Variational Approximation with Dirichlet Process

Daeyoung Lim*
Department of Statistics
Korea University

January 21, 2016

1 Dirichlet Process

A realization of the Dirichlet distribution is intuitively a pmf generated with parameter α. This suggests that the Dirichlet distribution is a probability distribution over pmfs. Put differently, the Dirichlet distribution samples from a bag filled with dice and each time it is sampled, it returns a die, so to speak. However, it also implies that the Dirichlet distribution is no more than a distribution over pmfs with finite sample spaces. We want this to cover infinite sample spaces.

The collection of all probability distributions over an infinite sample space is beyond control. For such a reason, the Dirichlet process restricts itself to a more manageable subset of the collection: discrete probability distributions over the infinite sample space that can be written as an infinite sum of weighted indicator functions.

1.1 Stick-breaking process

$$\begin{aligned} v_k|\alpha &\sim \text{Beta}\left(1,\alpha\right) & \eta_k^* &\sim H \\ \pi_k &= v_k\prod_{\ell=1}^{k-1}\left(1-v_\ell\right) & G &= \sum_{k=1}^{\infty}\pi_k\delta_{\eta_k^*} \end{aligned}$$

where $\delta_{\eta_k^*}$ is the *Dirac-delta* function centered around η_k^*. Then $G \sim \text{DP}\left(\alpha, H\right)$.

2 DP mixture model

Using the stick-breaking representation of DP, DP mixture is represented as follows:

1. Draw $v_k \sim \text{Beta}\left(1,\alpha\right)$.
2. Draw $\eta_k^* \sim H$.
3. For the n^{th} data point:
   - Draw $Z_n \sim \text{Mult}\left(\pi_1, \pi_2, \ldots\right)$.
   - Draw $X_n \sim p\left(x_n|\eta_{z_n}^*\right)$.

[*] Prof. Taeryon Choi

Although the paper uses *Multi* to denote Multinoulli distribution, it becomes clearer if we use *Cat* for Categorical distribution as in machine learning literatures since the two are essentially the same with different names.

3 Gaussian DP mixtures

The observations have the following distributional form:

$$p(x_n|z_n, \eta_1^*, \eta_2^*, \ldots) = \prod_{i=1}^{\infty} \left(h(x_n) \exp\left\{\eta_i^{*\prime} x_n - a\left(\eta_i^*\right)\right\}\right)^{\mathbf{1}[z_n=i]}$$

It is a known fact that exponential families have the following form (multivariate form):

$$h(x) \exp\left\{\eta' T(x) - A(\eta)\right\}$$

where η is the vector of natural parameters and $T(x)$ is the vector of sufficient statistics. In the case of multivariate Gaussian, we can also transform the pdf into the form of an exponential family.

$$\begin{aligned} p(x) &= \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left\{-\frac{1}{2}\left(x-\mu\right)' \Sigma^{-1} \left(x-\mu\right)\right\} \\ &= \exp\left\{-\frac{1}{2}\log|2\pi\Sigma|\right\} \exp\left\{-\frac{1}{2}\left(x-\mu\right)' \Sigma^{-1} \left(x-\mu\right)\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\underbrace{x'\Sigma^{-1}x - 2\mu'\Sigma^{-1}x}_{\eta' T(x)} + \mu'\Sigma^{-1}\mu + \log|2\pi\Sigma|\right]\right\} \end{aligned}$$

Now we use the Frobenius product and vectorize operator to obtain the desired form.

$$\begin{aligned} x'\Sigma^{-1}x &= \Sigma^{-1} : xx' \\ &= \operatorname{vec}\left(\Sigma^{-1}\right)' \operatorname{vec}\left(xx'\right) \\ \mu'\Sigma^{-1}x &= \left(\Sigma^{-1}\mu\right)' x \\ \therefore x'\Sigma^{-1}x - 2\mu'\Sigma^{-1}x &= \begin{bmatrix} \operatorname{vec}\left(\Sigma^{-1}\right) \\ -2\Sigma^{-1}\mu \end{bmatrix}' \begin{bmatrix} \operatorname{vec}\left(xx'\right) \\ x \end{bmatrix} \end{aligned}$$

This yields

$$\begin{aligned} \eta &= \begin{bmatrix} -\frac{1}{2}\operatorname{vec}\left(\Sigma^{-1}\right) \\ \Sigma^{-1}\mu \end{bmatrix} \\ T(x) &= \begin{bmatrix} \operatorname{vec}\left(xx'\right) \\ x \end{bmatrix} \\ A(\eta) &= \frac{1}{2}\mu'\Sigma^{-1}\mu + \frac{1}{2}\log|2\pi\Sigma| \end{aligned}$$

Since the paper suggests a general form of Dirichlet process for all kinds of exponential families, we should specify on our own what distribution each observation x_n follows. The most usual case is Gaussian. Therefore, we will explicitly write the natural parameters and sufficient statistics of a multivariate Gaussian distribution. Then η_i^* becomes η and notationally abusive x_n becomes $T(x)$. Recall that η_i^* has to be sampled from a base measure H, or $\eta_i^* \sim H$. In this case, we have Σ^{-1} and μ inside η_i*. Usually, we impose a conjugate prior for $\begin{bmatrix}\Sigma^{-1} & \mu\end{bmatrix}$ which is in this case the Normal-Wishart distribution denoted with $\mathcal{NW}(\cdot)$.

3.1 Lower bound

$$\begin{aligned}\log p\left(\boldsymbol{x}|\alpha,\lambda\right) \geq &\ \mathrm{E}\left[\log p\left(\boldsymbol{V}|\alpha\right)\right]+\mathrm{E}\left[\log p\left(\boldsymbol{\eta}^{*}|\lambda\right)\right]\\ &+\sum_{n=1}^{N}\left(\mathrm{E}\left[\log p\left(Z_n|\boldsymbol{V}\right)\right]+\mathrm{E}\left[\log p\left(x_n|Z_n\right)\right]\right)\\ &-\mathrm{E}\left[\log q\left(\boldsymbol{V},\boldsymbol{\eta}^{*},\boldsymbol{Z}\right)\right].\end{aligned}$$

As always, the expectations are taken with respect to the variational distributions (optimal distributions).

3.1.1 $\mathrm{E}\left[\log p\left(\boldsymbol{V}|\alpha\right)\right]$

According to the setting, v_t are sampled from the Beta distribution with parameters 1 and α. Hence, it is reasonable to assume that the variational distribution $q\left(v_t\right)$ follows Beta as well.

$$q\left(v_i\right)\sim \mathrm{Beta}\left(\gamma_{i,1},\gamma_{i,2}\right)$$

Now we can compute the expectation.

$$\begin{aligned}p\left(\boldsymbol{V}|\alpha\right) &\sim \prod_{i}^{\infty}\mathrm{Beta}\left(v_i|1,\alpha\right)\\ &=\prod_{i=1}^{\infty}\alpha\left(1-v_i\right)^{\alpha-1}\\ &=\prod_{i=1}^{T-1}\alpha\left(1-v_i\right)^{\alpha-1}\\ \log p\left(\boldsymbol{V}|\alpha\right) &=\sum_{i=1}^{T-1}\left(1-\alpha\right)\log\left(1-v_i\right)+\log\alpha\\ \mathrm{E}\left[\log p\left(\boldsymbol{V}|\alpha\right)\right] &=\log\alpha+\sum_{i=1}^{T-1}\left(1-\alpha\right)\mathrm{E}\left[\log\left(1-v_i\right)\right]\end{aligned}$$

Since $v_i \sim \mathrm{Beta}\left(\gamma_{i,1}, \gamma_{i,2}\right)$, it is easy to show that $1 - v_i \sim \mathrm{Beta}\left(\gamma_{i,2}, \gamma_{i,1}\right)$. We prove the expectation of log-Beta as a side note.

$$
\begin{aligned}
X &\sim \mathrm{Beta}\left(\alpha, \beta\right) \\
\mathrm{E}\left[\log X\right] &= \int_0^1 \log x \frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)} x^{\alpha-1}\left(1-x\right)^{\beta-1}\,dx \\
&= \frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)} \int_0^1 \frac{\partial}{\partial\alpha} x^{\alpha-1}\left(1-x\right)^{\beta-1}\,dx \\
&= \frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)} \frac{\partial}{\partial\alpha} \int_0^1 x^{\alpha-1}\left(1-x\right)^{\beta-1}\,dx \\
&= \frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)} \frac{\partial}{\partial\alpha} \frac{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}{\Gamma\left(\alpha+\beta\right)} \\
&= \frac{1}{\mathrm{B}\left(\alpha,\beta\right)} \frac{\partial\,\mathrm{B}\left(\alpha,\beta\right)}{\partial\alpha} \\
&= \frac{\partial \log \mathrm{B}\left(\alpha,\beta\right)}{\partial\alpha} \\
&= \frac{\partial \log\Gamma\left(\alpha\right)\partial\alpha}{-}\frac{\partial\log\Gamma\left(\alpha+\beta\right)}{\partial\alpha} \\
&= \varphi\left(\alpha\right) - \varphi\left(\alpha+\beta\right)
\end{aligned}
$$

where φ is the digamma function. Therefore, $\mathrm{E}\left[\log\left(1-v_i\right)\right] = \varphi\left(\gamma_{i,2}\right) - \varphi\left(\gamma_{i,1}+\gamma_{i,2}\right)$. This completes the expectation:

$$
\mathrm{E}\left[\log p\left(\boldsymbol{V}|\alpha\right)\right] = \log\alpha + \sum_{i=1}^{T-1}\left(1-\alpha\right)\left\{\varphi\left(\gamma_{i,2}\right) - \varphi\left(\gamma_{i,1}+\gamma_{i,2}\right)\right\}.
$$

3.1.2 $\mathrm{E}\left[\log p\left(\boldsymbol{\eta}^*|\lambda\right)\right]$

Since $\boldsymbol{\eta}^*$ is the vector of natural parameters and we have assumed Gaussian observations, $\boldsymbol{\eta}^*$ is $\begin{bmatrix}\Sigma^{-1} & \mu\end{bmatrix}$. Now we will apply the conjugate prior to both Σ^{-1} and μ which is the Normal-Wishart distribution.

$$
\begin{aligned}
\Sigma^{-1} &\sim \mathcal{W}_p\left(W, n\right) \\
\mu|\Sigma &\sim \mathcal{N}\left(\mu_0, \rho\Sigma\right) \\
p\left(\Sigma^{-1}, \mu\right) &= \mathcal{W}\left(\Sigma^{-1}|W, n\right)\mathcal{N}\left(\mu|\mu_0, \rho\Sigma\right) \\
&= \frac{\left|\Sigma^{-1}\right|^{\frac{n-p-1}{2}}}{2^{np/2}\left|W\right|^{n/2}\Gamma_p\left(\frac{n}{2}\right)} \exp\left(-\frac{1}{2}\mathrm{Tr}\left(W^{-1}\Sigma^{-1}\right)\right) \cdot \frac{1}{\sqrt{\left|2\pi\rho\Sigma\right|}} \exp\left\{-\frac{1}{2}\left(\mu-\mu_0\right)' \frac{1}{\rho}\Sigma^{-1}\left(\mu-\mu_0\right)\right\}
\end{aligned}
$$

Note that $p\left(\boldsymbol{\eta}^*\right) = \prod_{i=1}^T p\left(\Sigma_i^{-1}, \mu_i\right)$. Taking the logarithm,

$$
\begin{aligned}
\log p\left(\boldsymbol{\eta}^*\right) &= \sum_{i=1}^T \log p\left(\Sigma_i^{-1}, \mu_i\right) \\
&= \frac{n-p-1}{2}\log\left|\Sigma_i^{-1}\right| - \frac{np}{2}\log 2 - \frac{n}{2}\log\left|W\right| - \log\Gamma_p\left(\frac{n}{2}\right) - \frac{1}{2}\mathrm{Tr}\left(W^{-1}\Sigma^{-1}\right) \\
&\quad - \frac{1}{2}\log\left|2\pi\rho\Sigma_i\right| - \frac{1}{2}\left(\mu_i-\mu_0\right)'\frac{1}{\rho}\Sigma_i^{-1}\left(\mu_i-\mu_0\right)
\end{aligned}
$$

We set the variational distribution of Σ^{-1} and μ as follows:

$$\begin{aligned} q\left(\Sigma_i^{-1}\right) &\sim \mathcal{W}_p\left(V_i, d\right) \\ q\left(\mu_i\right) &\sim \mathcal{N}\left(m_i, S_i\right). \end{aligned}$$

Now taking the expectation,

$$\begin{aligned} \mathrm{E}\left[\log p\left(\boldsymbol{\eta}^*\right)\right] &= \sum_{i=1}^{T} \mathrm{E}\left[\log p\left(\Sigma_i^{-1}, \mu_i\right)\right] \\ &= \sum_{i=1}^{T} \left(\frac{n-p-1}{2} \mathrm{E}\left[\log\left|\Sigma_i^{-1}\right|\right] - \frac{np}{2}\log 2 - \frac{n}{2}\log|W| - \log\Gamma_p\left(\frac{n}{2}\right) - \frac{1}{2}\mathrm{Tr}\left(W^{-1}\mathrm{E}\left[\Sigma_i^{-1}\right]\right)\right. \\ &\left. -\frac{p}{2}\log(2\pi) + \frac{1}{2}\mathrm{E}\left[\log\left|\Sigma_i^{-1}\right|\right] - \frac{1}{2}\mathrm{E}\left[\left(\mu_i - \mu_0\right)' \frac{1}{\rho}\Sigma_i^{-1}\left(\mu_i - \mu_0\right)\right]\right) \end{aligned}$$

Here we should use the identity that

$$\mathrm{E}\left[\log\left|\Sigma_i^{-1}\right|\right] = \varphi_p\left(\frac{d}{2}\right) + p\log 2 + \log\left|V_i\right|.$$

$$\begin{aligned} \mathrm{E}\left[\left(\mu_i - \mu_0\right)' \frac{1}{\rho}\Sigma_i^{-1}\left(\mu_i - \mu_0\right)\right] &= \mathrm{E}\left[\mathrm{Tr}\left(\left(\mu_i - \mu_0\right)\frac{1}{\rho}\Sigma_i^{-1}\left(\mu_i - \mu_0\right)\right)\right] \\ &= \mathrm{E}\left[\mathrm{Tr}\left(\frac{1}{\rho}\Sigma_i^{-1}\left(\mu_i - \mu_0\right)\left(\mu_i - \mu_0\right)'\right)\right] \\ &= \mathrm{Tr}\left(\frac{1}{\rho}\mathrm{E}\left[\Sigma_i^{-1}\right]\left(S_i + \left(m_i - \mu_0\right)\left(m_i - \mu_0\right)'\right)\right) \\ &= \mathrm{Tr}\left(\frac{1}{\rho} d V_i \left(S_i + \left(m_i - \mu_0\right)\left(m_i - \mu_0\right)'\right)\right) \\ &= \frac{d}{\rho}\mathrm{Tr}\left(V_i S_i\right) + \frac{d}{\rho}\left(m_i - \mu_0\right)' V_i \left(m_i - \mu_0\right) \end{aligned}$$

Combining all the results,

$$\begin{aligned} \sum_{i=1}^{T} \mathrm{E}\left[\log p\left(\Sigma_i^{-1}, \mu_i\right)\right] &= \sum_{i=1}^{T} \left(\frac{n-p}{2}\left\{\varphi_p\left(\frac{d}{2}\right) + p\log 2 + \log\left|V_i\right|\right\} - \frac{np}{2}\log 2 - \frac{n}{2}\log|W| - \log\Gamma_p\left(\frac{n}{2}\right)\right. \\ &\left. -\frac{1}{2}\mathrm{Tr}\left(dW^{-1}V_i\right) - \frac{p}{2}\log(2\pi) - \frac{1}{2}\left[\frac{d}{\rho}\mathrm{Tr}\left(V_i S_i\right) + \frac{d}{\rho}\left(m_i - \mu_0\right)' V_i \left(m_i - \mu_0\right)\right]\right) \end{aligned}$$

Note that upper case V_i is the scale matrix of the Wishart distribution whereas the lower case v_i is the length of broken sticks sampled from Beta distribution. These notations are kept separate for identifiability.

3.1.3 $\mathrm{E}\left[\log p\left(Z_n|\boldsymbol{V}\right)\right]$

Recall that $Z_n|\boldsymbol{V} \sim \mathrm{Mult}\left(\pi\left(\boldsymbol{v}\right)\right)$ and that

$$\pi_k = v_k \prod_{\ell=1}^{k-1}\left(1 - v_\ell\right).$$

Looking carefully, π_k contains $1 - v_i$ term if $k > i$. Therefore, the pdf of $Z_n | \boldsymbol{V}$ is

$$p\left(Z_n | \boldsymbol{V}\right) = \prod_{i=1}^{\infty} \left(1 - v_i\right)^{\mathbf{1}[z_n > i]} v_i^{\mathbf{1}[z_n = i]}.$$

We should apply the truncation that the original author postulates: $q\left(z_n > T\right) = 0$. Then,

$$\begin{aligned}
\mathrm{E}\left[\log p\left(Z_n | \boldsymbol{V}\right)\right] &= \mathrm{E}\left[\log\left(\prod_{i=1}^{\infty} \left(1 - v_i\right)^{\mathbf{1}[z_n > i]} v_i^{\mathbf{1}[z_n = i]}\right)\right] \\
&= \sum_{i=1}^{\infty} q\left(z_n > i\right) \mathrm{E}\left[\log\left(1 - v_i\right)\right] + q\left(z_n = i\right) \mathrm{E}\left[\log v_i\right] \\
&= \sum_{i=1}^{T} q\left(z_n > i\right) \mathrm{E}\left[\log\left(1 - v_i\right)\right] + q\left(z_n = i\right) \mathrm{E}\left[\log v_i\right].
\end{aligned}$$

Borrowing the notations of the author, the variational distribution of Z_n is Categorical or Multinoulli with parameters $\phi_1, \phi_2, \ldots, \phi_N$. By definition, the probability of $Z_n = i$ is $q\left(Z_n = i\right) = \phi_{n,i}$. Incorporating all the results above,

$$\mathrm{E}\left[\log p\left(Z_n | \boldsymbol{V}\right)\right] = \sum_{i=1}^{T} \left\{\left[\sum_{j=i+1}^{T} \phi_{n,j}\right] \left(\varphi\left(\gamma_{i,2}\right) - \varphi\left(\gamma_{i,1} + \gamma_{i,2}\right)\right) + \phi_{n,i} \left(\varphi\left(\gamma_{i,1}\right) - \varphi\left(\gamma_{i,1} + \gamma_{i,2}\right)\right)\right\}$$

3.1.4 $\mathrm{E}\left[\log p\left(x_n | Z_n\right)\right]$

Wrong way

Now once Z_n is determined, it indicates which parameters $\Sigma_{z_n}^{-1}, \mu_{z_n}$ to use. This immediately defines the density:

$$p\left(x_n | Z_n\right) = \frac{1}{\sqrt{\left|2\pi \Sigma_{z_n}\right|}} \exp\left\{-\frac{1}{2}\left(x_n - \mu_{z_n}\right)' \Sigma_{z_n}^{-1} \left(x_n - \mu_{z_n}\right)\right\}$$

$$\begin{aligned}
\mathrm{E}\left[\log p\left(x_n | Z_n\right)\right] &= -\frac{p}{2} \log\left(2\pi\right) + \frac{1}{2} \mathrm{E}\left[\log\left|\Sigma_{z_n}^{-1}\right|\right] - \frac{1}{2} \mathrm{E}\left[\left(x_n - \mu_{z_n}\right)' \Sigma_{z_n}^{-1} \left(x_n - \mu_{z_n}\right)\right] \\
\mathrm{E}\left[\log\left|\Sigma_{z_n}^{-1}\right|\right] &= \varphi_p\left(\frac{d}{2}\right) + p \log 2 + \log\left|V_{z_n}\right| \\
\mathrm{E}\left[\left(x_n - \mu_{z_n}\right)' \Sigma_{z_n}^{-1} \left(x_n - \mu_{z_n}\right)\right] &= \mathrm{E}\left[\mathrm{Tr}\left(\left(x_n - \mu_{z_n}\right)' \Sigma_{z_n}^{-1} \left(x_n - \mu_{z_n}\right)\right)\right] \\
&= \mathrm{E}\left[\mathrm{Tr}\left(\Sigma_{z_n}^{-1} \left(x_n - \mu_{z_n}\right)\left(x_n - \mu_{z_n}\right)'\right)\right] \\
&= \mathrm{Tr}\left(\mathrm{E}\left[\Sigma_{z_n}^{-1}\right]\left(S_{z_n} + \left(x_n - m_{z_n}\right)\left(x_n - m_{z_n}\right)'\right)\right) \\
&= \mathrm{Tr}\left(dV_{z_n} S_{z_n} + dV_{z_n} \left(x_n - m_{z_n}\right)\left(x_n - m_{z_n}\right)'\right) \\
&= d\,\mathrm{Tr}\left(V_{z_n} S_{z_n}\right) + d\left(x_n - m_{z_n}\right)' V_{z_n} \left(x_n - m_{z_n}\right)
\end{aligned}$$

Coming back to the original expectation,

$$\begin{aligned}
\mathrm{E}\left[\log p\left(x_n | Z_n\right)\right] = &-\frac{p}{2} \log\left(2\pi\right) + \frac{1}{2}\left(\varphi_p\left(\frac{d}{2}\right) + p \log 2 + \log\left|V_{z_n}\right|\right) \\
&- \frac{1}{2}\left(d\,\mathrm{Tr}\left(V_{z_n} S_{z_n}\right) + d\left(x_n - m_{z_n}\right)' V_{z_n} \left(x_n - m_{z_n}\right)\right)
\end{aligned}$$

This is a wrong way to compute this expectation not because it is not a mathematically valid expression, but rather because it is mathematically intractable. Therefore, we propose another form of expression that is.

Correct way

When expressing the density $p\left(x_n|Z_n\right)$, we should avoid using the subscript z_n directly. Instead, we loop through the entirety of η_i^* and use the indicator variable that filters the one that is desired:

$$p\left(x_n|z_n\right)=\prod_{i=1}^{\infty}\left[\frac{1}{\sqrt{|2\pi\Sigma_i|}}\exp\left\{-\frac{1}{2}\left(x_n-\mu_i\right)'\Sigma_i^{-1}\left(x_n-\mu_i\right)\right\}\right]^{\mathbf{1}[z_n=i]}.$$

This way, we can utilize the expectation $\mathrm{E}\left[\mathbf{1}\left[z_n=i\right]\right]$ which is directly the probability of z_n becoming i: $q\left(z_n=i\right)=\phi_{n,i}$. Furthermore, the loop is truncated because we assumed $q\left(z_n>T\right)=0$

$$\begin{aligned}
\log p\left(x_n|z_n\right)&=\sum_{i=1}^{\infty}\mathbf{1}\left[z_n=i\right]\left\{-\frac{p}{2}\log\left(2\pi\right)+\frac{1}{2}\log\left|\Sigma_i^{-1}\right|-\frac{1}{2}\left(x_n-\mu_i\right)'\Sigma_i^{-1}\left(x_n-\mu_i\right)\right\}\\
\mathrm{E}\left[\log p\left(x_n|z_n\right)\right]&=\sum_{i=1}^{\infty}q\left(z_n=i\right)\left\{-\frac{p}{2}\log\left(2\pi\right)+\frac{1}{2}\left(\varphi_p\left(\frac{d}{2}\right)+p\log 2+\log\left|V_i\right|\right)\right.\\
&\qquad\left.-\frac{d}{2}\operatorname{Tr}\left(V_iS_i\right)-\frac{d}{2}\left(x_n-m_i\right)'V_i\left(x_n\right)-m_i)\right\}\\
&=\sum_{i=1}^{T}\phi_{n,i}\left\{-\frac{p}{2}\log\left(2\pi\right)+\frac{1}{2}\left(\varphi_p\left(\frac{d}{2}\right)+p\log 2+\log\left|V_i\right|\right)\right.\\
&\qquad\left.-\frac{d}{2}\operatorname{Tr}\left(V_iS_i\right)-\frac{d}{2}\left(x_n-m_i\right)'V_i\left(x_n-m_i\right)\right\}
\end{aligned}$$

3.1.5 $\mathrm{E}\left[\boldsymbol{V},\boldsymbol{\eta}^*,\boldsymbol{Z}\right]$

As stated in the paper, the variational distributions of the variational parameters are represented as follows:

$$q\left(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z}\right)=\prod_{t=1}^{T-1}q_{\gamma_t}\left(v_t\right)\prod_{t=1}^{T}q_{\tau_t}\left(\eta_t^*\right)\prod_{n=1}^{N}q_{\phi_n}\left(z_n\right).$$

And we already know the following:

$$\begin{aligned}
v_i&\sim\operatorname{Beta}\left(\gamma_{i,1},\gamma_{i,2}\right)\\
\eta_i^*&\sim\mathcal{NW}\left(\mu_i|m_i,S_i;\,\Sigma_i^{-1}|n,V_i\right)\\
z_n&\sim\operatorname{Cat}\left(\phi_{n,i},\ldots,\phi_{n,T}\right).
\end{aligned}$$

Therefore, we can compute the desired expectation.

$$
\begin{aligned}
\log q\left(\boldsymbol{V}, \boldsymbol{\eta}^{*}, \boldsymbol{Z}\right) = & \sum_{i=1}^{T-1} \log \left\{\frac{\Gamma\left(\gamma_{i,1}+\gamma_{i,2}\right)}{\Gamma\left(\gamma_{i,1}\right)+\Gamma\left(\gamma_{i,2}\right)} v_i^{\gamma_{i,1}-1}\left(1-v_i\right)^{\gamma_{i,2}-1}\right\} \\
& +\sum_{i=1}^{T} \log \left\{\frac{1}{\sqrt{\left|2 \pi S_i\right|}} \exp \left(-\frac{1}{2}\left(\mu_i-m_i\right)^{\prime} S_i^{-1}\left(\mu_i-m_i\right)\right)\right\} \\
& +\sum_{i=1}^{T} \log \left\{\frac{\left|\Sigma_i^{-1}\right|^{\frac{n-p-1}{2}}}{2^{\frac{np}{2}}\left|V_i\right|^{\frac{n}{2}} \Gamma_p\left(\frac{n}{2}\right)} \exp \left(-\frac{1}{2} \operatorname{Tr}\left(V_i^{-1} \Sigma_i^{-1}\right)\right)\right\} \\
& +\sum_{n=1}^{N} \log \left\{\prod_{i=1}^{T} \phi_{n,i}^{\mathbf{1}\left[z_n=i\right]}\right\} \\
= & \sum_{i=1}^{T-1}\left\{\log \Gamma\left(\gamma_{i,1}+\gamma_{i,2}\right)-\log \Gamma\left(\gamma_{i,1}\right)-\log \Gamma\left(\gamma_{i,2}\right)+\left(\gamma_{i,1}-1\right) \log v_i+\left(\gamma_{i,2}-1\right) \log \left(1-v_i\right)\right\} \\
& +\sum_{i=1}^{T}\left\{-\frac{1}{2} \log \left|2 \pi S_i\right|-\frac{1}{2}\left(\mu_i-m_i\right)^{\prime} S_i^{-1}\left(\mu_i-m_i\right)\right\} \\
& +\sum_{i=1}^{T}\left\{\frac{n-p-1}{2} \log \left|\Sigma_i^{-1}\right|-\frac{np}{2} \log 2-\frac{n}{2} \log \left|V_i\right|-\log \Gamma_p\left(\frac{n}{2}\right)-\frac{1}{2} \operatorname{Tr}\left(V_i^{-1} \Sigma_i^{-1}\right)\right\} \\
& +\sum_{n=1}^{N} \sum_{i=1}^{T} \mathbf{1}\left[z_{n=1}\right] \log \phi_{n,i}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}\left[\log q\left(\boldsymbol{V}, \boldsymbol{\eta}^{*}, \boldsymbol{Z}\right)\right] = & \sum_{i=1}^{T-1}\left\{\log \Gamma\left(\gamma_{i,1}+\gamma_{i,2}\right)-\log \Gamma\left(\gamma_{i,1}\right)-\log \Gamma\left(\gamma_{i,2}\right)+\left(\gamma_{i,1}-1\right)\left(\varphi\left(\gamma_{i,1}\right)-\varphi\left(\gamma_{i,1}+\gamma_{i,2}\right)\right)\right. \\
& \left.+\left(\gamma_{i,2}-1\right)\left(\varphi\left(\gamma_{i,2}\right)-\varphi\left(\gamma_{i,1}+\gamma_{i,2}\right)\right)\right\}+\sum_{i=1}^{T}\left\{-\frac{p}{2} \log (2 \pi)-\frac{1}{2} \log \left|S_i\right|-\frac{p}{2}\right\} \\
& +\sum_{i=1}^{T}\left\{\frac{n-p-1}{2}\left(\varphi_p\left(\frac{d}{2}\right)+p \log 2+\log \left|V_i\right|\right)-\frac{np}{2} \log 2-\frac{n}{2} \log \left|V_i\right|-\log \Gamma_p\left(\frac{n}{2}\right)-\frac{np}{2}\right\} \\
& +\sum_{n=1}^{N} \sum_{i=1}^{T} \phi_{n,i} \log \phi_{n,i}
\end{aligned}
$$

3.2 Updating via coordinate ascent

Free parameters (or variatioanl parameters) need updating. Mean-field assumption lends itself directly to obtaining the coordinate ascent algorithm. The variational distributions we need to compute are $\boldsymbol{v}, \boldsymbol{\eta}^{*}, \boldsymbol{z}$.

3.2.1 $\boldsymbol{v}$

Recall that mean-field assumption severs the dependencies between parameters so as to gain more degrees of freedom. Those who are not familiar with the methodology are referred to Wand and

Omerod(2010). Simply put, mean-field assumption converts function optimization into directly updating the parameters.

$$
\begin{aligned}
p\left(\boldsymbol{v}|\text{rest}\right) &= p\left(\boldsymbol{V}|\alpha\right)\prod_{i=1}^{N} p\left(z_n|\boldsymbol{V}\right)\\
&= \prod_{i=1}^{T-1}\alpha\left(1-v_i\right)^{\alpha-1}\prod_{n=1}^{N}\prod_{i=1}^{\infty}\left(1-v_i\right)^{\mathbf{1}[z_n>i]}v_i^{\mathbf{1}[z_n=i]}\\
\log p\left(\boldsymbol{v}|\text{rest}\right) &= \sum_{i=1}^{T-1}\left\{\log\alpha+\left(\alpha-1\right)\log\left(1-v_i\right)\right\}+\sum_{n=1}^{N}\sum_{i=1}^{T-1}\mathbf{1}\left[z_n>i\right]\log\left(1-v_i\right)+\mathbf{1}\left[z_n=i\right]\log v_i\\
\mathrm{E}_{-\boldsymbol{v}}\left[\log p\left(\boldsymbol{v}|\text{rest}\right)\right] &= \sum_{i=1}^{T-1}\left\{\left(\alpha+\sum_{n=1}^{N}q\left(z_n>i\right)-1\right)\log\left(1-v_i\right)+\sum_{n=1}^{N}q\left(z_n=i\right)\log v_i\right\}
\end{aligned}
$$

This is the log-density of Beta distribution. Each v_i follow different Beta distribution with different parameters $(\gamma_{i,1},\gamma_{i,2})$. Thus, we will not consider the first summation $\sum_{i=1}^{T-1}$.

$$
\begin{aligned}
\gamma_{i,1} &= 1+\sum_{n=1}^{N}\phi_{n,i}\\
\gamma_{i,2} &= \alpha+\sum_{n=1}^{N}\sum_{j=i+1}^{T}\phi_{n,j}
\end{aligned}
$$

3.2.2 $\boldsymbol{\eta}^*$

In exponential family notation, the vector of variational parameters is $\boldsymbol{\eta}^*$. However, in our model, since we assumed the observations follow Gaussian distribution, the variational parameters become μ_i and Σ_i^{-1}. We first go with $\boldsymbol{\mu}$.

$$
\begin{aligned}
\log p\left(\boldsymbol{\mu}|\text{rest}\right) &= \sum_{i=1}^{T}\log\mathcal{N}\left(\mu_i|\mu_0,\rho\Sigma_i\right)+\sum_{n=1}^{N}\sum_{i=1}^{T}\log\left[p\left(x_n|z_n\right)\right]^{\mathbf{1}[z_n=i]}\\
\mathrm{E}_{-\boldsymbol{\mu}}\left[\log p\left(\boldsymbol{\mu}|\text{rest}\right)\right] &= \sum_{i=1}^{T}\left\{-\frac{p}{2}\log\left(2\pi\rho\right)+\frac{1}{2}\left(\varphi_p\left(\frac{d}{2}\right)+p\log 2+\log\left|V_i\right|\right)-\frac{d}{2\rho}\left(\mu_i-\mu_0\right)'V_i\left(\mu_i-\mu_0\right)\right\}\\
&\quad+\sum_{i=1}^{T}\sum_{n=1}^{N}\left\{\phi_{n,i}\left[-\frac{p}{2}\log\left(2\pi\right)+\frac{1}{2}\left(\varphi_p\left(\frac{d}{2}\right)+p\log 2+\log\left|V_i\right|\right)-\frac{d}{2}\left(x_n-\mu_i\right)'V_i\left(x_n-\mu_i\right)\right]\right\}
\end{aligned}
$$

Since we are considering μ_i separately, we will not consider the outer summation $\sum_{i=1}^{T}$. Then,

$$\begin{aligned}
\mathrm{E}_{-\mu_i}\left[\log p\left(\mu_i|\text{rest}\right)\right] &\propto \sum_{n=1}^{N}\left\{-\frac{d}{2}\phi_{n,i}\left(x_n-\mu_i\right)' V_i\left(x_n-\mu_i\right)\right\}-\frac{d}{2\rho}\left(\mu_i-\mu_0\right)' V_i\left(\mu_i-\mu_0\right) \\
&\propto -\frac{1}{2}\left\{\left(\frac{d}{\rho}+\sum_{n=1}^{N} d\phi_{n,i}\right)\mu_i' V_i\mu_i-2\left(\sum_{n=1}^{N} d\phi_{n,i}x_n+\frac{d}{\rho}\mu_0\right)' V_i\mu_i\right\} \\
&\sim \mathcal{N}\left(\left(\frac{d}{\rho}+\sum_{n=1}^{N} d\phi_{n,i}\right)^{-1}\left(\sum_{n=1}^{N} d\phi_{n,i}x_n+\frac{d}{\rho}\mu_0\right), \left(\frac{d}{\rho}+\sum_{n=1}^{N} d\phi_{n,i}\right)^{-1} V_i^{-1}\right).
\end{aligned}$$

Therefore, the updating rule for m_i and S_i are

$$\begin{aligned}
m_i &= \left(\frac{d}{\rho}+\sum_{n=1}^{N} d\phi_{n,i}\right)^{-1}\left(\sum_{n=1}^{N} d\phi_{n,i}x_n+\frac{d}{\rho}\mu_0\right) \\
S_i &= \left(\frac{d}{\rho}+\sum_{n=1}^{N} d\phi_{n,i}\right)^{-1} V_i^{-1}.
\end{aligned}$$

Next is Σ_i^{-1} which follows Wishart distribution.

$$\begin{aligned}
p\left(\boldsymbol{\Sigma}^{-1}|\text{rest}\right) &= \prod_{i=1}^{T}\mathcal{N}\left(\mu_i|\mu_0,\rho\Sigma_i\right)\mathcal{W}\left(\Sigma_i^{-1}|W,n\right)\prod_{n=1}^{N}\prod_{i=1}^{\infty}\left[p\left(x_n|z_n\right)\right]^{\mathbf{1}\left[z_n=i\right]} \\
\log p\left(\boldsymbol{\Sigma}^{-1}|\text{rest}\right) &= \sum_{i=1}^{T}\left\{-\frac{p}{2}\log\left(2\pi\right)+\frac{1}{2}\log\left|\Sigma_i^{-1}\right|-\frac{1}{2}\left(\mu_i-\mu_0\right)'\frac{1}{\rho}\Sigma_i^{-1}\left(\mu_i-\mu_0\right)\right. \\
&\quad\left.+\frac{n-p-1}{2}\log\left|\Sigma_i^{-1}\right|-\frac{1}{2}\mathrm{Tr}\left(W^{-1}\Sigma_i^{-1}\right)-\frac{np}{2}\log 2-\frac{n}{2}\log\left|W\right|-\log\Gamma_p\left(\frac{n}{2}\right)\right\} \\
&\quad+\sum_{i=1}^{T}\sum_{n=1}^{N}\left[\mathbf{1}\left[z_n=i\right]\left\{-\frac{p}{2}\log\left(2\pi\right)+\frac{1}{2}\log\left|\Sigma_i^{-1}\right|-\frac{1}{2}\left(x_n-\mu_i\right)'\Sigma_i^{-1}\left(x_n-\mu_i\right)\right\}\right] \\
\mathrm{E}\left[\log p\left(\boldsymbol{\Sigma}^{-1}|\text{rest}\right)\right] &= \sum_{i=1}^{T}\left\{-\frac{p}{2}\log\left(2\pi\right)-\frac{1}{2\rho}\left(\mathrm{Tr}\left(\Sigma_i^{-1}S_i\right)+\left(m_i-\mu_0\right)'\Sigma_i^{-1}\left(m_i-\mu_0\right)\right)\right. \\
&\quad\left.+\frac{n-p}{2}\log\left|\Sigma_i^{-1}\right|-\frac{1}{2}\mathrm{Tr}\left(W^{-1}\Sigma_i^{-1}\right)-\frac{np}{2}\log 2-\frac{n}{2}\log\left|W\right|-\log\Gamma_p\left(\frac{n}{2}\right)\right\} \\
&\quad+\sum_{i=1}^{T}\sum_{n=1}^{N}\left[\phi_{n,i}\left\{-\frac{p}{2}\log\left(2\pi\right)+\frac{1}{2}\log\left|\Sigma_i^{-1}\right|-\frac{1}{2}\left(\mathrm{Tr}\left(\Sigma_i^{-1}S_i\right)+\left(m_i-x_n\right)'\Sigma_i^{-1}\left(m_i-x_n\right)\right)\right\}\right] \\
&\sim \mathcal{W}\left(n+1+\sum_{n=1}^{N}\phi_{n,i}, \left(\left(\frac{1}{\rho}+\sum_{n=1}^{N}\phi_{n,i}\right)S_i+W_i^{-1}\right)^{-1}\right)
\end{aligned}$$

So if we say the variational distribution of Σ_i^{-1} is

$$q\left(d,V_i\right)$$

then the updating of d and V_i are

$$
\begin{aligned}
d &= n + 1 + \sum_{n=1}^{N} \phi_{n,i} \\
V_i &= \left(\left(\frac{1}{\rho} + \sum_{n=1}^{N} \phi_{n,i} \right) S_i + W_i^{-1} \right)^{-1} .
\end{aligned}
$$

3.3 $\boldsymbol{Z}$

$$
\begin{aligned}
\log p\left(\boldsymbol{z}|\text{rest}\right) &= \sum_{n=1}^{N} \sum_{i=1}^{T} \left\{ \mathbf{1}\left[z_n > i\right] \log\left(1 - v_i\right) + \mathbf{1}\left[z_n = i\right] \log v_i \right\} \\
&\quad + \sum_{n=1}^{N} \sum_{i=1}^{T} \left[\mathbf{1}\left[z_n = i\right] \left\{ -\frac{p}{2} \log\left(2\pi\right) + \frac{1}{2} \log\left|\Sigma_i^{-1}\right| - \frac{1}{2} \left(x_n - \mu_i\right)' \Sigma_i^{-1} \left(x_n - \mu_i\right) \right\} \right] \\
\mathrm{E}_{-\boldsymbol{z}}\left[\log p\left(\boldsymbol{z}|\text{rest}\right)\right] &= \sum_{n=1}^{N} \sum_{i=1}^{T} \left\{ \mathbf{1}\left[z_n > i\right] \left(\varphi\left(\gamma_{i,2}\right) - \varphi\left(\gamma_{i,1} + \gamma_{i,2}\right)\right) + \mathbf{1}\left[z_n = i\right] \left(\varphi\left(\gamma_{i,1}\right) - \varphi\left(\gamma_{i,1} + \gamma_{i,2}\right)\right) \right\} \\
&\quad + \sum_{n=1}^{N} \sum_{i=1}^{T} \left[\mathbf{1}\left[z_n = i\right] \left\{ -\frac{p}{2} \log\left(2\pi\right) + \frac{1}{2} \left(\varphi_p \left(\frac{d}{2} \right) + p \log 2 + \log\left|V_i\right| \right) \right.\right. \\
&\quad \left.\left. - \frac{n}{2} \left(\operatorname{Tr}\left(V_i S_i\right) + \left(m_i - x_n\right)' V_n \left(m_i - x_n\right) \right) \right\} \right] \\
&\propto \mathbf{1}\left[z_n = i\right] \left\{ \varphi\left(\gamma_{i,1}\right) - \varphi\left(\gamma_{i,1} + \gamma_{i,2}\right) - \frac{p}{2} \log\left(2\pi\right) + \frac{1}{2} \left(\varphi_p \left(\frac{d}{2} \right) + p \log 2 + \log\left|V_i\right| \right) \right. \\
&\quad \left. + \sum_{j=1}^{i-1} \left(\varphi\left(\gamma_{j,2}\right) - \varphi\left(\gamma_{j,1} + \gamma_{j,2}\right)\right) - \frac{n}{2} \left(\operatorname{Tr}\left(V_i S_i\right) + \left(m_i - x_n\right)' V_i \left(m_i - x_n\right) \right) \right\}
\end{aligned}
$$

Therefore,

$$
\phi_{n,i} = \exp\left(\Phi\right)
$$

where

$$
\begin{aligned}
\Phi &= \varphi\left(\gamma_{i,1}\right) - \varphi\left(\gamma_{i,1} + \gamma_{i,2}\right) - \frac{p}{2} \log\left(2\pi\right) + \frac{1}{2} \left(\varphi_p \left(\frac{d}{2} \right) + p \log 2 + \log\left|V_i\right| \right) \\
&\quad + \sum_{j=1}^{i-1} \left(\varphi\left(\gamma_{j,2}\right) - \varphi\left(\gamma_{j,1} + \gamma_{j,2}\right)\right) - \frac{n}{2} \left(\operatorname{Tr}\left(V_i S_i\right) + \left(m_i - x_n\right)' V_i \left(m_i - x_n\right) \right) .
\end{aligned}
$$