Variational Bayes

Richard Xu

November 25, 2022

1 A bit of history . . .

This note started in 2010 when I was inspired to help people read Chapter 10 of Bishop [1] where I was trying to explain a few things in an oversimplified (hopefully!) way. I revamped it for the class. I also added exponential family distributions and an example on LDA when the model is fully conjugate [2]

2 The Variational Bayes Framework

2.1 what is Evidence Lowerbound?

2.2 use Jensen Inequality

$$\begin{aligned}
\log p(x) &= \log \int_z p(x,z) \\
&= \log \int_z \frac{p(x,z)}{q_\phi(z|x)} q_\phi(z|x) \\
&= \log \left[\mathbb{E}_{z \sim q_\phi(z|x)} \left(\frac{p(x,z)}{q_\phi(z|x)} \right) \right] \\
&\geq \mathbb{E}_{z \sim q_\phi(z|x)} \left[\log \left(\frac{p(x,z)}{q_\phi(z|x)} \right) \right] \qquad \text{by Jensen's inequality} \\
&= \mathbb{E}_{z \sim q_\phi(z|x)} [\log(p(x,z)] - \mathbb{E}_{z \sim q_\phi(z|x)} [\log(q_\phi(z|x)] \\
&= \text{ELBO}(q)
\end{aligned} \tag{1}$$

2.3 simple expansion

$$\begin{aligned}
\log (p(x)) &= \log \left(\frac{p(x,z)}{p(z|x)} \right) \\
&= \log (p(x,z)) - \log (p(z|x)) \\
&= [\log (p(x,z)) - q_\phi(z)] - [\log (p(z|x)) - q_\phi(z)] \qquad \because \pm q_\phi(z) \\
&= \log \left(\frac{p(x,z)}{q_\phi(z)} \right) - \log \left(\frac{p(z|x)}{q_\phi(z)} \right)
\end{aligned} \tag{2}$$

now, let's taking the expectation on both sides, given $q_\phi(z)$:

$$
\begin{aligned}
\log\left(p(x)\right) &= \int q_\phi(z)\log\left(\frac{p(x,z)}{q_\phi(z)}\right)\mathrm{dz} - \int q_\phi(z)\log\left(\frac{p(z|x)}{q_\phi(z)}\right)\mathrm{d}z \\
&= \int q_\phi(z)\log\left(\frac{p(x,z)}{q_\phi(z)}\right)\mathrm{dz} + \int q_\phi(z)\log\left(\frac{q_\phi(z)}{p(z|x)}\right)\mathrm{d}z \\
&= \mathrm{ELBO}(q) + \mathbb{KL}(q\|p)
\end{aligned}
\tag{3}
$$

2.3.1 name to both terms

$$
\begin{aligned}
\mathrm{ELBO}(q) &= \int q_\phi(z)\log\left(\frac{p(x,z)}{q_\phi(z)}\right)\mathrm{d}z \\
\mathbb{KL}(q\|p) &= \int q_\phi(z)\log\left(\frac{p(z|x)}{q_\phi(z)}\right)\mathrm{d}z
\end{aligned}
$$

the question of why we do not minimize $\mathbb{KL}$ term directly? The **key** is that the $\mathbb{KL}$ term contains $p(z|x)$ and ELBO term contains $p(x|z)p(z)$!

since we can choose any $q_\phi(z)$ we'd like, and since we want $\mathbb{KL}(\cdot)$ to be minimized, there it's ideal to make:

$$
q_\phi(z) \equiv q_\phi(z|x) \tag{4}
$$

i.e., it should also depend on x. Otherwise, it's highly unlikely that the $\mathbb{KL}\big(q\|p(z|x)\big)$ will be minimized:

$$
\mathbb{KL}(q\|p) = \int q_\phi(z|x)\log\left(\frac{q_\phi(z|x)}{p(z|x)}\right)\mathrm{d}z \tag{5}
$$

We know that $p(x) = \mathrm{ELBO}(q) + \mathbb{KL}(q\|p)$. We consider ELBO$(q)$ is the lower bound of $p(x)$. Minimizing $\mathbb{KL}(q\|p)$ is the same as maximizing the lower bound ELBO(q), since the addition of the two becomes $p(x)$

3 The choice of $q(\mathbf{z})$: mean-field approximation

Since any $q(\mathbf{z})$ will work, therefore, we will choose the most simple form. Suppose let's choose $q(\mathbf{z})$, such that:

$$q(\mathbf{z}) = \prod_{i=1}^{M} q_i(z_i) \tag{6}$$

this is called mean-filed approximation.

$$\begin{aligned}
\mathrm{ELBO}(q) &= \int q_\phi(z) \log\left(\frac{p(x,z)}{q_\phi(z)}\right) \mathrm{d}z \\
&= \int q_\phi(z) \log(p(x,z)) \mathrm{d}z - \int q_\phi(z) \log(q_\phi(z)) \mathrm{d}z \\
&= \underbrace{\int \prod_{i=1}^{M} q_i(z_i) \log\left(p(\mathbf{x},\mathbf{z})\right) \mathrm{d}\mathbf{z}}_{\text{part (1)}} - \underbrace{\int \prod_{i=1}^{M} q_i(z_i) \sum_{i=1}^{M} \log\left(q_i(z_i)\right) \mathrm{d}\mathbf{z}}_{\text{part (2)}}
\end{aligned} \tag{7}$$

Since you have the objective function for $\mathrm{ELBO}(q)$, a natural approach would be to optimize it repetitively using the parameters associated with each q.

3.1 Simplification of (Part 1):

$$\begin{aligned}
\text{(Part 1)} &= \int \prod_{i=1}^{M} q_i(z_i) \log\left(p(\mathbf{x},\mathbf{z})\right) \mathrm{d}\mathbf{z} \\
&= \int_{Z_1}\int_{Z_2} ... \int_{Z_M} \prod_{i=1}^{M} q_i(z_i) \log\left(p(\mathbf{x},\mathbf{z})\right) \mathrm{d}z_1, \mathrm{d}z_2, ...\mathrm{d}z_M
\end{aligned} \tag{8}$$

Rearrange the expression by taking a particular $q_j(z_j)$ out of the integral. Note that unlike (Part2), we are not treating any terms to const.:

$$\begin{aligned}
\text{(Part 1)}_{q_j} &\equiv \text{(Part 1)} \\
&= \int_{z_j} q_j(z_j) \Big(\int_{Z_{i\neq j}} \cdots \int \prod_{i\neq j}^{M} q_i(z_i) \log\left(p(\mathbf{x},\mathbf{z})\right) \prod_{i\neq j}^{M} \mathrm{d}z_i \Big) \mathrm{d}z_j \\
&= \int_{z_j} q_j(z_j) \Big(\int_{Z_{i\neq j}} \cdots \int \log\left(p(\mathbf{x},\mathbf{z})\right) \prod_{i\neq j}^{M} q_i(z_i) \mathrm{d}z_i \Big) \mathrm{d}z_j
\end{aligned} \tag{9}$$

or, even more meaningfully, it can be put into an expectation function, and since $\prod_{i\neq j}^{M} q_i(z_i)$ is a joint probability density

$$\text{(Part 1)}_{q_j} = \int_{z_j} q_j(z_j) \big[\mathbb{E}_{i\neq j}\big[\log\left(p(\mathbf{x},\mathbf{z})\right)\big]\big] \mathrm{d}z_j \tag{10}$$

note that one may consider $\log(\tilde{p}_j(\mathbf{x},\mathbf{z})) \equiv \mathbb{E}_{i\neq j}\big[\log\left(p(\mathbf{x},\mathbf{z})\right)\big]$. Obviously, note that

$$\begin{aligned}\tilde{p}_j(\mathbf{x}, \mathbf{z}) &\neq p(z_j|\mathbf{x}) \\ &\neq q(z_j|\mathbf{x})\end{aligned} \tag{11}$$

and we have:

$$\tilde{p}_j(\mathbf{x}, \mathbf{z}) = \exp\big(\mathbb{E}_{i\neq j}\big[\log\left(p(\mathbf{x}, \mathbf{z})\right)\big]\big) \tag{12}$$

3.2 Simplification of (Part 2):

$$\text{(Part 2)} = \int \prod_{i=1}^{M} q_i(z_i) \sum_{i=1}^{M} \log\left(q_i(z_i)\right) \mathrm{d}\mathbf{z} \tag{13}$$

Note that the above needs to integrate out all $\mathbf{z} = \{z_1, ..., z_M\}$, which is quite daunting. However, notice that each term in the sum, $\sum_{i=1}^{M} \log\left(q_i(z_i)\right)$ involves only a single i, therefore, we are able to simplify the above into the following:

$$\text{(Part 2)} = \sum_{i=1}^{M} \Bigg(\int\limits_{z_i} q_1(z_i) \log\left(q_i(z_i)\right) \mathrm{d}z_i \Bigg) \tag{14}$$

For a particular $p_j(z_j)$, the rest of the sum can be treated like a constant, therefore for $p_j(z_j)$ can be written as:

$$\text{(Part 2)}_{q_j} = \int\limits_{z_j} q_i(z_i) \log\left(q_i(z_i)\right) \mathrm{d}z_j + \text{const.} \tag{15}$$

where const. are the term does not involve z_j.

3.3 Putting Part (1) and Part (2) together:

write ELBO(q) in terms of q_j, i.e., ELBO(q_j), in which we try to optimize q_j. The rest of the terms would also need to be optimized $\{q_i\}$:

$$\begin{aligned}\text{ELBO}(q_j) &= \text{Part (1)}_{q_j} - \text{Part (2)}_{q_j} \\ &= \int\limits_{z_j} q_j(z_j) \mathbb{E}_{i\neq j}\big[\log\left(p(\mathbf{x}, \mathbf{z})\right)\big] \mathrm{d}z_j - \int\limits_{z_j} q_j(z_j) \log\left(q_j(z_j)\right) \mathrm{d}z_j + \text{const.}\end{aligned} \tag{16}$$

the key to realize is that we do not need to take derivative as one would normally do. All we need is to re-arrange the terms, and to realize it's the KL term, so we can just math the two distributions.

Note that $\mathbb{E}_{i\neq j}\left[\log\left(p(\mathbf{x}, \mathbf{z})\right)\right]$ would be some log probability of z, we name it $\log(\tilde{p}(\mathbf{x}, \mathbf{z}))$, i.e.,:

$$\log(\tilde{p}(\mathbf{x}, \mathbf{z})) = \mathbb{E}_{i\neq j}\left[\log\left(p(\mathbf{x}, \mathbf{z})\right)\right] \tag{17}$$

Or equivalently as:

$$\begin{aligned}
\mathrm{ELBO}(q) &= \int_{z_j} q_j(z_j) \log \left[\frac{\tilde{p}(\mathbf{x}, \mathbf{z})}{q_i(z_i)} \right] + \text{const.}. \\
&= -\mathbb{KL}\Big(\mathbb{E}_{i \neq j} \left[\log \left(p(\mathbf{x}, \mathbf{z}) \right) \right] \| q_i(z_i) \Big)
\end{aligned} \tag{18}$$

Now **this is the key**: We can maximize $\mathrm{ELBO}(q)$, by minimizing the KL divergence, where we can find approximate and optimal $q_i^*(z_i)$, such that:

$$\begin{aligned}
\log \left(q_i^*(z_i) \right) &= \log(\tilde{p}(\mathbf{x}, \mathbf{z})) \\
&= \mathbb{E}_{i \neq j} \big[\log \left(p(\mathbf{x}, \mathbf{z}) \right) \big] \\
\implies q_i^*(z_i)) &= \exp \left(\mathbb{E}_{i \neq j} \big[\log \left(p(\mathbf{x}, \mathbf{z}) \right) \big] \right)
\end{aligned} \tag{19}$$

4 Example: Gaussian-Gamma (Conjugate) posterior

4.1 model

4.1.1 likelihood

Let $\mathcal{D} = \{x_1, \dots x_n\}$:

$$
\begin{aligned}
p(\mathcal{D}|\mu, \tau) &= \prod_{i=1}^{n} \left(\frac{\tau}{2\pi}\right)^{\frac{1}{2}} \exp\left(\frac{-\tau}{2}(x_i - \mu)^2\right) \\
&= \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \exp\left(\frac{-\tau}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right)
\end{aligned}
\tag{20}
$$

4.1.2 prior

$$
\begin{aligned}
p(\mu|\tau) &= \mathcal{N}(\mu_0, (\lambda_0\tau)^{-1}) \propto \exp\left(\frac{-\lambda_0\tau}{2}(\mu - \mu_0)^2\right) \\
p(\tau) &= \text{Gamma}(\tau|a_0, b_0) \propto \tau^{a_0 - 1}\exp^{-b_0\tau}
\end{aligned}
\tag{21}
$$

4.1.3 posterior

Of course, due to conjugacy, the solution can be found exactly:

$$
\begin{aligned}
p(\mu, \tau|\mathcal{D}) &\propto p(\mathcal{D}|\mu, \tau)p(\mu|\tau)p(\tau) \\
&= \mathcal{N}(\mu_n, (\lambda_n\tau)^{-1})\text{Gamma}(\tau|a_n, b_n)
\end{aligned}
\tag{22}
$$

where:

$$
\begin{aligned}
\mu_n &= \frac{\lambda_0\mu_0 + n\bar{x}}{\lambda_0 + n} \\
\lambda_n &= \lambda_0 + n \\
a_n &= a_0 + n/2 \\
b_n &= b_0 + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{\lambda_0 n(\bar{x} - \mu_0)^2}{2(\lambda_0 + n)}
\end{aligned}
\tag{23}
$$

the exact derivation will be omitted and can be found from external sources easily.

4.2 mean-field Variational Inference algorithm

we let $q(\mathbf{z})$ to be:

$$
q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau) \tag{24}
$$

We use Variational Bayes formula:

4.2.1 $\log\left(q_\mu^*(\mu)\right) = \mathbb{E}_{q_\tau(\tau)}\left[\log\left(p(\mu,\tau,\mathcal{D})\right)\right]$

$$
\begin{aligned}
\log\left(q_\mu^*(\mu)\right) &= \mathbb{E}_{q_\tau}\Big[\log\left(p(\mu,\tau,\mathcal{D})\right)\Big] \\
&= \mathbb{E}_{q_\tau}\Big[\log(p(\mathcal{D}|\mu,\tau)) + \log p(\mu|\tau)\Big] + \text{const.} \qquad \text{leave out terms do NOT contain } \mu \\
&= \mathbb{E}_{q_\tau}\Big[\underbrace{\frac{n}{2}\log(\tau) - \frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2}_{\log(p(\mathcal{D}|\mu,\tau))} \underbrace{+\frac{\lambda_0\tau}{2}(\mu-\mu_0)^2}_{\log p(\mu|\gamma)}\Big] + \text{const.} \\
&= -\frac{\mathbb{E}_{q_\tau}[\tau]}{2}\left[\sum_{i=1}^{n}(x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] + \text{const.}
\end{aligned}
\tag{25}
$$

Completing the square for the μ terms:

$$
\begin{aligned}
\sum_{i=1}^{n}(x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2 &= n\mu^2 - 2n\mu\bar{x} + \lambda_0\mu^2 - 2\lambda_0\mu_0\mu + \text{const.} \\
&= (n+\lambda_0)\mu^2 - 2\mu(n\bar{\mathbf{x}} + \lambda_0\mu_0) \\
&= (n+\lambda_0)\left(\mu^2 - \frac{2\mu(n\bar{\mathbf{x}}+\lambda_0\mu_0)}{(n+\lambda_0)}\right) \\
&= (n+\lambda_0)\left(\mu - \frac{(n\bar{\mathbf{x}}+\lambda_0\mu_0)}{(n+\lambda_0)}\right)^2 + \text{const.}
\end{aligned}
\tag{26}
$$

Therefore, we have:

$$
\begin{aligned}
\log\left(q_\mu^*(\mu)\right) &= -\frac{\mathbb{E}_{q_\tau}[\tau]}{2}\left[\sum_{i=1}^{n}(x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] + \text{const.} \\
&= -\frac{\mathbb{E}_{q_\tau}[\tau](n+\lambda_0)}{2}\left(\mu - \frac{(n\bar{\mathbf{x}}+\lambda_0\mu_0)}{(n+\lambda_0)}\right)^2 + \text{const.} \\
\Longrightarrow\ q_\mu^*(\mu) &= \mathcal{N}\left(\frac{n\bar{\mathbf{x}}+\lambda_0\mu_0}{n+\lambda_0}, \mathbb{E}_{q_\tau}[\tau](n+\lambda_0)\right) \qquad \because -\frac{\tau}{2}(x-\mu)^2
\end{aligned}
\tag{27}
$$

4.3 Computing $\log\left(q_i^*(\tau)\right) = \mathbb{E}_{q_\mu(\mu)}\left[\log\left(p(\mu,\tau,\mathcal{D})\right)\right]$

$$
\begin{aligned}
\log\left(q_\tau^*(\tau)\right) &= \mathbb{E}_{q_\mu}\left[\log\left(p(\mu,\tau,\mathcal{D})\right)\right] \\
&= \mathbb{E}_{q_\mu}\left[\log(p(\mathcal{D}|\mu,\tau)) + \log p(\mu|\tau) + \log p(\tau)\right] + \text{const.} \\
&= \mathbb{E}_{q_\mu}\Big[\underbrace{\frac{n}{2}\log(\tau) - \frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2}_{\log(p(\mathcal{D}|\mu,\tau))} \underbrace{-\frac{\lambda_0\tau}{2}(\mu-\mu_0)^2}_{\log p(\mu|\gamma)} \underbrace{+(a_0-1)\log(\tau) - b_0\tau}_{\log p(\tau)}\Big] + \text{const.}
\end{aligned}
\tag{28}
$$

Bring terms without μ outside of the integral:

$$= \frac{n}{2}\log(\tau) + (a_0 - 1)\log(\tau) - b_0\tau - \frac{\tau}{2}\mathbb{E}_{q_\mu(\mu)}\left[\sum_{i=1}^{n}(x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] + \text{const.}$$

$$= \Big(\underbrace{\frac{n}{2} + a_0}_{a_n} - 1\Big)\log(\tau) - \tau\Big(\underbrace{b_0 + \frac{1}{2}\mathbb{E}_{q_\mu(\mu)}\Big[\sum_{i=1}^{n}(x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2\Big]}_{b_n}\Big) + \text{const.} \tag{29}$$

We can rewrite,

$$\begin{aligned} b_n &= b_0 + \frac{1}{2}\mathbb{E}_{q_\mu}\left[\sum_{i=1}^{n}(x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] \\ &= b_0 + \frac{1}{2}\mathbb{E}_{q_\mu}\left[-2\mu n\bar{x} + n\mu^2 + \lambda_0\mu^2 - 2\lambda_0\mu_0\mu\right] + \sum_{i=1}^{n}(x_i)^2 + \lambda_0\mu_0^2 \\ &= b_0 + \frac{1}{2}\left[(n+\lambda_0)\mathbb{E}_{q_\mu}[\mu^2] - 2\left(n\bar{x} + \lambda_0\mu_0\right)\mathbb{E}_{q_\mu}[\mu] + \sum_{i=1}^{n}(x_i)^2 + \lambda_0\mu_0^2\right] \end{aligned} \tag{30}$$

We will compute $\mathbb{E}_{q_\mu}[\mu]$ and $\mathbb{E}_{q_\mu}[\mu^2]$ since we know of $q_\mu(\mu)$ from previously.

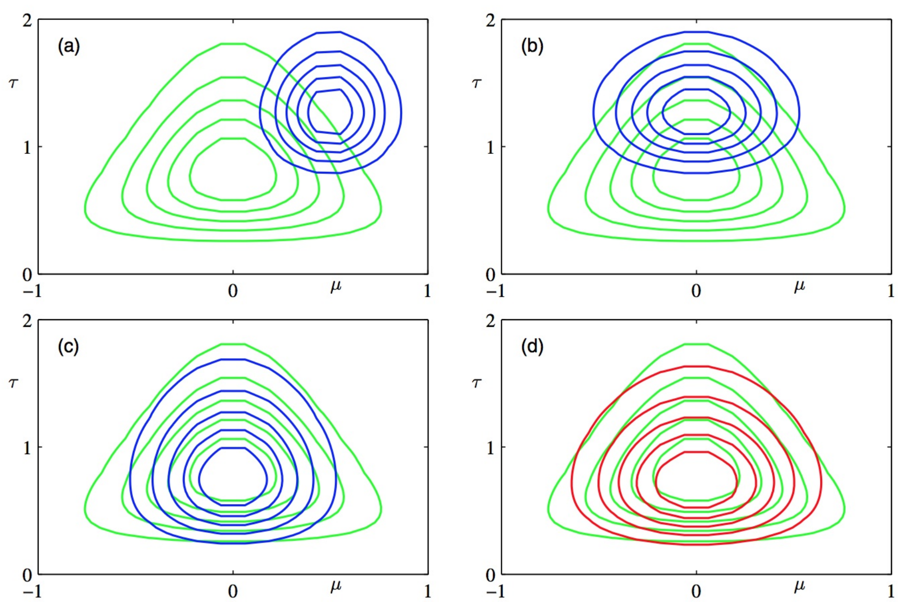


Figure 1: update for Normal Gamma: figure from [1]

5 Example of Gaussian Mixture Model Optional

5.1 The joint density

$$\begin{aligned} p(X, Z, \mu, \Lambda, \pi) &= p(X|Z, \mu, \Lambda, \pi)p(Z|\mu, \Lambda, \pi)p(\mu|\Lambda, \pi)p(\Lambda|\pi)p(\pi) \\ &= p(X|Z, \mu, \Lambda)p(Z|\pi)p(\mu|\Lambda)p(\Lambda)p(\pi) \end{aligned} \tag{31}$$

5.2 Definitions for each probabilities

5.2.1 Definition for $p(Z|\pi)$:

first, is the probability of mixture indices, $Z = \{z_1, ..., z_N\}$, given weights π.

$$\begin{aligned} p(Z|\pi) &= \prod_{i=1}^{N} p(z_n|\pi) \\ &= \prod_{i=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{ik}} \end{aligned} \tag{32}$$

The reason for which $\left(p(z_n|\pi) = \prod_{k=1}^{K} \pi_k^{z_{ik}}\right)$, or $\left(p(z|\pi) = \prod_{k=1}^{K} \pi_k^{z_k}\right)$, is because in Bishop, z is not represented in a scalar form, but rather in a vector of dimension K, which has a single element 1, and the rest are all 0s. For example, instead of using $p(z_n = 2|\pi = [0.2, 0.3, 0.5]) = 0.3$, Bishop uses $p(z_n = [0, 1, 0]|\pi = [0.2, 0.3, 0.5]) = 0.3$. In any case, this refers to the second element of π. Therefore, a more simpler and vocal representation for $p(z|\pi)$ is just the z^{th} value of π.

5.2.2 Definition for $p(X|Z, \mu, \Lambda)$:

$$p(X|Z, \mu, \Lambda) = \prod_{i=1}^{N} p(x_n|z_n, \mu, \Lambda)$$

In normal literatures, such as Bilmes, it is defined as:

$$= \prod_{i=1}^{N} \mathcal{N}(x_n|\mu_{z_n}, \Lambda_{z_n}^{-1})$$

However, due to the vector representation of Bishop, the above is defined as:

$$= \prod_{i=1}^{N}\prod_{k=1}^{K} \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})^{z_{ik}} \tag{33}$$

However, the above two represent the same thing:

5.2.3 Definition for $p(\pi)$:

This is just a straight Dirichlet probability:

$$
\begin{aligned}
p(\pi|\alpha_0) &= \mathrm{Dir}(\pi|\alpha_0) \propto C(\alpha_0) \prod_{k=1}^{K} \pi_k^{\alpha_{0k}-1} \\
&\implies \log(\pi|\alpha_0) \propto (\alpha_0 - 1) \sum_{k=1}^{K} \log \pi_k
\end{aligned} \tag{34}
$$

5.2.4 Definition for $p(\mu|\Lambda)p(\Lambda)$:

This is almost always a Gaussian-Wishart distribution:

$$
\begin{aligned}
p(\mu, \Lambda) &= p(\mu|\Lambda)p(\Lambda) \\
&= \prod_{k=1}^{K} \mathcal{N}(\mu_k|m_0, (\beta_0\Lambda_k)^{-1})\mathcal{W}(\Lambda_k|W_0, v_0)
\end{aligned} \tag{35}
$$

5.3 Begin VB of GMM

5.3.1 The expression for $q^*(Z)$:

$$
\begin{aligned}
\log q^*(Z) &= \mathbb{E}_{\pi,\mu,\Lambda}\left[\log p(X, Z, \pi, \mu, \Lambda)\right] + \text{const.} \\
&= \mathbb{E}_{\pi}\left[\log p(Z|\pi)\right] + \mathbb{E}_{\mu,\Lambda}\left[\log p(X|Z, \mu, \Lambda)\right] + \text{const.} \\
&= \mathbb{E}_{\pi}\left[\log \prod_{i=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{ik}}\right] + \mathbb{E}_{\mu,\Lambda}\left[\log \prod_{i=1}^{N}\prod_{k=1}^{K} \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})^{z_{ik}}\right] + \text{const.} \\
&= \mathbb{E}_{\pi}\left[\sum_{i=1}^{N}\sum_{k=1}^{K} \log \pi_k^{z_{ik}}\right] + \mathbb{E}_{\mu,\Lambda}\left[\sum_{i=1}^{N}\sum_{k=1}^{K} \log \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})^{z_{ik}}\right] + \text{const.} \\
&\text{given that:, } (\log a^b = b \log a) : \\
&= \mathbb{E}_{\pi}\left[\sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik} \log \pi_k\right] + \mathbb{E}_{\mu,\Lambda}\left[\sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik} \log \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})\right] + \text{const.} \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\mathbb{E}_{\pi}[\log \pi_k] + \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\mathbb{E}_{\mu,\Lambda}[\log \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})] + \text{const.} \\
&\text{Taking the common term to the left, } \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik} : \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\left(\mathbb{E}_{\pi}[\log \pi_k] + \mathbb{E}_{\mu,\Lambda}[\log \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})]\right) + \text{const.} \\
&\text{Bishop nominates a new term: } \log \rho_{ik} \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\left(\log \rho_{ik}\right) + \text{const.}
\end{aligned} \tag{36}
$$

Let's look at the expression for $\log \rho_{ik}$:

$$
\begin{aligned}
\log \rho_{ik} &= \mathbb{E}_{\pi}[\log \pi_k] + \mathbb{E}_{\mu_k,\Lambda_k}[\log \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})] \\
&= \mathbb{E}_{\pi}[\log \pi_k] + \mathbb{E}_{\mu_k,\Lambda_k}\left[\log\left(\frac{1}{(2\pi)^{(d/2)}}|\Lambda_k|^{1/2}\exp\left(-\frac{1}{2}(x_n-\mu_k)^\top\Lambda_k(x_n-\mu_k)\right)\right)\right] \\
&= \mathbb{E}_{\pi}[\log \pi_k] + \mathbb{E}_{\mu_k,\Lambda_k}\left[\log(2\pi)^{\frac{-d}{2}} + \frac{1}{2}\log|\Lambda_k| + \left(-\frac{1}{2}(x_n-\mu_k)^\top\Lambda_k(x_n-\mu_k)\right)\right] \\
&= \mathbb{E}_{\pi}[\log \pi_k] + \mathbb{E}_{\mu_k,\Lambda_k}\left[\frac{-d}{2}\log(2\pi) + \frac{1}{2}\log|\Lambda_k| - \left(\frac{1}{2}(x_n-\mu_k)^\top\Lambda_k(x_n-\mu_k)\right)\right] \\
&= \mathbb{E}_{\pi}[\log \pi_k] + \frac{-d}{2}\log(2\pi) + \frac{1}{2}\mathbb{E}_{\Lambda_k}[\log|\Lambda_k|] - \frac{1}{2}\mathbb{E}_{\mu_k,\Lambda_k}\left[(x_n-\mu_k)^\top\Lambda_k(x_n-\mu_k)\right]
\end{aligned}
\tag{37}
$$

Now, since $\log q^*(Z) = \log \rho_{ik}$

$$
\begin{aligned}
\log q^*(Z) &= \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\left(\log \rho_{ik}\right) + \text{const.} \implies \\
q^*(Z) &= \exp\left(\sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\left(\log \rho_{ik}\right) + \text{const.}\right) \\
&= C\prod_{i=1}^{N}\prod_{k=1}^{K}\exp(z_{ik}\left(\log \rho_{ik}\right)) = C\prod_{i=1}^{N}\prod_{k=1}^{K}\exp\left(\log \rho_{ik}^{z_{ik}}\right) = C\prod_{i=1}^{N}\prod_{k=1}^{K}\rho_{ik}^{z_{ik}}
\end{aligned}
\tag{38}
$$

Since $q^*(Z) = \prod_{i=1}^{N} q^*(z_n)$:

$$
q^*(Z) = \prod_{i=1}^{N} C\prod_{k=1}^{K}\rho_{ik}^{z_{ik}} \tag{39}
$$

In a way, $\rho_{ik}^{z_{ik}}$ plays the same role as π in $p(z_n|\pi)$, therefore, $\sum_{k=1}^{K}\pi_k = 1 \implies \sum_{k=1}^{K}\rho_{ik} = 1$:

$$
\begin{aligned}
q^*(Z) &= \prod_{i=1}^{N} q^*(z_i) = \prod_{i=1}^{N}\left(\frac{1}{\sum_{j=1}^{K}\rho_{nj}}\prod_{k=1}^{K}\rho_{ik}^{z_{ik}}\right) \\
&= \prod_{i=1}^{N}\prod_{k=1}^{K}\frac{\rho_{ik}^{z_{ik}}}{\sum_{j=1}^{K}\rho_{nj}} = \prod_{i=1}^{N}\prod_{k=1}^{K} r_{nk}^{z_{ik}}
\end{aligned}
\tag{40}
$$

This is a multinomial distribution, therefore, $\mathbb{E}[z_i = k] = r_{ik}$

5.3.2 The expression for $q^*(\pi, \mu, \Lambda)$:

$$
\begin{aligned}
\log q^*(\pi,\mu,\Lambda) &= \mathbb{E}_Z\left[\log p(X, Z, \pi, \mu, \Lambda)\right] + \text{const.} \\
&= \mathbb{E}_Z\left[\log p(X|Z,\mu,\Lambda)\right] + \mathbb{E}_Z\left[\log p(Z|\pi)\right] + \mathbb{E}_Z\left[\log p(\pi)\right] + \mathbb{E}_Z\left[\log p(\mu|\Lambda)\right] + \mathbb{E}_Z\left[\log p(\Lambda)\right] + \text{const.} \\
&= \mathbb{E}_Z\left[\log p(X|Z,\mu,\Lambda)\right] + \mathbb{E}_Z\left[\log p(Z|\pi)\right] + \log p(\pi) + \log p(\mu|\Lambda) + \log p(\Lambda) + \text{const.}
\end{aligned}
\tag{41}
$$

Combine the mean and precision together:

$$= \mathbb{E}_Z\left[\log p(X|Z,\mu,\Lambda)\right] + \mathbb{E}_Z\left[\log p(Z|\pi)\right] + \log p(\pi) + \log p(\mu,\Lambda) + \text{const.}$$

And since each (μ_k, Λ_k) are independent, therefore:

$$\begin{aligned}
&= \mathbb{E}_Z\left[\log \prod_{i=1}^{N}\prod_{k=1}^{K}\mathcal{N}(x_n|\mu_k,\Lambda_k^{-1})^{z_{ik}}\right] + \mathbb{E}_Z\left[\log p(Z|\pi)\right] + \log p(\pi) + \sum_{k=1}^{K}\log p(\mu_k,\Lambda_k) + \text{const.}\\
&= \mathbb{E}_Z\left[\sum_{i=1}^{N}\sum_{k=1}^{K}\log(z_{ik})\mathcal{N}(x_n|\mu_k,\Lambda_k^{-1})\right] + \mathbb{E}_Z\left[\log p(Z|\pi)\right] + \log p(\pi) + \sum_{k=1}^{K}\log p(\mu_k,\Lambda_k) + \text{const.}\\
&= \sum_{k=1}^{K}\sum_{i=1}^{N}\mathbb{E}_Z[\log(z_{ik})]\mathcal{N}(x_n|\mu_k,\Lambda_k^{-1}) + \mathbb{E}_Z\left[\log p(Z|\pi)\right] + \log p(\pi) + \sum_{k=1}^{K}\log p(\mu_k,\Lambda_k) + \text{const.}\\
&= \underbrace{\mathbb{E}_Z\left[\log p(Z|\pi)\right] + \log p(\pi)}_{\log q^*(\pi)} + \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{N}\mathbb{E}_Z[\log(z_{ik})]\mathcal{N}(x_n|\mu_k,\Lambda_k^{-1}) + \sum_{k=1}^{K}\log p(\mu_k,\Lambda_k)}_{\log q^*(\mu,\Lambda)} + \text{const.}
\end{aligned} \tag{42}$$

For the part of $\log q^*(\pi)$:

$$\begin{aligned}
\log q^*(\pi) &= \mathbb{E}_Z\left[\log p(Z|\pi)\right] + \log p(\pi)\\
&= \mathbb{E}_Z\left[\log \prod_{i=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{ik}}\right] + \log p(\pi)\\
&= \mathbb{E}_Z\left[\sum_{i=1}^{N}\sum_{k=1}^{K}z_{ik}\log \pi_k\right] + \log p(\pi)\\
&= \sum_{i=1}^{N}\sum_{k=1}^{K}\log \pi_k \mathbb{E}_Z[z_{ik}] + (\alpha_0 - 1)\sum_{k=1}^{K}\log \pi_k + \text{const.}\\
&= \sum_{k=1}^{K}\log \pi_k \sum_{i=1}^{N} r_{i,k} + (\alpha_0 - 1)\sum_{k=1}^{K}\log \pi_k + \text{const.}\\
&= \left(\underbrace{\sum_{i=1}^{N} r_{i,k} + \alpha_0}_{a_n} -1\right)\sum_{k=1}^{K}\log \pi_k + \text{const.} = \text{DIR}\left(\pi|a_n\right)
\end{aligned} \tag{43}$$

For the part of $\log q^*(\mu,\Lambda)$:

$$\log q^*(\mu,\Lambda) = \sum_{k=1}^{K}\sum_{i=1}^{N}\mathbb{E}_Z[\log(z_{ik})]\mathcal{N}(x_n|\mu_k,\Lambda_k^{-1}) + \sum_{k=1}^{K}\log p(\mu_k,\Lambda_k) \tag{44}$$

We only have the expression for $\mathbb{E}_{q^*(Z)}[Z]$, but not $\mathbb{E}_{q^*(Z)}[\log(Z)]$:

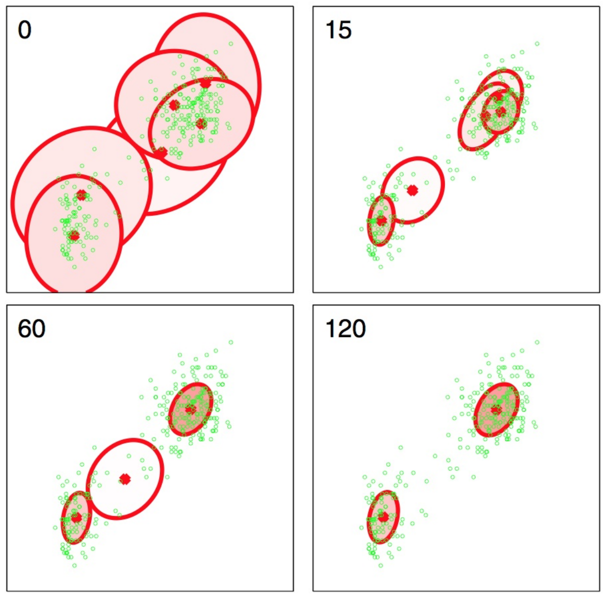


Figure 2: update for Gaussian Mixture Model: figure from [1]

6 Exponential Family distributions

6.1 Big picture

Given both the prior and likelihood are exponential family distributions and are in conjugacy, the variational inference (also mean-field approximation), i.e., $q(\mathbf{z}) = \prod_i q_i(z_i)$ can have the following update formula:

$$\eta_j = \mathbb{E}_{q(\mathbf{z} \setminus z_j | \cdot)}[\eta_{\text{post}}(\mathbf{z} \setminus z_j)] \tag{45}$$

where $\eta_{\text{post}}(\mathbf{z} \backslash z_j)$ is the natural parameter associated with posterior distribution $p(z_j | -)$. Of course it is expressed in terms of all other $\mathbf{z} \setminus z_j$,but z_j as part of its parameter.

Obviously, the corresponding $q(\cdot)$ must first exclude z_j.

compare this with the generic update formula:

$$\log\left(q_i^*(z_i)\right) = \mathbb{E}_{i \neq j}\big[\log\left(p(\mathbf{x}, \mathbf{z})\right)\big] \tag{46}$$

using exponential family update formula Eq.(45), the update is directly applied to the parameter.

Also note that using Eq.(48):

$$\begin{aligned} p(x) &= h(x) \exp\left(T(x)^\top \eta - A(\eta)\right) \\ \implies \log(p(x)) &\propto \eta \end{aligned} \tag{47}$$

6.2 Exponential Family

Most of the distributions we are going to look at are from **exponential family**. They are expressed in terms of its natural parameter η:

$$h(x) \exp\left(T(x)^\top \eta - A(\eta)\right) \tag{48}$$

$$\begin{aligned} &\underbrace{\exp(-A(\eta))}_{\text{normalization}} h(x) \exp\{T(x)^\top \eta\} \\ \implies &\exp(-A(\eta)) \int_x h(x) \exp\{T(x)^\top \eta\} = 1 \\ \implies &\int_x h(x) \exp\{T(x)^\top \eta\} = \exp(A(\eta)) \end{aligned} \tag{49}$$

6.3 example: 1-d Gaussian

$$
\begin{aligned}
\mathcal{N}(x;\mu,\sigma^2) &= (2\pi\sigma^2)^{-1/2}\exp^{-\frac{(x-\mu)^2}{2\sigma^2}} \\
&= \exp\left(-\frac{x^2-2x\mu+\mu^2}{2\sigma^2}-\frac{1}{2}\log(2\pi\sigma^2)\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}x^2+\frac{\mu}{\sigma^2}x-\frac{\mu^2}{2\sigma^2}-\frac{1}{2}\log(2\pi\sigma^2)\right) \\
&= \exp\left(\begin{bmatrix} x & x^2\end{bmatrix}\begin{bmatrix}\frac{\mu}{\sigma^2} & -\frac{1}{2\sigma^2}\end{bmatrix}^\top-\frac{\mu^2}{2\sigma^2}-\frac{1}{2}\log(2\pi\sigma^2)\right)
\end{aligned} \tag{50}
$$

$$
\begin{aligned}
T(\mathbf{x}) &= \begin{bmatrix} x & x^2\end{bmatrix} \\
\boldsymbol{\eta} &= \begin{bmatrix} \eta_1 & \eta_2\end{bmatrix} \\
&= \begin{bmatrix}\frac{\mu}{\sigma^2} & -\frac{1}{2\sigma^2}\end{bmatrix}
\end{aligned} \tag{51}
$$

1. for η_2:

$$
\eta_2 = -\frac{1}{2\sigma^2} \implies \sigma^2 = -\frac{1}{2\eta_2} \tag{52}
$$

2. for η_1:

$$
\begin{aligned}
\eta_1 = \frac{\mu}{\sigma^2} \implies \mu &= \eta_1\sigma^2 \\
&= \eta_1\frac{-1}{2\eta_2} \\
&= \frac{-\eta_1}{2\eta_2}
\end{aligned} \tag{53}
$$

summarize, we have:

$$
\theta = \begin{bmatrix}\mu \\ \sigma^2\end{bmatrix} = \begin{bmatrix}\frac{-\eta_1}{2\eta_2} \\ \frac{-1}{2\eta_2}\end{bmatrix} \tag{54}
$$

6.3.1 in natural parameter form

now we can remove μ and σ^2:

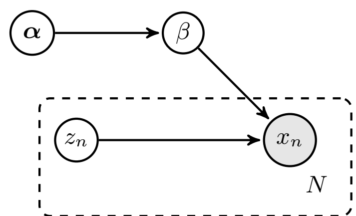
$$
\begin{aligned}
\mathcal{N}_{\text{nat}}(x,\boldsymbol{\eta}) &= \exp\left(\begin{bmatrix} x & x^2\end{bmatrix}\begin{bmatrix}\eta_1 & \eta_2\end{bmatrix}^\top-\frac{\mu^2}{2\sigma^2}-\frac{1}{2}\log(2\pi\sigma^2)\right) \\
&= \exp\left(\begin{bmatrix} x & x^2\end{bmatrix}\begin{bmatrix}\eta_1 & \eta_2\end{bmatrix}^\top-\frac{\left(\frac{-\eta_1}{2\eta_2}\right)^2}{2\left(\frac{-1}{2\eta_2}\right)}-\frac{1}{2}\log\left(2\pi\left(\frac{-1}{2\eta_2}\right)\right)\right) \\
&= \exp\left(T(x)^\top\boldsymbol{\eta}+\frac{\eta_1^2}{4\eta_2}-\frac{1}{2}\log\left(\frac{2\pi}{-2\eta_2}\right)\right) \\
&= \exp\left(T(x)^\top\boldsymbol{\eta}+\frac{\eta_1^2}{4\eta_2}+\frac{1}{2}\log(-2\eta_2)-\frac{1}{2}\log(2\pi)\right)
\end{aligned} \tag{55}
$$

now that the probability is fully in terms of the natural parameter

$$\mathcal{N}_{\text{nat}}(x, \boldsymbol{\eta}) = \exp\left(T(x)^\top \boldsymbol{\eta} - \underbrace{\left(\frac{-\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2)\right) - \frac{1}{2}\log(2\pi)}_{A(\boldsymbol{\eta})}\right) \tag{56}$$

6.4 Problem setting

It's always better to have a discussion with a concrete example setup. So we have the following problem setup, described in [2]:



joint density is of the form:

$$p(\mathbf{x}, \mathbf{z}, \boldsymbol{\beta}|\boldsymbol{\alpha}) = p(\boldsymbol{\beta}|\boldsymbol{\alpha}) \prod_{n=1}^{N} p(x_n, z_n|\boldsymbol{\beta}) \tag{57}$$

the conditionals are based on Exponential family:

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{x}, \mathbf{z}, \alpha) &= h(\boldsymbol{\beta}) \exp\left\{T(\boldsymbol{\beta})^\top \eta_{\text{post}}(\mathbf{x}, \mathbf{z}, \alpha) - A_{\text{post}}(\eta_{\text{post}}(\mathbf{x}, \mathbf{z}, \alpha))\right\} \\ p(z_{n,j}|x_n, z_{n,-j}, \boldsymbol{\beta}) &= h(z_{n,j}) \exp\left\{T(z_{n,j})\eta_{z_{n,j}}(x_n, z_{n,-j}, \boldsymbol{\beta}) - A_l\big(\eta_{z_{n,j}}(x_n, z_{n,-j}, \boldsymbol{\beta})\big)\right\} \end{aligned} \tag{58}$$

Think about why is this representation useful? Let's have look at a numerical example:

6.5 Conjugacy of exponential family distribution

Let's work through a concrete example of posterior $p(\boldsymbol{\beta}|x_n, z_n)$, instead of writing $\boldsymbol{\eta}_\beta$, we write $\boldsymbol{\beta}$ directly:

- prior:

$$p(\boldsymbol{\beta}|\boldsymbol{\alpha}) = h(\boldsymbol{\beta}) \exp\{T(\beta)^\top \boldsymbol{\alpha} - A_{\text{pri}}(\boldsymbol{\alpha})\} \tag{59}$$

suppose the sufficient statistics of the **prior** can be written as:

$$\begin{aligned} T(\boldsymbol{\beta}) &= \begin{bmatrix} \boldsymbol{\beta} \\ -A_l(\boldsymbol{\beta}) \end{bmatrix} \\ \implies \boldsymbol{\alpha} &= \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \end{aligned} \tag{60}$$

then the prior itself can be written as:

$$p(\boldsymbol{\beta}) = h(\boldsymbol{\beta}) \exp \left\{ \begin{bmatrix} \boldsymbol{\beta} \\ -A_l(\boldsymbol{\beta}) \end{bmatrix}^\top \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} - A_{\text{pri}}(\alpha) \right\} \tag{61}$$

- likelihood:

  and if the likelihood density (x_n, z_n) can be defined as:

$$p(x_n, z_n|\beta) = h(x_n, z_n) \exp \left\{ T(x_n, z_n)^\top \boldsymbol{\beta} - A_l(\boldsymbol{\beta}) \right\} \tag{62}$$

  then posterior condition on a single data point:

$$\begin{aligned} p(\boldsymbol{\beta}|x_n, z_n, \boldsymbol{\alpha}) &\propto \underbrace{h(\boldsymbol{\beta}) \exp\{T(\beta)^\top \boldsymbol{\alpha}\}} \underbrace{\exp\{T(x_n, z_n)^\top \boldsymbol{\beta} - A_l(\boldsymbol{\beta})\}} \\ &= h(\boldsymbol{\beta}) \exp \left\{ \boldsymbol{\beta}^\top \alpha_1 - \alpha_2 A_l(\boldsymbol{\beta}) + \boldsymbol{\beta}^\top T(x_n, z_n) - A_l(\boldsymbol{\beta}) \right\} \\ &= h(\boldsymbol{\beta}) \exp \left\{ \boldsymbol{\beta}^\top (\alpha_1 + T(x_n, z_n)) - \alpha_2 A_l(\boldsymbol{\beta}) - A_l(\boldsymbol{\beta}) \right\} \\ &= h(\boldsymbol{\beta}) \exp \left\{ \boldsymbol{\beta}^\top (\alpha_1 + T(x_n, z_n)) - (\alpha_2 + 1) A_l(\boldsymbol{\beta}) \right\} \\ &= h(\boldsymbol{\beta}) \exp \left\{ \begin{bmatrix} \boldsymbol{\beta} \\ -A_l(\boldsymbol{\beta}) \end{bmatrix}^\top \begin{bmatrix} \alpha_1 + T(x_n, z_n) \\ \alpha_2 + 1 \end{bmatrix} \right\} \\ &= h(\boldsymbol{\beta}) \exp \left\{ T(\beta)^\top \begin{bmatrix} \alpha_1 + T(x_n, z_n) \\ \alpha_2 + 1 \end{bmatrix} \right\} \end{aligned} \tag{63}$$

posterior on all data:

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{x}, \mathbf{z}, \boldsymbol{\alpha}) &\propto h(\boldsymbol{\beta}) \exp \left\{ \begin{bmatrix} \boldsymbol{\beta} \\ -A_l(\boldsymbol{\beta}) \end{bmatrix}^\top \begin{bmatrix} \hat{\alpha}_1 & \hat{\alpha}_2 \end{bmatrix} \right\} \\ &= h(\boldsymbol{\beta}) \exp \left\{ T(\beta)^\top \begin{bmatrix} \alpha_1 + \sum_{n=1}^N T(x_n, z_n) \\ \alpha_2 + N \end{bmatrix} \right\} \end{aligned} \tag{64}$$

6.5.1 Complete likelihood

$$\begin{aligned} p(\mathbf{x}, \mathbf{z}|\beta) &= \prod_{n=1}^N h(x_n, z_n) \exp\{\boldsymbol{\beta}^\top T(x_n, z_n) - A_l(\boldsymbol{\beta})\} \\ &= h(\mathbf{x}, \mathbf{z}) \exp \left\{ \sum_{n=1}^N \boldsymbol{\beta}^\top T(x_n, z_n) - N \times A_l(\boldsymbol{\beta}) \right\} \end{aligned} \tag{65}$$

6.5.2 Complete posterior

now, look at:

$$p(\boldsymbol{\beta}|\mathbf{x},\mathbf{z},\boldsymbol{\alpha}) \propto h(\boldsymbol{\beta}) \exp\left\{T(\beta)^\top \begin{bmatrix} \alpha_1 + \sum_{n=1}^N T(x_n, z_n) \\ \alpha_2 + N \end{bmatrix}\right\} \tag{66}$$

When we use the expression and use use η_{post} instead:

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{x},\mathbf{z},\boldsymbol{\alpha}) &= h(\boldsymbol{\beta}) \exp\left\{\eta_{\text{post}}(\mathbf{x},\mathbf{z},\boldsymbol{\alpha})^\top T(\beta) - A_{\text{post}}(\eta_{\text{post}}(\mathbf{x},\mathbf{z},\alpha))\right\} \\ \implies \eta_{\text{post}}(\mathbf{x},\mathbf{z},\boldsymbol{\alpha}) &= \begin{bmatrix} \alpha_1 + \sum_{n=1}^N t(x_n, z_n) \\ \alpha_2 + N \end{bmatrix} \\ \implies A_{\text{post}}(\eta_{\text{post}}(\mathbf{x},\mathbf{z},\boldsymbol{\alpha})) &= \int_{\boldsymbol{\beta}} h(\boldsymbol{\beta}) \exp\left\{\eta_{\text{post}}(\mathbf{x},\mathbf{z},\boldsymbol{\alpha})^\top T(\beta)\right\} \mathrm{d}\boldsymbol{\beta} \end{aligned} \tag{67}$$

6.6 Example: Posterior of Gaussian mean

6.6.1 likelihood

suppose data x_i come from unit variance Gaussian. Compare with Section (6.3), we saved one parameter:

$$\begin{aligned} p(x|\mu) &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-\mu)^2\right\} \\ &= \underbrace{\frac{\exp\left(-x^2/2\right)}{\sqrt{2\pi}}}_{h(x)} \exp\left\{\underbrace{\mu}_{\beta} \underbrace{x}_{T(x)} - \underbrace{\frac{\mu^2}{2}}_{A_l(\beta)}\right\} \end{aligned} \tag{68}$$

Therefore:

$$\begin{aligned} \beta &= \mu \\ T(x) &= x \\ A_l(\beta) &= \frac{\beta^2}{2} \\ h(x) &= \frac{\exp\left(-x^2/2\right)}{\sqrt{2\pi}} \end{aligned} \tag{69}$$

substitute into:

$$p(x|\beta) = \frac{\exp\left(-x^2/2\right)}{\sqrt{2\pi}} \exp\left\{\beta x + \underbrace{-\frac{\beta^2}{2}}_{A_l(\beta)}\right\} \tag{70}$$

6.6.2 what should the conjugate prior be?

A **conjugate prior** MUST be:

$$\begin{aligned} p(\beta|\alpha) &= h(\beta)\exp\left\{\alpha_1\beta + \alpha_2\underbrace{(-\beta^2/2)}_{A_l(\beta)} - A_g(\alpha)\right\} \\ &= h(\beta)\exp\left\{\begin{bmatrix}\alpha_1\\ \alpha_2\end{bmatrix}^\top \begin{bmatrix}\beta\\ -\frac{\beta_2}{2}\end{bmatrix} - A_g(\alpha)\right\} \end{aligned} \tag{71}$$

Wait, this doesn't look exactly in the form of Eq.(50), i.e.,:

$$\mathcal{N}(x;\mu,\sigma^2) = \exp\left(\begin{bmatrix}\frac{\mu}{\sigma^2}\\ -\frac{1}{2\sigma^2}\end{bmatrix}^\top \begin{bmatrix}x\\ x^2\end{bmatrix} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right) \tag{72}$$

We can arrange Eq.(71) to look like, but with parameter $\begin{bmatrix}\alpha_1 & -\frac{\alpha_2}{2}\end{bmatrix}^\top$:

$$p(\boldsymbol{\beta}|\alpha) = h(\boldsymbol{\beta})\exp\left\{\begin{bmatrix}\alpha_1\\ -\frac{\alpha_2}{2}\end{bmatrix}^\top \begin{bmatrix}\beta\\ \beta^2\end{bmatrix} - A_g(\alpha)\right\} \tag{73}$$

From our knowledge, a distribution with sufficient statistics $T(\beta) = \begin{bmatrix}\beta & \beta^2\end{bmatrix}$ is a Gaussian distribution.

Suppose the likelihood is an exponential family distribution. Every exponential family has a conjugate prior in theory. The natural parameter $\boldsymbol{\alpha} = \begin{bmatrix}\alpha_1 & \alpha_2\end{bmatrix}^\top$ has dimension dim(β) + 1. The sufficient statistics of the prior are $\begin{bmatrix}\beta & -A_l(\beta)\end{bmatrix}^\top$

6.7 For exponential family distribution: $\mathbb{E}_q[T(\beta)] = \nabla_\lambda A_g(\lambda)$

given that we have:

$$\begin{aligned} q(\beta|\lambda) &= h(\beta)\exp\{\lambda^\top T(\beta) - A_g(\lambda)\} \\ &= \frac{1}{\exp(A_g(\lambda))} h(\beta)\exp\{\lambda^\top T(\beta)\} \end{aligned} \tag{74}$$

why is it that we have:

$$\mathbb{E}_{q(\beta)}[T(\beta)] = \nabla_\lambda A_g(\lambda) \tag{75}$$

$$
\begin{aligned}
&\int_\beta q(\beta|\lambda)\mathrm{d}\beta = \int_\beta h(\beta)\exp\{\lambda^\top T(\beta) - A_g(\lambda)\}\mathrm{d}\beta = 1 \\
&\implies \nabla_\lambda \left(\int_\beta h(\beta)\exp\{\lambda^\top T(\beta) - A_g(\lambda)\}\mathrm{d}\beta\right) = 0 \\
&\implies \int_\beta \nabla_\lambda \left(h(\beta)\exp\{\lambda^\top T(\beta) - A_g(\lambda)\}\right)\mathrm{d}\beta = 0 \\
&\implies \int_\beta h(\beta)\exp\left\{\lambda^\top T(\beta) - A_g(\lambda)\right\}(T(\beta) - \nabla_\lambda A_g(\lambda)) = 0 \\
&\implies \int_\beta h(\beta)\exp\left\{\lambda^\top T(\beta) - A_g(\lambda)\right\}T(\beta) - \int_\beta h(\beta)\exp\left\{\lambda^\top T(\beta) - A_g(\lambda)\right\}\nabla_\lambda A_g(\lambda) = 0 \\
&\implies \mathbb{E}_{q(\beta)}[T(\beta)] - \nabla_\lambda A_g(\lambda) = 0
\end{aligned} \tag{76}
$$

6.8 The choice of $q(\beta, \mathbf{z})$

We choose $q(\beta, \mathbf{z})$ to decouple β and $\mathbf{z}$ completely:

$$
q(\beta, \mathbf{z}) = q(\beta|\lambda)\prod_{n=1}^{N}\prod_{j=1}^{J} q(z_{n,j}|\phi_{n,j}) \tag{77}
$$

- $q(\beta|\lambda)$ is the SAME distribution type as $p(\beta|\mathbf{x}, \mathbf{z}, \alpha)$, they only differ in parameter. This means they have the same sufficient statistics $T(\beta)$:

$$
\begin{aligned}
& q(\beta|\lambda) = h(\beta)\exp\{\lambda^\top T(\beta) - A_g(\lambda)\} \\
\text{compare with:}\quad & p(\beta|\mathbf{x}, \mathbf{z}, \alpha) = h(\beta)\exp\left\{\eta_{\text{post}}(\mathbf{x}, \mathbf{z}, \alpha)^\top T(\beta) - A_{\text{post}}(\eta_{\text{post}}(\mathbf{x}, \mathbf{z}, \alpha))\right\}
\end{aligned} \tag{78}
$$

- $q(z_{n,j}|\phi_{n,j})$ is the SAME distribution type as $p(z_{n,j}|x_n, z_{n,-j}, \beta)$, they only differ in parameter. This means they have the same sufficient statistics $T(z_{n,j})$:

$$
\begin{aligned}
& q(z_{n,j}|\phi_{n,j}) = h(z_{n,j})\exp\left\{\phi_{n,j}^\top T(z_{n,j}) - A_l(\phi_{n,j})\right\} \\
\text{compare with:}\quad & p(z_{n,j}|x_n, z_{n,-j}, \beta) = h(z_{n,j})\exp\left\{\eta_l(x_n, z_{n,-j}, \beta)^\top T(z_{n,j}) - A_l(\eta_l(x_n, z_{n,-j}, \beta))\right\}
\end{aligned} \tag{79}
$$

6.9 Proof for for ELBO(λ) for $q(\beta|\lambda)$ Optional

this section shows the proof for the update formula used in Eq.(45), i.e., $\eta_j = \mathbb{E}_{q(\mathbf{z}\setminus z_j|\cdot)}[\eta_{\text{post}}(\mathbf{z}\setminus z_j)]$, we will do so using an example from the setting described in this section.

Our goal is to maximize the ELBO, i.e.,

$$
\text{ELBO}(q) \triangleq \mathbb{E}_{q(\beta,\mathbf{z})}[\log p(\mathbf{x}, \mathbf{z}, \beta|\alpha)] - \mathbb{E}_{q(\beta,\mathbf{z})}[\log q(\mathbf{z}, \beta)] \tag{80}
$$

Note that q used here is $q(\beta, \mathbf{z})$ not just $q(\beta|\lambda)$

$$
\begin{aligned}
\text{ELBO}(\lambda) &= \mathbb{E}_{q(\beta,\mathbf{z})}[\log p(\beta|\mathbf{x},\mathbf{z},\alpha)] + \mathbb{E}_{q(\beta,\mathbf{z})}[\log p(\mathbf{x},\mathbf{z})] - \mathbb{E}_{q(\beta,\mathbf{z})}[\log q(\beta)] \\
&= \mathbb{E}_q\big[\log p(\beta|\mathbf{x},\mathbf{z},\alpha)\big] - \mathbb{E}_q\big[\log q(\beta)\big] + \text{const.} \\
&= \mathbb{E}_q\Big[\log\big(h(\beta)\exp\big\{\eta_{\text{post}}(\mathbf{x},\mathbf{z},\alpha)^\top T(\beta) - A_{\text{post}}(\eta_{\text{post}}(\mathbf{x},\mathbf{z},\alpha))\big\}\big)\Big] - \mathbb{E}_q[\log q(\beta)] + \text{const.} \\
&= \mathbb{E}_q[\log(h(\beta))] + \underbrace{\mathbb{E}_q[\eta_{\text{post}}(\mathbf{x},\mathbf{z},\alpha)^\top T(\beta)]} - \mathbb{E}_q[\log h(\beta)\exp\{\lambda^\top T(\beta) - A_{\text{pri}}(\lambda)\}] + \text{const.} \\
&= \mathbb{E}_q[\log(h(\beta))] + \underbrace{\mathbb{E}_{q(\mathbf{z}|\Phi)}[\eta_{\text{post}}(\mathbf{x},\mathbf{z},\alpha)]^\top \mathbb{E}_{q(\beta|\lambda)}[T(\beta)]} - \mathbb{E}_q[\log h(\beta)] - \mathbb{E}_q[\lambda^\top T(\beta)] + A_{\text{pri}}(\lambda) + \text{const.} \\
&= \mathbb{E}_{q(\mathbf{z}|\Phi)}[\eta_{\text{post}}(x,z,\alpha)]^\top \mathbb{E}_{q(\beta|\lambda)}[T(\beta)] - \lambda^\top \mathbb{E}_{q(\beta|\lambda)}[T(\beta)] + A_{\text{pri}}(\lambda) + \text{const.} \qquad \because A_{\text{pri}}(\lambda) \text{ contains } \lambda
\end{aligned} \tag{81}
$$

Substitute $\mathbb{E}_{q(\beta|\lambda)}[T(\beta)] = \nabla_\lambda A_{\text{pri}}(\lambda)$:

$$
\text{ELBO}(\lambda) = \mathbb{E}_{q(\mathbf{z}|\Phi)}[\eta_{\text{post}}(x,z,\alpha)]^\top \nabla_\lambda A_{\text{pri}}(\lambda) - \lambda^\top \nabla_\lambda A_{\text{pri}}(\lambda) + A_{\text{pri}}(\lambda) + \text{const.} \tag{82}
$$

Maximize $\text{ELBO}(\lambda)$ we get:

$$
\begin{aligned}
\nabla_\lambda \text{ELBO}(\lambda) &= \mathbb{E}_{q(\mathbf{z}|\Phi)}[\eta_g(\mathbf{x},\mathbf{z},\alpha)]^\top \nabla^2_\lambda A_{\text{pri}}(\lambda) - \nabla_\lambda A_{\text{pri}}(\lambda) - \lambda^\top \nabla^2_\lambda A_{\text{pri}}(\lambda) + \nabla_\lambda A_{\text{pri}}(\lambda) = 0 \\
&= \mathbb{E}_{q(\mathbf{z}|\Phi)}[\eta_g(\mathbf{x},\mathbf{z},\alpha)]^\top \nabla^2_\lambda A_{\text{pri}}(\lambda) - \lambda^\top \nabla^2_\lambda A_{\text{pri}}(\lambda) = 0 \\
&\implies \nabla^2_\lambda A_{\text{pri}}(\lambda)\big(\mathbb{E}_{q(\mathbf{z}|\Phi)}[\eta_{\text{post}}(\mathbf{x},\mathbf{z},\alpha)]^\top - \lambda^\top\big) = 0
\end{aligned} \tag{83}
$$

$$
\lambda = \mathbb{E}_{q(\mathbf{z}|\Phi)}[\eta_{\text{post}}(\mathbf{x},\mathbf{z},\alpha)] \tag{84}
$$

in words, when we try to update λ for $q(\beta|\lambda)$, it find the corresponding posterior $p(\beta|\mathbf{x},\mathbf{z},\alpha)$, and its natural parameter $\eta_{\text{post}}(\mathbf{x},\mathbf{z},\alpha)$, then computes the expectation with all the $q(\cdot)$ that its natural parameter has random variable for.

6.9.1 Update for ELBO($\phi_{n,j}$) for $q(z_{n,j}|\phi_{n,j})$

In a very similar fashion to $\mathcal{L}(\lambda)$, we can prove:

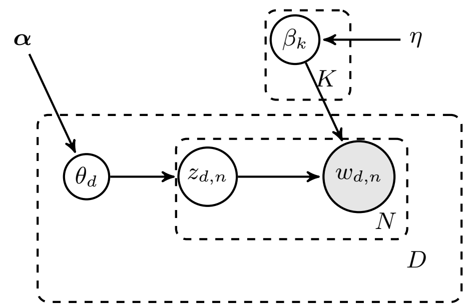
$$
\nabla_{\phi_{n,j}} \text{ELBO}(\phi_{n,j}) = \nabla^2_{\phi_{n,j}} A_l(\phi_{n,j})\big(\mathbb{E}_{q(\lambda)}[\eta_l(x_n, z_{n,-j}, \beta)]^\top - \phi_{n,j}^\top\big) = 0 \tag{85}
$$

$$
\phi_{n,j} = \mathbb{E}_{q(\lambda)}\big[\eta_l(x_n, z_{n,-j}, \beta)\big] \tag{86}
$$

in words, when we try to update $\phi_{n,j}$ for $q(z_{n,j}|\phi_{n,j})$, it find the corresponding posterior $p(z_{n,j}|x_n, z_{n,-j})$, and its natural parameter $\eta_l(x_n, z_{n,-j})$, then computes the expectation with all the $q(\cdot)$ that its natural parameter has random variable for.

7 Latent Dirichlet Allocation

let's visit Latent Dirichlet Allocation again [3]:



- $\beta_k \sim \text{Dir}(\eta, \dots \eta)$ for $k \in \{1, ..., K\}$.
- For each document d:
  $\theta \sim \text{Dir}(\alpha, \dots, \alpha)$
  For each word $w \in \{1, ..., N\}$:
  $z_{dn} \sim \text{Mult}(\theta_d)$
  $w_{dn} \sim \text{Mult}(\beta_{z_{dn}})$

7.1 define corresponding $q(\cdot)$

1. $q(z_{d,n})$

$$\begin{aligned} q(z_{d,n}) &= \text{Mult}(\phi_{d,n}) \\ \implies q(z_{d,n} = k) &= \phi_{d,n}^k \end{aligned} \tag{87}$$

2. $q(\beta_k)$

$$q(\beta_k) = \text{Dir}(\lambda_k) \tag{88}$$

3. $q(\theta_d)$

$$q(\theta_d) = \text{Dir}(\gamma_d) \tag{89}$$

7.1.1 Facts about Dirichlet Distribution

$$\begin{aligned} &\theta \sim \text{Dir}(\gamma_1, \dots \gamma_K) \\ \implies &\mathbb{E}[\log(\theta_k)|\gamma] = \Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K} \gamma_i\right) \quad \text{for component } k \end{aligned} \tag{90}$$

where:

$$\Psi(x) = \frac{\mathrm{d}}{\mathrm{d}x} \ln\big(\Gamma(x)\big) = \frac{\Gamma'(x)}{\Gamma(x)} \tag{91}$$

7.2 Updating $q(z_{d,n}|\phi_{d,n})$: optimize $\phi_{d,n}$

7.2.1 find natural parameter of posterior $p(z_{dn} = k|\theta_d, \beta_{1:K}, w_{d,n})$

$$
\begin{aligned}
p(z_{dn} = k|\theta_d, \beta_{1:K}, w_{d,n}) &\propto p(z_{d,n} = k|\theta_d) p(w_{d,n}|z_{d,n} = k, \beta_{1:K}) \\
&= \text{Mult}(\theta_{d,k}) \times \text{Mult}(\beta_{k,w_{d,n}}) \\
&\propto \exp\left(\underbrace{\log(\theta_{d,k}) + \log(\beta_{k,w_{d,n}})}_{\eta_l\left(\theta_d, \beta_{1:K}, w_{d,n}\right)} \times \underbrace{1}_{T(z_{d,n})} \right)
\end{aligned} \tag{92}
$$

7.2.2 optimize $\phi_{d,n}$

apply the update formula, in which we need the natural parameter for $p(z_{d,n}|\theta_d, \beta_{1:K}, w_{d,n})$ in the exception:

$$
\begin{aligned}
\eta(\phi_{d,n}^k) = \log(\phi_{d,n}^k) &\propto \mathbb{E}_{q(\theta_d)q(\beta_k)} \left[\eta_l\left(\theta_d, \beta_{1:K}, w_{d,n}\right)\right] \\
&= \mathbb{E}_{q(\theta_d, \beta_{1:K})} \left[\log(\theta_{d,k})\right] + \mathbb{E}_{q(\beta_k)} \left[\log(\beta_{k,w_{d,n}})\right] \\
&= \Psi(\gamma_{d,k}) - \Psi\Big(\sum_{k=1}^{K} \gamma_{d,k}\Big) + \Psi\Big(\lambda_{k,w_{d,n}}\Big) - \Psi\Big(\sum_{v} \lambda_{k,v}\Big)
\end{aligned} \tag{93}
$$

compare this with Eq.(45), i.e., $\eta_j = \mathbb{E}_{q(\mathbf{z} \setminus z_j)}[\eta_{\text{post}}(\mathbf{z} \setminus z_j)]$, you can see easily that:

$$
\mathbf{z} \setminus z_j \equiv \{\theta_d, \beta_{1:K}\} \tag{94}
$$

to obtain $\phi_{d,n}$:

$$
\begin{aligned}
\implies \phi_{d,n}^k &\propto \exp\left[\Psi(\gamma_{d,k}) - \underbrace{\Psi\Big(\sum_{k=1}^{K} \gamma_{d,k}\Big)}_{\text{irrelevant in proportionality}} + \Psi\Big(\lambda_{k,w_{d,n}}\Big) - \Psi\Big(\sum_{v} \lambda_{k,v}\Big)\right] \\
&\propto \exp\left[\Psi(\gamma_{d,k}) + \Psi\Big(\lambda_{k,w_{d,n}}\Big) - \Psi\Big(\sum_{v} \lambda_{k,v}\Big)\right]
\end{aligned} \tag{95}
$$

7.3 Updating $q(\theta_d|\gamma_d)$: optimize γ_d

7.3.1 find natural parameter of posterior $p(\theta_d|\mathbf{z}_d)$

$$
\begin{aligned}
p(\theta_d|\mathbf{z}_d) &= p(\theta_d|\alpha)\prod_{n=1}^{N} p(z_{d,n}|\theta_d) = \mathrm{Dir}(\alpha)\times\prod_{n=1}^{N}\mathrm{Mult}(z_{d,n}|\theta_d) \\
&= \prod_k \left(\theta_{d,k}^{\alpha_k-1}\prod_{n=1}^{N}\theta_{d,k}^{\mathbb{1}(z_{d,n}=k)}\right) \\
&= \exp\left[\log\left(\prod_k\left(\theta_{d,k}^{\alpha_k-1}\prod_{n=1}^{N}\theta_{d,k}^{\mathbb{1}(z_{d,n}=k)}\right)\right)\right] \\
&= \exp\left[\sum_k \log\left(\theta_{d,k}^{\alpha_k-1}\prod_{n=1}^{N}\theta_{d,k}^{\mathbb{1}(z_{d,n}=k)}\right)\right] \\
&= \exp\left[\sum_k\left(\log\theta_{d,k}^{\alpha_k-1}+\sum_{n=1}^{N}\log\left(\theta_{d,k}^{\mathbb{1}(z_{d,n}=k)}\right)\right)\right] \\
&= \exp\left[\sum_k\left((\alpha_k-1)\log\theta_{d,k}+\sum_{n=1}^{N}\mathbb{1}(z_{d,n}=k)\log\theta_{d,k}\right)\right] \\
&= \exp\left[\sum_k\left(\alpha_k-1+\sum_{n=1}^{N}\mathbb{1}(z_{d,n}=k)\right)\log\left(\theta_{d,k}\right)\right] \\
&= \exp\left(\underbrace{\begin{bmatrix}(\alpha_1-1+n_1)\\ \dots \\ (\alpha_K-1+n_K)\end{bmatrix}^{\top}}_{\eta_l(\alpha,z_d)}\underbrace{\begin{bmatrix}\log(\theta_{d,1})\\ \dots \\ \log(\theta_{d,K})\end{bmatrix}}_{T(\theta_d)}\right) \qquad \text{by letting } n_k=\sum_{n=1}^{N}\mathbb{1}(z_{d,n}=k)
\end{aligned}
\tag{96}
$$

7.3.2 optimize γ_d

$$
\begin{aligned}
\eta(\gamma_d) &= \mathbb{E}_{q(z_{d,n}|\phi_{d,n})}\left[\eta_l\left(\alpha,z_d\right)\right] \\
&= \mathbb{E}_{q(z_{d,n}|\phi_{d,n})}\left[(\alpha_1-1+n_1)\quad\dots\quad(\alpha_K-1+n_K)\right] \\
&= \left[(\alpha_1-1+n_1\phi_{d,n}^{1})\quad\dots\quad(\alpha_K-1+n_K)\phi_{d,n}^{K})\right] \\
&= \left[(\alpha_1-1+\textstyle\sum_{n=1}^{N}\mathbb{1}(z_{d,n}=1)\phi_{d,n}^{1})\quad\dots\quad(\alpha_K-1+\textstyle\sum_{n=1}^{N}\mathbb{1}(z_{d,n}=K)\phi_{d,n}^{K})\right]
\end{aligned}
\tag{97}
$$

compare this with Eq.(45), i.e., $\eta_j=\mathbb{E}_{q(\mathbf{z}\setminus z_j)}[\eta_{\text{post}}(\mathbf{z}\setminus z_j)]$, you can see easily that:

$$
\mathbf{z}\setminus z_j \equiv \{z_{d,n}\} \tag{98}
$$

to obtain γ_d:

$$
\begin{aligned}
\gamma_d &= \left[\left(\alpha_1+\textstyle\sum_{n=1}^{N}\mathbb{1}(z_{d,n}=1)\phi_{d,n}^{1}\right)\quad\dots\quad\left(\alpha_K+\textstyle\sum_{n=1}^{N}\mathbb{1}(z_{d,n}=K)\phi_{d,n}^{K}\right)\right] \\
&= \boldsymbol{\alpha}+\sum_{n=1}^{N}\phi_{d,n}
\end{aligned}
\tag{99}
$$

7.4 Updating $q(\beta_k|\lambda_k)$ optimize λ_k

7.4.1 find natural parameter of posterior $p(\beta_k|\mathbf{z},\mathbf{w})$

$$
\begin{aligned}
p(\beta_k|\mathbf{z},\mathbf{w}) &= p(\beta_k|\eta)\prod_{d=1}^{D}\prod_{n=1}^{N} p\left(w_{d,n}|\beta_k\right)^{\mathbb{1}(z_{d,n}=k)} = \mathrm{Dir}(\eta)\times\prod_{d=1}^{D}\prod_{n=1}^{N}\beta_k^{w_{d,n}\mathbb{1}(z_{d,n}=k)} \\
&\propto \exp\left(\underbrace{\left(\eta-1+\sum_{d=1}^{D}\sum_{n=1}^{N} w_{d,n}\mathbb{1}(z_{d,n}=k)\right)}_{\eta_l(\eta,Z,W)}\times\underbrace{\log(\beta_k)}_{t(\beta_k)}\right)
\end{aligned}
\tag{100}
$$

7.4.2 optimize λ_k

$$
\begin{aligned}
\eta(\lambda_k) &= \mathbb{E}_{\prod_{d=1}^{D}\prod_{n=1}^{N} q(z_{d,n}|\phi_{d,n}^k)}\left[\eta_l(\eta,\mathbf{z},\mathbf{w})\right] \\
&= \mathbb{E}_{\prod_{d=1}^{D}\prod_{n=1}^{N} q(z_{d,n}|\phi_{d,n}^k)}\left[\eta-1+\sum_{d=1}^{D}\sum_{n=1}^{N} w_{d,n}\mathbb{1}(z_{d,n}=k)\right] \\
&= \eta-1+\sum_{d=1}^{D}\sum_{n=1}^{N} w_{d,n}\phi_{d,n}^k
\end{aligned}
\tag{101}
$$

$$
\lambda_k = \eta+\sum_{d=1}^{D}\sum_{n=1}^{N} w_{d,n}\phi_{d,n}^k \tag{102}
$$

8 Collapsed Variational Inference Optional

$$q(z_{d,n}) = \text{Mult}(\phi_{d,n}) \text{ or } q(z_{d,n}=k) = \phi_{d,n}^k \qquad q(\beta_k) = \text{Dir}(\lambda_k) \qquad q(\theta_d) = \text{Dir}(\gamma_d) \tag{103}$$

$$\begin{aligned}\implies q(Z, \theta_1 \dots \theta_D, \beta_1 \dots \beta_K) &= \left(\prod_{d=1}^{d=D} \prod_{n=1}^{N} q(z_{d,n}|\phi_{d,n})\right) \prod_{d=1}^{D} q(\theta_d|\gamma_d) \prod_{k=1}^{K} q(\theta_k|\lambda_k) \\ \text{now change to: } &= \underbrace{\left(\prod_{d=1}^{d=D} \prod_{n=1}^{N} q(z_{d,n}|\phi_{d,n})\right)}_{q(Z)} q(\Theta, \beta|Z)\end{aligned} \tag{104}$$

Maximize ELOB, it becomes: (remove X for clarity)
Let $U = \{\Theta, \beta\}$:

$$\begin{aligned}\text{ELBO}(q) &\triangleq \mathbb{E}_{q(U,Z)}[\log p(Z,U)] - \mathbb{E}_{q(U,Z)}[\log q(Z,U)] \\ &= \mathbb{E}_{q(U,Z)}[\log p(Z,U)] - \mathbb{E}_{q(U,Z)}[\log q(U|Z) - \log q(Z)] \\ &= \mathbb{E}_{q(Z)}\left(\mathbb{E}_{q(U|Z)}[\log p(Z,U)]\right) - \mathbb{E}_{q(Z)}\left(\mathbb{E}_{q(U|Z)}[\log q(U|Z)]\right) - \mathbb{E}_{q(Z,U)}[\log q(Z)] \\ &= \mathbb{E}_{q(Z)}\left(\underbrace{\mathbb{E}_{q(U|Z)}\left([\log p(Z,U)] - [\log q(U|Z)]\right)}_{\mathcal{L}(q(U|Z))}\right) - \mathbb{E}_{q(Z)}[\log q(Z)]\end{aligned} \tag{105}$$

Think this as treating Z as X.
(removed X for clarity)

$$\begin{aligned}\underset{q(U|Z)}{\arg\max}(\text{ELBO}(q)) &= \underset{q(U|Z)}{\arg\max}\left[\mathbb{E}_{q(Z)}\left(\underbrace{\mathbb{E}_{q(U|Z)}\left([\log p_X(Z,U)] - [\log q(U|Z)]\right)}_{\mathcal{L}(q(U|Z))}\right) - \mathbb{E}_{q(Z)}[\log q(Z)]\right] \\ &= \mathbb{E}_{q(Z)}\left(\underbrace{\underset{q(U|Z)}{\arg\max}\left[\mathbb{E}_{q(U|Z)}\left([\log p(Z,U)] - [\log q(U|Z)]\right)\right]}\right) - \mathbb{E}_{q(Z)}[\log q(Z)] \\ &= \mathbb{E}_{q(Z)}[\underbrace{p(Z)}] - \mathbb{E}_{q(Z)}[\log q(Z)]\end{aligned} \tag{106}$$

$$\underset{q(U|Z)}{\arg\max}\left[\mathbb{E}_{q(U|Z)}\left([\log p(Z,U)] - [\log q(U|Z)]\right)\right] = p(Z) \tag{107}$$

maximum occur when $q(U|Z) = p(U|Z) \implies \mathbb{KL}\left(q(U|Z) \| p(U|Z)\right) = 0$

References

[1] Christopher M Bishop and Nasser M Nasrabadi, *Pattern recognition and machine learning*, vol. 4, Springer, 2006.

[2] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley, "Stochastic variational inference," *Journal of Machine Learning Research*, 2013.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.