DD2434/FDD3434 Machine Learning, Advanced Course
Module 3 Exercise

November 2023

Contents

3 Variational Inference – Exercises

3.1 Beta and Binomial model

Let $X = (X_1, .., X_N)$ be i.i.d. where $X_n | m, \theta \sim Binomial(m, \theta)$ and $\theta \sim Beta(\alpha, \beta)$.
a) Derive the CAVI updates for $q(\theta)$ using equation **??**.
b) How does this compare to the posterior in exercise 1.1 of Module 1? Describe qualitatively in one sentence why this is the case.

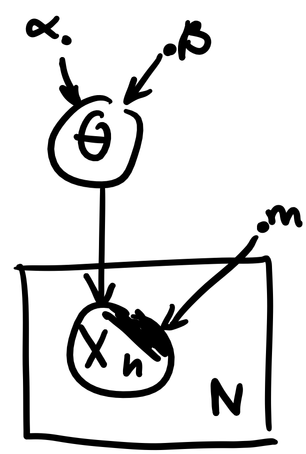
3.1.1 Solution

Exercises Module 3

Sunday, 15 October 2023   15:49

$X = (X_1, \dots, X_N)$, $\quad X_n \mid m, \theta \sim \text{Bin}(m, \theta)$

$\theta \sim \text{Beta}(\alpha, \beta)$



a) CAVI update

Solution:

Can use fact: $\mathbb{E}_{-\theta}\left[\log p(\theta, X)\right]$

1. Google $p(X_n \mid m, \theta)$ and $p(\theta)$ (Wikipedia page of Beta distribution and Binomial distribution usually works)

$$p(X_n = x_n \mid m, \theta) = \binom{m}{x_n} \theta^{x_n} (1-\theta)^{m - x_n}$$

$$p(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

2. Log form of $p(X, \theta)$:

$$\log p(X, \theta) = \log \prod_{n=1}^{N} \binom{m}{x_n} \theta^{x_n} \cdot (1-\theta)^{m - x_n} \cdot \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

Since this is the expected value w.r.t. all variables except theta,
But we have no other variables.

3. $$q^*(\theta) = \cancel{\mathbb{E}}_{-\theta}\left[\log \prod_{n=1}^{N} \binom{m}{x_n} \theta^{x_n} \cdot (1-\theta)^{m-x_n} \cdot \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\cancel{B(\alpha,\beta)}} \right]$$

Only interested in the expression up to additive constant

$$\stackrel{+}{=} \sum_{n=1}^{N} \log \binom{m}{x_n} \theta^{x_n} (1-\theta)^{m-x_n} \cdot \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

$$= \sum_{n=1}^{N} \log\binom{m}{x_n} + x_n \log\theta + (m-x_n)\log(1-\theta) + (\alpha-1)\log\theta$$

$$+ (\beta-1)\log(1-\theta) \stackrel{+}{=} \log\theta \cdot \left(\sum_{n=1}^{N} x_n + \alpha - 1\right)$$

$$+ \log(1-\theta) \cdot \left(N \cdot m - \sum x_n + \beta - 1\right)$$

Note: Same as for the true posterior in Module 1. exercise. Why?

Two explanations:

1. Since $q^*(\theta) = \mathbb{E}_{-\theta}[\log p(X,\theta)]$ simplifies to $\log p(X,\theta)$, which is the same as when showing the conjugate-prior relation in Module 1.

2. Our mean-field assumption is only over one variable, hence, we are not making any simplifying assumption at all.

3.2 Gaussian Mixture Model - light

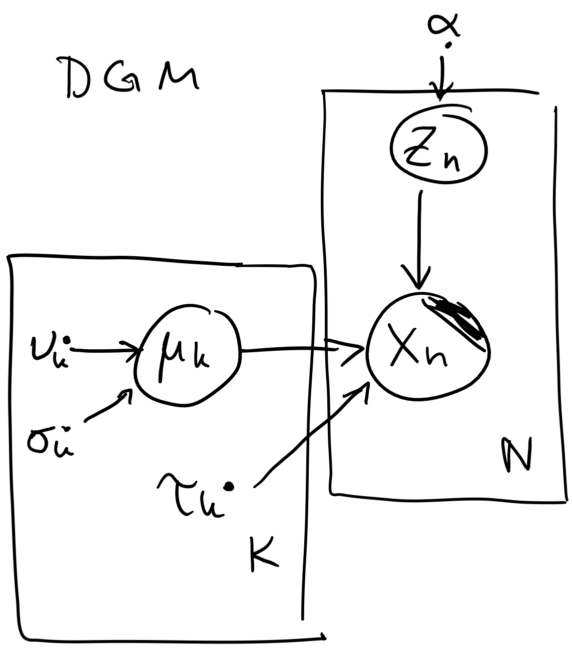
Here we will examine an simpler version of the Gaussian Mixture model. Still $p(X_n|Z_n = k, \mu_k, \tau_k) = Normal(\mu_k, \frac{1}{\tau_k})$, $p(Z_n|\pi) = Categorical(\pi)$, but we assume π and τ_k are given and let $p(\mu_k) = Normal(\nu_k, \sigma_k)$.
a) Write the DGM/Bayes net for the model.
b) Write out $\log p(X, Z, \mu)$.
c) Apply and state the mean-field approximation for Z and μ.
d) Derive the associated CAVI updates using **??**.
e) Implement the CAVI algorithm **??** and apply it to simulated data using the generative model (If you are unfamiliar with this, it will be shown in the Exercise session of module 3). Try simulating data for different K, N, ν_k, τ and π - under what circumstances does it have trouble finding all clusters?
f) In how many iterations does it converge?

3.2.1 Solution

GMM - light

den 13 november 2023 11:34

a) DGM



b) $$\log p(X, Z, \mu) = \sum_{n,k}^{N,K} \log p(x_n | z_n, \mu_k, \tau_k)^{\mathbb{1}\{z_n = k\}}$$

$$+ \log p(Z|\pi) + \log p(\mu)$$

$$\log p(X|Z,\mu) = \sum_{\substack{n=1\\k=1}}^{N,K} \mathbb{1}\{z_n = k\} \left(\log\left(\frac{\tau_k}{\sqrt{2\pi}} \right) - \frac{\tau_k}{2} (x_n - \mu_k)^2 \right)$$

$$\log p(Z|\pi) = \sum_{\substack{n=1,\\k=1}}^{N,K} z_n^k \log \pi_k$$

$$\log p(\mu) = \log \prod_{k=1}^{K} p(\mu_k) = \sum_{k=1}^{K} \log \frac{1}{\sigma_k \sqrt{2\pi}} - \frac{1}{2\sigma_k^2} (\mu_k - \nu_k)^2$$

c) Mean-field approximation: $q(z,\mu) = q(z)\,q(\mu)$

$$\log q^*(z) \overset{\text{eq.7.}}{=} \mathbb{E}_{-z}\left[\log p(X,z,\mu)\right] \overset{\pm}{=} \mathbb{E}_\mu\left[\log p(X|z,\mu)\right] \; ① $$

$$+ \underbrace{\mathbb{E}_\mu\left[\log p(z|\pi)\right]}_{= \log p(z|\pi) \; ②}$$

since $p(z|\pi)$ not dep. on μ

$$① = \mathbb{E}_\mu\left[\sum_{n=1,k=1}^{N,K} \mathbb{1}\{z_n=k\}\left(\log\frac{\tau_k}{\sqrt{2\pi}} - \frac{\tau_k}{2}(x_n-\mu_k)^2\right)\right]$$

$$= \sum_{n,k}^{N,K} \mathbb{1}\{z_n=k\}\left(\log\frac{\tau_k}{\sqrt{2\pi}} - \frac{\tau_k}{2}\mathbb{E}_\mu\left[(x_n-\mu_k)^2\right]\right)$$

$\Rightarrow$ ① + ② =

after we derive $q(\mu)$ update we can get expressions for these

$$\sum_{n,k}^{N,K} \mathbb{1}\{z_n=k\}\underbrace{\left(\log\frac{\tau_k}{\sqrt{2\pi}} - \frac{\tau_k}{2}\left(x_n^2 - 2x_n\mathbb{E}_\mu[\mu_k] + \mathbb{E}_\mu[\mu_k^2]\right) + \log\pi_k\right)}_{\equiv \log \rho_{nk}}$$

$$= \sum_{n,k}^{N,K} \mathbb{1}\{z_n=k\}\log\rho_{nk}$$

$$\Rightarrow q^*(z) \propto \prod_{n,k}^{N,K} \rho_{nk}^{\mathbb{1}\{z_n=k\}} \Rightarrow q^*(z) = \prod_{n,k}^{N,K} r_{nk}^{\mathbb{1}\{z_n=k\}},$$

i.e. $q(Z) = \prod_{n=1}^{N} q(z_n)$, where $r_{nk} = \dfrac{\rho_{nk}}{\sum_{k=1}^{K} \rho_{nk}}$

where each $q(z_n) = \mathrm{Cat}(r_{n1}, \dots, r_{nK})$

$$\log q(\mu)^* = \mathbb{E}_{-\mu}[\log p(X, Z, \mu)] \doteq \mathbb{E}_Z[\log p(X \mid Z, \mu)] + \log p(\mu)$$

$$= \sum_{n=1, k=1}^{N, K} \mathbb{E}_Z \left[1_{\{z_n = k\}} \left(\cancel{\log \frac{\tau_k}{\sqrt{2\pi}}} - \frac{\tau_k}{2} (x_n - \mu_k)^2 \right) \right] + \sum_{k=1}^{K} \cancel{\log \frac{1}{\sigma_k \sqrt{2\pi}}} - \frac{1}{2\sigma_k^2} (\mu_k - \nu_k)^2$$

$$\doteq \sum_{k=1}^{K} \sum_{n=1}^{N} \left\{ - \overbrace{\mathbb{E}_Z[1_{\{z_n = k\}}]}^{= q(z_n = k)} \frac{\tau_k}{2} \left(\cancel{x_n^2} - 2 \cdot \mu_k \cdot x_n + \mu_k^2 \right) \right\} - \frac{1}{2 \cdot \sigma_k^2} \left(\mu_k^2 - 2 \mu_k \nu_k + \cancel{\nu_k^2} \right)$$

$$= \sum_{k=1}^{K} -\frac{1}{2} \left(- \underbrace{\left(\sum_{n=1}^{N} q(z_n = k) \cdot x_n \right)}_{\equiv A_k^N} \tau_k \cdot 2 \cdot \mu_k + \underbrace{\left(\sum_{n=1}^{N} q(z_n = k) \right)}_{\equiv B_k^N} \tau_k \mu_k^2 \right) - \frac{1}{2} \left(\frac{-1}{\sigma_k^2} \cdot 2 \cdot \mu_k \nu_k + \frac{1}{\sigma_k^2} \mu_k^2 \right)$$

$$= \sum_{k=1}^{K} -\frac{1}{2} \left(-2 \cdot \mu_k \left(A_k^N \cdot \tau_k + \frac{\nu_k}{\sigma_k^2} \right) + \mu_k^2 \left(B_k^N \cdot \tau_k + \frac{1}{\sigma_k^2} \right) \right)$$

$$= \sum_{k=1}^{K} - \frac{1}{2 \underbrace{\left(B_k^N \cdot \tau_k + 1/\sigma_k^2 \right)}_{\sigma_k^{2*}}} \left(-2 \mu_k \underbrace{\frac{A_k^N \cdot \tau_k + \frac{\nu_k}{\sigma_k^2}}{B_k^N \cdot \tau_k + 1/\sigma_k^2}}_{\nu_k^*} + \mu_k^2 \right)$$

$$\Rightarrow \quad q(\mu) = \prod_{k=1}^{K} q(\mu_k), \quad q(\mu_k) = \mathcal{N}(\nu_k^*, \sigma_k^{2*})$$

$$\Rightarrow \begin{cases} \mathbb{E}_{\mu}[\mu_k] = \nu_k^* , \\ \mathbb{E}_{\mu_k}[\mu_k^2] = Var(\mu_k) - \mathbb{E}[\mu_k]^2 = \sigma_k^{2*} - \nu_k^{*2} \end{cases}$$

d) See Bernoulli Mixture Model solution for how to implement CAVI updates in VI-algorithm.

3.3 Mixture Model with Bernoulli observations

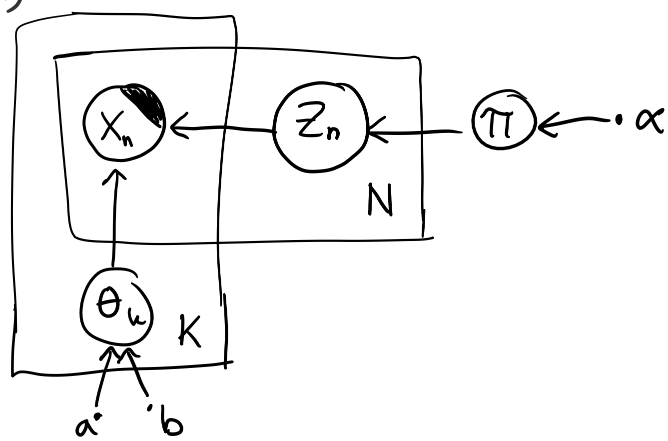
In the video lectures the CAVI updates for a Mixture model with Gaussian observational model is introduced, i.e., $p(X_n|Z_n = k, \mu_k, \tau_k) = Normal(\mu_k, \frac{1}{\tau_k})$, $p(Z_n|\pi) = Categorical(\pi)$, $p(\pi) = Dirichlet(\alpha)$.
In this exercise we examine a similar model, but with $X_n = \{X_{n1}, ..., X_{nD}\}$ with observational model $p(X_{nd}|Z_n = k, \theta_k) = Bernoulli(\theta_k)$ with prior $p(\theta_k) = Beta(a, b)$.
a) Write the DGM/Bayes net for the model.
b) Write out $\log p(X, Z, \pi, \theta)$.
c) Apply and state the mean-field approximation for Z, π and θ.
d) Derive the associated CAVI updates using **??**.
e) Implement the CAVI algorithm **??** and apply it to simulated data using the generative model (If you are unfamiliar with this, it will be shown in the Exercise session of module 3). Try simulating data for different K, N, θ_k and π - under what circumstances does it have trouble finding all clusters?
f) In how many iterations does it converge?

3.3.1 Solution

Bernoulli Mixture Model

den 8 november 2023 14:14

a)



b) $\log p(X,Z,\pi,\theta)$

Start with conditioning on X:

$$\log p(X,Z,\pi,\theta) = \log p(X|Z,\pi,\theta)\,p(Z,\pi,\theta)$$

$$= \log p(X|Z,\theta) + \log p(Z|\pi) + \log p(\pi) + \log p(\theta)$$

$$p(X|Z,\theta) = \prod_{n=1}^{N}\prod_{k=1}^{K} p(X_n=x_n|Z_n=k,\theta_k)^{\mathbb{1}\{Z_n=k\}} = \prod_{\substack{n=1,\\k=1}}^{N,K}\left(\prod_{d=1}^{D} p(X_{nd}=x_{nd}|Z_n=k,\theta_k)\right)^{\mathbb{1}\{Z_n=k\}}$$

$$= \prod_{n,k}^{N,K}\left(\prod_{d=1}^{D}\left(\theta_k^{x_{nd}}(1-\theta_k)^{1-x_{nd}}\right)\right)^{\mathbb{1}\{z_n=k\}}$$

$$\Rightarrow \log p(X|Z,\theta) = \sum_{n,k}^{N,K} z_n^k\left(\sum_{d=1}^{D} x_{nd}\log\theta_k + (1-x_{nd})\log(1-\theta_k)\right)$$

↳ shorter notation for $\mathbb{1}\{z_n=k\}$

$$= \sum_{n,k}^{N,K} z_n^k\left(\sum_{d=1}^{D} x_{nd}\log\theta_k + \Big(D-\underbrace{\sum_{d=1}^{D} x_{nd}}_{\equiv x_n^D}\Big)\log(1-\theta_k)\right)$$

$$\log p(Z|\pi) = \sum_{\substack{n=1,\\k=1}}^{N,K} z_n^k \log\pi_k$$

$$\log p(\pi) = \log\left(\frac{1}{B(\alpha)}\prod_{k=1}^{K}\pi_k^{\alpha_k-1}\right) \overset{+}{=} \cancel{-\log B(\alpha)} + \sum_{k=1}^{K}(\alpha_k-1)\log\pi_k$$

$$\log p(\theta) = \log\prod_{k=1}^{K} p(\theta_k) \overset{+}{=} \sum_{k=1}^{K}(a_k-1)\log\theta_k + (b_k-1)\log 1-\theta_k \cancel{-B(a_k,b_k)}$$

c) Mean-field approximation: $q(Z,\pi,\theta) = q(Z)\,q(\pi)\,q(\theta)$ (*)

Generally, we want to do as few assumptions as possible while still being able to solve the CAVI update equations.

That is why we avoid e.g. $q(Z,\pi,\theta) = \prod_{n=1}^{N} q(z_n)\, q(\pi) \prod_{k=1}^{K} q(\theta_k)$

However, $q(Z,\pi,\theta) = q(Z)\,q(\pi,\theta)$ would have been sufficient. For this exercise (*) is more clear.

d) Lets derive the CAVI updates for $q(Z)$, $q(\pi)$ and $q(\theta)$

$$\log q(\theta)^* = \mathbb{E}_{-\theta}[\log p(X,Z,\theta,\pi)] = \mathbb{E}_{-\theta}[\log p(X|Z,\theta) + \log p(\theta) + \cancel{\log p(Z|\pi)} + \cancel{\log p(\pi)}]$$

$$\overset{+}{=} \mathbb{E}_{-\theta}[\log p(X|Z,\theta) + \log p(\theta)] = \mathbb{E}_{-\theta}[\log p(X|Z,\theta)] + \mathbb{E}_{-\theta}[\log p(\theta)]$$

$$= \underbrace{\mathbb{E}_{-\theta}[\log p(X|Z,\theta)]}_{1} + \underbrace{\log p(\theta)}_{2}$$

$$= \mathbb{E}_{Z,\pi}\left[\sum_{n=1}^{N}\sum_{k=1}^{K} z_n^k \left(x_n^D \log \theta_k + (D - x_n^D) \log 1-\theta_k\right)\right]$$

$$= \sum_{n,k}^{N,K} \underbrace{\mathbb{E}_Z[z_n^k]}_{} \left(x_n^D \log \theta_k + (D - x_n^D) \log (1-\theta_k)\right)$$

$\mathbb{E}[\mathbb{1}(A=a)]$
$= \mathbb{P}(A=a) \rightarrow \mathbb{E}_Z[\mathbb{1}\{z_n = k\}]$
$\overset{N,K}{=} q(z_n = k)$

$$\Rightarrow 1 + 2 = \sum_{n,k}^{N,K} q(z_n=k)\left(x_n^D \log \theta_k + (D - x_n^D) \log (1-\theta_k)\right) + \sum_{k=1}^{K} (a-1)\log \theta_k + (b-1) \log 1-\theta_k$$

$$= \left\{ \begin{array}{l} \text{We suspect } q^*(\theta) \\ \text{in same form as} \\ \text{prior, so we rewrite} \\ \text{on same form as } 2 \end{array} \right\} = \sum_{k=1}^{K} \sum_{n=1}^{N} q(z_n=k)\, x_n^D \log \theta_k + (a-1) \log \theta_k$$

$$+ \sum_{n=1}^{N} q(z_n=k)(D-x_n^D)\log(1-\theta_k) + (b-1)\log(1-\theta_k)$$

=

$$= \sum_{k=1}^{K}\left(\underbrace{\sum_{n=1}^{N}\left[q(z_n=k)\,x_n^D\right]+a-1}_{a^*}\right)\log\theta_k + \left(\underbrace{\sum_{n=1}^{N}\left[q(z_n=k)(D-x_n^D)\right]+b-1}_{b^*}\right)\log(1-\theta_k)$$

By comparing to $\log p(\theta)$, we see that $\log q^*(\theta)$ is on the form of k independent Beta-distributions with parameters a^* and b^*,

i.e. $\quad q^*(\theta) = \prod_{k=1}^{K} q^*(\theta_k)$

where $\quad q^*(\theta_k) = \text{Beta}\left(\underbrace{\sum_{n=1}^{N}\left[q(z_n=k)\,x_n^D\right]+a-1}_{\equiv \hat{a}_k}, \underbrace{\sum_{n=1}^{N}\left[q(z_n=k)(D-x_n^D)\right]+b-1}_{\tilde{b}_k}\right)$

We will now do $q(z)$ and $q(\pi)$ in little less detail.

$$\log q^*(z) \overset{+}{=} \mathbb{E}_{-z}\left[\log p(X|z,\theta) + \log p(z|\pi)\right] =$$

$$= \underbrace{\mathbb{E}_{\theta}\left[\log p(X|z,\theta)\right]}_{1} + \underbrace{\mathbb{E}_{\pi}\left[\log p(z|\pi)\right]}_{2}$$

$$1 = \mathbb{E}_{\theta}\left[\sum_{n=1}^{N}\sum_{k=1}^{K} z_n^k\left(x_n^D\log\theta_k + (D-x_n^D)\log 1-\theta_k\right)\right]$$

$$= \sum_{n,k}^{N,K} z_n^k\left(x_n^D\,\mathbb{E}_{\theta}\left[\log\theta_k\right] + (D-x_n^D)\,\mathbb{E}_{\theta}\left[\log 1-\theta_k\right]\right)$$

$$\Rightarrow 1 + 2 = \sum_{n,k}^{N,K} z_n\left(x_n^D\,\mathbb{E}_{\theta}\left[\log\theta_k\right] + (D-x_n^D)\,\mathbb{E}_{\theta}\left[\log 1-\theta_k\right]\right) + \mathbb{E}_{\pi}\left[\sum_{n=1,\,k=1}^{N,K} z_n^k\log\pi_k\right]$$

only E_{θ_k} hits $\log\theta_k$

$$= \sum_{n,k}^{N,K} z_n^k\left(\underbrace{x_n^D\,\mathbb{E}_{\theta_k}\left[\log\theta_k\right] + (D-x_n^D)\,\mathbb{E}_{\theta_k}\left[\log 1-\theta_k\right] + \mathbb{E}_{\pi}\left[\log\pi_k\right]}_{\equiv \log\rho_{nk}}\right)$$

$$\Rightarrow \underset{\uparrow}{q^*(z)} \propto \underset{\uparrow}{\prod_{n,k}^{N,K}} \rho_{nk}^{z_{nk}} \quad \underset{\uparrow}{\Rightarrow} \quad q^*(z) = \prod_{\substack{n=1\\k=1}}^{N,K} r_{nk}^{z_{nk}}, \quad r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nj}}$$

{ See Bishop section 10.2.1 for more details on these steps }

Also, since we've already shown $q(\theta) = \mathrm{Beta}(\hat{a}, \tilde{b})$

digamma-function

we can evaluate $\mathbb{E}_{\theta_k}[\log \theta_k] = \left\{ \begin{matrix} \text{Wikipedia} \\ \text{Beta-dist.} \end{matrix} \right\} = \psi(\tilde{a}_k) - \psi(\tilde{a}_k + \tilde{b}_k)$

and $\mathbb{E}_{\theta_k}[\log(1-\theta_k)] = \psi(\tilde{b}_k) + \psi(\tilde{a}_k + \tilde{b}_k)$

$\mathbb{E}_\pi[\log \pi)$ we can only evaluate after analyzing $q(\pi)$:

$$\underline{\log q^*(\pi)} \overset{+}{=} \mathbb{E}_{-\pi}\left[\log p(z|\pi) + \log p(\pi)\right] = \mathbb{E}_z\left[\log p(z|\pi)\right] + \log p(\pi)$$

$$\overset{+}{=} \sum_{n,k}^{N,K} \mathbb{E}_z[z_{nk}] \log \pi_k + \sum_{k=1}^{K} (\alpha_k - 1) \log \pi_k = \sum_{k=1}^{K} \left(\sum_{n=1}^{N} q(z_n{=}k) + \alpha_k - 1 \right) \log \pi_k$$

$$\Rightarrow q^*(\pi) = \mathrm{Dir}\Big(\underbrace{\sum_{n=1}^{N} q(z_{nk}) + \alpha_k}_{\equiv \tilde{\alpha}} \Big)$$

Now, we can also calculate $\mathbb{E}_\pi[\log \pi_k] = \psi(\tilde{\alpha}_k) - \psi(\overset{\sum_{k=1}^{K} \tilde{\alpha}_i}{\tilde{\alpha}_0})$

In summary:

$$q(\theta) = \prod_{k=1}^{K} q(\theta_k), \quad q(\theta_k) = \mathrm{Beta}(\hat{a}_k, \tilde{b}_k), \quad \left\{ \begin{aligned} \hat{a}_k &= \sum_{n=1}^{N} \left[q(z_n{=}k)\, x_n^D \right] + a \\ \tilde{b}_k &= \sum_{n=1}^{N} \left[q(z_n{=}k)\, (D - x_n^D) \right] + b \end{aligned} \right.$$

$$q(z) = \prod_{n=1}^{N} q(z_n), \quad q(z_n) = \mathrm{Cat}(r_{n1}, \dots, r_{nK}), \quad r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nj}}$$

$$\log \rho_{nk} = x_n^D \left(\psi(\tilde{a}_k) - \psi(\tilde{a}_k + \tilde{b}_k) \right) + (D - x_n) \left(\psi(\tilde{b}_k) - \psi(\tilde{b}_k + \tilde{a}_k) \right) + \psi(\tilde{\alpha}_k) - \psi(\tilde{\alpha}_0)$$

$$q(\pi) = \mathrm{Dir}(\tilde{\alpha}), \quad \tilde{\alpha} = \sum_{n=1}^{N} \underbrace{q(z_n{=}k)}_{= r_{nk}} + \alpha_k$$

ELBO

$$\mathcal{L} = \mathbb{E}_{q(z)q(\theta)q(\pi)}\left[\log \frac{p(X,Z,\theta,\pi)}{q(Z,\theta,\pi)}\right] = \overset{①}{\mathbb{E}_{q(z)q(\theta)q(\pi)}\left[\log p(X|Z,\theta)\right]}$$

$$+ \overset{②}{\mathbb{E}_{q(z)q(\pi)}\left[\log p(Z|\pi)\right]} + \overset{③}{\mathbb{E}_{q(\pi)}\left[\log p(\pi)\right]} + \overset{④}{\mathbb{E}_{q(\theta)}\left[\log p(\theta)\right]}$$

$$- \overset{⑤}{\mathbb{E}_{q(z)}\left[\log q(Z)\right]} - \overset{⑥}{\mathbb{E}_{q(\pi)}\left[\log q(\pi)\right]} - \overset{⑦}{\mathbb{E}_{q(\theta)}\left[\log q(\theta)\right]}$$

← recall the expressions we derived in b)

$$① = \mathbb{E}_{q(z)q(\theta)}\left[\log \prod_{n,k}^{N,K}\left(\prod_{d=1}^{D}\left(\theta_k^{x_{nd}}(1-\theta_k)^{1-x_{nd}}\right)\right)^{1\{z_n=k\}}\right] =$$

$$= \sum_{n,k} \mathbb{E}_{q(z)q(\theta)}\left[1\{z_n=k\}\left(x_n^D \log \theta_k + (D-x_n^D)\log 1-\theta_k\right)\right] =$$

$$= \sum_{n,k} q(z_n=k)\left(x_n^D\, \mathbb{E}_\theta\left[\log \theta_k\right] + (D-x_n^D)\,\mathbb{E}_\theta\left[\log 1-\theta_k\right]\right)$$

$$② = \mathbb{E}_{q(z)q(\pi)}\left[\sum_{n=1,k=1}^{N,K} z_n^k \log \pi_k\right] = \sum_{n=1,k=1}^{N,K} q(z_n=k)\,\mathbb{E}_\pi\left[\log \pi_k\right]$$

$$③ = \mathbb{E}_{q(\pi)}\left[-\log B(\alpha) + \sum_{k=1}^{K}(\alpha_k-1)\log \pi_k\right] = -\log B(\alpha) + \sum_{k=1}^{K}(\alpha_k-1)\,\mathbb{E}_\pi\left[\log \pi_k\right]$$

$$④ = \mathbb{E}_{q(\theta)}\left[\sum_{k=1}^{K}(a_k-1)\log\theta_k + (b_k-1)\log 1-\theta_k - B(a_k,b_k)\right] = \sum_{k=1}^{K}(a_k-1)\,\mathbb{E}_\theta\left[\log\theta_k\right] + (b_k-1)\,\mathbb{E}_\theta\left[\log 1-\theta_k\right]$$
$$- B(a_k,b_k)$$

⑤, ⑥, ⑦ are entropies and we can look them up on Wikipedia (except ⑤)

$$⑤ = \mathbb{E}_{q(z)}\left[\sum_{n=1,k=1}^{N,K} \log r_{nk}^{1\{z_n=k\}}\right] = \sum_{n,k}^{N,K} q(z_n=k)\log q(z_n=k)$$

$$⑥ = \underset{\text{entropy}}{H_{q(\pi)}[\pi]} = \{\text{Wikipedia}\} = \log B(\alpha) + (\alpha_0 - K)\psi(\alpha_0) - \sum_{j=1}^{K}(\alpha_j-1)\psi(\alpha_j)$$

$$⑦ = H_{q(\theta)}[\theta] = \sum_{k=1}^{K} H_{q(\theta_k)}[\theta_k] = \{\text{wikipedia}\} = \sum_{k=1}^{K}\log B(a_k,b_k) - (a_k-1)\psi(a_k) - (b_k-1)\psi(b_k)$$
$$+ (a_k+b_k-2)\psi(a_k+b_k)$$

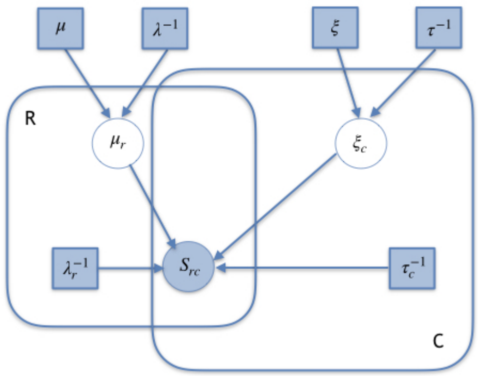


Figure 2: The graphical model (solution).

3.4 Cartesian Matrix Model (from assignment 1B, 2017)

The Cartesian Matrix Model (CMM) is defined as follows. There are R row distributions $\{N(\mu_r, \lambda_r^{-1}) : 1 \leq r \leq R\}$, each variance λ_r^{-1} is known and each μ_r has prior distribution $N(\mu, \lambda^{-1})$. There are also C column distributions $\{N(\xi_r, \tau_c^{-1}) : 1 \leq c \leq C\}$, each variance τ_c^{-1} is known and each ξ_c has prior distribution $N(\xi, \tau^{-1})$. All hyper-parameters are known. A matrix S is generated by, for each row $1 \leq r \leq R$ and each column $1 \leq c \leq C$, setting $S_{rc} = X_r + Y_c$ where X_r is sampled from $N(\mu_r, \lambda_r^{-1})$ and Y_c from $N(\xi_r, \tau_c^{-1})$. Use Variational Inference in order to obtained a variational distribution

$$q(\mu_1, \dots, \mu_R, \xi_1, \dots, \xi_C) = \prod_r q(\mu_r) \prod_c q(\xi_c)$$

that approximates $p(\mu_1, \dots, \mu_R, \xi_1, \dots, \xi_C|S)$. Tip: what distribution do you get from the sum of two Gaussian random variables? What is the relation between the means?

> **Question 15:** *Present the algorithm written down in a formal manner (using both text and mathematical notation, but not pseudo code).*

Figure 1: From Assignment 2, 2017

3.4.1 Solution

Solution:
Since each cell of the matrix denoted as S_{rc} is a sum of two Gaussians then as we know we can get the Gaussian result where the mean is the sum of the mean of X and the mean of

Y. The variance is also summed. So by having μ_r and ξ_c, the distribution of each cell can be expressed as $\mathcal{N}(\mu_r + \xi_c, \lambda_r^{-1} + \tau_c^{-1})$. In Fig. 2, the graphical model is illustrated. Below are the calculations of $q_l(\mu_l)$ and $q_t(\xi_t)$ according to variational approximation (l and t represent any relevant index for calculating all of the possible $q_l(\mu_l)$ and $q_t(\xi_t)$). The known parameters are represented with $\Theta = \{\lambda_{1:R}^{-1}, \lambda^{-1}, \mu, \tau_{1:C}^{-1}, \tau^{-1}, \xi\}$.

$$\log q(\mu_l) = E_{\substack{\xi_{1:C} \\ \mu_{1:R\neg l}}}[\log p(S, \mu_{1:R}, \xi_{1:C}|\Theta)] + const$$
$$\log q(\xi_t) = E_{\substack{\mu_{1:R} \\ \xi_{1:C\neg t}}}[\log p(S, \mu_{1:R}, \xi_{1:C}|\Theta)] + const$$

To calculate the above we first calculate log $q_l(\mu_l)$ which needs in return to first calculate $p(S, \mu_{1:R}, \xi_{1:C})$ (Note that we choose indexes l and later t to denote a specific index so that it is not mixed with the indexes in the sum and product terms):

$$\begin{aligned} p(S, \mu_{1:R}, \xi_{1:C}|\Theta) &= p(S|\mu_{1:R}, \xi_{1:C}, \Theta)p(\mu_{1:R}, \xi_{1:C}|\Theta) \\ &= \prod_{c=1}^{C}\prod_{r=1}^{R} p(S_{rc}|\mu_r + \xi_c, \lambda_r^{-1} + \tau_c^{-1}) \prod_{r=1}^{R} p(\mu_r|\mu, \lambda^{-1}) \prod_{c=1}^{C} p(\xi_c|\xi, \tau^{-1}) \end{aligned}$$

Notice that each μ_r and each ξ_c are independent in the above expression, resulting in factors. Now we calculate log $q_l(\mu_l)$ by taking the expectation over the log of the expression above:

$$\log q(\mu_l) = -\frac{1}{2}E_{\neg\mu_l}\left[\sum_{c=1}^{C}\sum_{r=1}^{R}\frac{(S_{rc} - (\mu_r + \xi_c))^2}{\lambda_r^{-1} + \tau_c^{-1}} + \frac{1}{\lambda^{-1}}\sum_{r=1}^{R}(\mu_r - \mu)^2 + \frac{1}{\tau^{-1}}\sum_{c=1}^{C}(\xi_c - \xi)^2\right] + const$$

All the components of r in $\sum_{r=1}^{R}$ are independent on μ_l except $r = l$; adding those independent terms in the above into constant results in:

$$-\frac{1}{2}E_{\xi_{1:C}}\left[\sum_{c=1}^{C}\frac{S_{lc}^2 - 2S_{lc}(\mu_l + \xi_c) + (\mu_l^2 + \xi_c^2 + 2\mu_l\xi_c)}{\lambda_l^{-1} + \tau_c^{-1}} + \lambda(\mu_l^2 + \mu^2 - 2\mu\mu_l) + \frac{1}{\tau^{-1}}\sum_{c=1}^{C}(\xi_c - \xi)^2\right]$$
$$+ const$$

The μ_l-independent terms in the above expression can be pushed into constant resulting in (the expectation goes through the terms as below):

$$\begin{aligned} \log q(\mu_l) &= -\frac{1}{2}E_{\xi_{1:C}}\left[\sum_{c=1}^{C}\frac{-2S_{lc}(\mu_l) + (\mu_l^2 + 2\mu_l\xi_c)}{\lambda_l^{-1} + \tau_c^{-1}} + \lambda(\mu_l^2 - 2\mu\mu_l)\right] + const \\ &= -\frac{1}{2}E_{\xi_{1:C}}\left[\left(\lambda + \sum_{c=1}^{C}\frac{1}{\lambda_l^{-1} + \tau_c^{-1}}\right)\mu_l^2 - 2\mu_l\left(\mu\lambda + \sum_{c=1}^{C}\frac{S_{lc} - \xi_c}{\lambda_l^{-1} + \tau_c^{-1}}\right)\right] + const \\ &= -\frac{1}{2}\left[\mu_l^2\left(\lambda + \sum_{c=1}^{C}\frac{1}{\lambda_l^{-1} + \tau_c^{-1}}\right) - 2\mu_l\left(\mu\lambda + \sum_{c=1}^{C}E_{\xi_c}\left[\frac{S_{lc} - \xi_c}{\lambda_l^{-1} + \tau_c^{-1}}\right]\right)\right] + const \\ &= -\frac{1}{2}\left[\mu_l^2\left(\lambda + \sum_{c=1}^{C}\frac{1}{\lambda_l^{-1} + \tau_c^{-1}}\right) - 2\mu_l\left(\mu\lambda + \sum_{c=1}^{C}\frac{S_{lc} - E_{\xi_c}[\xi_c]}{\lambda_l^{-1} + \tau_c^{-1}}\right)\right] + const \end{aligned}$$

By completing the square, the above expression becomes in a form of Gaussian with mean m_l and precision p_l, where

$$p_l = \lambda + \sum_{c=1}^{C} \frac{1}{\lambda_l^{-1} + \tau_c^{-1}}$$

$$m_l = \frac{\mu\lambda + \sum_{c=1}^{C} \frac{S_{lc} - E_{\xi_c}[\xi_c]}{\lambda_l^{-1} + \tau_c^{-1}}}{p_l}$$

$$q(\mu_l) = \mathcal{N}(\mu_l | m_l, p_l)$$

Similarly we can calculate $\log q_t(\xi_t)$ and so it results in a Gaussian with mean m_t and precision p_t, where

$$p_t = \tau + \sum_{r=1}^{R} \frac{1}{\tau_t^{-1} + \lambda_r^{-1}}$$

$$m_t = \frac{\xi\tau + \sum_{r=1}^{R} \frac{S_{rt} - E_{\mu_r}[\mu_r]}{\tau_t^{-1} + \lambda_r^{-1}}}{p_t}$$

$$q(\xi_t) = \mathcal{N}(\xi_t | m_t, p_t)$$

Since we have approximated $q(\mu_l)$, we have R number of estimations each having similar result as in m_l and p_l. Having $q(\xi_t)$ form, we have C number of estimations each having similar result as in m_t and p_t.
The only remaining parts to compute are $E_{\xi_c}[\xi_c]$ and $E_{\mu_r}[\mu_r]$. Since ξ_c and μ_r are Gaussian random variables, we can substitute the expression of the means

$$m_l = \frac{\mu\lambda + \sum_{c=1}^{C} \frac{S_{lc} - m_c}{\lambda_l^{-1} + \tau_c^{-1}}}{p_l}$$

$$m_t = \frac{\xi\tau + \sum_{r=1}^{R} \frac{S_{rt} - m_r}{\tau_t^{-1} + \lambda_r^{-1}}}{p_t}$$

We found the distributions of $q(\mu_l)$ and $q(\xi_t)$, therefore we can calculate $\prod_{r=1}^{R} q(\mu_r) \prod_{c=1}^{C} q(\xi_c)$.

3.5 Troll factories (from assignment, 1B 2022)

On a social media platform, K troll factories have posted N comments on a live news report from an ongoing war. A security agency wants to extract information on the troll factories, but due to integrity protection policies, the platform can only provide metadata of the posts, such as comment length X_n of each post as well as response time T_n. Together with the security agency's disinformation team, the newly employed ML expert develops a model which infers comment to factory assignment, Z_n, factory post volume fraction, π, troll factory specific response rate, λ_k, and average comment length, μ_k, and precision τ_k with the following distributions:

- $X_n|\mu_k, \tau_k, Z_n = k \sim Lognormal(\mu_k, \tau_k^{-1})$ - based on the assumptions that each troll factory has it's own strategy for comment length and variation in length and that comments are always of positive length.
- $\mu_k, \tau_k|\nu, \kappa, \alpha, \beta \sim NormalGamma(\nu, \kappa, \alpha, \beta)$
- $T_n|\lambda_k, Z_n = k \sim Exp(\lambda_k)$ - Comments are written as reactions to events with a factory specific response rate.
- $\lambda_k|a, b \sim Gamma(a, b)$ - The factory specific response rate is unkown, but the domain experts provide reasonable values for a and b.
- $Z_n|\pi \sim Categorical(\pi)$ - Each post is associated with K different factories.
- $\pi|\delta \sim Dirichlet(\delta)$

Note that comment length is a discrete entity, but we approximate the likelihood of observations with a continuous distribution in the model.
a) Provide a graphical model for the model described above.
b) Derive the CAVI update equations of each variational distribution.

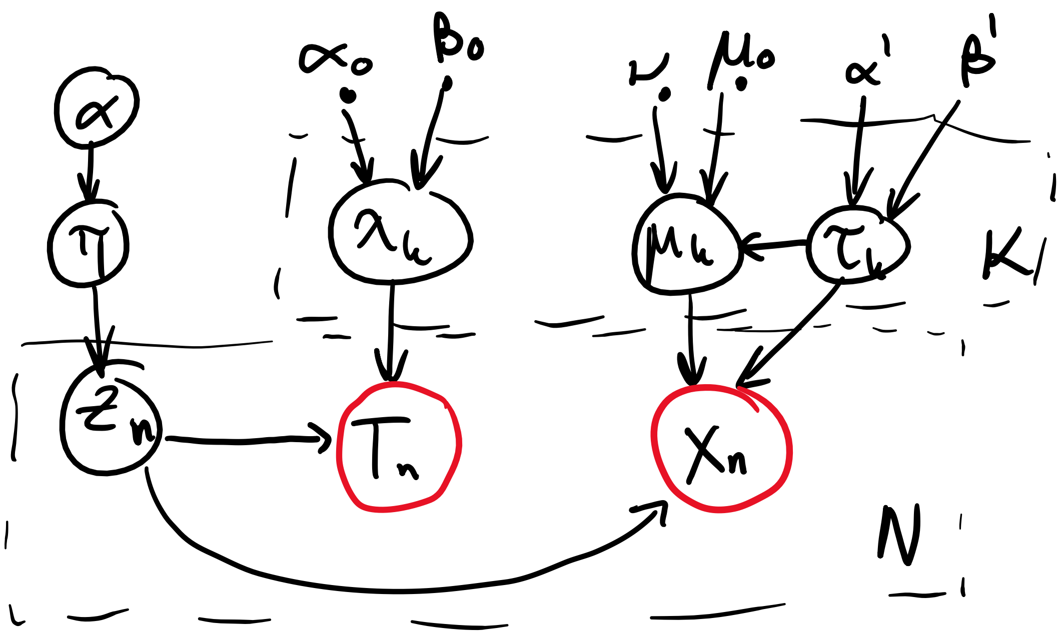
3.5.1 Solution

Troll factories

Friday, 10 November 2023 12:09

This is a problem from last years 1B.
All updates are not shown as they give too much information on this assignment.

A) Draw the DGM



B) Derive the CAVI updates

$$
\begin{aligned}
\ln \mathrm{LogNormal}(\mathbf{x}_n|\mu_{\mathbf{k}}) &= -\ln x_n + \frac{\ln \tau_k}{2} - \frac{\tau_k(\ln x_n - \mu_k)^2}{2} - \frac{\ln 2\pi}{2} \\
\ln Exp(\mathbf{t}_n|\lambda_k) &= \ln \lambda_k - \lambda_k t_n \\
\ln P(\pi|\alpha) &= \sum_{k=1}^{K} (\alpha - 1)\ln(\pi_k) + \text{Const.} \\
\ln P(\lambda_k|\alpha_0, \beta_0) &= (\alpha_0 - 1)\ln \lambda_k - \beta_0 \lambda_k + \text{Const.} \\
\ln P(\mu_k, \tau_k|\nu, \mu', \alpha', \beta') &= (\alpha' - \frac{1}{2})\ln \tau_k - \beta' \tau_k - \frac{\nu \tau_k(\mu_k - \mu')^2}{2} + \text{Const.}
\end{aligned}
$$

Denoting joint probability $\phi = P(\mathbf{X}, \mathbf{S}, \mathbf{Z}, \pi, \mathbf{\Lambda}, \mathbf{M}, \mathbf{T})$,

$$\begin{aligned}
\phi &= P(\mathbf{X}|\mathbf{Z}, \mathbf{M}, \mathbf{T})P(\mathbb{T}|\mathbf{Z}, \mathbf{\Lambda})P(\mathbf{Z}|\pi)P(\pi, \mathbf{\Lambda}, \mathbf{M}, \mathbf{T}) \\
\ln \phi &= \ln P(\mathbf{X}|\mathbf{Z}, \mathbf{M}, \mathbf{T}) + \ln P(\mathbb{T}|\mathbf{Z}, \mathbf{\Lambda}) + \ln P(\mathbf{Z}|\pi) + \ln P(\pi) + \ln P(\mathbf{\Lambda}) + \ln P(\mathbf{M}, \mathbf{T}) \\
&= (\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}[\ln \mathrm{LogNormal}(\mathbf{x}_n|\mu_\mathbf{k}, \tau_\mathbf{k}{}^{-1}) + \ln \mathit{Exp}(\mathbf{t}_n|\lambda_k) + \ln \pi_k]) + \\
&\quad \ln P(\pi|\alpha) + \sum_{k=1}^{K}(\ln P(\lambda_k|\alpha_0, \beta_0) + \ln P(\mu_k, \tau_k|\nu, \mu', \alpha', \beta'))
\end{aligned}$$

$$\begin{aligned}
\ln q_\mathbf{Z}(\mathbf{Z}) &= \mathbb{E}_{\pi,\mathbf{\Lambda},\mathbf{M},\mathbf{T}}[\ln \phi] \\
&= \mathbb{E}_{\pi,\mathbf{\Lambda},\mathbf{M},\mathbf{T}}[(\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}[\ln \mathrm{LogNormal}(\mathbf{x}_n|\mu_\mathbf{k}, \tau_\mathbf{k}{}^{-1}) + \ln \mathit{Exp}(\mathbf{t}_n|\lambda_k) + \ln \pi_k]) \\
&\qquad + \ln P(\pi, \mathbf{\Lambda}, \mathbf{M}, \mathbf{T})] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}[\mathbb{E}_{\mu_k,\tau_k}(\ln \mathrm{LogNormal}(\mathbf{x}_n|\mu_\mathbf{k}, \tau_\mathbf{k}{}^{-1})) + \mathbb{E}_{\lambda_k}(\ln \mathit{Exp}(\mathbf{t}_n|\lambda_k)) + \mathbb{E}_{\pi_k}(\ln \pi_k)] + \mathrm{Const}
\end{aligned}$$

Since the term $\ln P(\pi, \mathbf{\Lambda}, \mathbf{M}, \mathbf{T})$ is not a function of $\mathbf{Z}$, it is a constant. Moreover $q_\mathbf{Z}(\mathbf{Z})$ can be further broken down to $\prod_{n=1}^{N} q_{\mathbf{z_n}}(\mathbf{z_n})$. Thus,

$$q_{\mathbf{z_n}}(\mathbf{z_n}) = \sum_{k=1}^{K} z_{nk}[\mathbb{E}_{\mu_k,\tau_k}(\ln \mathrm{LogNormal}(\mathbf{x}_n|\mu_\mathbf{k}, \tau_\mathbf{k}{}^{-1})) + \mathbb{E}_{\lambda_k}(\ln \mathit{Exp}(\mathbf{t}_n|\lambda_k)) + \mathbb{E}_{\pi_k}(\ln \pi_k)] + \mathrm{Const}$$

Now evaluating each of the expectation terms using the above equations,

$$\begin{aligned}
\mathbb{E}_{\mu_k,\tau_k}(\ln \mathrm{LogNormal}(\mathbf{x}_n|\mu_\mathbf{k}, \tau_\mathbf{k}{}^{-1})) &= \frac{1}{2}\mathbb{E}_{\tau_k}[\ln \tau_\mathbf{k}] - \frac{1}{2}\mathbb{E}_{\mu_k,\tau_k}[\tau_k(\ln x_n - \mu_k)^2] + \mathrm{Const.} \\
\mathbb{E}_{\lambda_k}(\ln \mathit{Exp}(\mathbf{t}_n|\lambda_k)) &= \mathbb{E}_{\lambda_k}[\ln \lambda_k] - \mathbb{E}_{\lambda_k}[\lambda_k]t_n
\end{aligned}$$

Now for the term $q_{\pi,\mathbf{\Lambda},\mathbf{M},\mathbf{T}}(\pi, \mathbf{\Lambda}, \mathbf{M}, \mathbf{T})$

$$\begin{aligned}
\ln q_{\pi,\mathbf{\Lambda},\mathbf{M},\mathbf{T}}(\pi, \mathbf{\Lambda}, \mathbf{M}, \mathbf{T}) &= \mathbb{E}_\mathbf{Z}[\ln \phi] \\
&= \mathbb{E}_\mathbf{Z}[(\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}[\ln \mathrm{LogNormal}(\mathbf{x}_n|\mu_\mathbf{k}, \tau_\mathbf{k}{}^{-1}) + \ln \mathit{Exp}(\mathbf{t}_n|\lambda_k) + \ln \pi_k]) \\
&\qquad + \ln P(\pi, \mathbf{\Lambda}, \mathbf{M}, \mathbf{T})] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}_{z_{nk}}[z_{nk}][\ln \mathrm{LogNormal}(\mathbf{x}_n|\mu_\mathbf{k}, \tau_\mathbf{k}{}^{-1}) + \ln \mathit{Exp}(\mathbf{t}_n|\lambda_k) + \ln \pi_k] \\
&\qquad + \ln P(\pi) + \ln P(\mathbf{\Lambda}) + \ln P(\mathbf{M}, \mathbf{T}) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}[\ln \mathrm{LogNormal}(\mathbf{x}_n|\mu_\mathbf{k}, \tau_\mathbf{k}{}^{-1}) + \ln \mathit{Exp}(\mathbf{t}_n|\lambda_k) + \ln \pi_k] \\
&\qquad + \sum_{k=1}^{K}(\alpha - 1)\ln \pi_k + \sum_{k=1}^{K}[(\alpha_0 - 1)\ln \lambda_k - \beta_0\lambda_k] \\
&\qquad\quad + \sum_{k=1}^{K}[(\alpha' - \frac{1}{2})\ln \tau_k - \beta'\tau_k - \frac{\nu\tau_k(\mu_k - \mu')^2}{2}] + \mathrm{Const}
\end{aligned}$$

The above expression brings out the underlying independence, which is Π, Λ and $(\mathbf{M}, \mathbf{T})$, don't appear together or in pairs, and the fact that for $k \neq k\prime$, the terms corresponding to k and $k\prime$ don't appear together. Therefore it can be further factorised as $q_{\pi,\mathbf{\Lambda},\mathbf{M},\mathbf{T}}(\pi, \mathbf{\Lambda}, \mathbf{M}, \mathbf{T}) = q_\pi(\pi) q_{\mathbf{\Lambda}}(\mathbf{\Lambda}) q_{\mathbf{M},\mathbf{T}}(\mathbf{M}, \mathbf{T}) = q_\pi(\pi) \prod_{k=1}^{K} q_{\lambda_k}(\lambda_k) q_{\mu_\mathbf{k},\tau_\mathbf{k}}(\mu_\mathbf{k}, \tau_\mathbf{k})$. We can use the above equations to get each of the factorised units.

$$
\begin{aligned}
\ln q_\pi(\pi) &= \sum_{k=1}^{K} (\alpha - 1) \ln \pi_k + \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \ln \pi_k \\
&= \sum_{k=1}^{K} [(\alpha - 1) + N_k] \ln \pi_k \\
q_\pi(\pi) &= \mathrm{Dir}(\pi | \alpha + N_k)
\end{aligned}
$$

where $N_k = \sum_{n=1}^{N} r_{nk}$. For $q_{\lambda_k}(\lambda_k)$,

$$
\begin{aligned}
\ln q_{\lambda_k}(\lambda_k) &= (\alpha_0 - 1) \ln \lambda_k - \beta_0 \lambda_k + \sum_{n=1}^{N} r_{nk} [\ln \lambda_k - t_n \lambda_k] \\
&= \ln \lambda_k [\alpha_0 - 1 + N_k] - (\beta_0 + \sum_{n=1}^{N} r_{nk} t_n) \lambda_k + \mathrm{Const} \\
q_{\lambda_k}(\lambda_k) &= \mathrm{Gamma}(\alpha_0 + N_k, \beta_0 + \sum_{n=1}^{N} t_n r_{nk})
\end{aligned}
$$