Supplementary Material: Multiple Clustering Views from Multiple Uncertain Experts

Yale Chang Junxiang Chen Michael H. Cho
Peter J. Castaldi Edwin K. Silverman Jennifer G. Dy

1 Parameter Settings

All approaches need to specify K, the number of clusters.

When applying our MCVC to the synthetic and benchmark datasets:

- If there is only one ground-truth expert view, we set K to be the true number of clusters in that view.
- If there are more than one ground-truth expert view, we set K to be the maximal value among the true number of clusters in all expert views.

For SemiCrowd, ITML, MPCKMeans, CSPA, since we only apply them to one expert view, we set K to be the true number of clusters in that view.

For COPD data, we set $K = 4$ for all approaches according to a recent study on COPD [3].

Besides the number of clusters, the parameters that are specific to each approach are set as follows.

1.1 Proposed Approach: MCVC

For variational inference in our approach, we use the following parameter settings

1. G, the number of components in truncated Dirichlet Process, is set to be $M/2$, where M is the total number of experts. In this way, we try to enforce the constraint that on average, there should be at least two experts in each view. In all experiments, the number of expert views recovered by our approach is smaller than $G = M/2$. Therefore, the value of G we use is large enough to discover the true number of expert views.
2. For the parameters of prior distributions:
   - $p(\alpha_m), p(\beta_m)$: set the parameters of prior Beta distributions to be $(10, 1)$ to incorporate the prior knowledge that each expert's accuracy parameters should be far away from 0.5 (random guess) and close to 1. The choice can be illustrated from Figure 1. Under this setting, there is very small probability that the accuracy parameters can be close to 0.5.

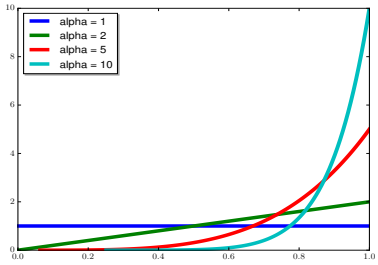

Figure 1: Probability density function of Beta distribution $Beta(\alpha, 1)$, $\alpha = 10$ can effectively make the accuracy parameters have very small probability to be close to 0.5.

   - $p(\nu_g)$: set its concentration parameter $\gamma = 1$, the recovered number of experts is stable across a range of different γ values (from 0.5 to 10).
   - $p(W_{ij}^{(g)}), p(b_i^{(g)})$: set the mean and standard deviation of prior Gaussian distributions to be 0 and 1 respectively.

3. Initializations of variational parameters:
   - the parameters of variational Beta distribution are initialized to be equal to those of the prior distribution;
   - the means of variational Gaussian distributions are initialized by randomly drawing samples from Gaussian distribution $\mathcal{N}(0, \frac{1}{D})$, the standard deviations of variational Gaussian distributions are initialized to be 0.001. This initialization strategy is similar to Xavier initialization [5] in order to avoid gradient saturation;
   - $\eta_{i,:}^{(g)}$ are initialized using $W^{(g)}, b^{(g)}$ according to the discriminative clustering model described in the main paper;
   - $\phi_{m,:}$ are initialized by sampling from a Dirichlet distribution with all its parameters being equal to 1, introducing most randomness for the initialization.
4. The number of random initializations for optimization is set to be 50 and the results are stable across different runs in all our experiments.
5. Cluster samples in the testing set:
   After learning our model using the training set, we can obtain the variational distributions of weight $q(W^{(g)})$ and offset $q(b^{(g)})$. We can cluster x_t, a sample from the testing set by integrating out $W^{(g)}, b^{(g)}$ through Monte Carlo approximation to the integration:

$$p(Z_t^{(g)} = k|x_t) = \int\int p(Z_t^{(g)}|W^{(g)}, b^{(g)}; x_t) q(W^{(g)}) q(b^{(g)}) dW^{(g)} db^{(g)} \tag{1}$$

$$\approx \frac{1}{L} \sum_{l=1}^{L} p(Z_t^{(g)}|\widehat{W^{(g)}}_l, \widehat{b^{(g)}}_l; x_t) \tag{2}$$

   where $\widehat{W^{(g)}}_l, \widehat{b^{(g)}}_l$ are the l-th sample from $q(W^{(g)})$ and $q(b^{(g)})$ respectively. We set $L = 100$ in all experiments. x_t is assigned to the cluster corresponding to the largest probability:

$$\hat{y}_t = \arg\max_{k=1,\cdots K} p(Z_t^{(g)} = k|x_t) \tag{3}$$

   where $\hat{y}_t$ is the predicted cluster label for x_t.

1.2 Meta Clustering

After computing the similarity matrix between multiple experts, we apply spectral clustering [6] to assign experts to different views. The number of views are automatically determined using the eigen-gap heuristic [7].

1.3 SemiCrowd

We follow parameter settings recommended by the author in [9]. In particular, d_0, d_1, two thresholds used to filter out uncertain sample pairs in the average similarity matrix, are set to be 0 and 0.8 respectively. Parameters of the ℓ_1 regularized matrix completion algorithm is set according to heuristics in the literature [8].

1.4 ITML

As is suggested by the author [4], we set lower and upper bounds associated with the constraint terms to be the 5^{th} and 95^{th} percentiles of the observed distribution of distances between pairs of points within the dataset. Other parameters related to the convergence of the algorithm are set to be default.

1.5 MPCKMeans

We specify the constraints according to the instructions listed on the author's website and directly run the author's Java implementation [1]

1.6 CSPA

After computing the average similarity matrix, we use spectral clustering [6] to obtain the cluster labels.

2 Variational Inference

2.1 Variational Distribution

Denote $h = \{\alpha_{1:M}, \beta_{1:M}, c_{1:M}, \nu_{1:G}, Z^{(1:G)}, W^{(1:G)}, b^{(1:G)}\}$ as the collection of latent variables, where G is the number of components of the truncated Dirichlet Process. We assume the variational distribution has the following factorization formula:

$$
\begin{aligned}
& q(\alpha_{1:M}, \beta_{1:M}, c_{1:M}, \nu_{1:G}, Z^{(1:G)}, W^{(1:G)}, b^{(1:G)}) && (4)\\
&= q(\alpha_{1:M}) \cdot q(\beta_{1:M}) \cdot q(c_{1:M}) \cdot q(\nu_{1:G}) \cdot q(Z^{(1:G)}) \cdot q(W^{(1:G)}) \cdot q(b^{(1:G)}) && (5)\\
&= \prod_{m=1}^{M} [q(\alpha_m) \cdot q(\beta_m) \cdot q(c_m)] \prod_{g=1}^{G} \left[q(\nu_g) q(Z^{(g)}) q(W^{(g)}) q(b^{(g)}) \right] && (6)\\
&= \prod_{m=1}^{M} [q(\alpha_m) \cdot q(\beta_m) \cdot q(c_m)] \prod_{g=1}^{G} \left[q(\nu_g) \prod_{i=1}^{n} q(Z_i^{(g)}) \prod_{i=1}^{d} \prod_{j=1}^{K} q(W_{ij}^{(g)}) \prod_{i=1}^{K} q(b_i^{(g)}) \right] && (7)
\end{aligned}
$$

where the marginal variational distribution of each random variable is

- $q(\alpha_m) = \text{Beta}(\tau_{\alpha^1}^{(m)}, \tau_{\alpha^2}^{(m)})$;
- $q(\beta_m) = \text{Beta}(\tau_{\beta^1}^{(m)}, \tau_{\beta^2}^{(m)})$;
- $q(c_m) = \text{Cat}(\phi_{m,:})$;
- $q(\nu_g) = \text{Beta}(\tau_{\nu^1}^{(g)}, \tau_{\nu^2}^{(g)})$;
- $q(Z_i^{(g)}) = \text{Cat}(\eta_{i,:}^{(g)})$;
- $q(W_{ij}^{(g)}) = \mathcal{N}(\mu_{W_{ij}}^{(g)}, \sigma_{W_{ij}}^{(g)2})$;
- $q(b_i^{(g)}) = \mathcal{N}(\mu_{b_i}^{(g)}, \sigma_{b_i}^{(g)2})$.

We use θ to denote all the variational parameters, which consist of the following

- For $m = 1 \cdots M$, consider $\{\tau_{\alpha^1}^{(m)}, \tau_{\alpha^2}^{(m)}, \tau_{\beta^1}^{(m)}, \tau_{\beta^2}^{(m)}, \phi_{m,:} \in \Delta^{G-1}\}$, where Δ^{G-1} is a simplex in G-dimensional space.
- For $g = 1 \cdots G$, consider $\tau_{\nu^1}^{(g)}, \tau_{\nu^2}^{(g)}, \eta_{i,:}^{(g)} \in \Delta^{K-1} (i = 1 \cdots n), \mu_{W_{ij}}^{(g)}, \sigma_{W_{ij}}^{(g)2} (i = 1 \cdots d, j = 1 \cdots K)$, $\mu_{b_i}^{(g)}, \sigma_{b_i}^{(g)2} (i = 1 \cdots K)$

Besides the simplex constraints on the parameters of categorical distribution, both the parameters of Beta distribution and the standard deviation of Gaussian distribution should have positive constraints.

2.2 Evidence Lower Bound

Given variational distribution $q(h; \theta)$, the log-likelihood $\log p(S^{(1:M)})$ can be decomposed as

$$
\begin{aligned}
\log p(S^{(1:M)}) &= \mathcal{L}_{q(h;\theta)} + KL \left[q(h;\theta) \,|\, p(h|S^{(1:M)}) \right] && (8)\\
&\geq \mathcal{L}_{q(h;\theta)} && (9)\\
&= \text{E}_{q(h;\theta)} \left[\log p(h, S^{(1:M)}) - \log q(h;\theta) \right] && (10)\\
&= \text{E}_{q(h;\theta)} \left[\log p(S^{(1:M)}|h) + \log p(h) - \log q(h;\theta) \right] && (11)\\
&= \text{E}_{q(h;\theta)} \left[\log p(S^{(1:M)}|h) \right] - KL \left[q(h;\theta) \,|\, p(h) \right] && (12)
\end{aligned}
$$

The first term prefers the variational distribution $q(h;\theta)$ that maximizes the expected conditional likelihood, the second term forces the variational distribution to be close to the prior distribution.

The overall objective of variational inference is to maximize the evidence lower bound (ELBO) $\mathcal{L}_{q(h;\theta)}$ w.r.t. θ.

Let $f(h;\theta) = \log p(S^{(1:M)}|h) + \log p(h) - \log q(h;\theta)$, the evidence lower bound can be expressed as

$$\mathcal{L}_{q(h;\theta)} = \mathrm{E}_{q(h;\theta)}\big[f(h;\theta)\big] \tag{13}$$

We first compute the three terms in $f(h;\theta)$ and then evaluate their expectations w.r.t. the variational distribution $q(h;\theta)$.

2.3 ELBO: The First Term $\log p(S^{(1:M)}|h)$

Consider the first term:

$$\log p(S^{(1:M)}|h) = \log\left(\prod_{m=1}^{M}\prod_{(i,j)\in E^{(m)}} p(S_{ij}^{(m)}|h)\right) \tag{14}$$

$$= \sum_{m=1}^{M}\sum_{(i,j)\in E^{(m)}} \log p(S_{ij}^{(m)}|h) \tag{15}$$

$$\begin{aligned} = \sum_{m=1}^{M}\Big\{&\Big(\sum_{(i,j)\in E^{(m)}} 1\Big)\log\beta_m + \Big(\sum_{(i,j)\in E^{(m)}} S_{ij}^{(m)}\Big)\log\Big(\frac{1-\beta_m}{\beta_m}\Big) \\ &+ \Big(\sum_{(i,j)\in E^{(m)}} S_{ij}^{(m)}A_{ij}^{(m)}\Big)\log\Big[\frac{\alpha_m\beta_m}{(1-\alpha_m)(1-\beta_m)}\Big] \\ &+ \Big(\sum_{(i,j)\in E^{(m)}} A_{ij}^{(m)}\Big)\log(\frac{1-\alpha_m}{\beta_m})\Big\} \end{aligned} \tag{16}$$

To compute $\mathrm{E}_{q(h;\theta)}[\log p(S^{(1:M)}\,|\,h)]$, the following formulas can be used:

$$\mathrm{E}_{q(h;\theta)}[\log\alpha_m] = \mathrm{E}_{q(\alpha_m)}[\log\alpha_m] = \psi(\tau_{\alpha^1}^{(m)}) - \psi(\tau_{\alpha^1}^{(m)} + \tau_{\alpha^2}^{(m)}) \tag{17}$$

$$\mathrm{E}_{q(h;\theta)}[\log(1-\alpha_m)] = \mathrm{E}_{q(\alpha_m)}[\log(1-\alpha_m)] = \psi(\tau_{\alpha^2}^{(m)}) - \psi(\tau_{\alpha^1}^{(m)} + \tau_{\alpha^2}^{(m)}) \tag{18}$$

$$\mathrm{E}_{q(h;\theta)}[\log\beta_m] = \mathrm{E}_{q(\beta_m)}[\log\beta_m] = \psi(\tau_{\beta^1}^{(m)}) - \psi(\tau_{\beta^1}^{(m)} + \tau_{\beta^2}^{(m)}) \tag{19}$$

$$\mathrm{E}_{q(h;\theta)}[\log(1-\beta_m)] = \mathrm{E}_{q(\beta_m)}[\log(1-\beta_m)] = \psi(\tau_{\beta^2}^{(m)}) - \psi(\tau_{\beta^1}^{(m)} + \tau_{\beta^2}^{(m)}) \tag{20}$$

$$\mathrm{E}_{q(h;\theta)}[A_{ij}^{(m)}] = \mathrm{E}_{q(c_m, Z^{(c_m)})}[\mathbb{I}(Z_i^{(c_m)} = Z_j^{(c_m)})] \tag{21}$$

$$= \mathrm{E}_{q(c_m)}\big[\mathrm{E}_{q(Z^{(c_m)})}[\mathbb{I}(Z_i^{(c_m)} = Z_j^{(c_m)})]\big] \tag{22}$$

$$= \mathrm{E}_{q(c_m)}\Big[\sum_{k=1}^{K}\eta_{ik}^{(c_m)}\eta_{jk}^{(c_m)}\Big] \tag{23}$$

$$= \sum_{g=1}^{G}\Big[\phi_{mg}\sum_{k=1}^{K}\eta_{ik}^{g}\eta_{jk}^{g}\Big] \tag{24}$$

2.4 ELBO: The Second Term $\log p(h)$

Consider the second term:

$$\log p(h) = \log \Bigg(\prod_{m=1}^{M} \Big(p(\alpha_m) p(\beta_m) p(c_m|\nu_{1:\infty}) \Big) \prod_{g=1}^{\infty} \Big(p(\nu_g) p(Z^{(g)}|W^{(g)}, b^{(g)}) p(W^{(g)}) p(b^{(g)}) \Big) \Bigg) \tag{25}$$

$$= \sum_{m=1}^{M} \Big(\log p(\alpha_m) + \log p(\beta_m) + \log p(c_m|\nu_{1:\infty}) \Big) \tag{26}$$

$$+ \sum_{g=1}^{\infty} \Big(\log p(\nu_g) + \log p(Z^{(g)}|W^{(g)}, b^{(g)}) + \log p(W^{(g)}) + \log p(b^{(g)}) \Big)$$

where

$$\log p(\alpha_m) = (\tau^{(m)}_{\alpha^1 0} - 1) \log \alpha_m + (\tau^{(m)}_{\alpha^2 0} - 1) \log(1 - \alpha_m) - \log \left(\frac{\Gamma(\tau^{(m)}_{\alpha^1 0}) \Gamma(\tau^{(m)}_{\alpha^2 0})}{\Gamma(\tau^{(m)}_{\alpha^1 0} + \tau^{(m)}_{\alpha^2 0})} \right) \tag{27}$$

$$\log p(\beta_m) = (\tau^{(m)}_{\beta^1 0} - 1) \log \beta_m + (\tau^{(m)}_{\beta^2 0} - 1) \log(1 - \beta_m) - \log \left(\frac{\Gamma(\tau^{(m)}_{\beta^1 0}) \Gamma(\tau^{(m)}_{\beta^2 0})}{\Gamma(\tau^{(m)}_{\beta^1 0} + \tau^{(m)}_{\beta^2 0})} \right) \tag{28}$$

$$\log p(c_m|\nu_{1:\infty}) = \sum_{g=1}^{\infty} c_{mg} \Big(\log \nu_g + \sum_{j=1}^{g-1} \log(1 - \nu_j) \Big) \tag{29}$$

$$\log p(\nu_g) = (\gamma - 1) \log(1 - \nu_g) - \log \left(\frac{\Gamma(\gamma)}{\Gamma(1 + \gamma)} \right) \tag{30}$$

$$= (\gamma - 1) \log(1 - \nu_g) + \log \gamma \tag{31}$$

$$\log p(Z^{(g)}|W^{(g)}, b^{(g)}) = \log \prod_{i=1}^{N} p(Z^{(g)}_i|W^{(g)}, b^{(g)}) \tag{32}$$

$$= \log \prod_{i=1}^{N} \prod_{k=1}^{K} p\big(Z^{(g)}_i = k|W^{(g)}, b^{(g)}\big)^{\mathbb{I}[Z^{(g)}_i = k]} \tag{33}$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{I}[Z^{(g)}_i = k] \log p(Z^{(g)}_i = k|W^{(g)}, b^{(g)}) \tag{34}$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{I}[Z^{(g)}_i = k] \Big((w^{(g)T}_k x_i + b^{(g)}_k) - \log \Big(\sum_{j=1}^{K} \exp(w^{(g)T}_j x_i + b^{(g)}_j) \Big) \Big) \tag{35}$$

$$\log p(W^{(g)}) = \sum_{i=1}^{d} \sum_{j=1}^{K} \Big(- \log(\sqrt{2\pi} \sigma^{(g)}_{W_{ij}0}) - \frac{(W^{(g)}_{ij} - \mu^{(g)}_{W_{ij}0})^2}{2\sigma^{(g)2}_{W_{ij}0}} \Big) \tag{36}$$

$$\log p(b^{(g)}) = \sum_{i=1}^{K} \Big(- \log(\sqrt{2\pi} \sigma^{(g)}_{b_i 0}) - \frac{(b^{(g)}_i - \mu^{(g)}_{b_i 0})^2}{2\sigma^{(g)2}_{b_i 0}} \Big) \tag{37}$$

To compute $\mathrm{E}_{q(h;\theta)}[\log p(h)]$, the following formulas can be used besides the formulas used to compute

$\mathrm{E}_{q(h;\theta)}[\log p(S^{(1:M)} \mid h)]$:

$$\mathrm{E}_{q(h;\theta)}[\log \nu_g] = \mathrm{E}_{q(\nu_g)}[\log \nu_g] = \psi(\tau_{\nu^1}^{(g)}) - \psi(\tau_{\nu^1}^{(g)} + \tau_{\nu^2}^{(g)}) \tag{38}$$

$$\mathrm{E}_{q(h;\theta)}[\log(1-\nu_g)] = \mathrm{E}_{q(\nu_g)}[\log(1-\nu_g)] = \psi(\tau_{\nu^2}^{(g)}) - \psi(\tau_{\nu^1}^{(g)} + \tau_{\nu^2}^{(g)}) \tag{39}$$

$$\mathrm{E}_{q(h;\theta)}[c_{mg}] = \mathrm{E}_{q(c_m)}[c_{mg}] = \phi_{mg} \tag{40}$$

$$\mathrm{E}_{q(h;\theta)}\left[\mathbb{I}[Z_i^{(g)} = k]\right] = \eta_{ik}^{(g)} \tag{41}$$

$$\mathrm{E}_{q(h;\theta)}[w_k^{(g)}] = \mathrm{E}_{q(W^{(g)})}[w_k^{(g)}] = \mu_{W_{:k}}^{(g)} \tag{42}$$

$$\mathrm{E}_{q(h;\theta)}[(W_{ij}^{(g)} - \mu_{W_{ij}0}^{(g)})^2] = \sigma_{W_{ij}}^{(g)2} + (\mu_{W_{ij}}^{(g)} - \mu_{W_{ij}0}^{(g)})^2 \tag{43}$$

$$\mathrm{E}_{q(h;\theta)}[(b_i^{(g)} - \mu_{b_i0}^{(g)})^2] = \sigma_{b_i}^{(g)2} + (\mu_{b_i}^{(g)} - \mu_{b_i0}^{(g)})^2 \tag{44}$$

2.4.1 Upper Bound of Log-sum Function

The computation of $\mathrm{E}_{q(h;\theta)}\left[\log\left(\sum_{j=1}^K \exp(w_j^{(g)T} x_i + b_j)\right)\right]$ does not have closed form and can only be approximated through sampling. However, we can use the upper bound of log-sum function based on the log concavity to derive a lower bound of ELBO [2].

$$\log\left(\sum_{j=1}^K \exp(w_j^{(g)T} x_i + b_j^{(g)})\right) \leq r_i^{(g)} \sum_{j=1}^K \exp(w_j^{(g)T} x_i + b_j^{(g)}) - \log(r_i^{(g)}) - 1 \tag{45}$$

We need introduce new variational parameters $r_i^{(g)} > 0$ to optimize. The expectation term can also be upper bounded as follows

$$\begin{aligned}
&\mathrm{E}_{q(h;\theta)}\left[\log\left(\sum_{j=1}^K \exp(w_j^{(g)T} x_i + b_j)\right)\right] \\
&\leq \mathrm{E}_{q(h;\theta)}\left[r_i^{(g)} \sum_{j=1}^K \exp(w_j^{(g)T} x_i + b_j^{(g)}) - \log(r_i^{(g)}) - 1\right] &(46)\\
&= r_i^{(g)} \sum_{j=1}^K \mathrm{E}_{q(h;\theta)}[\exp(\sum_{l=1}^d w_{lj}^{(g)} x_{il} + b_j^{(g)})] - \log(r_i^{(g)}) - 1 &(47)\\
&= r_i^{(g)} \sum_{j=1}^K \left(\mathrm{E}_{q(b_j^{(g)})}\left[e^{b_j^{(g)}}\right] \prod_{l=1}^d \mathrm{E}_{q(W_{lj}^{(g)})}\left[e^{w_{lj}^{(g)} x_{il}}\right]\right) - \log(r_i^{(g)}) - 1 &(48)
\end{aligned}$$

where the expectation can be evaluated using the mean of log-normal distribution.

$$\mathrm{E}_{q(b_j^{(g)})}\left[e^{b_j^{(g)}}\right] = \exp\left(\mu_{b_j}^{(g)} + \frac{\sigma_{b_j}^{(g)2}}{2}\right) \tag{49}$$

$$E_{q(W_{lj}^{(g)})}\left[e^{w_{lj}^{(g)} x_{il}}\right] = \exp\left(\mu_{W_{lj}}^{(g)} x_{il} + \frac{\sigma_{W_{lj}}^{(g)2} x_{il}^2}{2}\right) \tag{50}$$

Therefore, the upper bound can be written as

$$\begin{aligned}
&\mathrm{E}_{q(h;\theta)}\left[\log\left(\sum_{j=1}^K \exp(w_j^{(g)T} x_i + b_j)\right)\right] \\
&\leq r_i^{(g)} \sum_{j=1}^K \left(\exp\left(\mu_{b_j}^{(g)} + \frac{\sigma_{b_j}^{(g)2}}{2} + \sum_{l=1}^d \left(\mu_{W_{lj}}^{(g)} x_{il} + \frac{\sigma_{W_{lj}}^{(g)2} x_{il}^2}{2}\right)\right)\right) - \log(r_i^{(g)}) - 1 &(51)
\end{aligned}$$

The upper bound becomes minimal when

$$r_i^{(g)} = \frac{1}{\sum_{j=1}^{K}\left(\exp\left(\mu_{b_j}^{(g)} + \frac{\sigma_{b_j}^{(g)2}}{2} + \sum_{l=1}^{d}\left(\mu_{W_{lj}}^{(g)} x_{il} + \frac{\sigma_{W_{lj}}^{(g)2} x_{il}^2}{2}\right)\right)\right)} \tag{52}$$

Then the upper bound becomes

$$\mathrm{E}_{q(h;\theta)}\left[\log\Big(\sum_{j=1}^{K}\exp(w_j^{(g)T}x_i + b_j)\Big)\right] \tag{53}$$

$$\leq \log\sum_{j=1}^{K}\left(\exp\left(\mu_{b_j}^{(g)} + \frac{\sigma_{b_j}^{(g)2}}{2} + \sum_{l=1}^{d}\left(\mu_{W_{lj}}^{(g)} x_{il} + \frac{\sigma_{W_{lj}}^{(g)2} x_{il}^2}{2}\right)\right)\right) \tag{54}$$

To avoid numerical overflow when computing this bound due to the potential large exponent of the exponential function, we can use the logsumexp() function implemented in Scipy. The idea is to extract the maximal value in the sequence and compute its log, then each exponent in the sequence will be less than 1, leading to stable numerical behavior.

2.5 ELBO: The Third Term $\log q(h;\theta)$

Consider the third term:

$$\log q(h;\theta) = \sum_{m=1}^{M}\Big(\log q(\alpha_m) + \log q(\beta_m) + \log q(c_m)\Big) \tag{55}$$

$$+ \sum_{g=1}^{G}\left(\log q(\nu_g) + \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{I}[Z_i^{(g)} = k]\log q(Z_i^{(g)} = k) + \sum_{i=1}^{d}\sum_{j=1}^{K}\log q(W_{ij}^{(g)}) + \sum_{i=1}^{K}\log q(b_i^{(g)})\right) \tag{56}$$

where

$$\log q(\alpha_m) = (\tau_{\alpha^1}^{(m)} - 1)\log\alpha_m + (\tau_{\alpha^2}^{(m)} - 1)\log(1-\alpha_m) - \log\Big(\frac{\Gamma(\tau_{\alpha^1}^{(m)})\Gamma(\tau_{\alpha^2}^{(m)})}{\Gamma(\tau_{\alpha^1}^{(m)} + \tau_{\alpha^2}^{(m)})}\Big) \tag{57}$$

$$\log q(\beta_m) = (\tau_{\beta^1}^{(m)} - 1)\log\beta_m + (\tau_{\beta^2}^{(m)} - 1)\log(1-\beta_m) - \log\Big(\frac{\Gamma(\tau_{\beta^1}^{(m)})\Gamma(\tau_{\beta^2}^{(m)})}{\Gamma(\tau_{\beta^1}^{(m)} + \tau_{\beta^2}^{(m)})}\Big) \tag{58}$$

$$\log q(c_m) = \sum_{g=1}^{G} c_{mg}\log\phi_{mg} \tag{59}$$

$$\log q(\nu_g) = (\tau_{\nu^1}^{(g)} - 1)\log\nu_g + (\tau_{\nu^2}^{(g)} - 1)\log(1-\nu_g) - \log\Big(\frac{\Gamma(\tau_{\nu^1}^{(g)})\Gamma(\tau_{\nu^2}^{(g)})}{\Gamma(\tau_{\nu^1}^{(g)} + \tau_{\nu^2}^{(g)})}\Big) \tag{60}$$

$$\log q(Z_i^{(g)} = k) = \log\eta_{ik}^{(g)} \tag{61}$$

$$\log q(W_{ij}^{(g)}) = -\log(\sqrt{2\pi}\sigma_{W_{ij}}^{(g)}) - \frac{(W_{ij}^{(g)} - \mu_{W_{ij}}^{(g)})^2}{2\sigma_{W_{ij}}^{(g)2}} \tag{62}$$

$$\log q(b_i^{(g)}) = -\log(\sqrt{2\pi}\sigma_{b_i}^{(g)}) - \frac{(b_i^{(g)} - \mu_{b_i}^{(g)})^2}{2\sigma_{b_i}^{(g)2}} \tag{63}$$

The computation of $E_{q(h;\theta)}[\log q(h;\theta)]$ can be done using formulas from the previous two subsections.

3 Weights of Different Experts

As we can see from the derivation of the first-term $\log p(S^{(1:M)}|h)$ of ELBO, if we view Equation (16) as a function of $A_{ij}^{(m)}$, the weight of $S_{ij}^{(m)}$ (constraints provided by the m-th expert) is: $\log \frac{\alpha_m \beta_m}{(1-\alpha_m)(1-\beta_m)}$.

References

[1] M. Bilenko. Java implementation of mpckmeans, 2004.

[2] G. Bouchard. Efficient bounds for the softmax function and applications to approximate inference in hybrid models. In *NIPS 2007 workshop for approximate Bayesian inference in continuous/hybrid systems*. Citeseer, 2007.

[3] P. J. Castaldi, J. Dy, J. Ross, Y. Chang, G. R. Washko, D. Curran-Everett, A. Williams, D. A. Lynch, B. J. Make, J. D. Crapo, et al. Cluster analysis in the copdgene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema. *Thorax*, pages thoraxjnl–2013, 2014.

[4] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM, 2007.

[5] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*, volume 9, pages 249–256, 2010.

[6] A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

[7] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[8] J. Yi, R. Jin, A. K. Jain, and S. Jain. Crowdclustering with sparse pairwise labels: A matrix completion approach. In *AAAI Workshop on Human Computation*, volume 2. Citeseer, 2012.

[9] J. Yi, R. Jin, S. Jain, T. Yang, and A. K. Jain. Semi-crowdsourced clustering: Generalizing crowd labeling by robust distance metric learning. In *Advances in Neural Information Processing Systems*, pages 1772–1780, 2012.