Indian Institute of Technology Kanpur
CS698X: Topics in Probabilistic Modelling and Inference, 18–19

Student Name: Gurpreet Singh
Roll Number: 150259
Date: March 14, 2018

ASSIGNMENT 2

Question 1

Since $p(\mathbf{x} \,|\, \theta)$ is a Guassian with fixed covariance matrix, we can write

$$p(\mathbf{x} \,|\, \theta) \quad = \quad p(\mathbf{x} \,|\, \boldsymbol{\mu}) \quad = \quad \left(\frac{1}{(2\pi)^D \,|\, \boldsymbol{\Sigma} \,|} \right)^{1/2} \cdot \exp\left(\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

We can therefore compute $g(.)$ as follows

$$\begin{aligned} g(\mathbf{x}, \theta) \quad &= \quad \nabla_\theta \log\left(p(\mathbf{x} \,|\, \theta)\right) \quad = \quad \nabla_{\boldsymbol{\mu}} \log\left(p(\mathbf{x} \,|\, \boldsymbol{\mu})\right) \\ &= \quad \nabla_{\boldsymbol{\mu}} \left[\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) - \frac{D}{2} \log(2\pi) - \frac{1}{2} \log(|\, \boldsymbol{\Sigma} \,|) \right] \\ &= \quad \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \end{aligned}$$

For $\mathbf{F}$, we can write

$$\begin{aligned} \mathbf{F} \quad &= \quad \underset{p(\mathbf{x} \,|\, \boldsymbol{\mu})}{\mathbb{E}} \left[g(\mathbf{x}, \theta) g(\mathbf{x}, \theta)^{\mathrm{T}} \right] \\ &= \quad \int_{\mathbb{R}^D} g(\mathbf{x}, \theta) g(\mathbf{x}, \theta)^{\mathrm{T}} p(\mathbf{x} \,|\, \theta) \, \mathrm{d}\mathbf{x} \\ &= \quad \left(\frac{1}{(2\pi)^D \,|\, \boldsymbol{\Sigma} \,|} \right)^{1/2} \int_{\mathbb{R}^D} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \exp\left(\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right) \, \mathrm{d}\mathbf{x} \end{aligned}$$

Abusing the fact that $\boldsymbol{\Sigma}$ is a constant, using equation (61) from Petersen and Pedersen (2012), we write this as

$$\begin{aligned} \mathbf{F} \quad &= \quad \left(\frac{1}{(2\pi)^D \,|\, \boldsymbol{\Sigma} \,|} \right)^{1/2} \int_{\mathbb{R}^D} 2 \frac{\delta\left(\exp\left(\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right) \right)}{\delta(\boldsymbol{\Sigma})} \, \mathrm{d}\mathbf{x} \\ &= \quad 2 \left(\frac{1}{(2\pi)^D \,|\, \boldsymbol{\Sigma} \,|} \right)^{1/2} \frac{\delta\left(\int_{\mathbb{R}^D} \exp\left(\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right) \, \mathrm{d}\mathbf{x} \right)}{\delta(\boldsymbol{\Sigma})} \end{aligned}$$

The second equality is because the integral is over $\mathbf{x}$ which is independent of $\boldsymbol{\Sigma}$. Since the term within the integral is proportional to that of a gaussian, we can write it as

$$\begin{aligned} \mathbf{F} \quad &= \quad 2 \left(\frac{1}{(2\pi)^D \,|\, \boldsymbol{\Sigma} \,|} \right)^{1/2} \frac{\delta\left(\sqrt{(2\pi)^D \,|\, \boldsymbol{\Sigma} \,|} \right)}{\delta(\boldsymbol{\Sigma})} \\ &= \quad \boldsymbol{\Sigma}^{-1} \end{aligned}$$

The last equality follows from equation (49) from Petersen and Pedersen (2012)

Therefore, we can now write the final form of the kernel similarity

$$\kappa(\mathbf{x},\mathbf{x}') \quad = \quad g(\mathbf{x},\theta)^{\mathrm{T}}\mathbf{F}^{-1}g(\mathbf{x}',\theta)$$

Substituting from the terms of $g(.)$ and $\mathbf{F}$, we get

$$\kappa(\mathbf{x},\mathbf{x}') \quad = \quad (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{x}'-\boldsymbol{\mu}) \tag{a}$$

In case the covariance matrix is an identity matrix, we have

$$\kappa(\mathbf{x},\mathbf{x}') \quad = \quad (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\,(\mathbf{x}'-\boldsymbol{\mu}) \tag{b}$$

Intuitive Meaning

Since $\boldsymbol{\Sigma}^{-1}$ is a positive definitive square matrix, for some $\mathbf{U} \in \mathbb{R}^{N\times N}$, we can say $\boldsymbol{\Sigma}^{-1} = \mathbf{U}^{\mathrm{T}}\mathbf{U}$. Therefore, replacing this in the kernel term, we have

$$\begin{aligned}\kappa(\mathbf{x},\mathbf{x}') \quad &= \quad (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\,\mathbf{U}^{\mathrm{T}}\mathbf{U}\,(\mathbf{x}'-\boldsymbol{\mu}) \\ &= \quad \langle \mathbf{U}\,(\mathbf{x}-\boldsymbol{\mu})\,,\mathbf{U}\,(\mathbf{x}'-\boldsymbol{\mu})\rangle\end{aligned}$$

We wish to see when the value of this kernel is maximum. Suppose if we limit the norms of $(\mathbf{x}-\boldsymbol{\mu})$ and $(\mathbf{x}'-\boldsymbol{\mu})$ to 1, *i.e.* $\|\mathbf{x}-\boldsymbol{\mu}\| = 1 = \|\mathbf{x}'-\boldsymbol{\mu}\|$, then we can say that the dot product will be maximum when $\mathbf{U}(\mathbf{x}-\boldsymbol{\mu}) = \mathbf{U}(\mathbf{x}'-\boldsymbol{\mu}) \implies \mathbf{x} = \mathbf{x}'$.

From this, we can infer that if the norms of $\mathbf{x}-\boldsymbol{\mu}$ and $\mathbf{x}'-\boldsymbol{\mu}$ are fixed, then the kernel value will be maximum when the angle between the vectors $\mathbf{x}-\boldsymbol{\mu}$ and $\mathbf{x}'-\boldsymbol{\mu}$ is the least, and will be minimum when the angle is the largest.

Question 2

Let us write the expression for $f(.)$ alternatively

$$g_n(\mathbf{x}, y) \quad = \quad f(\mathbf{x} - \mathbf{x}_n, y - y_n) \quad = \quad \mathcal{N}\left(\begin{bmatrix} \mathbf{x} - \mathbf{x}_n \\ y - y_n \end{bmatrix} \;\middle|\; \mathbf{0}, \sigma^2 \mathbf{I}\right)$$

Using gaussian properties, we can say that this is equivalent to

$$g_n(\mathbf{x}, y) \quad = \quad \mathcal{N}\left(\begin{bmatrix} \mathbf{x} \\ y \end{bmatrix} \;\middle|\; \begin{bmatrix} \mathbf{x}_n \\ y_n \end{bmatrix}, \sigma^2 \mathbf{I}\right)$$

Hence, we can rewrite the joint distribution as

$$\mathbb{P}\left[\mathbf{x}, y\right] \quad = \quad \sum_{n=1}^{N} g_n(\mathbf{x}, y)$$

In order to compute the conditional probability $\mathbb{P}\left[y \,\middle|\, \mathbf{x}\right]$, we first compute the marginal probability $\mathbb{P}\left[\mathbf{x}\right]$, which can be expressed as

$$\begin{aligned} \mathbb{P}\left[\mathbf{x}\right] \quad &= \quad \int_{\mathbb{R}} \mathbb{P}\left[\mathbf{x}, y\right] \, \mathrm{d}y \\ &= \quad \int_{\mathbb{R}} \frac{1}{N} \sum_{n=1}^{N} g_n(\mathbf{x}, y) \, \mathrm{d}y \\ &= \quad \frac{1}{N} \sum_{n=1}^{N} \int_{\mathbb{R}} g_n(\mathbf{x}, y) \, \mathrm{d}y \end{aligned}$$

Since $g_n(.)$ is a multivariate gaussian, using properties of gaussian, we can say

$$\mathbb{P}\left[\mathbf{x}\right] \quad = \quad \frac{1}{N} \sum_{n=1}^{N} \mathcal{N}\left(\mathbf{x} \,\middle|\, \mathbf{x}_n, \sigma^2 \mathbf{I}\right)$$

Now, we can write the marginal as

$$\begin{aligned} \mathbb{P}\left[y \,\middle|\, \mathbf{x}\right] \quad &= \quad \frac{\mathbb{P}\left[y, \mathbf{x}\right]}{\mathbb{P}\left[\mathbf{x}\right]} \\ &= \quad \frac{1}{N \mathbb{P}\left[\mathbf{x}\right]} \sum_{n=1}^{N} g_n(\mathbf{x}, y) \end{aligned}$$

We can write $g(.)$ as

$$\begin{aligned} g(\mathbf{x}, y) \quad &= \quad \left(\frac{1}{2\pi\sigma^2}\right)^{(D+1)/2} \exp\left(-\frac{\sigma^2}{2} \begin{bmatrix} \mathbf{x} - \mathbf{x}_n \\ y - y_n \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathbf{x} - \mathbf{x}_n \\ y - y_n \end{bmatrix}\right) \\ &= \quad \left(\frac{1}{2\pi\sigma^2}\right)^{(D+1)/2} \exp\left(-\frac{(y - y_n)^2}{2\sigma^2}\right) \exp\left(-\frac{(\mathbf{x} - \mathbf{x}_n)^{\mathrm{T}} (\mathbf{x} - \mathbf{x}_n)}{2\sigma^2}\right) \\ &= \quad \mathcal{N}\left(y \,\middle|\, y_n, \sigma^2\right) \mathcal{N}\left(\mathbf{x} \,\middle|\, \mathbf{x}_n, \sigma^2 \mathbf{I}\right) \end{aligned}$$

Hence, we can write the marginal as a weighted sum of univariate gaussians, as

$$\mathbb{P}\left[\, y \,\middle|\, \mathbf{x} \,\right] \quad = \quad \sum_{n=1}^{N} \frac{\mathcal{N}\left(\mathbf{x} \,\middle|\, \mathbf{x}_n, \sigma^2 \mathbf{I}\right)}{\sum_{n'=1}^{N} \mathcal{N}\left(\mathbf{x} \,\middle|\, \mathbf{x}_{n'}, \sigma^2 \mathbf{I}\right)} \mathcal{N}\left(y \,\middle|\, y_n, \sigma^2\right)$$

Writing this in a simplified manner, we get

$$\mathbb{P}\left[\, y \,\middle|\, \mathbf{x} \,\right] \quad = \quad \sum_{n=1}^{N} \mu_n \, \mathcal{N}\left(y \,\middle|\, y_n, \sigma^2\right)$$

where

$$\mu_n \quad = \quad \frac{\mathcal{N}\left(\mathbf{x} \,\middle|\, \mathbf{x}_n, \sigma^2 \mathbf{I}\right)}{\sum_{n'=1}^{N} \mathcal{N}\left(\mathbf{x} \,\middle|\, \mathbf{x}_{n'}, \sigma^2 \mathbf{I}\right)}$$

are the mixture weights.

Mean of the conditional

$$\begin{aligned} \mathop{\mathbb{E}}_{y \,|\, \mathbf{x}}\left[\, y \,\right] \quad &= \quad \int_{\mathbb{R}} \sum_{n=1}^{N} \mu_n \, \mathcal{N}\left(y \,\middle|\, y_n, \sigma^2\right) y \; \mathrm{d}y \\ &= \quad \sum_{n=1}^{N} \mu_n \cdot \mathop{\mathbb{E}}_{\mathcal{N}(y \,|\, y_n, \sigma^2)}\left[\, y \,\right] \end{aligned}$$

$$\operatorname{mean}\left(y \,\middle|\, \mathbf{x}\right) \quad = \quad \sum_{n=1}^{N} \mu_n y_n$$

Variance of the conditional

$$\begin{aligned} \mathop{\mathbb{E}}_{y \,|\, \mathbf{x}}\left[\, y^2 \,\right] - \mathop{\mathbb{E}}_{y \,|\, y_n, \sigma^2}\left[\, y \,\right]^2 \quad &= \quad \int_{\mathbb{R}} \sum_{n=1}^{N} \mu_n \, \mathcal{N}\left(y \,\middle|\, y_n, \sigma^2\right) y^2 \; \mathrm{d}y - \left(\sum_{n=1}^{N} \mu_n y_n\right)^2 \\ &= \quad \sum_{n=1}^{N} \mu_n \left(\sigma^2 + y_n^2\right) - \left(\sum_{n=1}^{N} \mu_n y_n\right)^2 \end{aligned}$$

$$\operatorname{var}\left(y \,\middle|\, \mathbf{x}\right) \quad = \quad \sigma^2 + \sum_{n=1} \mu_n y_n^2 - \left(\sum_{n=1}^{N} \mu_n y_n\right)^2$$

Question 3

Part A

Following the procedure of Laplace Approximation, let us write the second order Taylor expansion of $\log\left(\mathbb{P}\left[\,x \,\middle|\, a, b\,\right]\right)$

Let $f(x) = \log\big(\text{Gamma}\left(x \,\middle|\, a, b\right)\big)$, then for some x_0

$$f(x) \quad \approx \quad f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2$$

As discussed in class, it is simpler to choose x_0 such that $f'(x_0)$ is 0 and $f''(x_0)$ is negative. A suitable option for this is the mode of the distribution. The mode of the gamma distribution is $(a-1)/b$. This exists iff $a > 1$.

Suppose the mode is represented by m, therefore, if $a > 1$, we set $x_0 = m = (a-1)/b$. Hence, we can write

$$f(x) \quad \approx \quad f(m) + \frac{1}{2}f''(m)(x - m)^2$$

We can compute the double derivative as follows

$$\begin{aligned}
f''(x) \quad &= \quad \frac{d^2}{dx^2}f(x) \\
&= \quad \frac{d}{dx}f'(x) \\
&= \quad \frac{d}{dx}\left(\log\left(b^a\right) - \log\left(\Gamma(a)\right) + (a-1)\log\left(x\right) - bx\right) \\
&= \quad \frac{d}{dx}\left(\frac{a-1}{x} - b\right) \\
&= \quad -\frac{a-1}{x^2}
\end{aligned}$$

Therefore,

$$\begin{aligned}
f''(m) \quad &= \quad -\frac{a-1}{m^2} \\
&= \quad -\frac{b^2}{a-1}
\end{aligned}$$

We can now write the approximation for the gamma distribution

$$\begin{aligned}
\text{Gamma}\left(x \,\middle|\, a, b\right) \quad &= \quad \exp\left(f(x)\right) \\
&\approx \quad \exp\left(f(m) + \frac{1}{2}f''(m)(x - m)^2\right) \\
&= \quad \exp\left(f(m)\right)\,\exp\left(-\frac{1}{2}\frac{b^2}{a-1}(x-m)^2\right) \\
&\propto \quad \exp\left(-\frac{1}{2}\left(\frac{b^2}{a-1}\right)(x-m)^2\right)
\end{aligned}$$

Therefore, we can approximate $\text{Gamma}\left(x \,\middle|\, a, b\right)$ with the gaussian $\mathcal{N}\left(x \,\middle|\, \mu, \sigma^2\right)$ where,

$$\begin{aligned} \mu \quad = \quad m \quad &= \quad \frac{a-1}{b} \\ \sigma^2 \quad &= \quad \frac{a-1}{b^2} \end{aligned}$$

Part B

Suppose we approximate $\mathrm{Gamma}\left(x \,\middle|\, a, b\right)$ with the gaussian $\mathcal{N}\left(x \,\middle|\, \mu', {\sigma'}^2\right)$ such that

$$\begin{aligned} \mu' \quad &= \quad \mathrm{mean}\left(x\right) \\ {\sigma'}^2 \quad &= \quad \mathrm{var}\left(x\right) \end{aligned}$$

Therefore,

$$\begin{aligned} \mu' \quad &= \quad \frac{a}{b} \\ {\sigma'}^2 \quad &= \quad \frac{a}{b^2} \end{aligned}$$

This, therefore, is equivalent to the gaussian obtained using Laplace Approximation if the value of a is large, $i.e.$ $a \gg 1$.

Part C

We can use the Laplace approximation of the Gamma distribution obtained in Part A to approximate the gaussian function as follows

$$\widetilde{\Gamma}\left(a\right) \quad \approx \quad \frac{b^a}{\mathcal{N}\left(x \,\middle|\, \mu, \sigma^2\right)} x^{a-1} e^{-bx}$$

Since the LHS is independent of both b and x, we can set arbitrary values for both. For simplicity, we assume $b = 1$ and $x = \mu = (a-1)/b$. Therefore,

$$\widetilde{\Gamma}\left(a\right) \quad = \quad \left(2\pi\sigma^2\right)^{1/2} \mu^{a-1} e^{-\mu}$$

A Comparision plot of the real Gamma function, $\Gamma(.)$ and the approximated Gamma function, $\widetilde{\Gamma}(.)$ is shown in Figure 1

It is clear from the figure that the approximation is indeed a good one, and there is small loss.

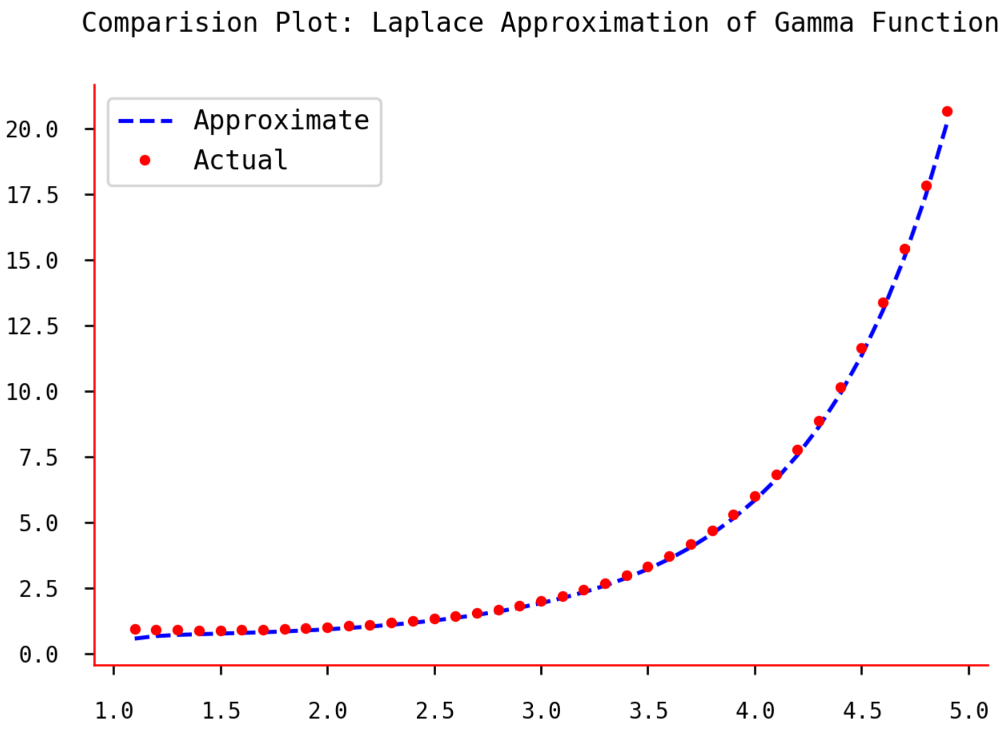

Figure 1: $\Gamma(.)$ vs $\widetilde{\Gamma}(.)$

Question 4

The poisson distribution is written as

$$\mathbb{P}\left[\, x \,\middle|\, \lambda \,\right] \quad = \quad e^{-\lambda} \frac{\lambda^{x}}{(x)!}$$

We can model a supervised learning problem as a Poisson distribution by mapping $\mathbf{x} \in \mathbb{R}^{D}$ to λ as $\lambda = \mathbf{w}^{\mathrm{T}}\mathbf{x}$ for some $\mathbf{w} \in \mathbb{R}^{D}$.

Therefore, we can write the likelihood as

$$\mathbb{P}\left[\, y \,\middle|\, \mathbf{x}, \mathbf{w} \,\right] \quad = \quad e^{-\mathbf{w}^{\mathrm{T}}\mathbf{x}} \frac{\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}\right)^{y}}{(y)!}$$

Maximum Likelihood Estimation

We need to find the Log Likelihood ($\log\left(\mathbb{P}\left[\, \mathbf{y} \,\middle|\, \mathbf{X}, \mathbf{w} \,\right]\right)$) of the data, which is

$$\begin{aligned}
\log\left(\mathbb{P}\left[\, \mathbf{y} \,\middle|\, \mathbf{X}, \mathbf{w} \,\right]\right) \quad &= \quad \log\left(\prod_{n=1}^{N} \mathbb{P}\left[\, y_n \,\middle|\, \mathbf{x}_n, \mathbf{w} \,\right]\right) \\
&= \quad \sum_{n=1}^{N} \log\left(\mathbb{P}\left[\, y_n \,\middle|\, \mathbf{x}_n, \mathbf{w} \,\right]\right) \\
&= \quad \sum_{n=1}^{N} y_n \log\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}\right) - \sum_{n=1}^{N} \log\left((y_n)!\right) - \mathbf{w}^{\mathrm{T}}\left(\sum_{n=1}^{N} \mathbf{x}_n\right)
\end{aligned}$$

Therefore the MLE estimate is given as

$$\begin{aligned}
\mathbf{w}_{\mathrm{MLE}} \quad &= \quad \arg\max_{\mathbf{w}} \ \log\left(\mathbb{P}\left[\, \mathbf{y} \,\middle|\, \mathbf{X}, \mathbf{w} \,\right]\right) \\
\implies 0 \quad &= \quad \left.\frac{d\left(\log\left(\mathbb{P}\left[\, \mathbf{y} \,\middle|\, \mathbf{X}, \mathbf{w} \,\right]\right)\right)}{d\left(\mathbf{w}\right)}\right|_{\mathbf{w}_{\mathrm{MLE}}} \\
\implies 0 \quad &= \quad \sum_{n=1}^{N} \frac{y_n}{\mathbf{w}_{\mathrm{MLE}}{}^{\mathrm{T}}\mathbf{x}_n} \mathbf{x}_n - \sum_{n=1}^{N} \mathbf{x}_n
\end{aligned}$$

This clearly does not have a closed solution. Therefore, we need iterative methods such as Gradient Descent or second order methods such as Newton's Method. Since this is an unconstrained optimization, we can use vanilla Gradient Descent for the MLE computation, which I have shown in Algorithm 1

Algorithm 1: Gradient Descent for MLE

Result: MLE Estimate of $\mathbf{w}$, *i.e.* $\mathbf{w}_{\mathrm{MLE}}$
Steps:

1. Initialize the value of $\mathbf{w}$ as $\mathbf{w}^{(0)}$
2. For $t = 1, 2 \ldots T$ or until convergence, do

$$\begin{aligned} \nabla L(\mathbf{w}^{(t-1)}) &= \sum_{n=1}^{N} \left(1 - \frac{y_n}{{\mathbf{w}^{(t-1)}}^{\mathrm{T}}\mathbf{x}_n}\right)\mathbf{x}_n \\ \mathbf{w}^{(t)} &= \mathbf{w}^{(t-1)} - \eta_t \nabla L(\mathbf{w}^{(t-1)}) \end{aligned}$$

3. Return $\mathbf{w}^{(t')}$ where t' corresponds to the last iteration.

Question 5

We follow the standard EM procedure. We first give the posterior probability of the latent variables z, represented as $\mathbb{P}\left[\, z_n \,\middle|\, y_n, \mathbf{x}_n, \mathbf{w} \,\right]$. It is however easier to separate the cases when $y_n = 1$ or $y_n = 0$.

$$\mathbb{P}\left[\, z_n \,\middle|\, y_n = 1, \mathbf{x}_n, \mathbf{w} \,\right] \quad = \quad \mathbb{P}\left[\, z_n \,\middle|\, z_n > 0, \mathbf{x}_n, \mathbf{w} \,\right]$$

This is similar to the truncated gaussian distribution (Truncated Gaussian Distribution, 2018). Hence, we can directly write the expression of the pdf for this

$$\mathbb{P}\left[\, z_n \,\middle|\, y_n = 1, \mathbf{x}_n, \mathbf{w} \,\right] \quad = \quad \mathbb{I}\left[\, z_n > 0 \,\right] \frac{\mathcal{N}\left(z_n \,\middle|\, \mathbf{w}^{\mathrm{T}}\mathbf{x}_n, 1\right)}{1 - \mathbf{\Phi}(-\mathbf{w}^{\mathrm{T}}\mathbf{x}_n)}$$

where $\mathbf{\Phi}(.)$ denotes the CDF function of the standard gaussian distribution.

Similarly, we can write the expression in case $y_n = 0$

$$\mathbb{P}\left[\, z_n \,\middle|\, y_n = 0, \mathbf{x}_n, \mathbf{w} \,\right] \quad = \quad \mathbb{I}\left[\, z_n < 0 \,\right] \frac{\mathcal{N}\left(z_n \,\middle|\, \mathbf{w}^{\mathrm{T}}\mathbf{x}_n, 1\right)}{\mathbf{\Phi}(-\mathbf{w}^{\mathrm{T}}\mathbf{x}_n)}$$

We can write this in a single expression as follows

$$\mathbb{P}\left[\, z_n \,\middle|\, y_n, \mathbf{x}_n, \mathbf{w} \,\right] \quad = \quad \left[y_n \frac{\mathbb{I}\left[\, z_n > 0 \,\right]}{1 - \mathbf{\Phi}(-\mathbf{w}^{\mathrm{T}}\mathbf{x}_n)} + (1 - y_n) \frac{\mathbb{I}\left[\, z_n \leq 0 \,\right]}{\mathbf{\Phi}(-\mathbf{w}^{\mathrm{T}}\mathbf{x}_n)} \right] \mathcal{N}\left(z_n \,\middle|\, \mathbf{w}^{\mathrm{T}}\mathbf{x}_n, 1\right)$$

Expectation Step

We can now write the Complete Data Likelihood as

$$\begin{aligned}
\mathbb{P}\left[\, \mathbf{y}, \mathbf{z} \,\middle|\, \mathbf{X}, \mathbf{w} \,\right] &= \mathbb{P}\left[\, \mathbf{z} \,\middle|\, \mathbf{y}, \mathbf{X}, \mathbf{w} \,\right] \mathbb{P}\left[\, \mathbf{y} \,\middle|\, \mathbf{X}, \mathbf{w} \,\right] \\
&= \prod_{n=1}^{N} \mathbb{P}\left[\, z_n \,\middle|\, y_n, \mathbf{x}_n, \mathbf{w} \,\right] \mathbb{P}\left[\, y_n \,\middle|\, x_n, \mathbf{w} \,\right] \\
&= \prod_{n=1}^{N} \mathcal{N}\left(z_n \,\middle|\, \mathbf{w}^{\mathrm{T}}\mathbf{x}_n, 1\right) \left[y_n \frac{\mathbb{I}\left[\, z_n > 0 \,\right]}{1 - \mathbf{\Phi}(-\mathbf{w}^{\mathrm{T}}\mathbf{x}_n)} + (1 - y_n) \frac{\mathbb{I}\left[\, z_n \leq 0 \,\right]}{\mathbf{\Phi}(-\mathbf{w}^{\mathrm{T}}\mathbf{x}_n)} \right] \\
&\qquad \left[y_n \left(1 - \mathbf{\Phi}(-\mathbf{w}^{\mathrm{T}}\mathbf{x}_n)\right) + (1 - y_n)\mathbf{\Phi}(-\mathbf{w}^{\mathrm{T}}\mathbf{x}_n) \right] \\
&= \prod_{n=1}^{N} \mathcal{N}\left(z_n \,\middle|\, \mathbf{w}^{\mathrm{T}}\mathbf{x}_n, 1\right) \left[y_n \mathbb{I}\left[\, z_n > 0 \,\right] + (1 - y_n) \mathbb{I}\left[\, z_n \leq 0 \,\right] \right]
\end{aligned}$$

Therefore, the CLL is

$$\begin{aligned}
\text{CLL} &= \sum_{n=1}^{N} \log\left(y_n \mathbb{I}\left[\, z_n > 0 \,\right] + (1 - y_n) \mathbb{I}\left[\, z_n \leq 0 \,\right]\right) - \frac{1}{2}\left(z_n - \mathbf{w}^{\mathrm{T}}\mathbf{x}_n\right)^2 - \frac{1}{2}\log\left(2\pi\right) \\
&= \sum_{n=1}^{N} y_n \log\left(\mathbb{I}\left[\, z_n > 0 \,\right]\right) + (1 - y_n) \log\left(\mathbb{I}\left[\, z_n \leq 0 \,\right]\right) - \frac{1}{2}\left(z_n - \mathbf{w}^{\mathrm{T}}\mathbf{x}_n\right)^2 - \frac{1}{2}\log\left(2\pi\right) \\
\Longrightarrow \underset{\mathbf{z} \,|\, \mathbf{y}}{\mathbb{E}}\left[\, \text{CLL} \,\right] &= \sum_{n=1}^{N} y_n \, \mathbb{E}\left[\, \log\left(\mathbb{I}\left[\, z_n > 0 \,\right]\right) \,\right] + (1 - y_n) \, \mathbb{E}\left[\, \log\left(\mathbb{I}\left[\, z_n \leq 0 \,\right]\right) \,\right] - \\
&\qquad \frac{1}{2}\left(\mathbb{E}\left[\, z_n^2 \,\right] - 2\mathbf{w}^{\mathrm{T}}\mathbf{x}_n \mathbb{E}\left[\, z_n \,\right] + \left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n\right)^2\right) - \frac{1}{2}\log\left(2\pi\right) \\
&= \sum_{n=1}^{N} \frac{1}{2}\mathbb{E}\left[\, z_n^2 \,\right] - \mathbf{w}^{\mathrm{T}}\mathbf{x}_n \mathbb{E}\left[\, z_n \,\right] + \frac{1}{2}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n\right)^2 - \frac{1}{2}\log\left(2\pi\right)
\end{aligned}$$

The last equality is obtained by analysing the expectations separately for the cases $y_n = 1$ and $y_n = 0$. We now compute the expectations required to completely express the $\mathbb{E}\left[\,\text{CLL}\,\right]$. These expectations, however, are computed with respected to a known (old) value of $\mathbf{w}$, and are expressed as follows

$$\begin{aligned}
\underset{z_n \,|\, y_n = 1}{\mathbb{E}}\left[\, z_n \,\right] \quad &= \quad {\mathbf{w}^{(\text{old})}}^{\text{T}}\mathbf{x}_n + \frac{\phi\left(-{\mathbf{w}^{(\text{old})}}^{\text{T}}\mathbf{x}_n\right)}{1 - \,\mathbf{\Phi}\left(-{\mathbf{w}^{(\text{old})}}^{\text{T}}\mathbf{x}_n\right)} \\
\underset{z_n \,|\, y_n = 0}{\mathbb{E}}\left[\, z_n \,\right] \quad &= \quad {\mathbf{w}^{(\text{old})}}^{\text{T}}\mathbf{x}_n - \frac{\phi\left(-{\mathbf{w}^{(\text{old})}}^{\text{T}}\mathbf{x}_n\right)}{\mathbf{\Phi}\left(-{\mathbf{w}^{(\text{old})}}^{\text{T}}\mathbf{x}_n\right)} \\
\implies \underset{z_n \,|\, y_n}{\mathbb{E}}\left[\, z_n \,\right] \quad &= \quad {\mathbf{w}^{(\text{old})}}^{\text{T}}\mathbf{x}_n + \phi\left(-{\mathbf{w}^{(\text{old})}}^{\text{T}}\mathbf{x}_n\right)\left(\frac{1}{1 - \,\mathbf{\Phi}\left(-{\mathbf{w}^{(old)}}^{\text{T}}\mathbf{x}_n\right)}\right)^{y_n}\left(\frac{-1}{\mathbf{\Phi}\left(-{\mathbf{w}^{(old)}}^{\text{T}}\mathbf{x}_n\right)}\right)^{1-y_n}
\end{aligned}$$

Note. We do not need $\mathbb{E}\left[\, z_n^2 \,\right]$ as it will not partake in the Maximization step (since it is constant with respect to $\mathbf{w}$). Therefore, I have skipped its computation

Therefore, we can write the final expression of $\mathbb{E}\left[\,\text{CLL}\,\right]$ as follows

$$\mathbb{E}\left[\,\text{CLL}\,\right] \quad = \quad \sum_{n=1}^{N} \frac{1}{2}\mathbb{E}\left[\, z_n^2 \,\right] - \mathbf{w}^{\text{T}}\mathbf{x}_n\mathbb{E}\left[\, z_n \,\right] + \frac{1}{2}\left(\mathbf{w}^{\text{T}}\mathbf{x}_n\right)^2 - \frac{1}{2}\log\left(2\pi\right)$$

where $\phi(.)$ and $\mathbf{\Phi}(.)$ are the pdf and the cdf functions (respectively) of the standard gaussian distribution, and the expectation terms are as given in the previous equations.

Maximization Step

Now, we can do MLE for the parameters by maximising the $\mathbb{E}\left[\,\text{CLL}\,\right]$ as in the standard EM Algorithm

$$\mathbf{w}_{\text{MLE}} \quad = \quad \underset{\mathbf{w}}{\arg\max}\; \mathbb{E}\left[\,\text{CLL}\,\right]$$

We need to differentiate this and equate to 0 for the desired MLE solution.

$$\begin{aligned}
0 \quad &= \quad \left.\frac{\delta\left(\mathbb{E}\left[\,\text{CLL}\,\right]\right)}{\delta\left(\mathbf{w}\right)}\right|_{\mathbf{w}=\mathbf{w}_{\text{MLE}}} \\
&= \quad \sum_{n=1}^{N}\left({\mathbf{w}_{\text{MLE}}}^{\text{T}}\mathbf{x}_n\right)\mathbf{x}_n - \mathbb{E}\left[\, z_n \,\right]\mathbf{x}_n \\
&= \quad \mathbf{X}^{\text{T}}\mathbf{X}\mathbf{w}_{\text{MLE}} - \mathbf{X}^{\text{T}}\,\mathbb{E}\left[\,\mathbf{z}\,\right]
\end{aligned}$$

Therefore, we can write

$$\mathbf{w}_{\text{MLE}} \quad = \quad \left(\mathbf{X}^{\text{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\text{T}}\,\mathbb{E}\left[\,\mathbf{z}\,\right]$$

where

$$\mathbb{E}\left[\,\mathbf{z}\,\right] \quad = \quad \left[\underset{z_1 \,|\, y_1}{\mathbb{E}}\left[\, z_1 \,\right],\; \underset{z_2 \,|\, y_2}{\mathbb{E}}\left[\, z_2 \,\right] \ldots \underset{z_N \,|\, y_N}{\mathbb{E}}\left[\, z_N \,\right]\right]^{\text{T}}$$

We can write the complete EM algorithm as given in Algorithm 2

```
Algorithm 2: EM Algorithm for Binary Classification as LVM

Result: Posterior Distribution $\mathbb{P}\left[\, \mathbf{z} \,\middle|\, \mathbf{y}, \mathbf{X}, \mathbf{w} \,\right]$, MLE estimate $\mathbf{w}_{\text{MLE}}$
Steps:

1. Initialize the value of $\mathbf{w}$ as $\mathbf{w}^{(0)}$

2. For t = 1, 2 ... T or until convergence, do

    (a) E Step:  Given $\mathbf{w}^{(t-1)}$, compute the posteriors as

        $\forall\, n \in [\, N \,], \quad \boldsymbol{\Phi}_n = \boldsymbol{\Phi}\left( -{\mathbf{w}^{(t-1)}}^{\mathrm{T}} \mathbf{x}_n \right)$
        $\forall\, n \in [\, N \,], \quad \mathbb{P}\left[\, z_n^{(t)} \,\middle|\, y_n, \mathbf{x}_n, \mathbf{w}^{(t-1)} \,\right] = \left( \frac{\mathbb{I}\left[\, z_n > 0 \,\right]}{1 - \boldsymbol{\Phi}_n} \right)^{y_n} \left( \frac{\mathbb{I}\left[\, z_n \leq 0 \,\right]}{\boldsymbol{\Phi}_n} \right)^{(1-y_n)} \mathcal{N}\left( z_n^{(t)} \,\middle|\, {\mathbf{w}^{(t-1)}}^{\mathrm{T}} \mathbf{x}_n, 1 \right)$

        Compute the expectation of $\mathbf{z}$ as

        $\forall\, n \in [\, N \,], \quad \underset{z_n \,|\, y_n}{\mathbb{E}}\left[\, z_n \,\right] = {\mathbf{w}^{(\text{t - 1})}}^{\mathrm{T}} \mathbf{x}_n + \phi\left( -{\mathbf{w}^{(\text{t - 1})}}^{\mathrm{T}} \mathbf{x}_n \right) \left( \frac{1}{1 - \boldsymbol{\Phi}_n} \right)^{y_n} \left( -\frac{1}{\boldsymbol{\Phi}_n} \right)^{1-y_n}$
        $\mathbb{E}\left[\, \mathbf{z} \,\right] = \left[\, \underset{z_1 \,|\, y_1}{\mathbb{E}}\left[\, z_1 \,\right], \underset{z_2 \,|\, y_2}{\mathbb{E}}\left[\, z_2 \,\right] \ldots \underset{z_N \,|\, y_N}{\mathbb{E}}\left[\, z_N \,\right] \,\right]^{\mathrm{T}}$

    (b) M Step:
        Compute the MLE solution of $\mathbf{w}$ as

        $\mathbf{w}^{(t)} = \left( \mathbf{X}^{\mathrm{T}} \mathbf{X} \right)^{-1} \mathbf{X}^{\mathrm{T}}\, \mathbb{E}\left[\, \mathbf{z} \,\right]$

3. Return $\mathbb{P}\left[\, \mathbf{z} \,\middle|\, \mathbf{y}, \mathbf{X}, \mathbf{w} \,\right] = \prod_{n=1}^{N} \mathbb{P}\left[\, z_n^{(t')} \,\middle|\, y_n, \mathbf{X}, \mathbf{w}^{(t'-1)} \,\right]$ and $\mathbf{w}_{\text{MLE}} = \mathbf{w}^{(t')}$
```

Note. It might not be possible to compute the values of $\boldsymbol{\Phi}\left(\mathbf{w}^{\mathrm{T}} \mathbf{x}_n \right)$ and therefore, we might need to approximate this as well using standard approximation techniques.

Question 6

It is difficult to model $\mathbb{P}\left[\,\mathbf{z}_n \,\middle|\, \mathbf{x}_n\,\right]$, not because it is intractable (since $\mathbf{z}$ is from a discrete distribution), but becausing computing the posterior (at least accurately) would take $\mathcal{O}\left(2^K\right)$ time. This is not practical for large values of K, therefore, we will try to approximate the expectation step. To be more precise, we will approximate $\mathbb{E}\left[\,\mathbf{z}_n \,\middle|\, \mathbf{x}_n\,\right]$. First, it is better to actually compute $\mathbb{E}\left[\,\mathrm{CLL}\,\right]$, assuming we know $\mathbb{E}\left[\,\mathbf{z}_n \,\middle|\, \mathbf{x}_n\,\right]$.

$$
\begin{aligned}
\mathbb{P}\left[\,\mathbf{X}, \mathbf{Z} \,\middle|\, \mathbf{W}, \beta\,\right] &= \mathbb{P}\left[\,\mathbf{X} \,\middle|\, \mathbf{Z}, \mathbf{W}, \beta\,\right] \cdot \mathbb{P}\left[\,\mathbf{Z} \,\middle|\, \boldsymbol{\pi}\,\right] \\
&= \prod_{n=1}^{N} \left\{ \mathbb{P}\left[\,\mathbf{x}_n \,\middle|\, \mathbf{z}_n, \mathbf{W}, \beta\,\right] \prod_{k=1}^{K} \mathbb{P}\left[\,{}_{nk} \,\middle|\, \pi_k\,\right] \right\}
\end{aligned}
$$

We know the forms of all these terms, and therefore write the CLL as follows,

$$
\implies \mathrm{CLL} \quad = \quad \sum_{n=1}^{N} \left\{ \frac{\beta}{2} \left(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n\right)^{\mathrm{T}} \left(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n\right) + \sum_{k=1}^{K} \left\{ z_{nk} \log\left(\pi_k\right) + \left(1 - z_{nk}\right) \log\left(1 - \pi_k\right) \right\} \right\} + \text{constants}
$$

We can now take the expectation of this ($\mathbb{E}\left[\,\mathrm{CLL}\,\right]$) with respect to the distribution $\mathbf{Z} \,\middle|\, \mathbf{X}, \mathbf{W}^{(\mathrm{old})}, \beta^{(\mathrm{old})}, \boldsymbol{\pi}^{(\mathrm{old})}$. The problem, as discussed earlier, is that the computation of the posterior probability will take $\mathcal{O}\left(2^K\right)$ time. Therefore, we choose to sample a set of values of $\mathbf{Z} \,\middle|\, \mathbf{X}$, and use Monte Carlo estimation for the expectations. My proposed solution for this is to use Gibbs Sampling, as there is conditional conjugacy in our case, which can be seen as below

$$
\begin{aligned}
\mathbb{P}\left[\,z_{nk} \,\middle|\, \mathbf{x}_n, \mathbf{z}_{-nk}, \mathbf{W}, \beta, \boldsymbol{\pi}\,\right] &= \mathbb{P}\left[\,\mathbf{x}_n \,\middle|\, z_{nk}, \mathbf{z}_{-nk}, \mathbf{W}, \beta\,\right] \mathbb{P}\left[\,z_{nk} \,\middle|\, \mathbf{z}_{-nk}, \boldsymbol{\pi}\,\right] \\
&= \mathbb{P}\left[\,\mathbf{x}_n \,\middle|\, z_{nk}, \mathbf{z}_{-nk}, \mathbf{W}, \beta\,\right] \mathbb{P}\left[\,z_{nk} \,\middle|\, \pi_k\,\right] \\
&= \mathcal{N}\left(\mathbf{x}_n \,\middle|\, \mathbf{W}\mathbf{z}_n, \beta^{-1}\mathbf{I}\right) \text{ Bernoulli}\left(z_{nk} \,\middle|\, \pi_k\right) \\
&\propto \exp\left(\frac{\beta}{2} \left(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n\right)^{\mathrm{T}} \left(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n\right) \right) \pi_k^{z_{nk}} \left(1 - \pi_k\right)^{\left(1 - z_{nk}\right)}
\end{aligned}
$$

We can ignore all terms that are independent of, as we only need to analyse the form of the distribution.

We can write $\mathbf{W}\mathbf{z}_n = \sum_{k'=1}^{K} \mathbf{w}_{k'} z_{nk'}$. Define $\mathbf{x}_{-nk} = \mathbf{x}_n - \sum_{\substack{k'=1 \\ k' \neq k}}^{K} z_{nk'} \mathbf{w}_{k'}$, then we can write

$$
\mathbb{P}\left[\,z_{nk} \,\middle|\, \mathbf{x}_n, \mathbf{z}_{-nk}, \mathbf{W}, \beta, \boldsymbol{\pi}\,\right] \quad \propto \quad \exp\left(\frac{\beta}{2} \left[\, z_{nk}^2 \mathbf{w}_k{}^{\mathrm{T}} \mathbf{w}_k - 2 z_{nk} \mathbf{x}_{-nk}{}^{\mathrm{T}} \mathbf{w}_k \,\right] \right) \left(\frac{\pi_k}{1 - \pi_k} \right)^{z_{nk}}
$$

Since $z_{nk} \in \{0, 1\}$, $z_{nk}^2 = z_{nk}$. Therefore,

$$
\mathbb{P}\left[\,z_{nk} \,\middle|\, \mathbf{x}_n, \mathbf{z}_{-nk}, \mathbf{W}, \beta, \boldsymbol{\pi}\,\right] \quad \propto \quad \left(\exp\left(\frac{\beta}{2} \left[\, \mathbf{w}_k{}^{\mathrm{T}} \mathbf{w}_k - 2\, \mathbf{x}_{-nk}{}^{\mathrm{T}} \mathbf{w}_k \,\right] \right) \frac{\pi}{1 - \pi_k} \right)^{z_{nk}}
$$

This represents a Bernoulli. Hence, we can say there is conditional conjugacy. Therefore, we can write the conditional posterior as follows

$$
\mathbb{P}\left[\,z_{nk} \,\middle|\, \mathbf{x}_n, \mathbf{z}_{-nk}, \mathbf{W}, \beta, \boldsymbol{\pi}\,\right] \quad = \quad {\pi'_k}^{z_{nk}} \left(1 - \pi'_k\right)^{\left(1 - z_{nk}\right)}
$$

where

$$
\pi'_k \quad = \quad \frac{\exp\left(\frac{\beta}{2} \left(\mathbf{w}_k - 2\mathbf{x}_{-nk}\right)^{\mathrm{T}} \mathbf{w}_k \right) \pi_k}{1 + \pi \left(\exp\left(\frac{\beta}{2} \left(\mathbf{w}_k - 2\mathbf{x}_{-nk}\right)^{\mathrm{T}} \mathbf{w}_k \right) - 1 \right)}
$$

```
Algorithm 3: Gibbs Sampling for Latent Variables, $\mathbf{Z}$

Result: Sample set, $\mathcal{S} = \left\{\mathbf{Z}^{(1)}, \mathbf{Z}^{2} \ldots \mathbf{Z}^{M}\right\}$
Steps:
  1. Initialize the zero^th sample as $\mathbf{Z}^{(0)}$
  2. For $m = 1, 2 \ldots M$
      (a) For all $k \in [\,K\,]$,
            $\pi'_k = \frac{\exp\left(\frac{\beta}{2}\left(\mathbf{w}_k - 2\mathbf{x}_{-nk}\right)^{\mathrm{T}}\mathbf{w}_k\right)\pi_k}{1 + \pi\left(\exp\left(\frac{\beta}{2}\left(\mathbf{w}_k - 2\mathbf{x}_{-nk}\right)^{\mathrm{T}}\mathbf{w}_k\right) - 1\right)}$
      (b) For all $n \in [\,N\,], k \in [\,K\,]$, sample $z^{m}_{nk}$ as
            $z^{(m)}_{nk} \sim \text{Bernoulli}\left(\pi'_k\right)$
  3. Return $\mathcal{S} = \left\{\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)} \ldots \mathbf{Z}^{(M)}\right\}$
```

Therefore, we can perform Gibbs Sampling to compute M samples of $\mathbf{Z}$ and approximate $\mathbb{E}\left[\,\text{CLL}\,\right]$ using (unweighted) Monte Carlo estimation. The Gibbs Sampling algorithm is given in Algorithm 3

Expectation Step

Suppose we can call the above algorithm (Algorithm 3) as $\mathsf{GibbsSampler}\left(\mathbf{W}^{(\text{old})}, \beta^{(\text{old})}, \boldsymbol{\pi}^{(\text{old})}\right)$. Then we can estimate $\mathbb{E}\left[\,\text{CLL}\,\right]$ as follows

$$\mathbb{E}\left[\,\text{CLL}\,\right] \approx \sum_{\mathbf{Z}\in\mathcal{S}} \sum_{n=1}^{N} \left\{ \frac{D}{2}\log\left(\beta\right) - \frac{\beta}{2}\left(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n\right)^{\mathrm{T}}\left(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n\right) + \ldots \right.$$

$$\left. \cdots + \sum_{k=1}^{K}\left\{z_{nk}\log\left(\pi_k\right) + \left(1 - z_{nk}\right)\log\left(1 - \pi_k\right)\right\} \right\} + \text{constants}$$

where $\mathcal{S} = \mathsf{GibbsSampler}\left(\mathbf{W}^{(\text{old})}, \beta^{(\text{old})}, \boldsymbol{\pi}^{(\text{old})}\right)$.

Maximization Step

We can now maximize the $\mathbb{E}\left[\,\text{CLL}\,\right]$ over the parameters, namely $\{\mathbf{W}, \boldsymbol{\pi}\}$. Hence, we can write this as

$$\left\{\mathbf{W}_{\text{MLE}}, \beta_{\text{MLE}}, \boldsymbol{\pi}_{\text{MLE}}\right\} = \underset{\{\mathbf{W}, \beta, \boldsymbol{\pi}\}}{\arg\max}\ \mathbb{E}\left[\,\text{CLL}\,\right]$$

Using standard MLE approach, we get

1. **MLE of $\boldsymbol{\pi}$**

$$\begin{aligned}
0 &= \frac{\delta\left(\mathbb{E}\left[\,\mathrm{CLL}\,\right]\right)}{\delta\left(\pi_k\right)} \\
\implies \pi_{\mathrm{k,\ MLE}} &= \frac{1}{N\cdot|\,\mathcal{S}\,|}\sum_{\mathbf{Z}\in\mathcal{S}}\ \sum_{n=1}^{N} z_{nk} \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\,z_{nk}\,\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\,\mathbf{z}_n\,\right]_k
\end{aligned}$$

2. **MLE of $\mathbf{W}$** Similarly, for $\mathbf{W}$

$$\begin{aligned}
0 &= \left.\frac{\delta\left(\mathbb{E}\left[\,\mathrm{CLL}\,\right]\right)}{\delta\left(\mathbf{W}\right)}\right|_{\mathbf{W}=\mathbf{W}_{\mathrm{MLE}}} \\
\implies 0 &= \sum_{\mathbf{Z}\in\mathcal{S}}\ \sum_{n=1}^{N}\left\{\mathbf{x}_n\mathbf{z}_n{}^{\mathrm{T}} - \mathbf{W}_{\mathrm{MLE}}\ \mathbf{z}_n\mathbf{z}_n{}^{\mathrm{T}}\right\} \\
\implies \mathbf{W}_{\mathrm{MLE}} &= \left(\sum_{\mathbf{Z}\in\mathcal{S}}\ \sum_{n=1}^{N}\mathbf{x}_n\mathbf{z}_n{}^{\mathrm{T}}\right)\left(\sum_{\mathbf{Z}\in\mathcal{S}}\ \sum_{n=1}^{N}\mathbf{z}_n\mathbf{z}_n{}^{\mathrm{T}}\right)^{-1} \\
\implies \mathbf{W}_{\mathrm{MLE}} &= \left(\sum_{n=1}^{N}\mathbf{x}_n\mathbb{E}\left[\,\mathbf{z}_n\,\right]^{\mathrm{T}}\right)\left(\sum_{n=1}^{N}\mathbb{E}\left[\,\mathbf{z}_n\mathbf{z}_n{}^{\mathrm{T}}\,\right]\right)^{-1} \\
\implies \mathbf{W}_{\mathrm{MLE}} &= \left(\sum_{n=1}^{N}\mathbf{x}_n\mathbb{E}\left[\,\mathbf{z}_n\,\right]^{\mathrm{T}}\right)\mathbb{E}\left[\,\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\,\right]^{-1}
\end{aligned}$$

3. **MLE of β**

$$\begin{aligned}
0 &= \left.\frac{\delta\left(\mathbb{E}\left[\,\mathrm{CLL}\,\right]\right)}{\delta\left(\beta\right)}\right|_{\beta=\beta_{\mathrm{MLE}}} \\
\implies 0 &= \sum_{\mathbf{Z}\in\mathcal{S}}\ \sum_{n=1}^{N}\left\{\frac{D}{2\beta} - \frac{1}{2}\left(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n\right)^{\mathrm{T}}\left(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n\right)\right\} \\
\implies \beta_{\mathrm{MLE}} &= ND\cdot|\,\mathcal{S}\,|\left(\sum_{\mathbf{Z}\in\mathcal{S}}\sum_{n=1}^{N}\left(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n\right)^{\mathrm{T}}\left(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n\right)\right)^{-1} \\
\implies \beta_{\mathrm{MLE}} &= ND\cdot\left(\sum_{n=1}^{N}\mathbb{E}\left[\left(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n\right)^{\mathrm{T}}\left(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n\right)\right]\right)^{-1} \\
\implies \beta_{\mathrm{MLE}} &= ND\cdot\left(\sum_{n=1}^{N}\left\{\mathbf{x}_n{}^{\mathrm{T}}\mathbf{x}_n + -2\mathbf{x}_n{}^{\mathrm{T}}\mathbf{W}_{\mathrm{MLE}}\mathbb{E}\left[\,\mathbf{z}_n\,\right] + \mathbb{E}\left[\,\mathbf{z}_n{}^{\mathrm{T}}\mathbf{W}_{\mathrm{MLE}}{}^{\mathrm{T}}\mathbf{W}_{\mathrm{MLE}}\mathbf{z}_n\,\right]\right\}\right)^{-1} \\
\implies \beta_{\mathrm{MLE}} &= ND\cdot\left(\sum_{n=1}^{N}\left\{\mathbf{x}_n{}^{\mathrm{T}}\mathbf{x}_n + -2\mathbf{x}_n{}^{\mathrm{T}}\mathbf{W}_{\mathrm{MLE}}\mathbb{E}\left[\,\mathbf{z}_n\,\right] + \operatorname{Tr}\left(\mathbb{E}\left[\,\mathbf{z}_n\mathbf{z}_n{}^{\mathrm{T}}\,\right]\mathbf{W}_{\mathrm{MLE}}{}^{\mathrm{T}}\mathbf{W}_{\mathrm{MLE}}\right)\right\}\right)^{-1}
\end{aligned}$$

I have generalised the complete algorithm in Algorithm 4

Algorithm 4: EM Algorithm for Subset Sum Model

Result: Posterior Distribution $\mathbb{P}\left[\mathbf{Z} \,\middle|\, \mathbf{X}, \mathbf{W}, \beta, \boldsymbol{\pi}\right]$, MLE estimates $\{\mathbf{w}_{\text{MLE}}, \beta_{\text{MLE}}, \boldsymbol{\pi}_{\text{MLE}}\}$
Steps:

1. Initialize the value of $\mathbf{w}$ as $\mathbf{w}^{(0)}$

2. For t = 1, 2 ... T or until convergence, do

   (a) **E Step:** Given $\{\mathbf{W}, \beta, \boldsymbol{\pi}\}^{(t-1)}$, approximate the posterior as a set of samples from the posterior using Algorithm 3 *i.e.*

   $$\mathcal{S}^{(t)} = \mathsf{GibbsSampler}\left(\mathbf{W}^{(t-1)}, \beta^{(t-1)}, \boldsymbol{\pi}^{(t-1)}\right) \sim \mathbb{P}\left[\mathbf{Z} \,\middle|\, \mathbf{W}^{(t-1)}, \beta^{(t-1)}, \boldsymbol{\pi}^{(t-1)}\right]$$

   Then approximate the expectations required as follows

   $$\begin{aligned} \forall\, n \in [\,N\,], \quad \mathbb{E}\left[\mathbf{z}_n^{(t)}\right] &= \frac{1}{|\,\mathcal{S}\,|} \sum_{\mathbf{Z} \in \mathcal{S}} \mathbf{z}_n \\ \forall\, n \in [\,N\,], \quad \mathbb{E}\left[\mathbf{z}_n^{(t)} {\mathbf{z}_n^{(t)}}^{\mathrm{T}}\right] &= \frac{1}{|\,\mathcal{S}\,|} \sum_{\mathbf{Z} \in \mathcal{S}} \mathbf{z}_n {\mathbf{z}_n}^{\mathrm{T}} \\ \forall\, n \in [\,N\,], \quad \mathbb{E}\left[\mathbf{Z}^{(t)} \mathbf{Z}^{\mathrm{T}(t)}\right] &= \sum_{n=1}^{N} \mathbb{E}\left[\mathbf{z}_n {\mathbf{z}_n}^{\mathrm{T}}\right] \end{aligned}$$

   (b) **M Step:** Compute the MLE solutions of $\{\mathbf{W}, \beta, \boldsymbol{\pi}\}$ as

   $$\begin{aligned} \forall\, k \in [\,K\,], \quad \pi_k^{(t)} &= \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}\left[\mathbf{z}_n^{(t)}\right]_k \\ \mathbf{W}^{(t)} &= \left(\sum_{n=1}^{N} \mathbf{x}_n \mathbb{E}\left[\mathbf{z}_n^{(t)}\right]^{\mathrm{T}}\right) \mathbb{E}\left[\mathbf{Z}^{(T)} {\mathbf{Z}^{(t)}}^{\mathrm{T}}\right]^{-1} \\ \left(\beta^{(t)}\right)^{-1} &= \frac{1}{ND} \sum_{n=1}^{N} \left\{ {\mathbf{x}_n}^{\mathrm{T}} \mathbf{x}_n + -2 {\mathbf{x}_n}^{\mathrm{T}} \mathbf{W}^{(t)} \mathbb{E}\left[\mathbf{z}_n^{(t)}\right] + \operatorname{Tr}\left(\mathbb{E}\left[\mathbf{z}_n^{(t)} {\mathbf{z}_n^{(t)}}^{\mathrm{T}}\right] {\mathbf{W}^{(t)}}^{\mathrm{T}} \mathbf{W}^{(t)}\right) \right\} \end{aligned}$$

3. Return $\mathbb{P}\left[\mathbf{Z} \,\middle|\, \mathbf{X}, \mathbf{W}, \beta, \boldsymbol{\pi}\right]$ as a set of samples $\mathcal{S}$, and $\{\mathbf{W}, \beta, \boldsymbol{\pi}\}_{\text{MLE}} = \{\mathbf{W}, \beta, \boldsymbol{\pi}\}^{(t')}$, where t' can represent the best value, computed over the average values or the last value (T), as per our choice.

References

Kaare Brandt Petersen and Michael Syskind Pedersen. *Matrix Cookbook*. 2012. URL `http://www.matrixcookbook.com`.

Truncated Gaussian Distribution. Wikipedia, The Free Encyclopedia, 2018. URL `https://en.wikipedia.org/wiki/Truncated_normal_distribution`.