QUESTION 1

Assignment Number: 2
Student Name: Gurpreet Singh
Roll Number: 150259
Date: November 8, 2017

Part 1

We know that Gaussian-Gaussian has the conjugacy property, and hence, we can say that the marginal probability of $\mathbf{x}$ *i.e.* $\mathbb{P}[\mathbf{x}]$

Therefore, we can write the marginal probability of $\mathbf{x}$ as a gaussian distribution with mean $\boldsymbol{\mu}'$ and covariance matrix Λ.

$$\begin{aligned}
\mathbb{P}\left[\mathbf{x} \,\middle|\, \boldsymbol{\mu}, W, \sigma\right] &= \int_{\mathbf{z}} \mathbb{P}\left[\mathbf{x} \,\middle|\, \mathbf{z}, \boldsymbol{\mu}, W, \sigma\right] \mathbb{P}\left[\mathbf{z} \,\middle|\, 0, \mathbf{I}\right] d\mathbf{z} \\
&= \mathcal{N}\left(\mathbf{x} \,\middle|\, \boldsymbol{\mu}', \Lambda\right)
\end{aligned}$$

Also, we know that $\mathbf{x} = W\mathbf{z} + \boldsymbol{\mu} + \epsilon$, and from the properties of Gaussian, we know $\boldsymbol{\mu}' = \mathbb{E}[\mathbf{x}]$ and $\Lambda = cov(\mathbf{x})$. Hence

$$\begin{aligned}
\boldsymbol{\mu}' &= \mathbb{E}[\mathbf{x}] \\
&= \mathbb{E}[W\mathbf{z} + \boldsymbol{\mu} + \epsilon] \\
&= W(\mathbb{E}[\mathbf{z}] = 0) + (\mathbb{E}[\boldsymbol{\mu}] = \boldsymbol{\mu}) + (\mathbb{E}[\epsilon] = 0) \\
&= \boldsymbol{\mu}
\end{aligned}$$

$$\begin{aligned}
\Lambda' &= \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T\right] \\
&= \mathbb{E}\left[(W\mathbf{z} + \epsilon)(W\mathbf{z} + \epsilon)^T\right] \\
&= \mathbb{E}\left[W\mathbf{z}\mathbf{z}^T W^T + W\mathbf{z}\epsilon^T + \epsilon W\mathbf{z}^T + \epsilon\epsilon^T\right] \\
&= W\mathbb{E}\left[\mathbf{z}\mathbf{z}^T\right] W^T + 0 + 0 + \mathbb{E}\left[\epsilon\epsilon^T\right] \\
&= W\mathbf{I}\mathbf{I}^T W^T + \sigma^2\mathbf{I} \\
&= WW^T + \sigma^2\mathbf{I}
\end{aligned}$$

Hence, we can write $\mathbb{P}[\mathbf{x}] = \mathcal{N}(\boldsymbol{\mu}, C)$ where $C = WW^T + \sigma^2\mathbf{I}$

Part 2

Note: I will use $C = WW^T + \sigma^2\mathbf{I}$ for the rest of the question
Since all the sample points are independent, we can write $\mathbb{P}[X] = \prod_{i=1}^{n} \mathbb{P}\left[\mathbf{x}^i\right]$. Therefore

$$
\begin{aligned}
\mathbb{P}\left[X \,\middle|\, \boldsymbol{\mu}, W, \sigma\right] &= \prod_{i=1}^{n} \mathbb{P}\left[\mathbf{x}^i \,\middle|\, \boldsymbol{\mu}, W, \sigma\right] \\
&= \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{x}^i \,\middle|\, \boldsymbol{\mu}, C\right) \\
\implies \mathbb{P}\left[X \,\middle|\, \boldsymbol{\mu}, W, \sigma\right] &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi \|C\|}} e^{\frac{-1}{2}\left(\mathbf{x}^i - \boldsymbol{\mu}\right) C^{-1} \left(\mathbf{x}^i - \boldsymbol{\mu}\right)^T}
\end{aligned}
$$

Part 3

We can write the MLE estimate of $\boldsymbol{\mu}$ as

$$
\begin{aligned}
\boldsymbol{\mu}^{MLE} &= \underset{\boldsymbol{\mu} \in \mathbb{R}^d}{\arg\max}\, \mathbb{P}\left[X \,\middle|\, \boldsymbol{\mu}, W, \sigma\right] \\
&= \underset{\boldsymbol{\mu} \in \mathbb{R}^d}{\arg\max}\, log\left(\mathbb{P}\left[X \,\middle|\, \boldsymbol{\mu}, W, \sigma\right]\right) \\
&= \underset{\boldsymbol{\mu} \in \mathbb{R}^d}{\arg\max}\, log\left(\prod_{i=1}^{n} \mathbb{P}\left[\mathbf{x}^i \,\middle|\, \boldsymbol{\mu}, W, \sigma\right]\right) \\
&= \underset{\boldsymbol{\mu} \in \mathbb{R}^d}{\arg\max} \sum_{i=1}^{n} log\left(\mathbb{P}\left[\mathbf{x}^i \,\middle|\, \boldsymbol{\mu}, W, \sigma\right]\right) \\
&= \underset{\boldsymbol{\mu} \in \mathbb{R}^d}{\arg\min} \sum_{i=1}^{n} \frac{1}{2}\left(\mathbf{x}^i - \boldsymbol{\mu}\right) C^{-1} \left(\mathbf{x}^i - \boldsymbol{\mu}\right)^T + \frac{1}{2} log\left(2\pi \|C\|\right) \\
&= \underset{\boldsymbol{\mu} \in \mathbb{R}^d}{\arg\min} \sum_{i=1}^{n} \frac{1}{2}\left(\mathbf{x}^i - \boldsymbol{\mu}\right) C^{-1} \left(\mathbf{x}^i - \boldsymbol{\mu}\right)^T
\end{aligned}
$$

We can simply differentiate the RHS term, and using differentials of matrices, we get

$$
\begin{aligned}
\frac{\delta\left(RHS\right)}{\delta\left(\boldsymbol{\mu}\right)} &= \frac{\delta\left(\sum_{i=1}^{n} \frac{1}{2}\left(\mathbf{x}^i - \boldsymbol{\mu}\right) C^{-1} \left(\mathbf{x}^i - \boldsymbol{\mu}\right)^T\right)}{\delta\left(\boldsymbol{\mu}\right)} = 0 \\
\implies \sum_{i=1}^{n} C^{-1}\left(\mathbf{x}^i - \boldsymbol{\mu}\right) &= 0 \\
\implies \boldsymbol{\mu}^{MLE} &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}^i
\end{aligned}
$$

Therefore, the MLE estimate of $\boldsymbol{\mu}$ is, as expected, the emperical mean of all the data points. Hence we use this to centralize for the PCA algorithm.