Indian Institute of Technology Kanpur
CS772: Probabilistic Machine Learning, 17–18

ASSIGNMENT 1

Student Name: Gurpreet Singh
Roll Number: 150259
Date: October 25, 2017

Question 1

The conditional distribution of $\mathbf{X}$ is given below

$$X_{nm} \quad \sim \quad \text{Poisson}\left(\mathbf{u}_n^T \mathbf{v}_m\right) \qquad \forall\, n \in [N], \forall\, m \in [M] \tag{1}$$

The priors on the latent variables are given as follows

$$u_{nk} \quad = \quad \text{Gamma}\left(a_u, b_u\right) \tag{2}$$

$$v_{mk} \quad = \quad \text{Gamma}\left(a_v, b_v\right) \tag{3}$$

Note *I have used the shape-rate representation as opposed to the question*

Using these, we can write the conditional posteriors for all latent variables. However, we require to define another set of latent variables

$\forall\, n, m$ and $\forall\, k \in [K]$, define X_{nmk} such that

$$X_{nmk} \sim \text{ Poisson}\left(u_{nk} v_{mk}\right) \tag{4}$$

$$\sum_{k=1}^{K} X_{nmk} = X_{nm}$$

Local Conditional Distributions

Distribution for X_{nmk}
We have already been provided with the conditional distrubution of this latent variable as a multinomial distribution

$$\left\{X_{nmk}\right\}_{k=1}^{K} \,\middle|\, X_{nm} \sim Multinomial\left(X_{nm}, \left\{\eta_{nmk}\right\}_{k=1}^{K}\right) \tag{5}$$

keeping all other latent variables fixed, it is easy to see that the best point estimate of η_{nmk} is

$$\eta_{nmk} = \frac{u_{nk} v_{mk}}{\mathbf{u}_n^T \mathbf{v}_m} \tag{6}$$

Distribution of u_{nk}
We can write the local conditional of u_{nk}

Note *I will be removing the terms that are independent of u_{nk} as they can be considered constants when we are considering proportionality*

$$
\begin{aligned}
p\left(u_{nk} \,\middle|\, \{X_{nmk}\}_{m=1}^{M}, \{v_{mk}\}_{m=1}^{M}\right) &\propto p\left(\{X_{nmk}\}_{m=1}^{M} \,\middle|\, u_{nk}, \{v_{mk}\}_{m=1}^{M}\right) p\left(u_{nk}\right) \\
&= \prod_{m=1}^{M} \left(p\left(X_{nmk} \,\middle|\, u_{nk}, v_{mk}\right)\right) p\left(u_{nk}\right) \\
&= \prod_{m=1}^{M} \left(e^{-u_{nk} v_{mk}} \frac{\left(u_{nk} v_{mk}\right)^{X_{nmk}}}{\left(X_{nmk}\right)!}\right) u_{nk}^{a_u - 1} e^{-b_u u_{nk}} \\
&= e^{-\left(b_u + \sum_{m=1}^{M} v_{mk}\right)} u_{nk}^{\left(a_u + \sum_{m=1}^{M} X_{nmk}\right) - 1}
\end{aligned}
$$

This has the same representation as that of a gamma distribution. Hence, upon noramlization, we get

$$
p\left(u_{nk} \,\middle|\, \mathbf{X}, \mathbf{v}\right) = \text{Gamma}\left(a_u + \sum_{m=1}^{M} X_{nmk}, b_u + \sum_{m=1}^{M} v_{mk}\right) \tag{7}
$$

Distribution of v_{mk}
This can be solved the exact same way as the distribution for the l.v. u_{nk}. Here, we get

$$
p\left(v_{mk} \,\middle|\, \mathbf{X}, \mathbf{u}\right) = \text{Gamma}\left(a_v + \sum_{n=1}^{N} X_{nmk}, b_v + \sum_{n=1}^{N} u_{nk}\right) \tag{8}
$$

Algorithm for Gibbs Sampling

Algorithm 1: Gibbs Sampling for Count Matrix Factorization

Let T_b be the number of samples to be burned and T, the number of samples required. Hence we iterate $T + T_b$ times and save the last T samples.

- Initialize $\{u_{nk}\}, \{v_{mk}\}$ (We can also initialize $\{X_{nmk}\}$, however it is redundant, and the samples will be replaced in the first iteration itself)
- Do till sufficient samples acquired
  1. $\forall\, n \in [N]$, $\forall\, m \in [M]$ and $\forall\, k \in [K]$ sample $\{X_{nmk}\}$ using equation 5
  2. $\forall\, n \in [N]$ and $\forall\, k \in [K]$ sample $\{u_{nk}\}$ using equation 7 and using the current samples of $\{X_{nmk}\}$
  3. $\forall\, m \in [M]$ and $\forall\, k \in [K]$ sample $\{v_{mk}\}$ using equation 8 and using the current samples of $\{X_{nmk}\}$ and $\{u_{nk}\}$

Sampling from Multinomial To find one sample from a multinomial distribution (say Multinomial $\left(x; \{\eta_k\}_{k \in L}\right)$), we can do the following

1. Sample $u \sim Unif\,(0, 1)$
2. Find min K' such that $\sum_{k=1}^{K'} \geq u$ and add 1 to $x_{K'}$

Question 2

Plots

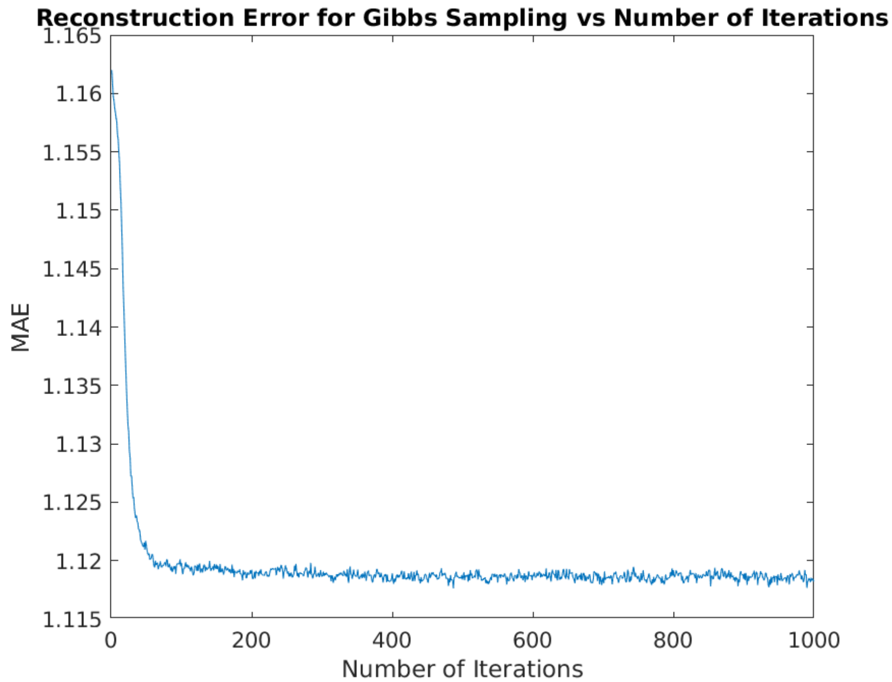

Figure 1: Reconstruction error using Gibbs Sampling

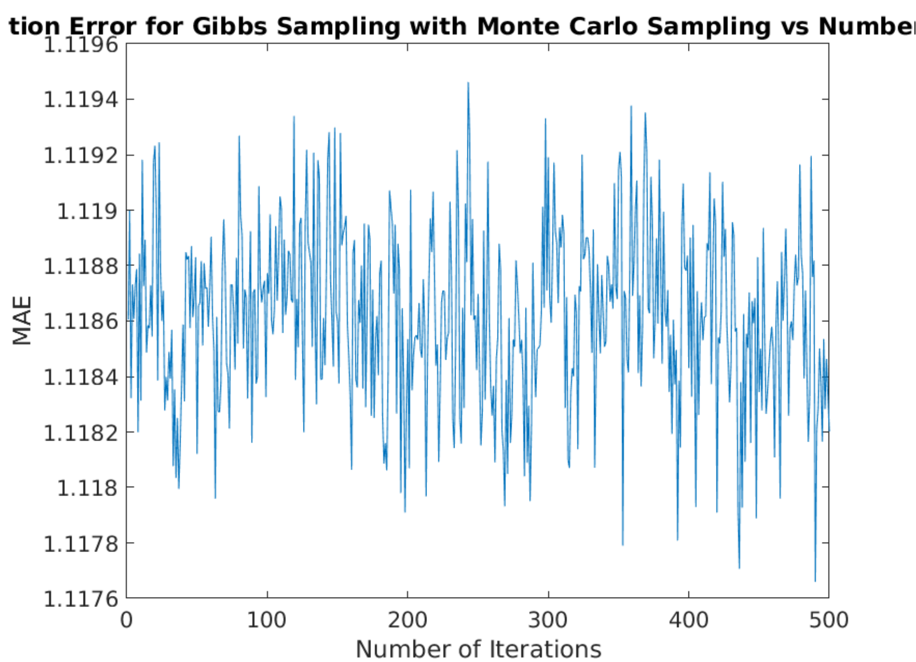

Figure 2: Reconstruction error using Gibbs Sampling along with Monte Carlo estimation

Observed Clusters of Words

Cluster ID	Words
1	windows window file problem program files run display running screen color set line work graphics version application image sun mit
2	people gun state law government bill writes rights fire guns article make police laws fbi crime don weapons clinton carry
3	car good bike dod cars article don engine buy front acs drive road ride driving ohio riding speed bmw miles
4	key system chip encryption keys government public security clipper information phone secure secret communications nsa data message escrow algorithm bit
5	writes pitt disease ve gordon medical blood read don banks steve pain food treatment medicine study normal patients effect effects
6	interested price sale offer send mail sell shipping call condition ll list email original excellent selling sound includes deal make
7	game year team play games win season hockey players league teams time baseball player san good red division won playing
8	card drive system disk computer video bit mac memory speed software dos hard board problem ram mb drives controller apple
9	writes article apr don cs david ca netcom org steve opinions cc deleted csd mine isc ve good disclaimer stanford
10	net wrote apple ve kent heard cheers sandvik don activities tom private ca great write cs lee dave find article
11	israel article writes israeli people org arab jews opinions don jewish number state peace research virginia palestinian war arabs israelis
12	information number list send mail time program read note data questions info include software code ftp general address free based
13	god people point fact true life jesus christian question things make time read christians word world man good bible person
14	university mail fax internet phone michael email computer bitnet information advance research gatech uk ac georgia tel department technology institute
15	space gov nasa high power earth work low cost sun toronto ground long access pat orbit moon henry small system
16	time back good make didn lot things don thing put people years long bad ve doesn ago point give ll
17	writes keith cwru caltech jon sgi usenet time freenet john cleveland good po find bob don start ago past internet
18	max cs hp part ca fi st de du ed ms ac tm ve bu ut ma al article se
19	writes article uiuc news andrew cso mike cmu uchicago frank hear robert post don curious day michael guy duke isn
20	people years world government country today american states history turkish state children president war military political united press national closed

Question 3

The acceptance rate for different variances are given as below

Variance	Acceptance Rate
0.01	0.973047
0.01	0.359208
0.01	0.007248

The means and variances approximated using the samples generated are as follows

Variance	Estimated Mean	Estimated Variance
0.01	4.1833, 4.0969	5.6557, 5.6262; 5.6262, 6.1487
0.01	4.0121, 4.0205	0.57950, 0.45151; 0.45151, 0.57677
0.01	4.0019, 4.0069	16.595, 16.487; 16.487, 16.632

Looking at the data, we can clearly say that $\sigma^2 = \mathbf{1}$ is the best value we can obtain, both from the perspective of estimated $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as well as the acceptance ratio. We do not need a very low acceptance ratio, as it reduces the diffusion rate a lot, also, very small variance causes a lot of error (Discussed in class).

Figure 3: Samples obtained using MH Sampling with Variance = 0.01

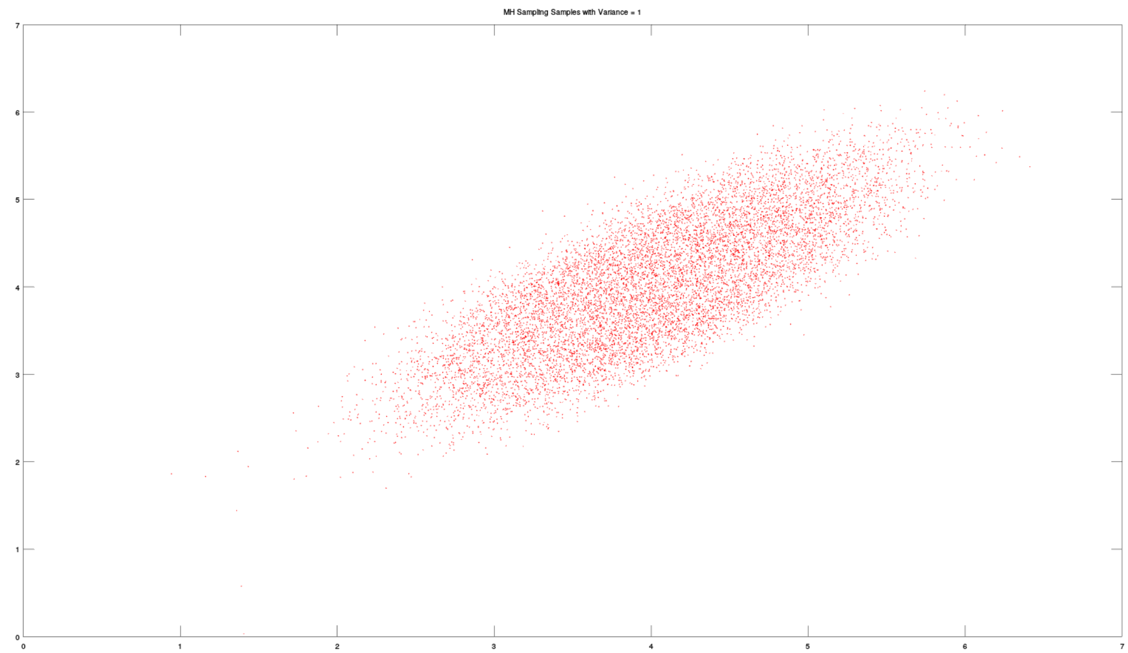

Figure 4: Samples obtained using MH Sampling with Variance = 1

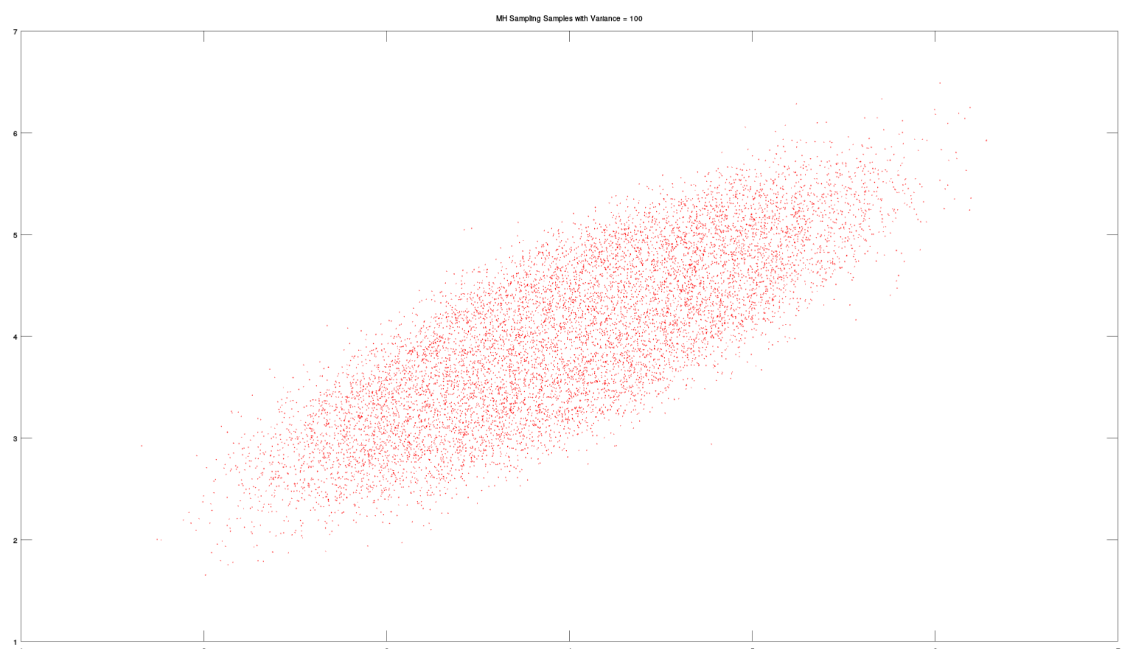

Figure 5: Samples obtained using MH Sampling with Variance = 100

Question 4

Notations

$\mathbf{X}$	Complete set of Data
$\mathbf{X}_{obs}$	Set of Observed Data
$\mathbf{X}_{miss}$	Set of Missing Data
$\mathbf{z}$	Set of Cluster Assignments for all samples
$\boldsymbol{\mu}$	Set of Cluster Means, with each cluster mean represented as $\boldsymbol{\mu}_k$
$\boldsymbol{\Sigma}$	Set of Cluster Covariances, with each cluster covariance represented as $\boldsymbol{\Sigma}_k$
$\boldsymbol{\pi}$	Set of Probabilities of Cluster Assignments

Complete Log Likelihood

Since we require to also model the missing data, we can treat these missing data samples as latent variables. Hence we need to compute the expectations with respect to both $\mathbf{X}_{miss}$ and $\mathbf{z}$.

In order to compute the CLL, we first need to find the joint distribution of $\mathbf{X}_{obs}, \mathbf{X}_{miss}, \mathbf{z}$

$$
\begin{aligned}
\mathbf{x}^n \,\big|\, z^n = k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k &\sim \mathcal{N}\left(\mathbf{x}^n \,\big|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \\
\mathbf{x}^n \,\big|\, z^n, \boldsymbol{\mu}, \boldsymbol{\Sigma} &\sim \prod_{k=1}^{K} \left(\mathcal{N}\left(\mathbf{x}^n \,\big|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right)^{\mathbb{I}[z^n=k]} \\
\mathbf{x}^n, z^n \,\big|\, \boldsymbol{\mu}, \boldsymbol{\Sigma} &\sim \prod_{k=1}^{K} \left(\pi_k \mathcal{N}\left(\mathbf{x}^n \,\big|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right)^{\mathbb{I}[z^n=k]} \\
\mathbf{X}, \mathbf{z} \,\big|\, \boldsymbol{\mu}, \boldsymbol{\Sigma} &\sim \prod_{n=1}^{N}\prod_{k=1}^{K} \left(\pi_k \mathcal{N}\left(\mathbf{x}^n \,\big|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right)^{\mathbb{I}[z^n=k]}
\end{aligned}
$$

Hence, we can write the CLL (removing terms independent of the latent variables) as

$$
\begin{aligned}
CLL &= log\left(p\left(\mathbf{X}, \mathbf{z} \,\big|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\right) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{I}\left[z^n = k\right] \left[log\left(\pi_k\right) - \frac{1}{2}\left(\mathbf{x}^n - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}^n - \boldsymbol{\mu}_k\right) - \frac{1}{2} log\left(\|\boldsymbol{\Sigma}_k\|\right)\right]
\end{aligned}
$$

We need to compute the expectation of CLL wrt to the joint posterior of $\mathbf{X}_{miss}$ and $\mathbf{z}$. Therefore, we can write the Expected CLL as

$$
\begin{aligned}
\mathbb{E}\left[CLL\right] &= \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\left(log\left(\pi_k\right) - \frac{1}{2}\left(\mathbf{x}^n - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}^n - \boldsymbol{\mu}_k\right) - \frac{1}{2} log\left(\|\boldsymbol{\Sigma}_k\|\right)\right)\right] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \left(\mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\right] log\left(\pi_k\right) - \frac{1}{2}\mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\left(\mathbf{x}^n - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}^n - \boldsymbol{\mu}_k\right)\right] - \frac{1}{2}\mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\right] log\left(\|\boldsymbol{\Sigma}\|\right)\right)
\end{aligned}
$$

where the second term can be written as

$$
\begin{aligned}
\left(\mathbf{x}^n - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}^n - \boldsymbol{\mu}_k\right) &= \mathrm{Tr}\left(\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}^n - \boldsymbol{\mu}_k\right)\left(\mathbf{x}^n - \boldsymbol{\mu}_k\right)^T\right) \\
\implies \mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\left(\mathbf{x}^n - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}^n - \boldsymbol{\mu}_k\right)\right] &= \mathrm{Tr}\left(\boldsymbol{\Sigma}_k^{-1}\,\mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\left(\mathbf{x}^n - \boldsymbol{\mu}_k\right)\left(\mathbf{x}^n - \boldsymbol{\mu}_k\right)^T\right]\right)
\end{aligned}
$$

We can further distribute the expectation as follows

$$
\left(\mathbf{x}^n - \boldsymbol{\mu}_k\right)\left(\mathbf{x}^n - \boldsymbol{\mu}_k\right)^T = \mathbf{x}^n \mathbf{x}^{nT} - 2\mathbf{x}^n \boldsymbol{\mu}_k^T + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T
$$

Using the trick from the midsem solutions, we can write $\mathbf{x}^n as \left[\mathbf{x}^n_{obs}, \mathbf{x}^n_{miss}\right]$, and following the same pattern, we can say that we require the following three expectations

1. $\mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\right]$
2. $\mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\mathbf{x}^n_{miss}\right]$
3. $\mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\mathbf{x}^n_{miss}{\mathbf{x}^n_{miss}}^T\right]$

E Step — Expectation Calculation

Expectation — $\mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\right]$

$$
\begin{aligned}
\mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\right] &= \sum_{k'=1}^{K} \int \mathbb{I}\left[k' = k\right]\, p\left(z^n = k', \mathbf{x}^n_{miss} \,\middle|\, \mathbf{x}^n_{obs}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) d\left(\mathbf{x}^n_{miss}\right) \\
&= \int p\left(z^n = k, \mathbf{x}^n_{miss} \,\middle|\, \mathbf{x}^n_{obs}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) d\left(\mathbf{x}^n_{miss}\right) \\
&= p\left(z^n = k \,\middle|\, \mathbf{x}^n_{obs}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \\
&\propto p\left(z^n = k\right)\, p\left(\mathbf{x}^n_{obs} \,\middle|\, z^n = k, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \\
&= \pi_k\, p\left(\mathbf{x}^n_{obs} \,\middle|\, z^n = k, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)
\end{aligned}
$$

Note *This is a marginal distribution wrt a multivariate Gaussian and can be easily computed*

$$
\begin{aligned}
\mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\right] &\propto \pi_k\, \mathcal{N}\left(\mathbf{x}^n_{obs} \,\middle|\, \boldsymbol{\mu}^n_k, \boldsymbol{\Sigma}^n_k\right) \\
&= \frac{\pi_k\, \mathcal{N}\left(\mathbf{x}^n_{obs} \,\middle|\, \boldsymbol{\mu}^n_k, \boldsymbol{\Sigma}^n_k\right)}{\sum_{k'=1}^{K} \pi_{k'}\, \mathcal{N}\left(\mathbf{x}^n_{obs} \,\middle|\, \boldsymbol{\mu}^n_{k'}, \boldsymbol{\Sigma}^n_{k'}\right)}
\end{aligned}
$$

Note $\boldsymbol{\mu}^n_k$ *represents the marginal mean for the observed data sample* $\mathbf{x}^n$*, and* $\boldsymbol{\mu}^{-n}_k$ *represents the marginal mean for the missing part. Similarly defined for* $\boldsymbol{\Sigma}$

Expectation — $\mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\mathbf{x}^n_{miss}\right]$

$$
\begin{aligned}
\mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\mathbf{x}^n_{miss}\right] &= \sum_{k'=1}^{K} \int \mathbb{I}\left[k' = k\right] \mathbf{x}^n_{miss}\, p\left(z^n = k', \mathbf{x}^n_{miss} \,\middle|\, \mathbf{x}^n_{obs}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) d\left(\mathbf{x}^n_{miss}\right) \\
&= \int \mathbf{x}^n_{miss}\, p\left(z^n = k, \mathbf{x}^n_{miss} \,\middle|\, \mathbf{x}^n_{obs}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) d\left(\mathbf{x}^n_{miss}\right) \\
&= p\left(z^n = k \,\middle|\, \mathbf{x}_{obs}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \int \mathbf{x}^n_{miss}\, p\left(\mathbf{x}^n_{miss} \,\middle|\, z^n = k, \mathbf{x}^n_{obs}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) d\left(\mathbf{x}^n_{miss}\right)
\end{aligned}
$$

This is the expected value of $\mathbf{x}^n_{obs}$ conditioned on $\mathbf{x}^n_{miss}$. This can be easily computed using the Gaussian properties. We represent this as $\boldsymbol{\mu}_k^{-n \,|\, n}$

$$
\mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\mathbf{x}^n_{miss}\right] = \mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\right] \mu_k^{-n \,|\, n}
$$

Expectation — $\mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\mathbf{x}^n_{miss}{\mathbf{x}^n_{miss}}^T\right]$

$$
\begin{aligned}
\mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\mathbf{x}^n_{miss}{\mathbf{x}^n_{miss}}^T\right] &= \sum_{k'=1}^{K} \int \mathbb{I}\left[k' = k\right] \mathbf{x}^n_{miss}{\mathbf{x}^n_{miss}}^T\, p\left(z^n = k', \mathbf{x}^n_{miss} \,\middle|\, \mathbf{x}^n_{obs}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) d\left(\mathbf{x}^n_{miss}\right) \\
&= p\left(z^n = k \,\middle|\, \mathbf{x}^n_{obs}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \int \mathbf{x}^n_{miss}{\mathbf{x}^n_{miss}}^T\, p\left(\mathbf{x}^n_{miss} \,\middle|\, z^n = k, \mathbf{x}^n_{obs}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) d\left(\mathbf{x}^n_{miss}\right) \\
&= \mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\right] \left(\boldsymbol{\mu}_k^{-n \,|\, n} {\boldsymbol{\mu}_k^{-n \,|\, n}}^T + \boldsymbol{\Sigma}_k^{-n \,|\, n}\right)
\end{aligned}
$$

M Step — Maximization Step

MLE estimate of $\boldsymbol{\pi}$

$$
\begin{aligned}
\hat{\boldsymbol{\pi}} &= \arg\max_{\boldsymbol{\pi}} \mathbb{E}\left[CLL\right] \\
&\quad \text{s.t.} \sum_{k=1}^{K} \pi_k \;=\; 1 \\
&= \arg\max_{\boldsymbol{\pi}} \mathbb{E}\left[CLL\right] + \lambda \left(1 - \sum_{k=1}^{K}\right) \\
\implies \frac{\delta\left(\mathbb{E}\left[CLL\right]\right)}{\delta\left(\pi_k\right)} &= \lambda \qquad \forall\, k \in [K] \\
\implies \hat{\pi_k} &= \frac{\sum_{n=1}^{N} \mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\right]}{\lambda}
\end{aligned}
$$

Using the constraint on $\boldsymbol{\pi}$, we get

$$
\hat{\pi_k} = \frac{\sum_{n=1}^{N} \mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\right]}{N}
$$

MLE estimate of $\boldsymbol{\mu}$

$$
\begin{aligned}
\hat{\boldsymbol{\mu}} &= \arg\max_{\boldsymbol{\mu}} \mathbb{E}\left[CLL\right] \\
\frac{\delta\left(\mathbb{E}\left[CLL\right]\right)}{\delta\left(\boldsymbol{\mu}_k\right)} &= -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}\left[\mathbb{I}\left[z^n = k\right] \frac{\delta\left(\left(\mathbf{x}^n - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}^n - \boldsymbol{\mu}_k\right)\right)}{\delta\left(\boldsymbol{\mu}_k\right)}\right] && = 0 \\
\implies -\frac{1}{2}\sum_{n=1}^{N} \mathbb{E}\left[\mathbb{I}\left[z^n = k\right] 2\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}^n - \boldsymbol{\mu}_k\right)\right] & && = 0 \\
\implies \hat{\boldsymbol{\mu}_k} &= \frac{\sum_{n=1}^{N} \mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\mathbf{x}^n\right]}{\sum_{n=1}^{N} \mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\right]}
\end{aligned}
$$

where

$$
\mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\mathbf{x}^n\right] = \begin{bmatrix} \mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\right]\mathbf{x}^n_{obs} \\ \mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\mathbf{x}_{miss}\right] \end{bmatrix}
$$

MLE estimate of $\boldsymbol{\Sigma}$

$$
\begin{aligned}
\hat{\boldsymbol{\Sigma}} &= \arg\max_{\boldsymbol{\Sigma}} \mathbb{E}\left[CLL\right] \\
\frac{\delta\left(\mathbb{E}\left[CLL\right]\right)}{\delta\left(\boldsymbol{\Sigma}\right)} &= \frac{1}{2}\sum_{n=1}^{N}\left(\boldsymbol{\Sigma}_k^{-2}\,\mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\left(\mathbf{x}^n - \boldsymbol{\mu}_k\right)\left(\mathbf{x}^n - \boldsymbol{\mu}_k\right)^T\right] - \boldsymbol{\Sigma}_k^{-1}\,\mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\right]\right) && = 0 \\
\implies \boldsymbol{\Sigma}_k &= \frac{\sum_{i=1}^{N} \mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\left(\mathbf{x}^n - \boldsymbol{\mu}_k\right)\left(\mathbf{x}^n - \boldsymbol{\mu}_k\right)^T\right]}{\sum_{n=1}^{N} \mathbb{E}\left[\mathbb{I}\left[z^n = k\right]\right]}
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbb{E}\left[\mathbb{I}\left[z^n=k\right]\left(\mathbf{x}^n-\boldsymbol{\mu}_k\right)\left(\mathbf{x}^n-\boldsymbol{\mu}_k\right)^T\right] &= \mathbb{E}\left[\mathbb{I}\left[z^n=k\right]\mathbf{x}^n\mathbf{x}^{nT}\right]-2\,\mathbb{E}\left[\mathbb{I}\left[z^n=k\right]\mathbf{x}^n\right]\mathbf{x}^{nT}+\boldsymbol{\mu}_k\boldsymbol{\mu}_k^T\,\mathbb{E}\left[\mathbb{I}\left[z^n=k\right]\right] \\
\mathbb{E}\left[\mathbb{I}\left[z^n=k\right]\mathbf{x}^n\mathbf{x}^{nT}\right] &= \begin{bmatrix} \mathbb{E}\left[\mathbb{I}\left[z^n=k\right]\right]\mathbf{x}^n_{obs}{\mathbf{x}^n_{obs}}^T & \mathbb{E}\left[\mathbb{I}\left[z^n=k\right]\mathbf{x}^n_{miss}\right]{\mathbf{x}^n_{obs}}^T \\ \mathbf{x}^n_{obs}\,\mathbb{E}\left[\mathbb{I}\left[z^n=k\right]\mathbf{x}^n_{miss}\right]^T & \mathbb{E}\left[\mathbb{I}\left[z^n=k\right]\mathbf{x}^n_{miss}{\mathbf{x}^n_{miss}}^T\right] \end{bmatrix}
\end{aligned}
$$