=1\right) = \frac{1}{2}\Phi\left(\frac{-\hat{\delta}^T\hat{\Sigma}^{-1}(\hat{\mu}-\mu_1)}{\sqrt{\hat{\delta}^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\delta}}}\right)$

Thus, the misclassification rate is:

$$R^* = \frac{1}{2}\Phi\left(\frac{\hat{\delta}^T\hat{\Sigma}^{-1}(\mu_1-\hat{\mu})}{\sqrt{\hat{\delta}^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\delta}}}\right) + \frac{1}{2}\Phi\left(-\frac{\hat{\delta}^T\hat{\Sigma}^{-1}(\mu_0-\hat{\mu})}{\sqrt{\hat{\delta}^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\delta}}}\right)$$

(e) We propose another classification rule that assigns X to class of $\theta = 0$ iff $\hat{\beta}^T(X-\hat{\mu}) \geq 0$

where $\hat{\mu} = (\hat{\mu}_0 + \hat{\mu}_1)/2$ and $\hat{\beta}$ solves $\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \ \frac{1}{2}\beta^T \hat{\Sigma} \beta - (\hat{\mu}_0 - \hat{\mu}_1)^T \beta + \lambda \sum_{j=1}^{p} |\beta_j|$

Derive the Majorization-Minimization algorithm for solving $\hat{\beta}$.

Give an explicit choice of step size and closed-form expressions on how iterations need to be done.

[MAJORIZATION]

Let $l(\beta) = \frac{1}{2}\beta^T \hat{\Sigma} \beta - (\hat{\mu}_0 - \hat{\mu}_1)^T \beta$

Then, $\nabla l(\beta) = \hat{\Sigma}\beta - (\hat{\mu}_0 - \hat{\mu}_1)$; $\nabla^2 l(\beta) = \hat{\Sigma}$

By 2nd-order Taylor Expansion of $l(\beta)$ around $\tilde{\beta}$:

$$l(\beta) = l(\tilde{\beta}) + \nabla l(\tilde{\beta})^T(\beta - \tilde{\beta}) + \frac{1}{2}(\beta - \tilde{\beta})^T \nabla^2 l(\tilde{\beta})(\beta - \tilde{\beta})$$

$$\leq \underbrace{l(\tilde{\beta}) + \nabla l(\tilde{\beta})^T(\beta - \tilde{\beta}) + c(\beta - \tilde{\beta})^T(\beta - \tilde{\beta})}_{\text{Call this } l_Q(\beta)}, \text{ where } c \geq \lambda_{max}(\hat{\Sigma})$$

[MINIMIZATION]

Now, let $p(\beta) = l_Q(\beta) + \lambda\|\beta\|_1$; we want to minimize $p(\beta)$ WRT β to find $\hat{\beta}^{(new)}$

$$p(\beta) = l_Q(\beta) + \lambda\|\beta\|_1 = l(\tilde{\beta}) + \nabla l(\tilde{\beta})^T(\beta - \tilde{\beta}) + c(\beta - \tilde{\beta})^T(\beta - \tilde{\beta}) + \lambda\|\beta\|_1$$

$$\frac{dp(\beta)}{d\beta} = \nabla l(\tilde{\beta}) + 2c(\beta - \tilde{\beta}) + \lambda \frac{d\|\beta\|_1}{d\beta}$$

- If $\beta \neq 0$, then $\frac{d\|\beta\|_1}{d\beta} = \frac{d}{d\beta}\left(\sqrt{\beta^2}\right) = \frac{1}{2}(\beta^2)^{-1/2} \cdot 2\beta = \frac{\beta}{\|\beta\|_1}$
- If $\beta = 0$, then we have previously shown $\frac{d\|\beta\|_1}{d\beta} = [-1, 1]$

So, $\frac{d\|\beta\|_1}{d\beta} = \begin{cases} 1 & \text{if } \beta > 0 \\ [-1,1] & \text{if } \beta = 0 \\ -1 & \text{if } \beta < 0 \end{cases}$ (and we know $0 \in \left.\frac{dp(\beta)}{d\beta}\right|_{\tilde{\beta}^{new}}$ by KKT)

→ We will look @ all 3 cases:

- When $\beta > 0 \Rightarrow \tilde{\beta}^{(new)} > 0$

$$\frac{d\rho(\beta)}{d\beta} = \hat{\Sigma}\tilde{\beta} - (\hat{\mu}_0 - \hat{\mu}_1) + 2c(\beta - \tilde{\beta}) + \lambda \overset{set}{=} 0$$

$$-\hat{\Sigma}\tilde{\beta} + (\hat{\mu}_0 - \hat{\mu}_1) + 2c\tilde{\beta} - \lambda = 2c\tilde{\beta}^{(new)}$$

$$\Rightarrow \tilde{\beta}^{(new)} = \underbrace{\tilde{\beta} - \frac{1}{2c}\left(\hat{\Sigma}\tilde{\beta} - (\hat{\mu}_0 - \hat{\mu}_1) + \lambda \right)}_{\text{as long as this is } > 0}$$

- When $\beta < 0 \Rightarrow \tilde{\beta}^{(new)} < 0$

Similarly, $$\tilde{\beta}^{(new)} = \underbrace{\tilde{\beta} - \frac{1}{2c}\left(\hat{\Sigma}\tilde{\beta} - (\hat{\mu}_0 - \hat{\mu}_1) + \lambda \right)}_{\text{as long as this is } < 0}$$

- When $\beta = 0 \Rightarrow \tilde{\beta}^{(new)} = 0$

$$\frac{d\rho(\beta)}{d\beta} = \hat{\Sigma}\tilde{\beta} - (\hat{\mu}_0 - \hat{\mu}_1) + 2c(\beta - \tilde{\beta}) + \lambda[-1,1] \Big|_{\beta = 0}$$

$$= \Big[\underbrace{\hat{\Sigma}\tilde{\beta} - (\hat{\mu}_0 - \hat{\mu}_1) - 2c\tilde{\beta} - \lambda}_{\text{must be } < 0},\ \underbrace{\hat{\Sigma}\tilde{\beta} - (\hat{\mu}_0 - \hat{\mu}_1) - 2c\tilde{\beta} + \lambda}_{\text{must be } > 0} \Big]$$

as long as $$\left(\hat{\Sigma}\tilde{\beta} - (\hat{\mu}_0 - \hat{\mu}_1) - 2c\tilde{\beta} \right) \in [-\lambda, \lambda]$$

$$\tilde{\beta} - \frac{1}{2c}\left(\hat{\Sigma}\tilde{\beta} - (\hat{\mu}_0 - \hat{\mu}_1) \right) \in \left[-\frac{1}{2c}\lambda, \frac{1}{2c}\lambda \right]$$

So, $$\tilde{\beta}^{(new)} = S\Big(\underbrace{\tilde{\beta} - \frac{1}{2c}\left(\hat{\Sigma}\tilde{\beta} - (\hat{\mu}_0 - \hat{\mu}_1) \right)}_{A}, \frac{1}{2c}\lambda \Big)$$

$$= \begin{cases} A - \lambda & \text{if } |A| > \lambda \text{ and } A > 0 \\ 0 & \text{if } |A| \le \lambda \\ A + \lambda & \text{if } |A| > \lambda \text{ and } A < 0 \end{cases}$$

We first initialize β at $\beta^{(0)} \in \mathbb{R}^p$

Iterate until convergence: $\beta^{(k)} = S(A, \frac{1}{2c}\lambda)$

Stop if $\|\beta^{(k)} - \beta^{(k-1)}\|_2 < \varepsilon$ for some pre-defined stopping threshold ε.

(f) R_n is misclassification rate of rule in (e).

Suppose we can show $\hat{\beta} \xrightarrow{p} \Sigma^{-1}(\mu_0 - \mu_1)$ as $n \to \infty$. Show $R_n \xrightarrow{p} R^*$

From (b), we know $R^* = \Phi\left(-\frac{1}{2}(\delta^T \Sigma^{-1} \delta)^{1/2}\right) = \Phi\left(-\frac{1}{2}\left((\mu_0-\mu_1)^T \Sigma^{-1}(\mu_0-\mu_1)\right)^{1/2}\right)$

The rule in (e) assigns X to class $\theta = 0$ iff $\hat{\beta}^T(X - \hat{\mu}) \geq 0$

$$\text{So, } R_n = 1 \cdot P(\hat{\beta}^T(X-\hat{\mu}) \geq 0 | \theta = 1) P(\theta=1) + 1 \cdot P(\hat{\beta}^T(X-\hat{\mu}) < 0 | \theta = 0) P(\theta = 0)$$

$$= \frac{1}{2} \underbrace{P(\hat{\beta}^T(X-\hat{\mu}) \geq 0 | \theta = 1)} + \frac{1}{2} P(\hat{\beta}^T(X-\hat{\mu}) < 0 | \theta = 0)$$

$$P(\hat{\beta}^T(X-\hat{\mu}) \geq 0 | \theta = 1) = P(\hat{\beta}^T(X - \mu_1 - (\hat{\mu} - \mu_1)) \geq 0 | \theta = 1)$$

$$= P(\hat{\beta}^T(X-\mu_1) \geq \hat{\beta}^T(\hat{\mu} - \mu_1) | \theta = 1) \; ; \quad X|(\theta=1) \sim N_p(\mu_1, \Sigma)$$

$$\Rightarrow \hat{\beta}^T(X-\mu_1) | (\theta = 1) \sim N_p(0, \hat{\beta}^T \Sigma \hat{\beta})$$

$$= P\left(\frac{\hat{\beta}^T(X-\mu_1)}{\sqrt{\hat{\beta}^T \Sigma \hat{\beta}}} \geq \frac{\hat{\beta}^T(\hat{\mu}-\mu_1)}{\sqrt{\hat{\beta}^T \Sigma \hat{\beta}}} \middle| \theta = 1\right) \leftarrow \Rightarrow \frac{\hat{\beta}^T(X-\mu_1)}{\sqrt{\hat{\beta}^T \Sigma \hat{\beta}}} \middle| \theta = 1 \sim N_p(0, I)$$

$$= \Phi\left(\frac{-\hat{\beta}^T(\hat{\mu}-\mu_1)}{\sqrt{\hat{\beta}^T \Sigma \hat{\beta}}}\right)$$

Similarly, $P(\hat{\beta}^T(X-\hat{\mu}) < 0 | \theta = 0) = \Phi\left(\frac{\hat{\beta}^T(\hat{\mu}-\mu_0)}{\sqrt{\hat{\beta}^T \Sigma \hat{\beta}}}\right)$

So, $R_n = \frac{1}{2}\Phi\left(\frac{-\hat{\beta}^T(\hat{\mu}-\mu_1)}{\sqrt{\hat{\beta}^T \Sigma \hat{\beta}}}\right) + \frac{1}{2}\Phi\left(\frac{\hat{\beta}^T(\hat{\mu}-\mu_0)}{\sqrt{\hat{\beta}^T \Sigma \hat{\beta}}}\right)$

Since $\hat{\beta} \to_p \Sigma^{-1}(\mu_0 - \mu_1)$, then $R_n \to \frac{1}{2}\Phi\left(\frac{(\mu_0-\mu_1)^T \Sigma^{-1}(\hat{\mu}-\mu_1)}{((\mu_0-\mu_1)^T\Sigma^{-1}(\mu_0-\mu_1))^{1/2}}\right) + \frac{1}{2}\Phi\left(\frac{(\mu_0-\mu_1)^T \Sigma^{-1}(\hat{\mu}-\mu_0)}{((\mu_0-\mu_1)^T\Sigma^{-1}(\mu_0-\mu_1))^{1/2}}\right)$

$\hat{\mu} - \mu_1 = \frac{\mu_0 + \mu_1}{2} - \mu_1 = \frac{1}{2}\delta$

$\hat{\mu} - \mu_0 = -\frac{1}{2}\delta$

$\hat{\mu}_1 \to \mu_1$, $\hat{\mu}_0 \to \mu_0$

so $\hat{\delta} \to (\mu_0 - \mu_1)$

$$\Rightarrow = \frac{1}{2}\Phi\left(\frac{-\frac{1}{2}(\mu_0-\mu_1)^T\Sigma^{-1}\hat{\delta}}{((\mu_0-\mu_1)^T\Sigma^{-1}(\mu_0-\mu_1))^{1/2}}\right) + \frac{1}{2}\Phi\left(\frac{-\frac{1}{2}(\mu_0-\mu_1)^T\Sigma^{-1}\hat{\delta}}{((\mu_0-\mu_1)^T\Sigma^{-1}(\mu_0-\mu_1))^{1/2}}\right)$$

$$= \frac{1}{2}\Phi\left(\frac{-\frac{1}{2}(\mu_0-\mu_1)^T\Sigma^{-1}(\mu_0-\mu_1)}{((\mu_0-\mu_1)^T\Sigma^{-1}(\mu_0-\mu_1))^{1/2}}\right) + \frac{1}{2}\Phi\left(\frac{-\frac{1}{2}(\mu_0-\mu_1)^T\Sigma^{-1}(\mu_0-\mu_1)}{((\mu_0-\mu_1)^T\Sigma^{-1}(\mu_0-\mu_1))^{1/2}}\right)$$

$$= \Phi\left(-\frac{1}{2}((\mu_0-\mu_1)^T\Sigma^{-1}(\mu_0-\mu_1))^{1/2}\right) = R^* \checkmark$$