Chapter 4. Gaussian Models

4.1.

$$
\begin{aligned}
& p_X(x) = \frac{1}{2}\mathbf{1}_{[-1,1]} \\
& P(Y \le y) = P(x^2 \le y) = \sqrt{y} \Rightarrow p_Y(y) = \frac{1}{2\sqrt{y}} \\
& \mu_x = 0, \mu_y = \int_0^1 \frac{\sqrt{y}}{2} dy = \frac{1}{3} \\
& \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = \mathbb{E}[X(Y-\frac{1}{3})] = \mathbb{E}[X^3 - \frac{1}{3}X] = \int_{-1}^{1} (x^3 - \frac{1}{3}x) \cdot \frac{1}{2} dx = 0. \\
& \Rightarrow \rho(X,Y) = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_{XY}} = 0.
\end{aligned} \tag{1}
$$

4.2.
a. $P(Y \le x) = \mathbb{E}[P(Y \le x|W)] = P(X \le x)P(W = 1) + P(-X \le x)P(W = -1) = \Phi(x)$, where $\Phi(x)$ is the cumulative density function of standard normal distribution.
b. $\text{Cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[X^2W] = \mathbb{E}[X^2]\mathbb{E}[W] = 0$.

4.3.
$\text{Cov}[X,Y]^2 = \mathbb{E}[(X-\mu)(Y-\nu)]^2 \le \mathbb{E}[(X-\mu)]^2\mathbb{E}[(Y-\nu)]^2 = \text{Var}[X]\text{Var}[Y]$.
where Cauchy-Schwarz inequality was used.
Therefore, we have

$$
\frac{\text{Cov}[X,Y]^2}{\text{Var}[X]\text{Var}[Y]} \le 1 \Rightarrow -1 \le \rho(X,Y) \le 1.
$$

4.4.

$\text{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[aX^2 + bX] - \mathbb{E}[X]\mathbb{E}[aX + b]$.
$\mathbb{E}[X^2] = \sigma_x^2 + \mu_x^2 \Rightarrow \text{Cov}[X, Y] = a\sigma_x^2 + a\mu_x^2 - \mu_x(a\mu_x + b) + b\mu_x = a\mu_x^2$.
Therefore,

$$\frac{\text{Cov}[X, Y]}{\sigma_x \sigma_y} = \frac{a\sigma_x^2}{\sigma_x |a| \sigma_x} = \text{sgn}(a).$$

4.5.
Let $\mathbf{\Sigma} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$ be the orthonormal eigendecomposition of $\mathbf{\Sigma}$, then we have $\mathbf{\Sigma}^{-1} = \mathbf{Q}^T\mathbf{\Lambda}^{-1}\mathbf{Q}$. Setting $\mathbf{y} = \mathbf{Q}(\mathbf{x} - \boldsymbol{\mu})$ simplifies the exponent as follows:

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = -\frac{1}{2}\mathbf{y}^T \mathbf{\Lambda}^{-1} \mathbf{y}.$$

Changing of variables in the integral gives

$$\int e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})} d\mathbf{x} = \left| \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right| \int e^{-\frac{1}{2}\mathbf{y}^T \mathbf{\Lambda}^{-1} \mathbf{y}} d\mathbf{y}.$$

$$\mathbf{x} = \mathbf{Q}^T \mathbf{y} + \boldsymbol{\mu} \Rightarrow \left| \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right| = \left| \mathbf{Q}^T \right| = 1,$$

as $\mathbf{Q}$ is orthonormal. Now it suffices to calculate the second integral. Using the fact that $\mathbf{\Lambda}$ is a diagonal matrix whose entries are exactly the eigenvalues of $\mathbf{\Sigma}$ (For convenience, denote them by λ_i), we know that the exponent matrix factorizes over its row vectors. Therefore,

$$\int e^{-\frac{1}{2}\mathbf{y}^T \mathbf{\Lambda}^{-1} \mathbf{y}} d\mathbf{y} = \prod_{i=1}^{D} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\lambda_i^{-1} y_i^2} dy_i = (2\pi)^{\frac{D}{2}} \prod_{i=1}^{D} \lambda_i^{\frac{1}{2}} = (2\pi)^{\frac{D}{2}} \left| \mathbf{\Sigma} \right|^{\frac{1}{2}}.$$

4.6.
The probability density function of normal distribution is

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} \left| \mathbf{\Sigma} \right|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}.$$

In $D = 2$ case, the formula would be

$$p(x_1, x_2) = \frac{1}{(2\pi)^{\frac{2}{2}} \left| \sigma_1^2 \sigma_2^2 (1 - \rho^2) \right|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}.$$

The exponent is

$$
\begin{aligned}
&-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\
&= -\frac{1}{2}\begin{pmatrix} x_1-\mu_1 & x_2-\mu_2 \end{pmatrix}\frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)}\begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix} \\
&= -\frac{1}{2\sigma_1^2\sigma_2^2(1-\rho^2)}\begin{pmatrix} x_1-\mu_1 & x_2-\mu_2 \end{pmatrix}\begin{pmatrix} \sigma_2^2(x_1-\mu_1)-\rho\sigma_1\sigma_2(x_2-\mu_2) \\ -\rho\sigma_1\sigma_2(x_1-\mu_1)+\sigma_1^2(x_2-\mu_2) \end{pmatrix} \\
&= -\frac{1}{2\sigma_1^2\sigma_2^2(1-\rho^2)}(\sigma_2^2(x_1-\mu_1)^2+\sigma_1^2(x_2-\mu_2)^2-2\rho\sigma_1\sigma_2(x_1-\mu_1)(x_2-\mu_2))
\end{aligned} \tag{2}
$$

Therefore, the probability density function is

$$
p(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)}(\frac{(x_1-\mu_1)^2}{\sigma_1^2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}-2\rho\frac{(x_1-\mu_1)}{\sigma_1}\frac{(x_2-\mu_2)}{\sigma_2})}.
$$

4.7.
a.

$$
\begin{aligned}
&P(X_2|x_1) = \mathcal{N}(x_2|\boldsymbol{\mu}_2+\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1-\boldsymbol{\mu}_1),\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}) \\
&= \frac{1}{\sqrt{2\pi(\sigma_2^2-\rho^2\sigma_1^2)}}e^{-\frac{1}{2}(x_2-(\mu_2+\frac{\sigma_{21}}{\sigma_1^2}(x_1-\mu_1)))} \\
&= \frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_2}e^{-\frac{1}{2}(x_2-(\mu_2+\frac{\sigma_2}{\sigma_1}(x_1-\mu_1)))}.
\end{aligned} \tag{3}
$$

b. If $\sigma_1 = \sigma_2 = 1$,

$$
P(X_2|x_1) = \frac{1}{\sqrt{2\pi(1-\rho^2)}}e^{-\frac{1}{2}((x_2-\mu_2)-(x_1-\mu_1)))}. \tag{4}
$$

4.9.

$$P(\mu|\mathcal{D}) = \mathcal{N}(\mu|\frac{\frac{n_1}{v_1}\bar{y}^{(1)} + \frac{n_2}{v_2}\bar{y}^{(2)}}{\frac{n_1}{v_1} + \frac{n_2}{v_2}}, \frac{1}{\frac{n_1}{v_1} + \frac{n_2}{v_2}})$$
$$= \mathcal{N}(\mu|\frac{n_1 v_2 \bar{y}^{(1)} + n_2 v_1 \bar{y}^{(2)}}{n_1 v_2 + n_2 v_1}, \frac{v_1 v_2}{n_1 v_2 + n_2 v_1}) \tag{5}$$

4.10.
Recall that the probability density functions of marginalized and conditioned normal distributions are
$p(\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_2|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$,
$p(\mathbf{x}_1|\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$.

To get the parameters in the information form, we use the relations $\boldsymbol{\xi} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ and

$$\begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \end{pmatrix}. \tag{6}$$

Mean of marginalized distribution:
$\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2 = \boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{\Sigma}_{21}\boldsymbol{\xi}_1 + \boldsymbol{\Sigma}_{22}\boldsymbol{\xi}_2) = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\xi}_1 + \boldsymbol{\xi}_2$.
$\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = -(\boldsymbol{\Lambda}/\boldsymbol{\Lambda}_{11})(\boldsymbol{\Lambda}/\boldsymbol{\Lambda}_{11})^{-1}\boldsymbol{\Lambda}_{21}\boldsymbol{\Lambda}_{11}^{-1} = \boldsymbol{\Lambda}_{21}\boldsymbol{\Lambda}_{11}^{-1}$ where $\boldsymbol{\Lambda}/\boldsymbol{\Lambda}_{11}$ denotes the Schur complement.
$\Rightarrow \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2 = \boldsymbol{\xi}_2 - \boldsymbol{\Lambda}_{21}\boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\xi}_1$.

Variance of marginalized distribution:
$\boldsymbol{\Sigma}_{22}^{-1} = \boldsymbol{\Lambda}_{22} - \boldsymbol{\Lambda}_{21}\boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}$.

Mean of conditioned distribution:
$\boldsymbol{\Lambda}_{11}(\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2))$
$= \boldsymbol{\Lambda}_{11}((\boldsymbol{\Sigma}_{11}\boldsymbol{\xi}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\xi}_2) - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - (\boldsymbol{\Sigma}_{21}\boldsymbol{\xi}_1 + \boldsymbol{\Sigma}_{22}\boldsymbol{\xi}_2)))$
$= \boldsymbol{\Lambda}_{11}(\boldsymbol{\Sigma}_{11}\boldsymbol{\xi}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\xi}_2 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{x}_2 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\xi}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\xi}_2)$
$= \boldsymbol{\Lambda}_{11}((\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})\boldsymbol{\xi}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{x}_2)$
$= \boldsymbol{\Lambda}_{11}(\boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\xi}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{x}_2)$
$= \boldsymbol{\xi}_1 + \boldsymbol{\Lambda}_{11}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{x}_2 = \boldsymbol{\xi}_1 + \boldsymbol{\Lambda}_{11}(-\boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12})\mathbf{x}_2$

$= \boldsymbol{\xi}_1 - \mathbf{\Lambda}_{12}\mathbf{x}_2.$

Variance of conditioned distribution:
$(\mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21})^{-1} = \mathbf{\Lambda}_{11}$

Therefore, the probability density functions of marginalized and conditioned normal distributions in information form are
$p(\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_2|\boldsymbol{\xi}_2 - \mathbf{\Lambda}_{21}\mathbf{\Lambda}_{11}^{-1}\boldsymbol{\xi}_1, \mathbf{\Lambda}_{22} - \mathbf{\Lambda}_{21}\mathbf{\Lambda}_{11}^{-1}\mathbf{\Lambda}_{12}),$
$p(\mathbf{x}_1|\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_1|\boldsymbol{\xi}_1 - \mathbf{\Lambda}_{12}\mathbf{x}_2, \mathbf{\Lambda}_{11}).$

4.11.

$$\begin{aligned}
&p(\boldsymbol{\mu}, \mathbf{\Sigma}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\mu}, \mathbf{\Sigma})p(\boldsymbol{\mu}, \mathbf{\Sigma}) \\
&\propto e^{-\frac{1}{2}\sum_{i=1}^{N}[(\mathbf{x}_i-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})]} \cdot e^{-\frac{1}{2}[\kappa_0(\boldsymbol{\mu}-\mathbf{m}_0)^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}-\mathbf{m}_0)+\mathrm{tr}(\mathbf{\Sigma}^{-1}S_0)]} \\
&\cdot |\mathbf{\Sigma}|^{-\frac{v_0+D+2}{2}} \cdot |\mathbf{\Sigma}|^{-\frac{N}{2}} \\
&\propto e^{-\frac{1}{2}[N(\boldsymbol{\mu}-\bar{\mathbf{x}})^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}-\bar{\mathbf{x}})+\mathrm{tr}(\mathbf{\Sigma}^{-1}S_{\bar{\mathbf{x}}})+\kappa_0(\boldsymbol{\mu}-\mathbf{m}_0)^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}-\mathbf{m}_0)+\mathrm{tr}(\mathbf{\Sigma}^{-1}S_0)]} \\
&\cdot |\mathbf{\Sigma}|^{-\frac{v_0+D+N+2}{2}}.
\end{aligned} \tag{7}$$

To simplify the exponent term, it is needed to calculate the following term first:

$$\begin{aligned}
&N(\boldsymbol{\mu} - \bar{\mathbf{x}})(\boldsymbol{\mu} - \bar{\mathbf{x}})^T + \kappa_0(\boldsymbol{\mu} - \mathbf{m}_0)(\boldsymbol{\mu} - \mathbf{m}_0)^T \\
&= N(\bar{\mathbf{x}}\bar{\mathbf{x}}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T - \bar{\mathbf{x}}\boldsymbol{\mu}^T - \boldsymbol{\mu}\bar{\mathbf{x}}^T) + \kappa_0(\boldsymbol{\mu}\boldsymbol{\mu}^T + \mathbf{m}_0\mathbf{m}_0^T - \boldsymbol{\mu}\mathbf{m}_0^T - \mathbf{m}_0\boldsymbol{\mu}^T) \\
&= (\kappa_0 + N)\boldsymbol{\mu}\boldsymbol{\mu}^T - (\kappa_0 + N)(\boldsymbol{\mu}(\frac{\kappa_0}{\kappa_0 + N}\mathbf{m}_0)^T + (\frac{\kappa_0}{\kappa_0 + N}\mathbf{m}_0)\boldsymbol{\mu}^T) \\
&- (\kappa_0 + N)(\boldsymbol{\mu}(\frac{N}{\kappa_0 + N}\bar{\mathbf{x}})^T + (\frac{N}{\kappa_0 + N}\bar{\mathbf{x}})\boldsymbol{\mu}^T) + \kappa_0\mathbf{m}_0\mathbf{m}_0^T + N\bar{\mathbf{x}}\bar{\mathbf{x}}^T \\
&= (\kappa_0 + N)(\boldsymbol{\mu}\boldsymbol{\mu}^T - \boldsymbol{\mu}\mathbf{m}_N^T - \mathbf{m}_N\boldsymbol{\mu}^T) + \kappa_0\mathbf{m}_0\mathbf{m}_0^T + N\bar{\mathbf{x}}\bar{\mathbf{x}}^T \\
&= (\kappa_0 + N)(\boldsymbol{\mu} - \mathbf{m}_N)(\boldsymbol{\mu} - \mathbf{m}_N)^T - (\kappa_0 + N)\mathbf{m}_N\mathbf{m}_N^T + \kappa_0\mathbf{m}_0\mathbf{m}_0^T + N\bar{\mathbf{x}}\bar{\mathbf{x}}^T \\
&= (\kappa_0 + N)(\boldsymbol{\mu} - \mathbf{m}_N)(\boldsymbol{\mu} - \mathbf{m}_N)^T + \kappa_0\mathbf{m}_0\mathbf{m}_0^T + N\bar{\mathbf{x}}\bar{\mathbf{x}}^T \\
&- (\frac{\kappa_0^2}{\kappa_0 + N}\mathbf{m}_0\mathbf{m}_0^T + \frac{\kappa_0 N}{\kappa_0 + N}(\mathbf{m}_0\bar{\mathbf{x}}^T + \bar{\mathbf{x}}\mathbf{m}_0^T) + \frac{N^2}{\kappa_0 + N}\bar{\mathbf{x}}\bar{\mathbf{x}}^T)
\end{aligned} \tag{8}$$

$$
\begin{aligned}
&= (\kappa_0 + N)(\boldsymbol{\mu} - \mathbf{m}_N)(\boldsymbol{\mu} - \mathbf{m}_N)^T + \kappa_0\mathbf{m}_0\mathbf{m}_0^T + N\bar{\mathbf{x}}\bar{\mathbf{x}}^T \\
&- (\kappa_0\mathbf{m}_0\mathbf{m}_0^T + N\bar{\mathbf{x}}\bar{\mathbf{x}}^T) - (\frac{\kappa_0 N}{\kappa_0 + N})(\mathbf{m}_0\bar{\mathbf{x}}^T + \bar{\mathbf{x}}\mathbf{m}_0^T - \mathbf{m}_0\mathbf{m}_0^T - \bar{\mathbf{x}}\bar{\mathbf{x}}^T) \\
&= (\kappa_0 + N)(\boldsymbol{\mu} - \mathbf{m}_N)(\boldsymbol{\mu} - \mathbf{m}_N)^T + \frac{\kappa_0 N}{\kappa_0 + N}(\bar{\mathbf{x}} - \mathbf{m}_0)(\bar{\mathbf{x}} - \mathbf{m}_0)^T.
\end{aligned}
\tag{9}
$$

Using the trace identity $\text{tr}(ABC) = \text{tr}(BCA)$ and plugging in the calculation above, the exponent becomes

$$
\begin{aligned}
&N(\boldsymbol{\mu}-\bar{\mathbf{x}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\bar{\mathbf{x}})+\text{tr}(\boldsymbol{\Sigma}^{-1}S_{\bar{\mathbf{x}}})+\kappa_0(\boldsymbol{\mu}-\mathbf{m}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\mathbf{m}_0)+\text{tr}(\boldsymbol{\Sigma}^{-1}S_0) \\
&= \text{tr}(N(\boldsymbol{\mu}-\bar{\mathbf{x}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\bar{\mathbf{x}})+\kappa_0(\boldsymbol{\mu}-\mathbf{m}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\mathbf{m}_0))+\text{tr}(\boldsymbol{\Sigma}^{-1}(S_{\bar{\mathbf{x}}}+S_0)) \\
&= \text{tr}(\boldsymbol{\Sigma}^{-1}((\kappa_0+N)(\boldsymbol{\mu}-\mathbf{m}_N)(\boldsymbol{\mu}-\mathbf{m}_N)^T + \frac{\kappa_0 N}{\kappa_0+N}(\bar{\mathbf{x}}-\mathbf{m}_0)(\bar{\mathbf{x}}-\mathbf{m}_0)^T)) \\
&+ \text{tr}(\boldsymbol{\Sigma}^{-1}(S_{\bar{\mathbf{x}}} + S_0)) \\
&= (\kappa_0 + N)(\boldsymbol{\mu} - \mathbf{m}_N)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \mathbf{m}_N) + \text{tr}(\boldsymbol{\Sigma}^{-1}S_N).
\end{aligned}
\tag{10}
$$

Hence, the posterior distribution is

$$
p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{D}) \propto |\boldsymbol{\Sigma}|^{-\frac{v_N+2}{2}} e^{-\frac{1}{2}((\kappa_0+N)(\boldsymbol{\mu}-\mathbf{m}_N)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\mathbf{m}_N)+\text{tr}(\boldsymbol{\Sigma}^{-1}S_N))}.
$$

4.12.

$$
\begin{aligned}
&p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{N}{2}} e^{-\frac{1}{2}\text{tr}(\boldsymbol{\Sigma}^{-1}S_{\boldsymbol{\mu}})} \\
&\Rightarrow \log p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{N}{2}\log|\hat{\boldsymbol{\Sigma}}| - \frac{N}{2}\text{tr}(\hat{\boldsymbol{\Sigma}}^{-1}\frac{S_{\bar{\mathbf{x}}}}{N}) = -\frac{N}{2}\log|\hat{\boldsymbol{\Sigma}}| - \frac{N}{2}\text{tr}(\hat{\boldsymbol{\Sigma}}^{-1}\hat{S}).
\end{aligned}
\tag{11}
$$

a. Since the information is completely determined by $\hat{\boldsymbol{\Sigma}}$, which is a symmetric $N \times N$ matrix here, the degree of freedom is $d = \frac{N(N+1)}{2}$.

b. In this case $\hat{\boldsymbol{\Sigma}}$ is diagonal, so $d = N$.

4.13.

$$p(\mu|\mathcal{D}) \propto p(\mathcal{D}|\mu)p(\mu) = \mathcal{N}(\bar{x}|\mu, \frac{4}{n})\mathcal{N}(\mu|\mu_0, 9) = \mathcal{N}(\mu|\mu_n, \sigma_n^2)$$

where

$$\frac{1}{\sigma_n^2} = \frac{1}{9} + \frac{n}{4} \Rightarrow \sigma_n^2 = \frac{36}{4+9n}$$

$$\mu_n = \sigma_n^2(\frac{n}{4}\bar{x} + \frac{\mu}{9}).$$

To make width of the credible interval less than 1,

$$2 \cdot \frac{6}{\sqrt{4+9n}} < 1 \Rightarrow n > 61.$$

4.14.

a.

$$p(\mu|\mathcal{D}) \propto p(\mathcal{D}|\mu)p(\mu) = \mathcal{N}(\mu|\mu_n, \sigma_n^2)$$

where

$$\frac{1}{\sigma_n^2} = \frac{1}{s^2} + \frac{n}{\sigma^2} \Rightarrow \sigma_n^2 = \frac{s^2 n^2}{\sigma^2 + ns^2}$$

$$\mu_n = \frac{s^2\sigma^2}{\sigma^2+ns^2}(\frac{n}{\sigma^2}\bar{x} + \frac{m}{s^2}).$$

The log likelihood is

$$\log p(\mu|\mathcal{D}) = -\frac{1}{2}\log\sigma_n^2 - \frac{(\mu-\mu_n)^2}{2\sigma_n^2}$$

$$\Rightarrow \frac{\partial}{\partial\mu}p(\mu|\mathcal{D}) = -\frac{\mu-\mu_n}{\sigma_n^2}$$

$$\Rightarrow \hat{\mu} = \mu_n.$$

$$\Rightarrow \hat{\mu}_{MAP} = \frac{s^2\sigma^2}{\sigma^2+ns^2}(\frac{n}{\sigma^2}\bar{x} + \frac{m}{s^2}).$$

b.

$$\lim_{n\to\infty}\hat{\mu}_{MAP} = \frac{s^2\sigma^2}{s^2}\frac{\bar{x}}{\sigma^2} = \bar{x}.$$

c.

$$\lim_{s\to\infty}\hat{\mu}_{MAP} = \lim_{s\to\infty}\frac{\sigma^2}{n+\frac{\sigma^2}{s^2}}(\frac{n}{\sigma^2}\bar{x}+\frac{m}{s^2}) = \frac{\sigma^2}{n}(\frac{n}{\sigma^2}\bar{x}) = \bar{x}.$$

d.

$$\lim_{s\to 0+}\hat{\mu}_{MAP} = \lim_{s\to 0+}\frac{\sigma^2}{ns^2+\sigma^2}(\frac{s^2 n}{\sigma^2}\bar{x}+m) = m.$$

4.15.

a.

$$\begin{aligned}
&\mathbf{C}_{n+1} - \frac{n-1}{n}\mathbf{C}_n \\
&= \frac{1}{n}\sum_{i=1}^{n+1}(\mathbf{x}_i - \mathbf{m}_{n+1})(\mathbf{x}_i - \mathbf{m}_{n+1})^T - \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{m}_n)(\mathbf{x}_i - \mathbf{m}_n)^T.
\end{aligned} \tag{12}$$

Meanwhile,

$$\begin{aligned}
&\mathbf{m}_{n+1} - \mathbf{m}_n = \frac{1}{n+1}\sum_{i=1}^{n+1}\mathbf{x}_i - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = -\frac{1}{n(n+1)}\sum_{i=1}^{n}\mathbf{x}_i + \frac{1}{n+1}\mathbf{x}_{n+1} \\
&= -\frac{\mathbf{m}_n}{n+1} + \frac{\mathbf{x}_{n+1}}{n+1} = \frac{1}{n+1}(\mathbf{x}_{n+1} - \mathbf{m}_n).
\end{aligned} \tag{13}$$

If we denote $\mathbf{x}_{n+1} - \mathbf{m}_n = \mathbf{u}_n$,

$$\begin{aligned}
&(\mathbf{x}_i - \mathbf{m}_{n+1})(\mathbf{x}_i - \mathbf{m}_{n+1})^T = (\mathbf{x}_i - \mathbf{m}_n - \frac{\mathbf{u}_n}{n+1})(\mathbf{x}_i - \mathbf{m}_n - \frac{\mathbf{u}_n}{n+1})^T \\
&= (\mathbf{x}_i - \mathbf{m}_n)(\mathbf{x}_i - \mathbf{m}_n)^T - \frac{1}{n+1}((\mathbf{x}_i - \mathbf{m}_n)\mathbf{u}_n^T + \mathbf{u}_n(\mathbf{x}_i - \mathbf{m}_n)^T) + \frac{\mathbf{u}_n\mathbf{u}_n^T}{(n+1)^2}.
\end{aligned} \tag{14}$$

Using this relationship to the original formula, we get

$$
\begin{aligned}
&\mathbf{C}_{n+1} - \frac{n-1}{n}\mathbf{C}_n \\
&= \frac{1}{n}(\mathbf{x}_{n+1} - \mathbf{m}_n - \frac{\mathbf{u}_n}{n+1})(\mathbf{x}_{n+1} - \mathbf{m}_n - \frac{\mathbf{u}_n}{n+1})^T \\
&+ \frac{1}{n}\sum_{i=1}^{n}(-\frac{1}{n+1}((\mathbf{x}_i - \mathbf{m}_n)\mathbf{u}_n^T + \mathbf{u}_n(\mathbf{x}_i - \mathbf{m}_n)^T) + \frac{\mathbf{u}_n\mathbf{u}_n^T}{(n+1)^2}) \\
&= \frac{1}{n}(\frac{n}{n+1}\mathbf{u}_n)(\frac{n}{n+1}\mathbf{u}_n)^T + \frac{n\mathbf{u}_n\mathbf{u}_n^T}{(n+1)^2} \\
&- \frac{1}{n(n+1)}\sum_{i=1}^{n}((\mathbf{x}_i - \mathbf{m}_n)\mathbf{u}_n^T + \mathbf{u}_n(\mathbf{x}_i - \mathbf{m}_n)^T) \\
&= \frac{\mathbf{u}_n\mathbf{u}_n^T}{n+1} + \frac{1}{n(n+1)}((\sum_{i=1}^{n}\mathbf{x}_i - n\mathbf{m}_n)\mathbf{u}_n^T + \mathbf{u}_n(\sum_{i=1}^{n}\mathbf{x}_i - n\mathbf{m}_n)^T) \\
&= \frac{(\mathbf{x}_{n+1} - \mathbf{m}_n)(\mathbf{x}_{n+1} - \mathbf{m}_n)^T}{n+1} + \frac{1}{n(n+1)}(\mathbf{0}\mathbf{u}_n^T + \mathbf{u}_n\mathbf{0}^T).
\end{aligned}
\tag{15}
$$

$$
\Rightarrow \mathbf{C}_{n+1} = \frac{n-1}{n}\mathbf{C}_n + \frac{1}{n+1}(\mathbf{x}_{n+1} - \mathbf{m}_n)(\mathbf{x}_{n+1} - \mathbf{m}_n)^T.
$$

b. The time complexity is equal to the time complexity of computing $(\mathbf{x}_{n+1} - \mathbf{m}_n)(\mathbf{x}_{n+1} - \mathbf{m}_n)^T$, which is $O(d^2)$.

c.

$$
\begin{aligned}
\mathbf{C}_{n+1}^{-1} &= (\frac{n-1}{n}\mathbf{C}_n + \frac{1}{n+1}\mathbf{u}_n\mathbf{u}_n^T)^{-1} \\
&= \frac{n}{n-1}\mathbf{C}_n^{-1} - \frac{(\frac{n}{n-1}\mathbf{C}_n^{-1})(\frac{1}{n+1})\mathbf{u}_n\mathbf{u}_n^T(\frac{n}{n-1})\mathbf{C}_n^{-1}}{1+(\frac{1}{n+1})\mathbf{u}_n^T(\frac{n}{n-1})\mathbf{C}_n^{-1}\mathbf{u}_n} \\
&= \frac{n}{n-1}\mathbf{C}_n^{-1} - \frac{n^2\mathbf{C}_n^{-1}\mathbf{u}_n\mathbf{u}_n^T\mathbf{C}_n^{-1}}{(n-1)^2(n+1)+(n-1)n\mathbf{u}_n^T\mathbf{C}_n^{-1}\mathbf{u}_n} \\
&= \frac{n}{n-1}(\mathbf{C}_n^{-1} - \frac{n\mathbf{C}_n^{-1}\mathbf{u}_n\mathbf{u}_n^T\mathbf{C}_n^{-1}}{(n^2-1)+n\mathbf{u}_n^T\mathbf{C}_n^{-1}\mathbf{u}_n}) \\
&= \frac{n}{n-1}(\mathbf{C}_n^{-1} - \frac{\mathbf{C}_n^{-1}(\mathbf{x}_{n+1}-\mathbf{m}_n)(\mathbf{x}_{n+1}-\mathbf{m}_n)^T\mathbf{C}_n^{-1}}{\frac{n^2-1}{n}+n(\mathbf{x}_{n+1}-\mathbf{m}_n)^T\mathbf{C}_n^{-1}(\mathbf{x}_{n+1}-\mathbf{m}_n)})
\end{aligned}
\tag{16}
$$

d. Same with problem b, the time complexity is $O(d^2)$.

4.16.

$$
\begin{aligned}
\frac{p(\mathbf{x}|y=1)}{p(\mathbf{x}|y=0)} &= \frac{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)}{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)} = \sqrt{\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|}}e^{-\frac{1}{2}((\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)-(\mathbf{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\mathbf{x}-\boldsymbol{\mu}_0))} \\
\log\frac{p(\mathbf{x}|y=1)}{p(\mathbf{x}|y=0)} &= -\frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|} - \frac{1}{2}((\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)-(\mathbf{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\mathbf{x}-\boldsymbol{\mu}_0))
\end{aligned}
\tag{17}
$$

If the covariance is shared across classes ($\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$), the form becomes

$$
\begin{aligned}
&-\frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|} - \frac{1}{2}((\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)-(\mathbf{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)) \\
&= -\frac{1}{2}(\mathrm{tr}(\boldsymbol{\Sigma}^{-1}((\mathbf{x}-\boldsymbol{\mu}_1)^T(\mathbf{x}-\boldsymbol{\mu}_1)-(\mathbf{x}-\boldsymbol{\mu}_0)^T(\mathbf{x}-\boldsymbol{\mu}_0))).
\end{aligned}
\tag{18}
$$

If the shared covariance $\mathbf{\Sigma}$ has only diagonal entries, the scatter matrix factorizes over rows, therefore

$$\begin{aligned}
&-\frac{1}{2}(\mathrm{tr}(\mathbf{\Sigma}^{-1}((\mathbf{x}-\boldsymbol{\mu}_1)^T(\mathbf{x}-\boldsymbol{\mu}_1)-(\mathbf{x}-\boldsymbol{\mu}_0)^T(\mathbf{x}-\boldsymbol{\mu}_0))) \\
&= -\frac{1}{2}\sum_{i=1}^{d}(\frac{1}{\sigma_i^2}(2x_i-(\boldsymbol{\mu}_0+\boldsymbol{\mu}_1)_i)(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)_i).
\end{aligned} \tag{19}$$

If the diagonal, shared covariance matrix is multiple of identity matrix ($\sigma^2 = \sigma_1^2 = \cdots = \sigma_d^2$), then

$$\begin{aligned}
&-\frac{1}{2}\sum_{i=1}^{d}(\frac{1}{\sigma_i^2}(2x_i-(\boldsymbol{\mu}_0+\boldsymbol{\mu}_1)_i)(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)_i) \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{d}((2x_i-(\boldsymbol{\mu}_0+\boldsymbol{\mu}_1)_i)(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)_i).
\end{aligned} \tag{20}$$

4.18.
a.

$$\begin{aligned}
&p(y|x_1=0,x_2=0) = \frac{p(x_1=0,x_2=0|y)p(y)}{p(x_1=0,x_2=0)} \\
&p(x_1=0,x_2=0|y)p(y) = p(x_1=0|y)p(x_2=0|y)p(y) \\
&= (1-\theta_c)(\frac{1}{\sqrt{2\pi}\sigma_c}e^{-\frac{\mu_c^2}{2\sigma_c^2}})\pi_c
\end{aligned} \tag{21}$$

Therefore, $p(y|x_1=0,x_2=0)$ is normalized form of

$$
\begin{aligned}
&(0.5 \cdot 0.5 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}}, 0.25 \cdot 0.5 \cdot \frac{1}{\sqrt{2\pi}}, 0.25 \cdot 0.5 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}}) \\
&\sim (\frac{2}{\sqrt{e}+3}, \frac{\sqrt{e}}{\sqrt{e}+3}, \frac{1}{\sqrt{e}+3}).
\end{aligned} \tag{22}
$$

b.

$$
\begin{aligned}
&p(y|x_1 = 0) = \frac{p(x_1 = 0|y)p(y)}{p(x_1 = 0)} \\
&\sim (0.5 \cdot 0.5, 0.25 \cdot 0.5, 0.25 \cdot 0.5) \sim (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})
\end{aligned} \tag{23}
$$

c.

$$
\begin{aligned}
&p(y|x_2 = 0) = \frac{p(x_2 = 0|y)p(y)}{p(x_2 = 0)} \\
&\sim (0.5 \cdot \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}}, 0.25 \cdot \frac{e^{0}}{\sqrt{2\pi}}, 0.25 \cdot \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}}) \\
&\sim (\frac{2}{\sqrt{e}+3}, \frac{\sqrt{e}}{\sqrt{e}+3}, \frac{1}{\sqrt{e}+3}).
\end{aligned} \tag{24}
$$

d. We can observe $p(y|x_2 = 0) = p(y|x_1 = 0, x_2 = 0)$.
This happens because $x_1|y$ has uniform density.

4.19.

$$
\begin{aligned}
&p(y = 1|\mathbf{x}, \boldsymbol{\theta}) = \pi_0 |2\pi k^d \boldsymbol{\Sigma}_0|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_0^{-1} (\mathbf{x}-\boldsymbol{\mu}_1) \cdot \frac{1}{k}} \\
&\propto e^{\frac{1}{k}(\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1) + \log \frac{\pi_1}{\sqrt{k}^d} - \frac{1}{2k} \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}} \\
&= \frac{1}{1 + e^{(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_1)^T \mathbf{x} + (\gamma_0 - \gamma_1) + \delta}}
\end{aligned} \tag{25}
$$

where

$$
\begin{aligned}
&(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_1)^T = (\boldsymbol{\mu}_0 - \frac{1}{k}\boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}, \\
&\gamma_0 - \gamma_1 = -\frac{1}{2}(\boldsymbol{\mu}_0 - \frac{1}{\sqrt{k}}\boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \frac{1}{\sqrt{k}}\boldsymbol{\mu}_1), \\
&\delta = e^{\frac{1-k}{2k}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}}.
\end{aligned}
\tag{26}
$$

4.20.
a. GaussI $\leq$ LinLog.
Both have logistic posteriors, but LinLog optimizes log probabilities.
b. GaussX $\leq$ QuadLog.
Both have logistic posteriors with quadratic features, but QuadLog optimizes log probabilities.
c. LinLog $\leq$ QuadLog.
Logistic regression with linear features are a subclass of logistic regression with quadratic features.
d. GaussI $\leq$ QuadLog.
By a. and c.
e. No. Different log-likelihood can provide same classification result.

4.21.
a.

$$p(x|\mu_1, \sigma_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, p(x|\mu_2, \sigma_2) = \frac{1}{\sqrt{2\pi \cdot 10^6}} e^{-\frac{(x-1)^2}{2 \cdot 10^6}}$$

$$p(x|\mu_1, \sigma_1) = p(x|\mu_2, \sigma_2) \Rightarrow x = \pm 3.717$$

$\Rightarrow p(x|\mu_1, \sigma_1) \geq p(x|\mu_2, \sigma_2)$ for $x \in [-3.717, 3.717]$.

b.

$$p(x|\mu_1, \sigma_1) = p(x|\mu_2, \sigma_2) \Rightarrow x = 0.5$$

$\Rightarrow p(x|\mu_1, \sigma_1) \geq p(x|\mu_2, \sigma_2)$ for $x \leq 0.5$

4.22.

a.

$$
\begin{aligned}
p(Y=1|\mathbf{x}) &= \frac{1}{2\pi\cdot\sqrt{0.49}}\exp(-\frac{1}{2}\begin{pmatrix}-0.5 & 0.5\end{pmatrix}\begin{pmatrix}\frac{1}{0.7} & 0\\ 0 & \frac{1}{0.7}\end{pmatrix}\begin{pmatrix}-0.5\\ 0.5\end{pmatrix}) \approx 0.1591\\
p(Y=2|\mathbf{x}) &= \frac{1}{2\pi\cdot\sqrt{0.6}}\exp(-\frac{1}{2}\begin{pmatrix}-1.5 & -0.5\end{pmatrix}\begin{pmatrix}\frac{0.8}{0.6} & -\frac{0.2}{0.6}\\ -\frac{0.2}{0.6} & \frac{0.8}{0.6}\end{pmatrix}\begin{pmatrix}-1.5\\ -0.5\end{pmatrix}) \approx 0.0498\\
p(Y=3|\mathbf{x}) &= \frac{1}{2\pi\cdot\sqrt{0.6}}\exp(-\frac{1}{2}\begin{pmatrix}0.5 & -0.5\end{pmatrix}\begin{pmatrix}\frac{0.8}{0.6} & -\frac{0.2}{0.6}\\ -\frac{0.2}{0.6} & \frac{0.8}{0.6}\end{pmatrix}\begin{pmatrix}0.5\\ -0.5\end{pmatrix}) \approx 0.1354
\end{aligned}
\tag{27}
$$

Therefore, $\mathbf{x}$ is most likely to be in class 1.

b.

$$
\begin{aligned}
p(Y=1|\mathbf{x}) &= \frac{1}{2\pi\cdot\sqrt{0.49}}\exp(-\frac{1}{2}\begin{pmatrix}0.5 & 0.5\end{pmatrix}\begin{pmatrix}\frac{1}{0.7} & 0\\ 0 & \frac{1}{0.7}\end{pmatrix}\begin{pmatrix}0.5\\ 0.5\end{pmatrix}) \approx 0.1591\\
p(Y=2|\mathbf{x}) &= \frac{1}{2\pi\cdot\sqrt{0.6}}\exp(-\frac{1}{2}\begin{pmatrix}-0.5 & -0.5\end{pmatrix}\begin{pmatrix}\frac{0.8}{0.6} & -\frac{0.2}{0.6}\\ -\frac{0.2}{0.6} & \frac{0.8}{0.6}\end{pmatrix}\begin{pmatrix}-0.5\\ -0.5\end{pmatrix}) \approx 0.1600\\
p(Y=3|\mathbf{x}) &= \frac{1}{2\pi\cdot\sqrt{0.6}}\exp(-\frac{1}{2}\begin{pmatrix}1.5 & -0.5\end{pmatrix}\begin{pmatrix}\frac{0.8}{0.6} & -\frac{0.2}{0.6}\\ -\frac{0.2}{0.6} & \frac{0.8}{0.6}\end{pmatrix}\begin{pmatrix}1.5\\ -0.5\end{pmatrix}) \approx 0.0302
\end{aligned}
\tag{28}
$$

Therefore, $\mathbf{x}$ is most likely to be in class 2.

4.23.
a. $\mu_m \approx 72.33$, $\sigma_m^2 \approx 24.89$, $\pi_m = 0.5$
$\mu_f = 65$, $\sigma_f^2 \approx 12.67$, $\pi_f = 0.5$

b. $p(y = m|x, \hat{\theta}) \approx 0.831$.

c. Using bivariate normal distribution can be a better alternative.