Gibbs Sampling for parameter estimation of Multivariate GMM

Hrishikesh Pable

February 2023

Contents

Chapter 1

Introduction

This paper contains the derivation of the conditional posteriors for Gibbs Sampling of a Multivariate Gaussian Mixture Model with k components. We consider a Multivariate Normal-Wishart prior over the mean and variance of each gaussian, and a dirichlet prior over the mixture proportions π.

Chapter 2

Notations

- T : Total number of dimensions of the data.
- N : Total number of data points.
- K : Total number of components of the Mixture Model.
- μ : $\mu_1, \mu_2, \ldots, \mu_K$
- Λ : $\Lambda_1, \Lambda_2, \ldots, \Lambda_K$
- μ_k : The T mean vector of the kth Mixture Component.
- Λ_k : The $T * T$ precision matrix (i.e. the inverse of the covariance matrix) of the kth Mixture Component.
- X : Matrix of N data points. size $= N * T$.
- x_i : ith data point i.e. a vector of length T.
- m_k : The mean vector (length=T) of the gaussian part of the Multivariate Normal-Wishart prior over the mean and variance of the kth mixture component.
- G_k : The $T * T$ precision matrix of the gaussian part of the Multivariate Normal-Wishart prior over the mean and variance of the kth mixture component. $G_k = p_k * \Lambda_k$.
- q_k, V_k : Hyperparameters of the Wishart part of the Multivariate Normal-Wishart prior over the mean and variance of the kth component. q_k is a vector of length T and V_k is a matrix of size $T * T$
- Z : Vector of length N consisting of cluster assignments of the N data points.
- z_i : The mixture component to which the ith data point belongs.
- $z_{i,k}$: The indicator variable $\mathrm{I}(z_i = k)$.
- n_k : The number of data points belonging to the kth cluster. $n_k = \sum_{i=1}^{n} z_{i,k}$.
- π : $\pi_1, \pi_2, \ldots, \pi_K$ The mixture proportions corresponding each of the K clusters.
- β : The hyperparameter of the Dirichlet prior over the mixing proportions π.

Chapter 3

Graphical Model of Gaussian Mixture Model

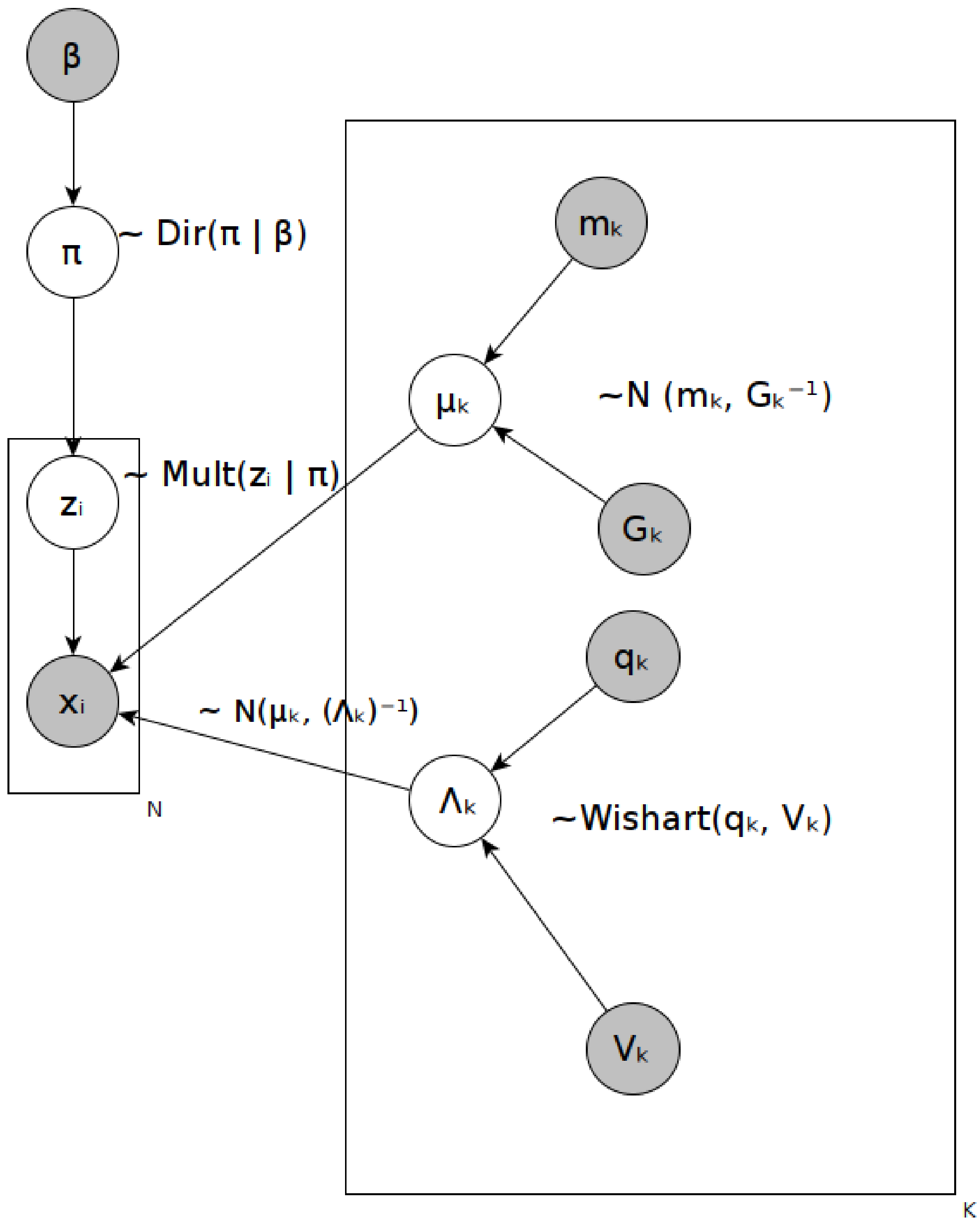

Figure 3.1: Gaussian Mixture Model

Chapter 4

Gibbs Sampling Derivation

We need the following 4 conditional distributions:

1. $P(Z|X, \mu, \Lambda, \pi)$
2. $P(\mu|X, Z, \Lambda, \pi)$
3. $P(\Lambda|X, Z, \mu, \pi)$
4. $P(\pi|X, Z, \mu, \Lambda)$

First we find $P(Z|X, \mu, \Lambda, \pi)$.

$$P(Z|X, \mu, \Lambda, \pi) = \frac{P(Z, X, \mu, \Lambda, \pi)}{P(X, \mu, \Lambda, \pi)}$$

Now, we can factorise the joint distribution by using the above bayesian network as follows:

$$\begin{aligned} P(\pi, Z, X, \mu, \Lambda) &= P(\pi|\beta)P(Z|\pi)P(X|Z, \mu, \Lambda)P(\mu, \Lambda) \\ &= P(\pi|\beta)P(Z|\pi)P(X|Z, \mu, \Lambda)P(\mu|m, G)P(\Lambda|q, V) \\ &= Dir(\pi|\beta)\prod_{i=1}^{N} Mult(z_i|\pi)\prod_{i=1}^{N}\mathcal{N}\left(x_i|\mu_{z_i}, \Lambda_{z_i}\right)\prod_{k=1}^{K}\mathcal{N}\left(\mu_k|m_k, G_k\right)\prod_{k=1}^{K}\mathcal{W}_T(\Lambda_k|q_k, V_k) \end{aligned} \tag{4.1}$$

We can simplify these terms seperately as follows:

$$\begin{aligned} P(Z|\pi)P(\pi|\beta) &= \left(\prod_{i=1}^{N} Mult(z_i|\pi)\right)\Big(Dir(\pi|\beta)\Big) \\ &\propto \left(\prod_{i=1}^{N}\prod_{k=1}^{K}(\pi_k)^{z_{i,k}}\right)\left(\prod_{k=1}^{K}{\pi_k}^{\beta_k - 1}\right) \\ &\propto \left(\prod_{k=1}^{K}{\pi_k}^{\left(\sum_{i=1}^{N} z_{i,k}\right)}\right)\left(\prod_{k=1}^{K}{\pi_k}^{(\beta_k - 1)}\right) \\ &\propto \prod_{k=1}^{K}{\pi_k}^{\left(\sum_{i=1}^{N} z_{i,k} + \beta_k - 1\right)} \\ &\propto \prod_{k=1}^{K}{\pi_k}^{(n_k + \beta_k - 1)} \end{aligned}$$

$$
\begin{aligned}
P(X|Z,\mu,\Lambda) &= \prod_{i=1}^{N} P(x_i|\mu_{z_i},\Lambda_{z_i}) \\
&\propto \prod_{i=1}^{N} |\Lambda_{z_i}^{-1}|^{-1/2} \exp\left(\frac{-1}{2}(x-\mu_{z_i})^T\Lambda_{z_i}(x-\mu_{z_i})\right) \\
&\propto \left(\prod_{i=1}^{N} |\Lambda_{z_i}^{-1}|^{-1/2}\right)\left(\prod_{i=1}^{N} \exp\left(\frac{-1}{2}(x-\mu_{z_i})^T\Lambda_{z_i}(x-\mu_{z_i})\right)\right) \\
&\propto \left(\prod_{k=1}^{K} |\Lambda_{k}^{-1}|^{-n_k/2}\right)\left(\prod_{k=1}^{K}\prod_{i:z_{i,k}=1} \exp\left(\frac{-1}{2}(x-\mu_{z_i})^T\Lambda_{z_i}(x-\mu_{z_i})\right)\right) \\
&\propto \left(\prod_{k=1}^{K} |\Lambda_{k}^{-1}|^{-n_k/2}\right)\left(\prod_{k=1}^{K} \exp\left(\sum_{i:z_{i,k}=1}\frac{-1}{2}(x-\mu_{z_i})^T\Lambda_{z_i}(x-\mu_{z_i})\right)\right) \\
&\propto \left(\prod_{k=1}^{K} |\Lambda_{k}|^{n_k/2} \exp\left(\sum_{i:z_{i,k}=1}\frac{-1}{2}(x-\mu_{z_i})^T\Lambda_{z_i}(x-\mu_{z_i})\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
P(\mu,\Lambda) &= \prod_{k=1}^{K} P(\mu_k,\Lambda_k|m_k,G_k,q_k,V_k) \\
&= \prod_{k=1}^{K} P(\mu_k|m_k,G_k)P(\Lambda_k|q_k,V_k) \\
&\propto \prod_{k=1}^{K} |G_k|^{1/2} \exp\left(\frac{-1}{2}(\mu_k-m_k)^T G_k(\mu_k-m_k)\right)|\Lambda_k|^{(q_k-t-1)/2}\exp\left(-tr(V_k^{-1}\Lambda_k)/2\right) \\
&\propto \prod_{k=1}^{K} |p_k\Lambda_k|^{1/2} \exp\left(\frac{-1}{2}(\mu_k-m_k)^T p_k\Lambda_k(\mu_k-m_k)\right)|\Lambda_k|^{(q_k-t-1)/2}\exp\left(-tr(V_k^{-1}\Lambda_k)/2\right) \\
&\propto \prod_{k=1}^{K} |\Lambda_k|^{(q_k-t)/2} \exp\left(\frac{-1}{2}\left((\mu_k-m_k)^T p_k\Lambda_k(\mu_k-m_k)+tr(V_k^{-1}\Lambda_k)\right)\right)
\end{aligned}
$$

We can find $P(Z|\mu,\Lambda,\pi)$ by dividing the joint distribution in equation 4.1 by the same joint distribution marginalised over z. This would lead to the $Dir(\pi|\beta)$, $\mathcal{N}(\mu_k|m_k,G_k)$ and $\mathcal{W}(\Lambda_k|q_k,V_k)$ terms getting cancelled from the numerator and the denominator. So, we get,

$$
\begin{aligned}
P(Z|X,\mu,\Lambda,\pi) &= \frac{P(Z,X,\mu,\Lambda,\pi)}{\sum\limits_Z P(Z,X,\mu,\Lambda,\pi)} \\
&= \frac{\prod_{i=1}^{N} Mult(z_i|\pi)\prod_{i=1}^{N}\mathcal{N}(x_i|\mu_{z_i},\Lambda_{z_i})}{\sum\limits_Z\prod_{i=1}^{N} Mult(z_i|\pi)\prod_{i=1}^{N}\mathcal{N}(x_i|\mu_{z_i},\Lambda_{z_i})} \\
&= \frac{\left(\prod_{i=1}^{N}\prod_{k=1}^{K}(\pi_k)^{z_{i,k}}\right)\left(\prod_{k=1}^{K}|\Lambda_k|^{n_k/2}\exp\left(\sum_{i:z_{i,k}=1}\frac{-1}{2}(x-\mu_{z_i})^T\Lambda_{z_i}(x-\mu_{z_i})\right)\right)}{\sum_Z\left(\prod_{i=1}^{N}\prod_{k=1}^{K}(\pi_k)^{z_{i,k}}\right)\left(\prod_{k=1}^{K}|\Lambda_k|^{n_k/2}\exp\left(\sum_{i:z_{i,k}=1}\frac{-1}{2}(x-\mu_{z_i})^T\Lambda_{z_i}(x-\mu_{z_i})\right)\right)} \\
&\propto \frac{\left(\prod_{k=1}^{K}(\pi_k)^{\sum_{i=1}^{N} z_{i,k}}\right)\left(\prod_{k=1}^{K}|\Lambda_k|^{n_k/2}\exp\left(\sum_{i:z_{i,k}=1}\frac{-1}{2}(x-\mu_{z_i})^T\Lambda_{z_i}(x-\mu_{z_i})\right)\right)}{\sum_Z\left(\prod_{k=1}^{K}(\pi_k)^{\sum_{i=1}^{N} z_{i,k}}\right)\left(\prod_{k=1}^{K}|\Lambda_k|^{n_k/2}\exp\left(\sum_{i:z_{i,k}=1}\frac{-1}{2}(x-\mu_{z_i})^T\Lambda_{z_i}(x-\mu_{z_i})\right)\right)} \\
&\propto \frac{\left(\prod_{k=1}^{K}(\pi_k)^{\sum_{i=1}^{N} z_{i,k}}|\Lambda_k|^{n_k/2}\exp\left(\sum_{i:z_{i,k}=1}\frac{-1}{2}(x-\mu_{z_i})^T\Lambda_{z_i}(x-\mu_{z_i})\right)\right)}{\sum_Z\left(\prod_{k=1}^{K}(\pi_k)^{\sum_{i=1}^{N} z_{i,k}}|\Lambda_k|^{n_k/2}\exp\left(\sum_{i:z_{i,k}=1}\frac{-1}{2}(x-\mu_{z_i})^T\Lambda_{z_i}(x-\mu_{z_i})\right)\right)}
\end{aligned}
$$

Now, we will find the other three posterior conditionals. For that, we need to simplify the joint probability expression in equation 4.1 by putting together the individual simplifications that we made to the parts of the joint probability as follows:

$$
\begin{aligned}
P(\pi, Z, X, \mu, \Lambda) &= P(\pi|\beta)P(Z|\pi)P(X|\mu,\Lambda)P(\mu,\Lambda) \\
&= P(\pi|\beta)P(Z|\pi)P(X|Z,\mu,\Lambda)P(\mu|m,G)P(\Lambda|q_k,V_k) \\
&= Dir(\pi|\beta)\prod_{i=1}^{N} Mult(z_i|\pi)\prod_{i=1}^{N}\mathcal{N}\left(x_i|\mu_{z_i},\Lambda_{z_i}\right)\prod_{k=1}^{K}\mathcal{N}\left(\mu_k|m_k,G_k\right)\prod_{k=1}^{K}\mathcal{G}(\Lambda_k|q_k,V_k) \\
&\propto \left(\prod_{k=1}^{K}\pi_k{}^{(n_k+\beta_k-1)}\right)\left(\prod_{k=1}^{K}|\Lambda_k|^{n_k/2}\exp\left(\sum_{i:z_{i,k}=1}\frac{-1}{2}(x-\mu_{z_i})^T\Lambda_{z_i}(x-\mu_{z_i})\right)\right) \\
&\quad \left(\prod_{k=1}^{K}|\Lambda_k|^{(q_k-t)/2}\exp\left(\frac{-1}{2}\left((\mu_k-m_k)^T p_k\Lambda_k(\mu_k-m_k)+tr(V_k^{-1}\Lambda_k)\right)\right)\right) \\
&\propto \left(\prod_{k=1}^{K}\pi_k{}^{(n_k+\beta_k-1)}|\Lambda_k|^{(q_k+n_k-t)/2}\right) \\
&\quad \left(\prod_{k=1}^{K}\exp\left(\frac{-1}{2}\left(tr(V_k^{-1}\Lambda_k)+\sum_{i:z_{i,k}=1}(x-\mu_{z_i})^T\Lambda_{z_i}(x-\mu_{z_i})+(\mu_k-m_k)^T p_k\Lambda_k(\mu_k-m_k)\right)\right)\right)
\end{aligned}
$$

Now, let $I = tr(V_k^{-1}\Lambda_k)+\sum_{i:z_{i,k}=1}(x-\mu_{z_i})^T\Lambda_{z_i}(x-\mu_{z_i})+(\mu_k-m_k)^T p_k\Lambda_k(\mu_k-m_k)$

We need to bring it into the form $I = tr(V_k^{*-1}\Lambda)+(\mu_k-m_k^*)^T p_k^*\Lambda_k(\mu_k-m_k^*)$ to get the parameters of the posterior Multivariate Normal-Wishart distribution.

$$
\begin{aligned}
I &= tr(V_k^{-1}\Lambda_k)+\sum_{i:z_{i,k}=1}(x-\mu_{z_i})^T\Lambda_{z_i}(x-\mu_{z_i})+(\mu_k-m_k)^T p_k\Lambda_k(\mu_k-m_k) \\
&= tr(V_k^{-1}\Lambda_k)+\sum_{i:z_{i,k}=1}\left(x_i^T\Lambda_{z_i}x_i-\mu_{z_i}^T\Lambda_{z_i}x_i+\mu_{z_i}^T\Lambda_{z_i}\mu_{z_i}-x_i^T\Lambda_{z_i}\mu_{z_i}\right)+p_k\left(\mu_k^T{}_k\mu_k-m_k^T\Lambda_k\mu_k-\mu_k^T\Lambda_k m_k+m_k^T\Lambda_k m_k\right)
\end{aligned}
$$

Now, let the cholesky decomposition of Λ_k be:

$$\Lambda_k = U_k^T U_k$$

where U_k is an upper-triangular matrix. So, I can we rewritten as:

$$
\begin{aligned}
I &= tr(V_k^{-1}\Lambda_k)+\sum_{i:z_{i,k}=1}\left(x_i^T U_{z_i}^T U_{z_i}x_i-\mu_{z_i}^T\Lambda_{z_i}x_i+\mu_{z_i}^T U_{z_i}^T U_{z_i}\mu_{z_i}-x_i^T\Lambda_{z_i}\mu_{z_i}\right) \\
&\quad +p_k\left(\mu_k^T U_k^T U_k\mu_k-m_k^T\Lambda_k\mu_k-\mu_k^T\Lambda_k m_k+m_k^T U_k^T U_k m_k\right) \\
&= tr(V_k^{-1}\Lambda_k)+\sum_{i:z_{i,k}=1}\left(||U_{z_i}x_i||^2-\langle\Lambda_{z_i}x_i,\mu_{z_i}\rangle+||U_{z_i}\mu_{z_i}||^2-\langle\Lambda_{z_i}\mu_{z_i},x_i\rangle\right) \\
&\quad +p_k\left(||U_k\mu_k||^2-\langle\Lambda_k\mu_k,m_k\rangle-\langle\Lambda_k m_k,\mu_k\rangle+||U_k m_k||^2\right) \\
&= tr(V_k^{-1}\Lambda_k)+\sum_{i:z_{i,k}=1}\left(||U_k x_i||^2-2\langle\Lambda_k\mu_k,x_i\rangle+||U_k\mu_k||^2\right)+p_k\left(||U_k\mu_k||^2-2\langle\Lambda_k\mu_k,m_k\rangle+||U_k m_k||^2\right) \\
&= tr(V_k^{-1}\Lambda_k)+\sum_{i:z_{i,k}=1}||U_k x_i||^2-2\langle\Lambda_k\mu_k,\sum_{i:z_{i,k}=1}x_i\rangle+n_k||U_k\mu_k||^2+p_k||U_k\mu_k||^2-2\langle\Lambda_k\mu_k,p_k m_k\rangle+p_k||U_k m_k||^2 \\
&= tr(V_k^{-1}\Lambda_k)+\sum_{i:z_{i,k}=1}(||U_k x_i||^2)+p_k||U_k m_k||^2+(p_k+n_k)||U_k\mu_k||^2-2\langle\Lambda_k\mu_k,\sum_{i:z_{i,k}=1}(x_i)+p_k m_k\rangle \\
&= tr(V_k^{-1}\Lambda_k)+\sum_{i:z_{i,k}=1}(||U_k x_i||^2)+p_k||U_k m_k||^2+(p_k+n_k)\left(||U_k\mu_k||^2-2\langle\Lambda_k\mu_k,\frac{\bar{x_k}K+p_k m_k}{p_k+n_k}\rangle\right) \\
&= tr(V_k^{-1}\Lambda_k)+\sum_{i:z_{i,k}=1}(||U_k x_i||^2)+p_k||U_k m_k||^2+p_k^*\left(||U_k\mu_k||^2-2\langle\Lambda_k\mu_k,m_k^*\rangle+||U_k m_k^*||^2\right)-p_k^*||U_k m_k^*||^2 \\
&= tr(V_k^{-1}\Lambda_k)+\sum_{i:z_{i,k}=1}(||U_k x_i||^2)+p_k||U_k m_k||^2-p_k^*||U_k m_k^*||^2+p_k^*\left((\mu_k-m_k^*)^T\Lambda_k(\mu_k-m_k^*)\right)
\end{aligned}
$$

So, we have,

$$
\begin{aligned}
p_k^* &= p_k+n_k \\
m_k^* &= \frac{\bar{x_k}K+p_k m_k}{p_k+n_k} \\
G_k^* &= p_k^*\Lambda_k
\end{aligned}
$$

Also, by comparing the above expression with $I = tr(V_k^{*-1}\Lambda) + (\mu_k - m_k^*)^T p_k^* \Lambda_k (\mu_k - m_k^*)$, we get,

$$
\begin{aligned}
tr(V^{*-1}\Lambda_k) &= tr(V_k^{-1}\Lambda_k) + \sum_{i:z_{i,k}=1} (||U_k x_i||^2) + p_k ||U_k m_k||^2 - p_k^* ||U_k m_k^*||^2 \\
&= tr(V_k^{-1}\Lambda_k) + \sum_{i:z_{i,k}=1} (\langle \Lambda_k x_i, x_i \rangle) + p_k \langle \Lambda_k m_k, m_k \rangle - p_k^* \langle \Lambda_k m_k^*, m_k^* \rangle \\
&= tr(V_k^{-1}\Lambda_k) + \sum_{i:z_{i,k}=1} tr(x_i(\Lambda_k x_i)^T) + p_k tr(m_k(\Lambda_k m_k)^T) - p_k^* tr(m_k^*(\Lambda_k m_k^*)^T) \\
&= tr(V_k^{-1}\Lambda_k) + \sum_{i:z_{i,k}=1} tr(x_i x_i^T \Lambda_k^T) + p_k tr(m_k m_k^T \Lambda_k^T) - p_k^* tr(m_k^* m_k^{*T} \Lambda_k^T) \\
&= tr\left((V_k^{-1} + \sum_{i:z_{i,k}=1} (x_i x_i^T) + p_k m_k m_k^T + p_k^* m_k^* m_k^{*T})\Lambda_k \right)
\end{aligned}
$$

Hence, $V_k^{*-1} = V_k^{-1} + \sum_{i:z_{i,k}=1} (x_i x_i^T) + p_k(m_k m_k^T) + p_k^*(m_k^* m_k^{*T})$

We can now write the joint distribution as follows:

$$
\begin{aligned}
P(\pi, Z, X, \mu, \Lambda) \propto & \left(\prod_{k=1}^K \pi_k{}^{(n_k+\beta_k-1)} |\Lambda_k|^{(q_k+n_k-t)/2} \right) \\
& \left(\prod_{k=1}^K \exp\left(\frac{-1}{2} \left(tr(V_k^{-1}\Lambda_k) + \sum_{i:z_{i,k}=1} (x - \mu_{z_i})^T \Lambda_{z_i} (x - \mu_{z_i}) + (\mu_k - m_k)^T p_k \Lambda_k (\mu_k - m_k) \right) \right) \right) \\
\propto & \left(\prod_{k=1}^K \pi_k{}^{(n_k+\beta_k-1)} |\Lambda_k|^{(q_k+n_k-t)/2} \right) \\
& \left(\prod_{k=1}^K \exp\left(\frac{-1}{2} \left(tr(V_k^{*-1}\Lambda) + (\mu_k - m_k^*)^T p_k^* \Lambda_k (\mu_k - m_k^*) \right) \right) \right) \\
\propto & \prod_{k=1}^K \pi_k{}^{(\beta_k^*-1)} \times |p_k^* \Lambda_k|^{1/2} \exp\left(\frac{-1}{2} \left((\mu_k - m_k^*)^T p_k^* \Lambda_k (\mu_k - m_k^*) \right) \right) \times \\
& |\Lambda_k|^{(q_k^*-t-1)/2} \exp\left(tr(V_k^{*-1}\Lambda) \right) \\
\propto & \left(\prod_{k=1}^K \pi_k{}^{(\beta_k^*-1)} \right) \times \left(\prod_{k=1}^K |p_k^* \Lambda_k|^{1/2} \exp\left(\frac{-1}{2} \left((\mu_k - m_k^*)^T p_k^* \Lambda_k (\mu_k - m_k^*) \right) \right) \right) \times \\
& \left(\prod_{k=1}^K |\Lambda_k|^{(q_k^*-t-1)/2} \exp\left(tr(V_k^{*-1}\Lambda) \right) \right) \\
\propto & \left(Dir(\pi|\beta^*) \right) \left(\prod_{k=1}^K \mathcal{N}(\mu_k | m_k^*, G_k^{*-1}) \right) \left(\prod_{k=1}^K \mathcal{W}_T(\Lambda_k | q_k^*, V_k^*) \right)
\end{aligned}
$$

So, we finally have,

$$
\begin{aligned}
P(\pi|Z, X, \mu, \Lambda) &= \frac{P(\pi, Z, X, \mu, \Lambda)}{P(Z, X, \mu, \Lambda)} \\
&= \frac{P(\pi, Z, X, \mu, \Lambda)}{\sum_\pi P(\pi, Z, X, \mu, \Lambda)} \\
&= Dir(\pi|\beta^*)
\end{aligned}
$$

$$
\begin{aligned}
P(\mu|Z, X, \Lambda, \pi) &= \frac{P(\pi, Z, X, \mu, \Lambda)}{P(\pi, Z, X, \Lambda)} \\
&= \frac{P(\pi, Z, X, \mu, \Lambda)}{\sum_\mu P(\pi, Z, X, \mu, \Lambda)} \\
&= \prod_{k=1}^K \mathcal{N}(\mu_k | m_k^*, G_k^{*-1})
\end{aligned}
$$

$$
\begin{aligned}
P(\Lambda|Z, X, \mu, \pi) &= \frac{P(\pi, Z, X, \Lambda, \mu)}{P(\pi, Z, X, \mu)} \\
&= \frac{P(\pi, Z, X, \Lambda, \mu)}{\sum_\Lambda P(\pi, Z, X, \Lambda, \mu)} \\
&= \prod_{k=1}^K \mathcal{W}_T(\Lambda_k | q_k^*, V_k^*)
\end{aligned}
$$