Project: Supervised Mixed-membership Model on Leaf Classification Problem

Yang Kang, Jing Wu and Ji Xu
December 18, 2016

1 Introduction

Automating plant recognition is an interesting and important problem. It has many applications including species population tracking and preservation, plant-based medicinal research and crop, and food supply management. Leaf classification is a part of the automating plant recognition problem.

At first glance, the problem seems to be irrelevant to graphical model. However, there are some variables that effect the appearances of leaves and can help us to identify the species. For example, the geometric location, defence mechanism against their natural enemies like edged shape, colors that will attract certain insects that will help them reproduce. Unfortunately, these variables are not available in our data set which means they are latent variables over the features that are actually given. Therefore, our idea is to use Mixed-membership Model to solve this problem.

2 Mixed-membership Model for Gaussian Distribution

Our data set comes from one of the Kaggle competition projects. In the given data set, we need to classify 99 different species among $N = 990$ leaves samples with 10 samples for each species. We take out 99 samples with 1 samples for each species as test set and the remaining data is the training set. Each sample has $M = 192$ features to help us to classify their categories. In summary, the data has a big number of clusters, 99, and high dimension of variables, 192, with only small number of samples, 10 for each cluster. Therefore, we want to use graphical model methods to make use of all potential properties among clusters. The idea is the following:

2.1 Intuition:

We want to borrow the idea from Mixed-membership Model, specifically, the LDA model. Since we have discussed a lot about the LDA model in class, we

only list out the generative process to clarify the notations for the later use.

1. Draw components $\beta_k \sim \text{Dir}(\eta)$.
2. For each group i:
   (a) Draw proportions $\theta_i \sim \text{Dir}(\alpha)$
   (b) For each data point j within the group:
      i. Draw a mixture assignment $z_{ij}|\theta_i \sim \text{Cat}(\theta_i)$.
      ii. Draw the data point $x_{ij}|\beta_k, z_{ij} \sim g(\cdot|\beta_{z_{ij}}) = \text{Mult}(\beta_{z_{ij}})$.

Now, to mimic the LDA model, we first assume that each plant consists of several hidden labels or patterns, referred as "HL". For example, a pine tree is classified as "Pinaceae" in "class", which usually have needle-shaped leaves. Also plants will share some common properties within the same geographic location even they are in different classes. In other words, these "HL" are not exclusive to each other in general and therefore can be considered as topics in the LDA model. Then we assume the distribution of the "HL"s/topics $\boldsymbol{\theta}_i$ are drawn from $\text{Dir}(\alpha)$ for some α. Next, we assume the "HL"s for each leaf is drawn from a categorical distribution $\text{Cat}(\boldsymbol{\theta}_i)$. Finally, we assume for each leaf, it will contain certain type of features which is drawn from $g(\cdot|\beta_k)$ and these features correspond to the words in the LDA model.

Yet, there are three differences between our model and the LDA model. The first difference is that we don't know what are the topics exactly, so we have to assume the total number of topics as K and make it as an input.

The second difference is that $g(\cdot|\beta_k)$ is a categorical distribution in the LDA model because words are discrete. But in our case, since each feature is a continuous variable, we assume $g(\cdot|\beta_k)$ has a continuous pdf, or more precisely, it is a Gaussian distribution. In fact, we need to model one more layer, i.e, we assume $x_{ij} \sim N(\mu_{z_{ij},j}, \sigma^2)$, where $\mu_{k,j}$ is the jth component of $\boldsymbol{\mu}_k$ for "HL"/topic k. Further, we assume $\boldsymbol{\mu}_k \sim N(\boldsymbol{\beta}, \sigma_0^2 I)$.

The third difference is that our data is supervised instead of unsupervised. Hence, we need a classification label for the prediction not any parameter estimates. To solve this issue, we need a function h which maps our parameter estimates to the name of the species. To simplify the problem, it is natural to assume that each species can be identified by the distribution over the "HL"s/topics. Hence, we only need to find a function h on the domains of $\boldsymbol{\theta}_i$. We will discuss the choice of h at the end of this section and in Section 4.

To adjust all these differences, we have the following generative process and Figure 1 shows the graph representation of the model:

1. For each leaf sample $i \in [N]$:
    - Draw "HL"/topic proportions $\boldsymbol{\theta}_i \sim Dir(\alpha)$.
2. For each "HL"/topic $k \in [K]$:
    - Draw Gaussian mean $\boldsymbol{\mu}_k \sim N(\boldsymbol{\beta}, \sigma_0^2 I)$.
3. For each jth feature/word for ith sample:
    - Draw assignment $z_{ij}|\boldsymbol{\theta}_i \sim \text{Cat}(\boldsymbol{\theta}_i)$.
    - Draw $x_{ij}|\boldsymbol{\mu}, z_{ij} \sim N(\mu_{z_{ij},j}, \sigma^2)$.
4. Determine class label $y_i = h(\text{Cat}(\boldsymbol{\theta}_i))$.

Constant $\boldsymbol{\beta}$ is in $\mathbb{R}^{\tilde{M}}$ where $\tilde{M}$ is the dimension of the entire feature set after an independent dimension reduction procedure.

Once we obtain $\boldsymbol{\mu}_k$ for each "HL"/topic, then use the following procedure, we can have an estimator of $\boldsymbol{\theta}_{new}$, the parameter for the distribution over the topics for each data point $\boldsymbol{x}_{new}$ in the test set.

1. For $j \in [\tilde{M}]$, calculate

$$p(z_{new,j}|x_{new,j}) \quad \propto \quad exp(-\frac{(x_{new,j} - \mu_{z_{new,j},j})^2}{2\sigma^2}) \tag{1}$$

2. Our estimate for $\boldsymbol{\theta}_{new}$ is the following:

$$\hat{\boldsymbol{\theta}}_{new} = \frac{1}{\tilde{M}} \sum_j p(z_{new,j}|x_{new,j}).$$

Now, we just need to find a suitable h to get the label $y_{new} = h(\text{Cat}(\hat{\boldsymbol{\theta}}_{new}))$ for each data point $\boldsymbol{x}_{new}$ in the test set. One naive way to construct h as the following:

$$\begin{aligned} i^* &= \arg\min_i \ d(\boldsymbol{\theta}_i, \hat{\boldsymbol{\theta}}_{new}), \\ y_{new} &= y_{i^*}, \end{aligned}$$

i.e, we sign the new point with the same cluster label of the leaf $i*$ which has the smallest distance $d(\cdot, \cdot)$ in the training set. $d(\cdot, \cdot)$ can be set to KL-divergence.

3 Variational Inference:

Besides MCMC, mean-field variational inference method is a powerful tool solving the problem of computing the posterior. We summarized main steps of our implementation of variational inference in the following:

First, we compute the joint log-likelihood of our model:

$$\log p(\boldsymbol{X}, z, \boldsymbol{\theta}, \boldsymbol{\mu}) = \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{K} \mathbb{1}(z_{ij}=k)\log\theta_{ik} + \mathbb{1}(z_{ij}=k)[-\frac{1}{2\sigma^2}(x_{ij}-\mu_{k,j})^2] + \sum_{k=1}^{K}[-\frac{1}{2\sigma_0^2}\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k] + \sum_{i=1}^{N}\sum_{k=1}^{K}(\alpha-1)\log\theta_{ik}$$

where $\boldsymbol{\theta}_{1:N}, z_{1:N,1:M}, \boldsymbol{\mu}_{1:K}$ are the parameters and $\boldsymbol{\beta}=\mathbf{0}, \sigma_0=0.5, \sigma=0.5, \alpha=0.2$ are the value we choose for the prior parameters. Then we choose

$$q(z_{ij}) \sim \text{Dir}(\phi_{ij}), \quad q(\boldsymbol{\theta}_i) \sim \text{Dir}(\alpha_{i1}, \cdots, \alpha_{ik}) \quad \text{and} \quad q(\mu_{kj}) \sim N(\tilde{\mu}_{kj}, \tilde{\sigma}_{kj}^2)$$

Therefore, we have

- $q(z_{ij}) \propto \prod_{k=1}^{K}\Big(\exp\{\mathbb{E}_q[\log\theta_{ik}] + \mathbb{E}_q[-\frac{1}{2\sigma^2}(x_{ij}-\mu_{k,j})^2]\}\Big)^{\mathbb{1}(z_{ij}=k)}$.
  Hence,
  $$\phi_{ij}(k) = \frac{\exp\{\mathbb{E}_q[\log\theta_{ik}] + \mathbb{E}_q[-\frac{1}{2\sigma^2}(x_{ij}-\mu_{k,j})^2]\}}{\sum_{l=1}^{K}\exp\{\mathbb{E}_q[\log\theta_{il}] + \mathbb{E}_q[-\frac{1}{2\sigma^2}(x_{ij}-\mu_{l,j})^2]\}}$$
- $q(\boldsymbol{\theta}_i) \propto \prod_{k=1}^{K}\theta_{ik}^{\alpha+\sum_{j=1}^{M}\phi_{ij}(k)-1}$.
  Hence,
  $$\alpha_{ik} = \alpha + \sum_{j=1}^{M}\phi_{ij}(k).$$
  By property of exponential family, we have
  $$\mathbb{E}\log\theta_{ik} = \psi(\alpha_{ik}) - \psi(\sum_{l=1}^{K}\alpha_{il}),$$
  where $\psi(\cdot)$ is the derivative of $\log\Gamma(\cdot)$ and $\Gamma(\cdot)$ is the Gamma function.
- $q(\boldsymbol{\mu}_k) \propto \prod_{j=1}^{M}\exp\{-\frac{1}{2\sigma_0^2}\mu_{k,j}^2 - \sum_{i=1}^{N}\big(-\frac{1}{2\sigma^2}(x_{ij}-\mu_{k,j})^2\phi_{ij}(k)\big)\}$.
  Hence,
  $$\tilde{\mu}_{kj} = \frac{\tilde{\sigma}_{kj}^2}{\sigma^2}\sum_{i=1}^{N}x_{ij}\phi_{ij}(k) \quad \text{and} \quad \frac{1}{\tilde{\sigma}_{kj}^2} = \frac{1}{\sigma_0^2} + \frac{\sum_{i=1}^{N}\phi_{ij}(k)}{\sigma^2},$$
  and
  $$\mathbb{E}[(x_{ij}-\mu_{k,j})^2] = x_{ij}^2 - 2x_{ij}\tilde{\mu}_{kj} + \tilde{\mu}_{kj}^2 + \tilde{\sigma}_{kj}^2.$$

Finally, the ELBO can be calculated as the following:

$$\text{ELBO} = \mathbb{E}_q[\log P(\boldsymbol{X}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\mu})] - \mathbb{E}_q[\log q(\boldsymbol{\mu})] - \mathbb{E}_q[\log q(\boldsymbol{\theta})] - \mathbb{E}_q[\log q(\boldsymbol{z})],$$

where

$$\mathbb{E}_q[\log P(\boldsymbol{X}, z, \boldsymbol{\theta}, \boldsymbol{\mu})] = \sum_{i,j,k} \Big(\phi_{ij}(k) \big(\mathbb{E}[\log \theta_{ik}] - \frac{1}{2\sigma^2} \mathbb{E}[(x_{ij} - \mu_{k,j})^2] \big) \Big)$$

$$- \sum_{k,j} \big(\frac{\tilde{\mu}_{kj}^2}{2\sigma_0^2} + \frac{\tilde{\sigma}_{kj}^2}{2\sigma_0^2} + \frac{1}{2} \log 2\pi\sigma_0^2 \big) + (\alpha - 1) \sum_{i,k} \mathbb{E}[\log \theta_{ik}]$$

$$\mathbb{E}_q[\log q(\boldsymbol{z})] = \sum_{i,j,k} \phi_{ij}(k) \log \phi_{ij}(k), \qquad \mathbb{E}_q[\log q(\boldsymbol{\theta})] = \sum_{i,k} (\alpha_{ik} - 1) \mathbb{E}_q[\log \theta_{ik}]$$

$$\mathbb{E}_q[\log q(\boldsymbol{\mu})] = - \sum_{k,j} (\log \tilde{\sigma}_{kj} + \frac{1}{2} \log 2\pi + \frac{1}{2})$$

For prediction, we just need to change (1) to:

$$p(z_{new,j} | x_{new,j}) \quad \propto \quad \mathbb{E}_q exp(-\frac{(x_{new,j} - \mu_{z_{new,j},j})^2}{2\sigma^2}) \propto \frac{exp(-\frac{(x_{new,j} - \tilde{\mu}_{z_{new,j},j})^2}{2(\sigma^2 + \tilde{\sigma}^2_{z_{new,j},j})})}{\sqrt{\sigma^2 + \tilde{\sigma}^2_{z_{new,j},j}}}$$

4 Results:

We have tried $K = 2, 4, 6, 8$ "HL"s/topics. The ELBOs are increasing and presented in Figure 2. The final correct rate achieves over 32 percent which is much better than the pure guess $1/99$. Yet it is still not good enough comparing to other supervised methods such as random forest. One possible reason is that our choice of mapping function h is not good enough. In fact, possibly due to small sample size, $\boldsymbol{\theta}_i$s coming from the same species are not that closer to each other than other $\boldsymbol{\theta}_i$s in KL-divergence. One idea to adjust this issue is to train a neural net that can mapping $\boldsymbol{\theta}_i$ close to others that come from the same species. Besides this, another possible reason for the low accuracy is that the features are not drawn from spherical Gaussian with variance σ^2 but more general ones. Yet, we may require more data for more general case.

References

[1] Mcauliffe, Jon D., and Blei, D. (2008). Supervised topic models. *Advances in neural information processing systems*.

[2] Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3: 993-1022.

[3] Airoldi, E., Blei, D., Fienberg, S. E., & Xing, E. (2008). Mixed membership stochastic blockmodels.*Journal of Machine Learning Research*, 9: 1981-2014.

[4] Mixed-membership Models
`http://www.cs.columbia.edu/~blei/fogm/2016F/doc/mixed_membership.pdf`

[5] Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999). Introduction to variational methods for graphical models. *Machine Learning*, 37:183âĂŞ233.

[6] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*.

5 Appendix

5.1 Figures

Figure 1 shows the graphical representation. Figure 2 shows the ELBO for different choices of K.

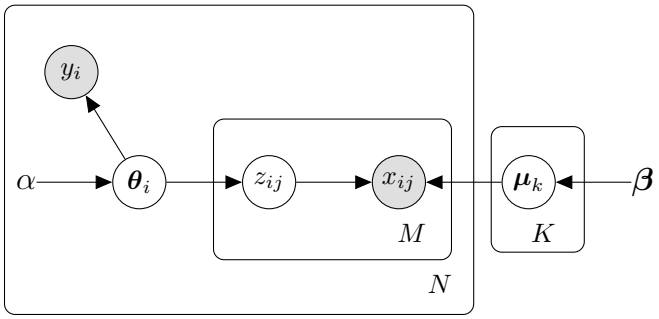

Figure 1: The graph representation for our model

5.2 Detailed derivation for Section 3:

To implement variational inference on our proposed model, the road map is as following:

1. Derive the joint likelihood,
2. Using the mean-field assumption, pick the variation family, and update each variational distribution,
3. Compute ELBO to access convergence, where ELBO is the evidence of lower bound and is a monotonically increasing function of time of iteration.

Follow the road map, we have

1. Joint likelihood of our model As shown in the Section 2, the parameters are $\boldsymbol{\theta}_{1:N}, z_{1:N,1:M}, \boldsymbol{\mu}_{1:K}$. We will assume the prior parameter $\boldsymbol{\beta} = \mathbf{0}, \sigma_0 = 0.5, \sigma = 0.1, \alpha = 0.1$. The reason that we assign a relative small value $(<< 1)$ for the pior parameter on Dirichlet distribution is to force the allocation of latent class to be sparse, therefore to avoid identifiability issue.

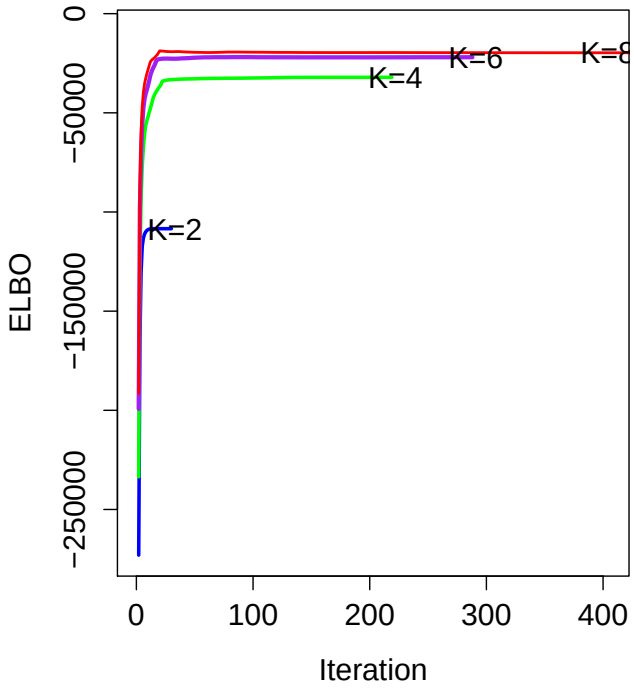

Figure 2: ELBOs for different values of K. As we can see, ELBO will be generally higher if we choose larger K.

Then, the joint likelihood will be the following:

$$p(\boldsymbol{X}, z, \boldsymbol{\theta}, \boldsymbol{\mu}) = p(\boldsymbol{\mu}) \prod_{i=1}^{N} p(x_i, z_i, \boldsymbol{\theta}_i | \boldsymbol{\mu})$$
$$= [\prod_{k=1}^{K} p(\boldsymbol{\mu}_k)] [\prod_{i=1}^{N} p(\boldsymbol{\theta}_i)] [\prod_{i=1}^{N} \prod_{j=1}^{M} p(x_{ij} | \boldsymbol{\mu}, z_{ij}) p(z_{ij} | \boldsymbol{\theta}_i)]$$

where each term follows:

$$p(\boldsymbol{\mu}_k) \propto \exp\{-\frac{1}{2\sigma_0^2} \boldsymbol{\mu}_k^T \boldsymbol{\mu}_k\}$$
$$p(\boldsymbol{\theta}_i) \propto \prod_{k=1}^{K} \theta_{ik}^{\alpha - 1}$$
$$p(z_{ij} | \boldsymbol{\theta}_i) \propto \prod_{k=1}^{K} \theta_{ik}^{\mathbb{1}(z_{ij}=k)}$$
$$p(x_{ij} | \boldsymbol{\mu}, z_{ij}) \propto \exp\{-\frac{1}{2\sigma^2}(x_{ij} - \mu_{z_{ij},j})^2\}$$
$$= \prod_{k=1}^{K} \Big[\exp\{-\frac{1}{2\sigma^2}(x_{ij} - \mu_{k,j})^2\} \Big]^{\mathbb{1}(z_{ij}=k)}$$

In conclusion, the log of joint likelihood can be written as:

$$\log p(x, z, \boldsymbol{\theta}, \boldsymbol{\mu}) = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{K} \mathbb{1}(z_{ij} = k) \log \theta_{ik} + \mathbb{1}(z_{ij} = k)[-\frac{1}{2\sigma^2}(x_{ij} - \mu_{k,j})^2]$$
$$+ \sum_{k=1}^{K} [-\frac{1}{2\sigma_0^2} \boldsymbol{\mu}_k^T \boldsymbol{\mu}_k] + \sum_{i=1}^{N} \sum_{k=1}^{K} (\alpha - 1) \log \theta_{ik}$$

2. Variational distribution

   According to mean-field assumption, we need to pick a factorization of $q(z, \boldsymbol{\theta}, \boldsymbol{\mu}) \approx p(z, \boldsymbol{\theta}, \boldsymbol{\mu} | x)$. We split these variables according to how they are generated in the prior. This makes learning the variational distribution q much easier.

$$q(\boldsymbol{\mu}, \boldsymbol{\theta}, z) = [\prod_{k=1}^{K} q(\boldsymbol{\mu}_k)] [\prod_{i=1}^{N} q(\boldsymbol{\theta}_i)] [\prod_{i=1}^{N} \prod_{j=1}^{M} q(z_{ij})]$$

   . To find the q distribution, we can focus only on terms in the log joint likelihood involving the term we are looking at. We will show the derivation of each q distribution as following.

- $q(z_{ij})$: indicator of which hidden label feature x_{ij} came from.

  $$\begin{aligned} q(z_{ij}) &\propto \exp\{\mathbb{E}_q[\log P(x, z, \boldsymbol{\theta}, \boldsymbol{\mu})]\} \\ &\propto \exp\{\mathbb{E}_q[\sum_{k=1}^{K} \mathbb{1}(z_{ij}=k)\log\theta_{ik} + \mathbb{1}(z_{ij}=k)[-\frac{1}{2\sigma^2}(x_{ij}-\mu_{k,j})^2]]\} \\ &= \prod_{k=1}^{K}\Big(\exp\{\mathbb{E}_q[\log\theta_{ik}] + \mathbb{E}_q[-\frac{1}{2\sigma^2}(x_{ij}-\mu_{k,j})^2]\}\Big)^{\mathbb{1}(z_{ij}=k)} \end{aligned}$$

  Notice that this is simply a categorical distribution.

  $$q(z_{ij}) = \text{Cat}(\phi_{ij}), \text{ where } \phi_{ij}(k) = \frac{\exp\{\mathbb{E}_q[\log\theta_{ik}] + \mathbb{E}_q[-\frac{1}{2\sigma^2}(x_{ij}-\mu_{k,j})^2]\}}{\sum_{l=1}^{K}\exp\{\mathbb{E}_q[\log\theta_{il}] + \mathbb{E}_q[-\frac{1}{2\sigma^2}(x_{ij}-\mu_{l,j})^2]\}}$$

- $q(\boldsymbol{\theta}_i)$: the distribution on hidden label for sample i.

  $$\begin{aligned} q(\boldsymbol{\theta}_i) &\propto \exp\{\mathbb{E}_q[\sum_{j=1}^{M}\sum_{k=1}^{K}\mathbb{1}(z_{ij}=k)\log\theta_{ik} + (\alpha-1)\sum_{k=1}^{K}\log\theta_{ik}]\} \\ &= \prod_{k=1}^{K}\exp\{(\alpha-1)\log\theta_{ik} + \sum_{j=1}^{M}\mathbb{E}_q[\mathbb{1}(z_{ij}=k)]\log\theta_{ik}\} \\ &= \prod_{k=1}^{K}\exp\{(\alpha + \sum_{j=1}^{M}\phi_{ij}(k) - 1)\log\theta_{ik}\} \\ &= \prod_{k=1}^{K}\theta_{ik}^{\alpha+\sum_{j=1}^{M}\phi_{ij}(k)-1} \end{aligned}$$

  Then,

  $$q(\boldsymbol{\theta}_i) = \text{Dir}(\alpha_{i1}, ..., \alpha_{iK}) \text{ ,where } \alpha_{ik} = \alpha + \sum_{j=1}^{M}\phi_{ij}(k)$$

- $q(\boldsymbol{\mu}_k)$: the distribution on hidden labels.

  $$\begin{aligned} q(\boldsymbol{\mu}_k) &\propto \exp\{\mathbb{E}_q[\sum_{i=1}^{N}\sum_{j=1}^{M}\mathbb{1}(z_{ij}=k)[-\frac{1}{2\sigma^2}(x_{ij}-\mu_{k,j})^2] - \frac{1}{2\sigma_0^2}\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\} \\ &= \exp\{-\frac{1}{2\sigma_0^2}\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\}\prod_{i=1}^{N}\prod_{j=1}^{M}\exp\{-\frac{1}{2\sigma^2}(x_{ij}-\mu_{k,j})^2\phi_{ij}(k)\} \\ &= \prod_{j=1}^{M}\exp\{-\frac{1}{2\sigma_0^2}\mu_{k,j}^2 - \sum_{i=1}^{N}\big(\frac{1}{2\sigma^2}(x_{ij}-\mu_{k,j})^2\phi_{ij}(k)\big)\} \end{aligned}$$

  Thus,

  $$q(\mu_{kj}) = N(\tilde{\mu}_{kj}, \tilde{\sigma}_{kj}^2) \text{ ,where } \tilde{\mu}_{kj} = \frac{\tilde{\sigma}_{kj}^2}{\sigma^2}\sum_{i=1}^{N}x_{ij}\phi_{ij}(k),\ \frac{1}{\tilde{\sigma}_{kj}^2} = \frac{1}{\sigma_0^2} + \frac{\sum_{i=1}^{N}\phi_{ij}(k)}{\sigma^2}$$

- Side note for updating on $\phi(k)$:

  We still need to find the form of $\mathbb{E}[\log \theta_{ik}]$ and $\mathbb{E}[(x_{ij} - \mu_{k,j})^2]$ in order to know $q(z_{ij})$.

  To derive $\mathbb{E}[\log \theta_{ik}]$, because $q(\boldsymbol{\theta}_i)$ follows Dirichlet distribution, which is in exponential family. Also, $\log \theta_{ik}$ is the sufficient statistics after we writing the distribution in canonical form. From the result on exponential family, we know that $\mathbb{E}[\log \theta_{ik}] = \frac{\partial A}{\partial \alpha_{ik}}$, where $A = \sum_{k=1}^{K} \log \Gamma(\alpha_{ik}) - \log \Gamma(\sum_{l=1}^{K} \alpha_{il})$ is the log normalizer. Then, we have

  $$\mathbb{E}[\log \theta_{ik}] = \psi(\alpha_{ik}) - \psi(\sum_{l=1}^{K} \alpha_{il}),$$

  where $\psi(\cdot)$ is the derivative of $\log \Gamma(\cdot)$ and can be evaluated in programming language using a built-in function.

  For $\mathbb{E}[(x_{ij} - \mu_{k,j})^2]$, it is relative easy to compute, because $\mu_{k,j}$ follows normal distribution. Thus, we have

  $$\mathbb{E}[(x_{ij} - \mu_{k,j})^2] = x_{ij}^2 - 2x_{ij}\tilde{\mu}_{kj} + \tilde{\mu}_{kj}^2 + \tilde{\sigma}_{kj}^2$$

  .

3. ELBO

$$\text{ELBO} = \mathbb{E}_q[\log P(x, z, \boldsymbol{\theta}, \boldsymbol{\mu})] - \mathbb{E}_q[\log q(\boldsymbol{\mu})] - \mathbb{E}_q[\log q(\boldsymbol{\theta})] - \mathbb{E}_q[\log q(z)]$$

where each term is as following,

$$\begin{aligned}
\mathbb{E}_q[\log P(x, z, \boldsymbol{\theta}, \boldsymbol{\mu})] &= \sum_{i,j,k} \Big(\phi_{ij}(k) \big(\mathbb{E}[\log \theta_{ik}] - \frac{1}{2\sigma^2} \mathbb{E}[(x_{ij} - \mu_{k,j})^2] \big) \Big) \\
&\quad - \sum_{k,j} (\tilde{\mu}_{kj}^2 + \tilde{\sigma}_{kj}^2) + (\alpha - 1) \sum_{i,k} \mathbb{E}[\log \theta_{ik}] \\
\mathbb{E}_q[\log q(\boldsymbol{\mu})] &= -\sum_{k,j} (\log \tilde{\sigma}_{kj} + \frac{1}{2} \log 2\pi + \frac{1}{2}) \\
\mathbb{E}_q[\log q(\boldsymbol{\theta})] &= \sum_{i,k} (\alpha_{ik} - 1) \mathbb{E}_q[\log \theta_{ik}] \\
\mathbb{E}_q[\log q(z)] &= \sum_{i,j,k} \phi_{ij}(k) \log \phi_{ij}(k)
\end{aligned}$$