1) Recall from Gaussian mixture distribution that it can be written as $p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$ Let's also introduce K-dimensional binary random variable z in which particular element $z_k = 1$ and all others are equal to zero. Values z_k satisfy $z_k \in \{0,1\}$ and $\sum_k z_k = 1$. After this, we define marginal distribution over z in terms of mixing coefficients π_k where $p(z_k = 1) = \pi_k$ such that $\sum_{k=1}^{K} \pi_k = 1$ with $0 \le \pi_k \le 1$

Since z uses 1-of-K representation, we can also rewrite this distribution as $p(z) = \prod_{k=1}^{K} \pi_k^{z_k}$ $\Rightarrow$ So, $p(x|z_k = 1) = \mathcal{N}(x|\mu_k, \Sigma_k)$ where $p(x|z) = \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k}$. By summing up joint distribution over all possible states of z

$$p(x) = \sum_z p(z)p(x|z) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$ $\rightsquigarrow$ Gaussian mixture involving explicit latent variables

Another quantity $\rightsquigarrow \gamma(z_k)$ to denote $p(z_k = 1|x)$ where from Bayes

$$\gamma(z_k) \equiv p(z_k = 1|x) = \frac{p(z_k = 1)p(x|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(x|z_j = 1)} = \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)}$$

We will view π_k as prior probability of $z_k = 1$, and quantity $\gamma(z_k)$ as corresponding posterior probability once we have observed x. In fact, $\gamma(z_k)$ is viewed as the responsibility that component k takes for explaining the observation x.

In general, EM Algorithm can be regarded as a soft-version of K-means for Gaussian mixture model. The EM algorithm uses responsibilities to make a soft assignment of each data point to one of the clusters. When ε is fixed, responsibility of data point i assigning to cluster k is given by the following relation:

$$r_i^{(k)} = \frac{\exp\left(-\frac{1}{2\varepsilon^2}\|y_i-\mu_k\|^2\right)}{\sum_{p=1}^{K}\exp\left(-\frac{1}{2\varepsilon^2}\|y_i-\mu_p\|^2\right)}$$

It's observable fact that as $\varepsilon \to 0$, $r_i^{(k)} \to 1$ for the cluster k that is closest to y_i and $r_i^{(k)} \to 0$ for other clusters, indicating how K-means algorithm is recovered as $\varepsilon \to 0$

More rigorous: Notice $r_i^{(k)} = \frac{\exp\left(\frac{-1}{2\varepsilon^2}\|y_i-\mu_k\|^2\right)}{\sum_{p=1}^{K}\exp\left(\frac{-1}{2\varepsilon^2}\|y_i-\mu_p\|^2\right)} =$

$$= \frac{1}{\sum_{p=1}^{K}\exp\left(\frac{-1}{2\varepsilon^2}\left(\|y_i-\mu_p\|^2-\|y_i-\mu_k\|^2\right)\right)}$$

and that

$$\exp\left(\frac{-1}{2\varepsilon^2}\left(\|y_i-\mu_p\|^2-\|y_i-\mu_k\|^2\right)\right) \to \begin{cases} 1, & \|y_i-\mu_p\| = \|y_i-\mu_k\| \\ 0, & \|y_i-\mu_p\| > \|y_i-\mu_k\| \\ \infty, & \|y_i-\mu_p\| < \|y_i-\mu_k\| \end{cases}$$

Thus, it means if 1 is the unique closest cluster to y_i, then the denominator of $r_i^{(1)}$ will have all terms tending to zero, except for 1 term tending to 1, so the limit is $\frac{1}{1+0+\cdots+0} = 1$ while the denominator of $r_i^{(k)}$ for $k \neq 1$ will have at least 1 term ending to ∞, making the limit 0 instead.

It's also worthwhile to mention that if there are multiple "closest clusters" to a point y_i, then the limiting probability will be uniform among these clusters.

Now, coming to the more rigorous approach where we initially defined $\gamma(z_{nk})$ as $\frac{\pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n | \mu_j, \Sigma_j)}$ since we are considering GMM with covariance matrices defined as εI (where without loss of generality, we assume it's variance parameter)

in which the normal distribution becomes $p(x | \mu_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^M \varepsilon^M}} e^{-\frac{1}{2\varepsilon}\|x-\mu_k\|^2}$ where $\gamma(z_{nk})$ is accompanied in the formula as

$$\frac{\pi_k e^{-\frac{\|x_n-\mu_k\|^2}{2\varepsilon}}}{\sum_{j=1}^{K} \pi_j e^{-\frac{\|x_n-\mu_j\|^2}{2\varepsilon}}}$$

Now, if we also define

$$r_{nk} = \begin{cases} 1, & \text{if } k = \arg\min_j \|x_n - \mu_j\|^2 \\ 0, & \text{in the other case} \end{cases}$$

Taking limit $\varepsilon \to 0$, we'll showcase how responsibilities $\gamma(z_{nk})$ for data pts x_n will go to unity except for specific term

Denote the distance by δ_k where $\delta_k = \|x_n - \mu_k\|^2$

If we substitute those notations into the original formula, one can get that $\frac{\pi_k e^{\frac{-\|x_n-\mu_k\|^2}{2\varepsilon}}}{\sum_{j=1}^{K} \pi_j e^{-\frac{\|x_n-\mu_j\|^2}{2\varepsilon}}} = \frac{\pi_k e^{-\frac{\delta_k}{2\varepsilon}}}{\sum_{j=1}^{K} \pi_j e^{-\frac{\delta_j}{2\varepsilon}}}$

If one defines the minimum value of all values δ_i's, let us denote it as ρ: $\underline{\min(\delta_1, \delta_2, \dots, \delta_K) = \rho}$, then $\frac{\pi_k e^{-\frac{\delta_k}{2\varepsilon}}}{\sum_{j=1}^{K} \pi_j e^{-\frac{\delta_j}{2\varepsilon}}} = \frac{\pi_k e^{\frac{\rho-\delta_k}{2\varepsilon}}}{\sum_{j=1}^{K} \pi_j e^{\frac{\rho-\delta_j}{2\varepsilon}}}$

it's clear that $\rho \leq \delta_i$ for all i, and if $\rho = \delta_t$ for some t, then taking k different from those t indices yield that

$\lim_{\varepsilon \to 0} \frac{\pi_k e^{\frac{\rho-\delta_k}{2\varepsilon}}}{\sum_{j=1}^{K} \pi_j e^{\frac{\rho-\delta_j}{2\varepsilon}}} = \lim_{\varepsilon \to 0} \frac{\pi_k e^{\frac{\rho-\delta_k}{2\varepsilon}}}{\sum_{t} \pi_t + \sum_{j \neq t} \pi_j e^{\frac{\rho-\delta_j}{2\varepsilon}}}$ with

$\rho \neq \delta_k$ giving $\lim_{\varepsilon \to 0} e^{\frac{\rho-\delta_k}{2\varepsilon}} = \lim_{\varepsilon \to 0} \frac{1}{e^{\frac{\delta_k-\rho}{2\varepsilon}}} = \lim_{\varepsilon \to 0} \frac{1}{\infty} = 0$ ✓

However, choosing the values k in which δ_k becomes the minimum (set of values t)

then, $\lim_{\varepsilon \to 0} \frac{\pi_t e^{\frac{\rho-\delta_t}{2\varepsilon}}}{\sum_{j=1}^{K} \pi_j e^{\frac{\rho-\delta_j}{2\varepsilon}}} = \lim_{\varepsilon \to 0} \frac{\pi_t}{\pi_t + \sum_{j \neq t} \pi_j e^{\frac{\rho-\delta_j}{2\varepsilon}}} = 1$ ✓ ⊕

which concludes that limiting case is equivalent to k-means clustering

Anar Rzayev MA3473 Homework #03 ID: 20190788

3) We first prove that $(A+UWV^T)^{-1} = A^{-1} - A^{-1}U(W^{-1}+V^TA^{-1}U)^{-1}V^TA^{-1}$ where $A \in \mathbb{R}^{n\times n}$, $W \in \mathbb{R}^{m\times m}$, $U, V \in \mathbb{R}^{n\times m}$.

The formula could be shown by checking that $A+UWV^T$ times the desired inverse on the RHS gives identity

$$(A+UWV^T)\left[A^{-1} - A^{-1}U(W^{-1}+V^TA^{-1}U)^{-1}V^TA^{-1}\right] =$$

$$= I + UWV^TA^{-1} - AA^{-1}U(W^{-1}+V^TA^{-1}U)^{-1}V^TA^{-1} -$$

$$- UWV^TA^{-1}U(W^{-1}+V^TA^{-1}U)^{-1}V^TA^{-1} = \{I+UWV^TA^{-1}\} -$$

$$- \{U(W^{-1}+V^TA^{-1}U)^{-1}V^TA^{-1} + UWV^TA^{-1}U(W^{-1}+V^TA^{-1}U)^{-1}V^TA^{-1}\} =$$

$$= \{I+UWV^TA^{-1}\} - \left[U(W^{-1}+V^TA^{-1}U)^{-1} + UWV^TA^{-1}U(W^{-1}+V^TA^{-1}U)^{-1}\right]\cdot V^TA^{-1} =$$

$$= \{I+UWV^TA^{-1}\} - \left[U(W^{-1}+V^TA^{-1}U)^{-1} + UWV^TA^{-1}U(W^{-1}+V^TA^{-1}U)^{-1}\right]V^TA^{-1} =$$

$$= \{I+UWV^TA^{-1}\} - \left[U+UWV^TA^{-1}U\right](W^{-1}+V^TA^{-1}U)^{-1}V^TA^{-1} =$$

$$= I + UWV^TA^{-1} - UW\underbrace{\left[W^{-1}+V^TA^{-1}U\right](W^{-1}+V^TA^{-1}U)^{-1}}V^TA^{-1} =$$

$$= I + UWV^TA^{-1} - UWV^TA^{-1} = I$$ since the squirly lines are multiplied into identity matrix I. Hence, the property holds. ☑

Now, we'll prove the first task of given question. The key factor is to consider that A, B - symmetric matrices since they are covariance matrices, with sizes $A, B \in \mathbb{R}^{D \times D}$ and if we take into account $a, b \in \mathbb{R}^D$, then column-vector c and covariance matrix C also become in dimension $\mathbb{R}^{D \times 1}$ and $\mathbb{R}^{D \times D}$, respectively. ($c \in \mathbb{R}^D$, $C \in \mathbb{R}^{D \times D}$)

Initially, we apply the given formulas for Gaussian distributions and check whether they satisfy or not.

$$\mathcal{N}(x|a, A) = \frac{1}{\sqrt{(2\pi)^D \det(A)}} e^{-\frac{1}{2}(x-a)^T A^{-1}(x-a)}$$

$$\mathcal{N}(x|b, B) = \frac{1}{\sqrt{(2\pi)^D \det(B)}} e^{-\frac{1}{2}(x-b)^T B^{-1}(x-b)}$$

$$\mathcal{N}(x|c, C) = \frac{1}{\sqrt{(2\pi)^D \det(C)}} e^{-\frac{1}{2}(x-c)^T C^{-1}(x-c)}$$

$$\mathcal{I}^{-1} = \frac{1}{\sqrt{(2\pi)^D \det(A+B)}} e^{-\frac{1}{2}(a-b)^T (A+B)^{-1}(a-b)}$$

$\mathcal{I}^{-1} \mathcal{N}(x|c, C)$ $\mathcal{N}(x|a, A)\mathcal{N}(x|b, B)$ and

Since the $(2\pi)^D$ terms are canceled out whenever we multiply

It's sufficient to show two properties: $\det(A)\det(B) = \det(C) \cdot \det(A+B)$

and $\frac{1}{2}(x-a)^T A^{-1}(x-a) + \frac{1}{2}(x-b)^T B^{-1}(x-b) = \frac{1}{2}(a-b)^T (A+B)^{-1}(a-b)$

$$+ \frac{1}{2}(x-c)^T C^{-1}(x-c)$$

Claim: $\det(A)\det(B) = \det(C)\det(A+B)$

Proof: $C = (A^{-1}+B^{-1})^{-1} \Rightarrow C^{-1} = A^{-1}+B^{-1}$ where we know that $\det(C^{-1}) = \frac{1}{\det(C)} = \det(A^{-1}+B^{-1})$. $\boxed{\det(C) = \frac{1}{\det(A^{-1}+B^{-1})}}$

$AB(A^{-1}+B^{-1}) = ABA^{-1}+A$, It's enough to show $\det(A)\det(B) = \frac{1}{\det(A^{-1}+B^{-1})}\det(A+B)$ or $\det(A+B) = \det(A^{-1}+B^{-1})\det(A)\det(B)$

$A(A^{-1}+B^{-1})B = AA^{-1}B + AB^{-1}B = B+A$, meaning that of determinants $\det(A(A^{-1}+B^{-1})B) = \det(A+B)$. From the multiplication rule we get $\det(A+B) = \det(A)\det(A^{-1}+B^{-1})\det(B)$ ✓ ■ as desired

Claim: $-\frac{1}{2}(x-a)^T A^{-1}(x-a) + \frac{-1}{2}(x-b)^T B^{-1}(x-b) = \frac{-1}{2}(a-b)^T(A+B)^{-1}\cdot(a-b)$

$+ -\frac{1}{2}(x-c)^T C^{-1}(x-c)$

Proof: It's enough to show $(x-a)^T A^{-1}(x-a) + (x-b)^T B^{-1}(x-b) = (a-b)^T(A+B)^{-1}(a-b) + (x-c)^T C^{-1}(x-c)$

$(x^T-a^T)A^{-1}(x-a) = x^TA^{-1}x - a^TA^{-1}x - x^TA^{-1}a + a^TA^{-1}a$

$(a-b)^T(A+B)^{-1}(a-b) = (a^T-b^T)(A+B)^{-1}(a-b) = a^T(A+B)^{-1}a - b^T(A+B)^{-1}a - a^T(A+B)^{-1}b + b^T(A+B)^{-1}b$. Plugging those values yield ~~$x^TA^{-1}x$~~ $- a^TA^{-1}x - x^TA^{-1}a + a^TA^{-1}a +$ $+$ ~~$x^TB^{-1}x$~~ $- b^TB^{-1}x - x^TB^{-1}b + b^TB^{-1}b =$

$\stackrel{?}{=} a^T(A+B)^{-1}a - c^T(A+B)^{-1}a - a^T(A+B)^{-1}c + c^T(A+B)^{-1}c +$

$+ x^TC^{-1}x - c^TC^{-1}x - x^TC^{-1}c + c^TC^{-1}c =$

$= a^T(A+B)^{-1}a - c^T(A+B)^{-1}a - a^T(A+B)^{-1}c + c^T(A+B)^{-1}c +$

$+ x^TA^{-1}x + x^TB^{-1}x - (A^{-1}a + B^{-1}c)^TC^TC^{-1}x -$

$- x^TC^{-1}C(A^{-1}a + B^{-1}c) + (A^{-1}a + B^{-1}c)^TC^TC^{-1}c =$

$= a^T(A+B)^{-1}a - c^T(A+B)^{-1}a - a^T(A+B)^{-1}c + c^T(A+B)^{-1}c +$

$+ \cancel{x^TA^{-1}x} + \cancel{x^TB^{-1}x} - (A^{-1}a + B^{-1}c)^Tx - x^T(A^{-1}a + B^{-1}c) +$

$+ (A^{-1}a + B^{-1}c)^TC(A^{-1}a + B^{-1}c)$ It's enough to show

$- a^TA^{-1}x - \cancel{x^TA^{-1}a} + a^TA^{-1}a - c^TB^{-1}x - \cancel{x^TB^{-1}c} + c^TB^{-1}c$

$\stackrel{?}{=} a^T(A+B)^{-1}a - c^T(A+B)^{-1}a - a^T(A+B)^{-1}c + c^T(A+B)^{-1}c$

$- (A^{-1}a + B^{-1}c)^Tx - x^T\cancel{(A^{-1}a + B^{-1}c)} + (A^{-1}a + B^{-1}c)^TC \cdot$

$\cdot (A^{-1}a + B^{-1}c)$

$(A^{-1}a + B^{-1}c)^T = (A^{-1}a)^T + (B^{-1}c)^T = a^T(A^{-1})^T + c^T(B^{-1})^T$

Since $(A^{-1})^T = (A^T)^{-1} = A^{-1}$ as A, B - symmetric $\Rightarrow$

$\cancel{-a^TA^{-1}x} + a^TA^{-1}a - \cancel{c^TB^{-1}x} + c^TB^{-1}c \stackrel{?}{=} a^T(A+B)^{-1}a - c^T(A+B)^{-1}a$

$- a^T(A+B)^{-1}c + c^T(A+B)^{-1}c - \cancel{a^TA^{-1}x} - \cancel{c^TB^{-1}x} +$

$+ (A^{-1}a + B^{-1}c)^T\left(A^{-1} + B^{-1}\right)^{-1}(A^{-1}a + B^{-1}c)$, so it's left as follows

$a^TA^{-1}a + b^TB^{-1}b \stackrel{?}{=} a^T(A+B)^{-1}a - b^T(A+B)^{-1}a - a^T(A+B)^{-1}b +$

$+ b^T(A+B)^{-1}b + (a^TA^{-1} + b^TB^{-1})(A^{-1}+B^{-1})^{-1}(A^{-1}a + B^{-1}b)$

Since we proved that $A(A^{-1}+B^{-1})B = B+A$ and in the similar case, $B(B^{-1}+A^{-1})A = A+B \Rightarrow$ taking inverses of both sides guarantees that

$(A+B)^{-1} = B^{-1}(A^{-1}+B^{-1})^{-1}A^{-1} = A^{-1}(A^{-1}+B^{-1})^{-1}B^{-1}$

Computing the straight line $\rightsquigarrow a^TA^{-1}(A^{-1}+B^{-1})^{-1}A^{-1}a +$

$+ b^TB^{-1}(A^{-1}+B^{-1})^{-1}A^{-1}a + a^TA^{-1}(A^{-1}+B^{-1})^{-1}B^{-1}b +$

$+ b^TB^{-1}(A^{-1}+B^{-1})^{-1}B^{-1}b = a^TA^{-1}(A^{-1}+B^{-1})^{-1}A^{-1}a +$

$+ b^TB^{-1}(A^{-1}+B^{-1})^{-1}B^{-1}b + b^T(A+B)^{-1}a + a^T(A+B)^{-1}b$

$a^TA^{-1}a + b^TB^{-1}b \stackrel{?}{=} a^T(A+B)^{-1}a - b^T(A+B)^{-1}a - a^T(A+B)^{-1}b +$

$+ b^T(A+B)^{-1}b + b^T(A+B)^{-1}a + a^T(A+B)^{-1}b +$

$+ a^TA^{-1}(A^{-1}+B^{-1})^{-1}A^{-1}a + b^TB^{-1}(A^{-1}+B^{-1})^{-1}B^{-1}b$

$a^TA^{-1}a + b^TB^{-1}b \Rightarrow a^TA^{-1}a + b^TB^{-1}b \stackrel{?}{=} a^T(A+B)^{-1}a + b^T(A+B)^{-1}b +$

$+ a^TA^{-1}(A^{-1}+B^{-1})^{-1}A^{-1}a + b^TB^{-1}(A^{-1}+B^{-1})^{-1}B^{-1}b$

$(CD)^{-1} = D^{-1}C^{-1} \Rightarrow (A^{-1}+B^{-1})^{-1}A^{-1} = (A(A^{-1}+B^{-1}))^{-1} = (I+AB^{-1})^{-1}$

Since $(A+B)^{-1} = B^{-1}(A^{-1}+B^{-1})^{-1}A^{-1} \Rightarrow B(A+B)^{-1}A =$

$= (A^{-1}+B^{-1})^{-1}$, where $(A^{-1}+B^{-1})^{-1}A^{-1} = B(A+B)^{-1}$ and similarly

$(A^{-1}+B^{-1})^{-1}B^{-1}$ can be computed by $\rightsquigarrow (A+B)^{-1} = A^{-1}(A^{-1}+B^{-1})^{-1}B^{-1}$

$A(A+B)^{-1}B = (A^{-1}+B^{-1})^{-1}$ with $\underline{(A^{-1}+B^{-1})^{-1}B^{-1} = A(A+B)^{-1}}$

$a^TA^{-1}a + b^TB^{-1}b \overset{?}{=} a^T(A+B)^{-1}a + b^T(A+B)^{-1}b +$ second expression is

$+ a^TA^{-1}B(A+B)^{-1}a + b^TB^{-1}A(A+B)^{-1}b$ where the

$\left[a^T(A+B)^{-1}a + a^TA^{-1}B(A+B)^{-1}a\right] + \left[b^T(A+B)^{-1}b + b^TB^{-1}A(A+B)^{-1}b\right]$

$= (a^T + a^TA^{-1}B)(A+B)^{-1}a + (b^T + b^TB^{-1}A)(A+B)^{-1}b =$

$= a^T(I + A^{-1}B)(A+B)^{-1}a + b^T(I + B^{-1}A)(A+B)^{-1}b.$ So

$a^TA^{-1}a + b^TB^{-1}b \overset{?}{=} a^T(I+A^{-1}B)(A+B)^{-1}a + b^T(I+B^{-1}A)(A+B)^{-1}b$

Claim: $A^{-1} = (I + A^{-1}B)(A+B)^{-1}$ and $B^{-1} = (I+B^{-1}A)(A+B)^{-1}$

Proof: $A\left((I + A^{-1}B)(A+B)^{-1}\right) = (A+B)(A+B)^{-1} = I$, therefore

$\boxed{A^{-1} = (I + A^{-1}B)(A+B)^{-1}}$ Similarly, $B\left((I+B^{-1}A)(A+B)^{-1}\right) =$

$= \left(B(I+B^{-1}A)\right)(A+B)^{-1} = (B+A)(B+A)^{-1} = I \Rightarrow$ Therefore,

$\boxed{B^{-1} = (I+B^{-1}A)(A+B)^{-1}}$ ✓ Following the claim results, it's easily seen that final equality indeed holds. Therefore, the topmost claim would also satisfy.

In conclusion, we obtain that $\boxed{\mathcal{N}(x|a,A)\,\mathcal{N}(x|b,B) = z^{-1}\mathcal{N}(x|c,C)}$ ✱ holds correctly ☑ • ⊤

4) a) In order to solve the problem, we'll prove the general result:

Claim: Let $p(x) = \mathcal{N}(x|\mu, \Lambda)$ and $p(y|x) = \mathcal{N}(y|Ax+b, L)$ then $p(y) = \mathcal{N}(y|A\mu+b, L+A\Lambda A^T)$ and $p(x|y) = \mathcal{N}(x|\Sigma\{A^TL^{-1}(y-b)+\Lambda^{-1}\mu\}, \Sigma)$ where Σ is defined as $\Sigma = (\Lambda^{-1}+A^TL^{-1}A)^{-1}$

Proof: $y|x \sim \mathcal{N}(Ax+b, L)$ and $x \sim \mathcal{N}(\mu, \Lambda)$

$$y^TL^{-1}y - 2y^TL^{-1}(Ax+b) + (Ax+b)^TL^{-1}(Ax+b) + x^T\Lambda^{-1}x - 2x^T\Lambda^{-1}\mu + \mu^T\Lambda^{-1}\mu =$$

$$= y^TL^{-1}y - 2y^TL^{-1}Ax - 2y^TL^{-1}b + x^TA^TL^{-1}Ax + 2x^TA^TL^{-1}b + b^TL^{-1}b + x^T\Lambda^{-1}x - 2x^T\Lambda^{-1}\mu + \mu^T\Lambda^{-1}\mu$$

The key idea is to make quadratic forms out of the x terms to integrate them out: By grouping with straight line terms and squirly terms, with the fact that $(y^TL^{-1}Ax)^T = x^TA^TL^{-1}y$ due to the $y^TL^{-1}Ax$ being scalar and L-symmetric $\Rightarrow L^{-1}$-symmetric

$$x^T(\Lambda^{-1}+A^TL^{-1}A)x - 2x^T(\Lambda^{-1}\mu + A^TL^{-1}y - A^TL^{-1}b) + \mu^T\Lambda^{-1}\mu + b^TL^{-1}b$$

Let $V = [A^TL^{-1}A + \Lambda^{-1}]^{-1}$

$u = [\Lambda^{-1}\mu + A^TL^{-1}y - A^TL^{-1}b]$

$h = Vu$

$\Big\rangle$ $h^TV^{-1}h = u^TVu$

$C = \mu^T\Lambda^{-1}\mu + b^TL^{-1}b$

x is Gaussian and integrates to its normalizing constant. The term $c-K$, $K=u^T V u$ factors into the y exponential giving, $y^T L^{-1} y - 2y^T L^{-1}\beta + c - K$ and

$$u^T V u = \left[\Lambda^{-1}\mu + A^T L^{-1} y - A^T L^{-1}\beta\right]^T V \left[\Lambda^{-1}\mu + A^T L^{-1} y - A^T L^{-1}\beta\right]$$

$$= \mu^T \Lambda^{-1} V \Lambda^{-1}\mu + y^T L^{-1} A V A^T L^{-1} y + \beta^T L^{-1} A V A^T L^{-1}\beta$$
$$+ 2y^T L^{-1} A V \Lambda^{-1}\mu - 2\mu^T \Lambda^{-1} V A^T L^{-1}\beta - 2y^T L^{-1} A V A^T L^{-1}\beta$$

In together, this becomes the following:

$$y^T\left[L^{-1} - L^{-1}AVA^T L^{-1}\right]y - 2y^T\left[L^{-1}\beta + L^{-1}AV\Lambda^{-1}\mu - L^{-1}AVA^T L^{-1}\beta\right] + \mu^T\left[\Lambda^{-1} - \Lambda^{-1}V\Lambda^{-1}\right]\mu + \beta^T\left[L^{-1} - L^{-1}AVA^T L^{-1}\right]\beta - 2\beta^T L^{-1}AV\Lambda^{-1}\mu$$

which simplifies in the following style:

$$y^T R^{-1} y - 2y^T g + \beta^T R^{-1}\beta + \mu^T\left[\Lambda^{-1} - \Lambda^{-1}V\Lambda^{-1}\right]\mu - 2\beta^T L^{-1}AV\Lambda^{-1}\mu$$

where we substituted

$$R^{-1} = L^{-1} - L^{-1}AVA^T L^{-1} = \left[L + A\Lambda A^T\right]^{-1}$$
$$g = L^{-1}\beta + L^{-1}AV\Lambda^{-1}\mu - L^{-1}AVA^T L^{-1}\beta$$

Moreover, $Rg = A\mu + b$:

$$Rg = R\left[L^{-1}b + L^{-1}AV\Lambda^{-1}\mu - L^{-1}AVA^TL^{-1}b\right] =$$

$$= R\left[\left(L^{-1} - L^{-1}AVA^TL^{-1}\right)b + L^{-1}AV\Lambda^{-1}\mu\right] =$$

$$= R\left(L^{-1} - L^{-1}AVA^TL^{-1}\right)b + RL^{-1}AV\Lambda^{-1}\mu = b + RL^{-1}AV\Lambda^{-1}\mu$$

Since $R^{-1} = L^{-1} - L^{-1}AVA^TL^{-1} \Rightarrow Rg = RL^{-1}AV\Lambda^{-1}\mu + b$

After canceling out the b term on both sides, it's sufficient to solve for the other term, equating the desired coefficients:

$$RL^{-1}AV\Lambda^{-1}\mu = \left[L + A\Lambda A^T\right]L^{-1}AV\Lambda^{-1}\mu = A\mu$$

$$= A + A\Lambda A^TL^{-1}A = A\left[\Lambda^{-1} + A^TL^{-1}A\right]\Lambda$$

(using the definition of $V \rightsquigarrow \left[A^TL^{-1}A + \Lambda^{-1}\right]^{-1}$)

$$I + \Lambda A^TL^{-1}A = I + A^TL^{-1}A\Lambda$$

$$\Lambda A^TL^{-1}A = \Lambda A^TL^{-1}A \quad \checkmark$$

And therefore, $Rg = A\mu + b$ with $R^{-1}A\mu = g - R^{-1}b = L^{-1}b + L^{-1}AV\Lambda^{-1}\mu - L^{-1}AVA^TL^{-1}b$

$$-\left(L^{-1} - L^{-1}AVA^TL^{-1}\right)b = L^{-1}b + L^{-1}AV\Lambda^{-1}\mu - L^{-1}AVA^TL^{-1}b - L^{-1}b$$

$$+ L^{-1}AVA^TL^{-1}b = L^{-1}AV\Lambda^{-1}\mu$$

Now, simplifying the previous equation, we get the following equation for y:

$$y^T R^{-1} y - 2y^T R^{-1}(A\mu + b) + b^T R^{-1} b + \mu^T \left[\Lambda^{-1} - \Lambda^{-1} V \Lambda^{-1}\right]\mu$$

$$-2b^T L^{-1} A V \Lambda^{-1}\mu$$

Taking into account the identity proved in problem 3, it's easily seen that

$$\Lambda^{-1} - \Lambda^{-1} I \left(F^{-1} + I^T \Lambda^{-1} I\right)^{-1} I^T \Lambda^{-1} = \left(\Lambda + I F I^T\right)^{-1}$$

So, $\Lambda^{-1} - \Lambda^{-1} I \left(F^{-1} + I \Lambda^{-1} I\right)^{-1} I \Lambda^{-1} = \left(\Lambda + F\right)^{-1}$

Substitute $F^{-1} = A^T L^{-1} A \Rightarrow V = \left[F^{-1} + \Lambda^{-1}\right]^{-1}$ and $\left(\Lambda + F\right)^{-1} = \Lambda^{-1} - \Lambda^{-1} V \Lambda^{-1}$ where $R^{-1} = L^{-1} - L^{-1} A V A^T L^{-1}$

$A^T R^{-1} A = A^T \left(L^{-1} - L^{-1} A V A^T L^{-1}\right) A$ and $V = \left[A^T L^{-1} A + \Lambda^{-1}\right]^{-1} \Rightarrow$ Our claim is to prove that

$A^T R^{-1} A = \Lambda^{-1} - \Lambda^{-1} V \Lambda^{-1} \Rightarrow A^T \left(L^{-1} - L^{-1} A V A^T L^{-1}\right) A =$

$= A^T L^{-1} A - A^T L^{-1} A V A^T L^{-1} A = \Lambda^{-1} - \Lambda^{-1} \left(A^T L^{-1} A + \Lambda^{-1}\right)^{-1}$

$A^T L^{-1} A - A^T L^{-1} A \left[A^T L^{-1} A + \Lambda^{-1}\right]^{-1} A^T L^{-1} A = \Lambda^{-1} -$

$-\Lambda^{-1} \left(A^T L^{-1} A + \Lambda^{-1}\right)^{-1} \Lambda^{-1}$

$\left(A^T L^{-1} A + \Lambda^{-1}\right)\left(A^T L^{-1} A - A^T L^{-1} A \left[A^T L^{-1} A + \Lambda^{-1}\right]^{-1}\right)$

$F^{-1} = A^T L^{-1} A \Rightarrow F^{-1} - F^{-1} \left(F^{-1} + \Lambda^{-1}\right)^{-1} F^{-1} = \Lambda^{-1} - \Lambda^{-1} \left(F^{-1} + \Lambda^{-1}\right)^{-1} \Lambda^{-1}$

Claim: $A^TR^{-1}A = \Lambda^{-1} - \Lambda^{-1}V\Lambda^{-1}$

Since $F^{-1} = A^TL^{-1}A$ and $V = \left[A^TL^{-1}A + \Lambda^{-1}\right]^{-1} = \left[F^{-1} + \Lambda^{-1}\right]^{-1}$

$A^TR^{-1}A = A^T\left[L^{-1} - L^{-1}AVA^TL^{-1}\right]A =$

$= A^T\left[L^{-1} - L^{-1}A\left[F^{-1} + \Lambda^{-1}\right]^{-1}A^TL^{-1}\right]A = A^TL^{-1}A -$

$- A^TL^{-1}A\left[F^{-1} + \Lambda^{-1}\right]^{-1}A^TL^{-1}A = F^{-1} - F^{-1}\left[F^{-1} + \Lambda^{-1}\right]^{-1}F^{-1}$

and $\Lambda^{-1} - \Lambda^{-1}\left[F^{-1} + \Lambda^{-1}\right]^{-1}\Lambda^{-1}$ is the second term $\Rightarrow$

Since we proved that $(\Lambda + F)^{-1} = \Lambda^{-1} - \Lambda^{-1}\left(F^{-1} + \Lambda^{-1}\right)^{-1}\Lambda^{-1}$ (inspired from problem 3) $\Rightarrow$ it's enough to show $F^{-1} - F^{-1}\left[F^{-1} + \Lambda^{-1}\right]^{-1}F^{-1}$ is equal to $(\Lambda + F)^{-1} \Rightarrow (\Lambda + F)\left(F^{-1} - F^{-1}\left[F^{-1} + \Lambda^{-1}\right]^{-1}F^{-1}\right) =$

$= \Lambda F^{-1} + I - \Lambda F^{-1}\left[F^{-1} + \Lambda^{-1}\right]^{-1}F^{-1} - \left[F^{-1} + \Lambda^{-1}\right]^{-1}F^{-1} \stackrel{?}{=} I$

$\left[F^{-1} + \Lambda^{-1}\right]^{+1}\left[\Lambda F^{-1} - \Lambda F^{-1}\left[F^{-1} + \Lambda^{-1}\right]^{-1}F^{-1} - \left[F^{-1} + \Lambda^{-1}\right]^{-1}F^{-1}\right]$

$\left(\Lambda - \Lambda F^{-1}\left[F^{-1} + \Lambda^{-1}\right]^{-1} - \left[F^{-1} + \Lambda^{-1}\right]^{-1}\right)F^{-1} \stackrel{?}{=} 0$

$\Lambda - \Lambda F^{-1}\left[F^{-1} + \Lambda^{-1}\right]^{-1} - \left[F^{-1} + \Lambda^{-1}\right]^{-1} \stackrel{?}{=} 0$

$\left(\Lambda - \Lambda F^{-1}\left[F^{-1} + \Lambda^{-1}\right]^{-1} - \left[F^{-1} + \Lambda^{-1}\right]^{-1}\right)\left(F^{-1} + \Lambda^{-1}\right) \stackrel{?}{=} 0$

$\Lambda F^{-1} + I - \Lambda F^{-1} - I = 0$ which is true ✓ In conclusion, claim is true $\Rightarrow$ $A^TR^{-1}A = \Lambda^{-1} - \Lambda^{-1}V\Lambda^{-1}$ ✓

Since we proved $A^TR^{-1}A = \Delta^{-1} - \Delta^{-1}V\Delta^{-1}$ where in the other claim we also had $(\Delta + F)^{-1} = \Delta^{-1} - \Delta^{-1}V\Delta^{-1}$ (using problem 8 result) we conclude $\boxed{A^TR^{-1}A = (\Delta + F)^{-1}}$ $\rightsquigarrow$ Taking inverse, it yields

$A^{-1}R(A^T)^{-1} = \Delta + F$, since $R = L + A\Delta A^T \rightsquigarrow A\Delta A^T = R - L$ where $F^{-1} = A^TL^{-1}A$, $A\left(A^{-1}R(A^T)^{-1}\right)A^T =$

$= A(\Delta + F)A^T = (AA^{-1})R(A^T)^{-1}A^T = R\left[(A^T)^{-1}A^T\right] =$

$= R = A\Delta A^T + AFA^T$, with $F^{-1} = A^TL^{-1}A$ indicates

$F = A^{-1}L(A^T)^{-1}$ and $AFA^T = AA^{-1}L(A^T)^{-1}A^T =$

$= (AA^{-1})L(A^T)^{-1}A^T = L\left[(A^T)^{-1}A^T\right] = L \rightsquigarrow \underline{AFA^T = L}$ so

$\boxed{R = A\Delta A^T + L}$ So, the key expressions in the red star ★ turn to be $\rightsquigarrow \underline{\Delta^{-1} - \Delta^{-1}V\Delta^{-1} = A^TR^{-1}A}$ and

$R^{-1}A = \left(L^{-1} - L^{-1}AVA^TL^{-1}\right)A = L^{-1}A - L^{-1}AVA^TL^{-1}A =$

$= L^{-1}A - L^{-1}A\left[A^TL^{-1}A + \Delta^{-1}\right]^{-1}A^TL^{-1}A \overset{?}{=} L^{-1}AV\Delta^{-1}$

It's suppicient to show $I - \left[A^TL^{-1}A + \Delta^{-1}\right]^{-1}A^TL^{-1}A \overset{?}{=} V\Delta^{-1}$

$I - V\,A^TL^{-1}A \overset{?}{=} V\Delta^{-1} \rightsquigarrow I \overset{?}{=} V\left(A^TL^{-1}A + \Delta^{-1}\right)$ which is obviously true ✓

In the end, we receive $R^{-1}A = L^{-1}AV\Lambda^{-1}$ Now, coming back into the equation red star

$$y^T R^{-1} y - 2y^T R^{-1}(A\mu + b) + b^T R^{-1} b + \mu^T \left[\Lambda^{-1} - \Lambda^{-1} V \Lambda^{-1}\right]\mu$$

$$-2b^T L^{-1} A V \Lambda^{-1}\mu, \text{ if we substitute the new results, we obtain}$$

$$y^T R^{-1} y - 2y^T R^{-1}(A\mu + b) + b^T R^{-1} b + \mu^T A^T R^{-1} A\mu = 2b^T R^{-1} A\mu$$

which gives the result that $\boxed{y \text{ is distributed normally}} \rightsquigarrow$

$$\boxed{y \sim \mathcal{N}(A\mu + b, R)} \text{ ★ Considering } R = A\Lambda A^T + L \Rightarrow$$

$$\boxed{y \sim \mathcal{N}(A\mu + b, A\Lambda A^T + L)} \text{ ★}$$

The remaining part will be inspired from those results, in which it turns to be easier

Initially, in the ★★ equation at the begining of solution problem 4, that long equation given for y can be simplified when conditioning on X since everything not involving X can be subsumed into the constant of proportionality. This will leave only

$$-2y^T L^{-1} A X + X^T A^T L^{-1} A X + 2X^T A^T L^{-1} b + X^T \Lambda^{-1} X - 2X^T \Lambda^{-1}\mu$$

$$= -2X^T A^T L^{-1} y + X^T A^T L^{-1} A X + 2X^T A^T L^{-1} b + X^T \Lambda^{-1} X - 2X^T \Lambda^{-1}\mu$$

After this, refactoring this term will yield

$$X^T \left[A^T L^{-1} A + \Lambda^{-1}\right] X - 2X^T \left[A^T L^{-1}(y - b) + \Lambda^{-1}\mu\right] + K$$

where K-constant (specifically, $y^T L^{-1} y - 2y^T L^{-1} b + b^T L^{-1} b + \mu^T \Lambda^{-1}\mu$)

If we define $A^T L^{-1} A + \Lambda^{-1} = \Sigma^{-1}$, then it's easily seen that

$$x|y \sim \mathcal{N}\left(\Sigma\left[A^T L^{-1}(y-b) + \Lambda^{-1}\mu\right], \Sigma\right)$$

Finally, we proved that $p(y) = \mathcal{N}(y|A\mu+b, L+A\Lambda A^T)$

and $p(x|y) = \mathcal{N}\left(x|\Sigma\left[A^T L^{-1}(y-b)+\Lambda^{-1}\mu\right], \Sigma\right)$

where Σ is defined $\Sigma = (\Lambda^{-1} + A^T L^{-1} A)^{-1}$

Sidenote: As an easily understandable logic for proving $p(y)$, we can mention the following ideas as well: (Idea is to just present bigger picture what happens)
Let's write the r.v.'s X, y as $X = \mu + \varepsilon_x$ and $y = AX + b + \varepsilon_y$ where $\varepsilon_x \sim \mathcal{N}(0, \Lambda)$ and $\varepsilon_y \sim \mathcal{N}(0, L)$. Plugging in X in the second equation gives us $y = A\mu + b + A\varepsilon_x + \varepsilon_y$. This's linear combination of normal distributed random variables and as such itself normal distributed with expectation $A\mu+b$ and covariance matrix $A\Lambda A^T + L$ since $\text{var}(AX) = A\,\text{var}(X)A^T$. From this, $p(y) = \mathcal{N}(y|A\mu+b, A\Lambda A^T + L)$ (precise calculations were indicated in the main solution, this's just main logic).
For the 2nd part, $p(x|y) \propto p(y|x)p(x)$ from Bayes, therefore we see $\propto \exp\left((y-Ax-b)^T L^{-1}(y-Ax-b) + (x-\mu)^T \Lambda^{-1}(x-\mu)\right)$. Using rigorous calculations and factoring out x, we obtain proportional to $\exp\left(\left(x-(\Lambda^{-1}+A^T L^{-1}A)^{-1}A^T L^{-1}(y-b)\right)^T (\Lambda^{-1}+A^T L^{-1}A)\cdot\right.$
$\left.\cdot\left(x-(\Lambda^{-1}+A^T L^{-1}A)^{-1}AL^{-1}(y-b)\right)\right)$ which's proportional to given normal

Combining the logic presented in sidenote as a big picture and how detailed rigorous proofs were presented before that, we successfully proved that

$p(x) = \mathcal{N}(x|\mu, \Lambda)$ and $p(y|x) = \mathcal{N}(y|Ax+b, L) \Rightarrow$

$p(y) = \mathcal{N}(y|A\mu+b, L + A\Lambda A^T)$ and

$p(x|y) = \mathcal{N}(x|\Sigma[A^T L^{-1}(y-b) + \Lambda^{-1}\mu], \Sigma)$ where

$\Sigma = (\Lambda^{-1} + A^T L^{-1} A)^{-1}$

Plugging those variables into original problem where it's explicitly said

$p(z) = \mathcal{N}(z|0, \underline{I}) \Rightarrow \mu = 0, \Lambda = \underline{I}$

$p(x|z) = \mathcal{N}(x|Wz + \mu, \sigma^2 \underline{I}) \Rightarrow A = W, b = \mu, L = \sigma^2 \underline{I}$

It's evident that $L^{-1} = \frac{1}{\sigma^2}\underline{I}$ and $\Lambda^{-1} = \underline{I} \Rightarrow$

$p(x) = \mathcal{N}(x|W \cdot 0 + \mu, \sigma^2 \underline{I} + W \underline{I} W^T) = \mathcal{N}(x|\mu, \sigma^2 \underline{I} + WW^T)$

$\boxed{p(x) = \mathcal{N}(x|\mu, \sigma^2 \underline{I} + WW^T)}$ ~ answer Similarly, we've.

$p(z|x) = \mathcal{N}(z|\Sigma[W^T \cdot \frac{1}{\sigma^2}\underline{I}(y-\mu) + \underline{I} \cdot 0], \Sigma)$

where $\Sigma = (\underline{I} + W^T \cdot \frac{1}{\sigma^2}\underline{I} W)^{-1} = (\underline{I} + \frac{1}{\sigma^2} W^T W)^{-1}$

$\boxed{p(z|x) = \mathcal{N}(z|\Sigma W^T \frac{1}{\sigma^2}(x-\mu), \Sigma)}$ ~ answer.

$\Sigma = (\underline{I} + \frac{1}{\sigma^2} W^T W)^{-1}$

V.

6) Using the first part from problem 4, we successfully proved that $p(x) = \mathcal{N}(x|\mu, \sigma^2 I + WW^T)$ Let us denote $C = \sigma^2 I + WW^T$ where we consider the determination of the model parameters using maximum likelihood. Given a data set $\mathcal{D} = \{x_i \in \mathbb{R}^D : 1 \le i \le N\}$ corresponding log likelihood function is given by $\ln p(X|\mu, W, \sigma^2) = \sum_{n=1}^{N} \ln p(x_n|W, \mu, \sigma^2)$ where $X = \{x_n\}$ clear that

So, using the formula for multivariate Gaussian distribution, it is

$$p(x_i|W, \mu, \sigma^2) = \frac{1}{(2\pi)^{D/2}} \frac{1}{\sqrt{\det(\text{covariance matrix})}} \exp\left\{-\frac{1}{2}(x-\mu)^T \cdot (\text{covariance matrix})^{-1}(x-\mu)\right\}$$

Since we had $p(x) = \mathcal{N}(x|\mu, C)$ where $C \rightsquigarrow$ covariance matrix $= \sigma^2 I + WW^T$

$$p(x_i|W, \mu, \sigma^2) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|C|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)\right\}$$

$$\Rightarrow \ln p(X|\mu, W, \sigma^2) = \sum_{i=1}^{N}\left[-\frac{D}{2}\ln(2\pi) + -\frac{1}{2}\ln|C| - \frac{1}{2}(x_i-\mu)^T C^{-1}(x-\mu)\right]$$

$$= -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|C| - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T C^{-1}(x_i-\mu)$$

Setting the derivative of the log likelihood wrt μ equal to zero gives the expected result μ: $\rightsquigarrow$ maximum likelihood estimator for μ

Our goal is to evaluate $\frac{\partial}{\partial \mu} \ln p(x|\mu, W, \sigma^2)$ in which the terms $-\frac{ND}{2}\ln(2\pi)$ and $-\frac{N}{2}\ln|C|$ don't play any role $\Rightarrow$ We just need to compute $\frac{\partial}{\partial \mu}\left(-\frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T C^{-1}(x_i-\mu)\right)$

Assume $x_i - \mu = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iD} \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_D \end{bmatrix} = \begin{bmatrix} x_{i1}-\mu_1 \\ x_{i2}-\mu_2 \\ \vdots \\ x_{iD}-\mu_D \end{bmatrix}$, $C^{-1} = \begin{bmatrix} c_{11} \dots c_{1D} \\ \vdots \\ c_{D1} \dots c_{DD} \end{bmatrix}$

$$(x_i-\mu)^T C^{-1}(x_i-\mu) = \begin{bmatrix} x_{i1}-\mu_1 & x_{i2}-\mu_2 \dots & x_{iD}-\mu_D \end{bmatrix} \begin{bmatrix} c_{11} \dots c_{1D} \\ \vdots \\ c_{D1} \dots c_{DD} \end{bmatrix} \begin{bmatrix} x_{i1}-\mu_1 \\ x_{i2}-\mu_2 \\ \vdots \\ x_{iD}-\mu_D \end{bmatrix}$$

$$= \sum_{t=1}^{D} \sum_{k=1}^{D} (x_{it}-\mu_t)(x_{ik}-\mu_k) c_{kt}$$

where $C = \sigma^2 I + WW^T$ is symmetric as it's covariance matrix

C^{-1}-symmetric

$C^T = \sigma^2 I + WW^T = C \Rightarrow \boxed{c_{kt} = c_{tk}}$ Taking derivative of $(x_i-\mu)^T C^{-1}(x_i-\mu)$ wrt each term $\mu_j \Rightarrow \frac{\partial}{\partial \mu_j}(x_i-\mu)^T C^{-1}(x_i-\mu) = \frac{\partial}{\partial \mu_j}(x_{ij}-\mu_j)^2 c_{jj}$

$$+ \frac{\partial}{\partial \mu_j}\left[\sum_{\substack{k=1 \\ k\neq j}}^{D} (x_{ij}-\mu_j)(x_{ik}-\mu_k) c_{kj} + \sum_{\substack{t=1 \\ t\neq j}}^{D} (x_{it}-\mu_t)(x_{ij}-\mu_j) c_{jt} \right]$$

$\boxed{c_{kj} = c_{jk}}$

$$= -2(x_{ij}-\mu_j) c_{jj} + \frac{\partial}{\partial \mu_j}\left[2 \sum_{\substack{k=1 \\ k\neq j}}^{D} (x_{ik}-\mu_k)(x_{ij}-\mu_j) c_{kj} \right] =$$

$$= -2(x_{ij}-\mu_j) c_{jj} + -2 \sum_{\substack{k=1 \\ k\neq j}}^{D} (x_{ik}-\mu_k)(c_{kj}) = -2 \sum_{k=1}^{D} (x_{ik}-\mu_k) c_{kj}$$

So, $\frac{\partial}{\partial \mu_j}(x_i-\mu)^T C^{-1}(x_i-\mu) = -2 \sum_{k=1}^{D} (x_{ik}-\mu_k) c_{jk}$ since C^{-1}-symmetric

Finally, if we consider $C^{-1}(x_i-\mu) = \begin{bmatrix} c_{11} \dots c_{1D} \\ \vdots \\ c_{D1} \dots c_{DD} \end{bmatrix} \begin{bmatrix} x_{i1}-\mu_1 \\ \vdots \\ x_{iD}-\mu_D \end{bmatrix}$ and thus,

$$C^{-1}(x_i-\mu)=\begin{bmatrix} C_{11}(x_{i1}-\mu_1)+\dots+C_{1D}(x_{iD}-\mu_D) \\ \vdots \\ C_{D1}(x_{i1}-\mu_1)+\dots+C_{DD}(x_{iD}-\mu_D) \end{bmatrix}$$

where the j^{th} entry is equivalent to $\sum_{k=1}^{D} C_{jk}(x_{ik}-\mu_k)$

which yields that $\frac{\partial}{\partial \mu_j}(x_i-\mu)^T C^{-1}(x_i-\mu) = -2\sum_{k=1}^{D} C_{jk}(x_{ik}-\mu_k)$

and $\left[C^{-1}(x_i-\mu)\right]_j = \sum_{k=1}^{D} C_{jk}(x_{ik}-\mu_k)$, therefore

$$\frac{\partial}{\partial \mu}\left(\frac{-1}{2}\sum_{i=1}^{N}(x_i-\mu)^T C^{-1}(x_i-\mu)\right) = \sum_{i=1}^{N} C^{-1}(x_i-\mu)$$ due to the fact that

$$\frac{\partial}{\partial \mu_j}\left(\frac{-1}{2}\sum_{i=1}^{N}(x_i-\mu)^T C^{-1}(x_i-\mu)\right) = \frac{-1}{2}\sum_{i=1}^{N}(-2)\sum_{k=1}^{D} C_{jk}(x_{ik}-\mu_k)$$

$= \sum_{i=1}^{N}\sum_{k=1}^{D} C_{jk}(x_{ik}-\mu_k)$ where j^{th} entry of $C^{-1}(x_i-\mu)$ is equal to $\sum_{k=1}^{D} C_{jk}(x_{ik}-\mu_k)$ $\rightsquigarrow$ derivative of log likelihood wrt μ is given by $\sum_{i=1}^{N} C^{-1}(x_i-\mu)$

Setting this derivative to zero $\Rightarrow C^{-1}\left(\sum_{i=1}^{N} x_i - n\mu\right)=0$ where $C = \sigma^2 I + VV^T$ with nonzero elements inside entries of matrix C^{-1}

Thus, $\mu = \frac{\sum_{i=1}^{N} x_i}{n} = \bar{X}$ $\rightsquigarrow$ soln for maximum likelihood estimate $\rightsquigarrow$ μ_{ML}

↳ mean of observed set of data points

If we differentiate $\sum_{i=1}^{N} C^{-1}(x_i-\mu)$ once again, we obtain $-\sum_{i=1}^{N} C^{-1} = -NC^{-1}$

A quadratic form $-y^T C^{-1} y$ has unique maximum at $y=0$ if C^{-1} is p.d.

C^{-1} is clearly p.d. since $C = \sigma^2 I + WW^T \Rightarrow y^T C y = y^T \sigma^2 I y + y^T WW^T y = \sigma^2 y^T y + (W^T y)^T W^T y > 0$ whenever $y \neq 0$

As we know inverse of p.d. matrix is also p.d. $\Rightarrow -y^T C^{-1} y < 0$ except for $y=0$. Therefore, maximum likelihood solution μ_{ML} indicates unique maximum

$\rightsquigarrow \boxed{\mu_{ML} = \bar{X}}$ (V).

c) As we showed in the previous problem, $\ln p(X|\mu, W, \sigma^2) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|C| - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T C^{-1}(x_i-\mu)$ where $C = \sigma^2 I + WW^T$, in order to find the maximum likelihood estimator of W, it's required to set the derivative of log likelihood wrt W as zero: $\frac{\partial}{\partial W}\ln p(X|\mu, W, \sigma^2) =$

$$= -\frac{N}{2}\frac{\partial}{\partial W}\log\det(C) - \frac{1}{2}\sum_{i=1}^{N}\frac{\partial}{\partial W}(x_i-\mu)^T C^{-1}(x_i-\mu)$$

where we use the fact $\frac{\partial}{\partial A}(y^T A^{-1} y) = -(A^{-1})^T y y^T (A^{-1})^T$ (from class notes)

$\frac{\partial}{\partial a_{ij}}\det(A) = \det(A)(A^{-1})_{ji}$ Note that $\frac{\partial C}{\partial W} = \frac{\partial}{\partial W}(WW^T)$

↳ Consequently, $\frac{\partial}{\partial A}\log\det(A) = (A^{-1})^T$ The key idea is to take derivatives wrt C, and then chain Rule

$\frac{\partial}{\partial W} \ln p(X|\mu, W, \sigma^2) = -\frac{N}{2} \frac{\partial}{\partial C} \log\det(C) \cdot \frac{\partial C}{\partial W} -$

$-\frac{1}{2}\sum_{i=1}^{N} \frac{\partial}{\partial C} \cdot (x_i-\mu)^T C^{-1}(x_i-\mu) \cdot \frac{\partial C}{\partial W}$ where C - covariance matrix (symmetric)

a symmetrix matrix.

Thus, $\frac{\partial}{\partial C} \log\det(C) = \frac{1}{\det(C)} \det(C)\, C^{-1} = C^{-1}$ where C^{-1} is

As mentioned, $\frac{\partial}{\partial A}\left(y^T A^{-1} y\right) = -(A^{-1})^T y y^T (A^{-1})^T$ where

$\frac{\partial}{\partial C}(x_i-\mu)^T C^{-1}(x_i-\mu) = -(C^{-1})^T (x_i-\mu)(x_i-\mu)^T (C^{-1})^T$

$(C^{-1})^T = (C^T)^{-1} = C^{-1} \Rightarrow \frac{\partial}{\partial W} \ln p(X|\mu, W, \sigma^2) = -\frac{N}{2} C^{-1} \frac{\partial C}{\partial W} =$

$-\frac{1}{2}\sum_{i=1}^{N} -C^{-1}(x_i-\mu)(x_i-\mu)^T C^{-1} \cdot \frac{\partial C}{\partial W}$, where $C = \sigma^2 I + WW^T$

It's easily seen that $\boxed{\frac{\partial C}{\partial W} = 2W}$ and if we set derivative into zero, we obtain

$0 = -\frac{N}{2} C^{-1} \cdot 2W + \frac{1}{2}\sum_{i=1}^{N} C^{-1}(x_i-\mu)(x_i-\mu)^T C^{-1} \cdot 2W$

$NC^{-1}W = \sum_{i=1}^{N} C^{-1}(x_i-\mu)(x_i-\mu)^T C^{-1} W$ where $C = \sigma^2 I + WW^T$

$\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T = NS$, $\sum_{i=1}^{N} C^{-1}(x_i-\mu)(x_i-\mu)^T = C^{-1}NS =$

$= NC^{-1}S$ where

$NC^{-1}W = NC^{-1}SC^{-1}W$, since N-scalar $N \geq 1 \Rightarrow$ $C^{-1}W = C^{-1}SC^{-1}W$

Considering C - covariance matrix (positive semi-definite matrix), it turns out to be symmetric and C^{-1} symmetric where $\det(C^{-1})$ is nonzero $\Rightarrow$ we can eliminate C^{-1} from $C^{-1}(W - \beta C^{-1} W) = 0$ and obtain $W = \beta C^{-1} W = \beta(\sigma^2 I + WW^T)^{-1} W$. In conclusion

$\beta(WW^T + \sigma^2 I)^{-1} W = W$ as desired ∎

2) It's important to define some crucial definitions and characteristics for Latent class analysis. We assume that data points are generated by different Gaussian distributions as $p(x_i) = \sum_{k=1}^{K} \pi_k \, p(x_i | \mu_k, \Sigma_k)$ where data points are $\{x_1, x_2, \dots, x_N\}$ and $\sum_{k=1}^{K} \pi_k = 1$ is satisfied. After that, we define a latent random variable $z_i = (z_{i1}, z_{i2}, \dots, z_{iK})$ with $z_{ik} \in \{0,1\}$, $\sum_{k=1}^{K} z_{ik} = 1$ and $p(z_{ik} = 1) = \pi_k$, with $1 \le k \le K$. In alternative manner, we obtain $p(z_i) = \prod_{k=1}^{K} \pi_k^{z_{ik}}$ ($\pi_1, \pi_2, \dots$, or π_K).

We also notice that $p(x_i | z_{ik} = 1) = p(x_i | \mu_k)$, and

$$p(x_i | z_i) = \prod_{k=1}^{K} p(x_i | \mu_k, \pi_k)^{z_{ik}}$$

$$p(x_i, z_i) = p(x_i | z_i) \, p(z_i) = \prod_{k=1}^{K} \pi_k^{z_{ik}} \, p(x_i | \mu_k, \pi_k)^{z_{ik}}$$

So, $p(x_i) = \sum_{z_i} p(x_i, z_i) = \sum_{k=1}^{K} \pi_k \, p(x_i | \mu_k)$ (independently generated)

Now, we consider the joint probability of all data points $X = \{x_1, x_2, \dots, x_N\}$, $Z = \{z_1, z_2, \dots, z_N\} \Rightarrow p(X, Z) = \prod_{i=1}^{N} p(x_i | z_i) p(z_i)$

$$= \prod_{i=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{ik}} \, p(x_i | \mu_k)^{z_{ik}}$$ where variables $z_i \rightsquigarrow$ are latent variables $\pi \rightsquigarrow$ assignment probabilities. It's also important to get $p(z_{ik} = 1 | x_i)$

Also we have to notice that $p(\mathcal{J}_{ik}=1|x_i) = \frac{p(\mathcal{J}_{ik}=1, x_i)}{p(x_i)} =$

$$= \frac{p(\mathcal{J}_{ik}=1, x_i)}{\sum_{\rho=1}^{K} p(x_i, \mathcal{J}_{i\rho}=1)} = \frac{p(x_i|\mathcal{J}_{ik}=1)\, p(\mathcal{J}_{ik}=1)}{\sum_{\rho=1}^{K} p(x_i|\mathcal{J}_{i\rho}=1)\, p(\mathcal{J}_{i\rho}=1)}$$

$$= \frac{\pi_k\, p(x_i|\mu_k)}{\sum_{\rho=1}^{K} \pi_\rho\, p(x_i|\mu_\rho)}$$

← using the current parameters, we compute the E-step ✓

Regarding the M-step, we have to update parameters μ_k and π_k

Considering $\sum_{k=1}^{K} \pi_k = 1$ and Lagrange multiplier λ, we formulate

$$\mathcal{L}(\Theta) := \log p(x|\Theta) + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right) = J(\Theta) + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)$$

$$\frac{\partial}{\partial \pi}\mathcal{L}(\Theta) = 0 \Rightarrow \sum_{i=1}^{N} \frac{p(x_i|\mu_k)}{\sum_{\rho=1}^{K} \pi_\rho\, p(x_i|\mu_\rho)} + \lambda = 0, \quad 1 \le k \le K$$

In the end, we get $\frac{1}{\pi_k} \sum_{i=1}^{N} \frac{\pi_k\, p(x_i|\mu_k)}{\sum_{\rho=1}^{K} \pi_\rho\, p(x_i|\mu_\rho)} + \lambda = 0$

Recall that $p(\mathcal{J}_{ik}=1|x_i) = \frac{\pi_k\, p(x_i|\mu_k)}{\sum_{\rho=1}^{K} \pi_\rho\, p(x_i|\mu_\rho)}$ holds true

Therefore, it follows that $\pi_k = \frac{-1}{\lambda} \sum_{i=1}^{N} p(z_{ik}=1 \mid x_i)$

Considering that $\sum_{k=1}^{K} \pi_k = 1$, λ satisfies the following relation

$$\lambda = -\sum_{k=1}^{K} \sum_{i=1}^{N} p(z_{ik}=1 \mid x_i) = -\sum_{i=1}^{N} \sum_{k=1}^{K} p(z_{ik}=1 \mid x_i) = -N$$

In conclusion, we obtain $\boxed{\pi_k = \frac{1}{N} \sum_{i=1}^{N} p(z_{ik}=1 \mid x_i)}$ updating the parameters for π_k during M-step ✓

For updating the μ_k parameters, we'll use the terminologies and techniques applied in problem 1 (defining responsibilities $\gamma(z_{nk})$, and similar approaches) ✓

```
Inertia for K-means clustering with all four features is 97.225
```