Supplement

1 Additional Experimental Results

IBTM with different features and parameter settings In this section, we present additional experimental results using the LabelMe dataset with image pairs which is similar to Section 4.2.1.1 in the original paper. Different features and different parameter settings are used to compare the performance.

Topic Size	Conv3_1 V512	Conv3_1 V1024	Conv5_1 V512	Conv5_1 V1024	fc8 V512	fc8 V1024
$T = 5, K = 5, S = 5$	51.25%	53.87%	62.88%	62%	66.37%	66.13%
$T = 15, K = 15, S = 15$	66.00%	68.13%	86.25%	89.75%	90.50%	91.87%
$T = 30, K = 30, S = 30$	68.63%	68.38%	87.13%	87.50%	91%	91.50%

Table 1: Performance comparison of LabelMe paired image experiments using different features, different vocabulary sizes and different number of topics. "V512" indicates the vocabulary size is 512 and "V1024" indicates the vocabulary size is 1024. "Conv3_1" means that the features are extracted using the Conv3_1 layer from the VGG 16 layer CNN [5], which is a middle layer of the deep net, while the Conv5_1 is the first layer from the last convolutional layer group. "fc8" means that the features are extracted using the fc8 layer which is the last fully connected layer. Each row of the table presents experiments with different numbers of topics which are marked in the first column of the table(K is the number of shared topics. T is the number of private topics for the first view and S is the number of private topics for the second view).

Full fc8	PCA 15 fc8	IBTM15 fc8
89.63%	88.125%	**91.87**%

Table 2: Performance comparison of LabelMe image experiments using features extracted from the last fully connected layer fc8 from VGG 16 layer CNN [5].

BoW SVM [4]	SPM SVM [4]	Object Bank [4]	IBTM15 SIFT
55%	58%	68%	**71.8**%

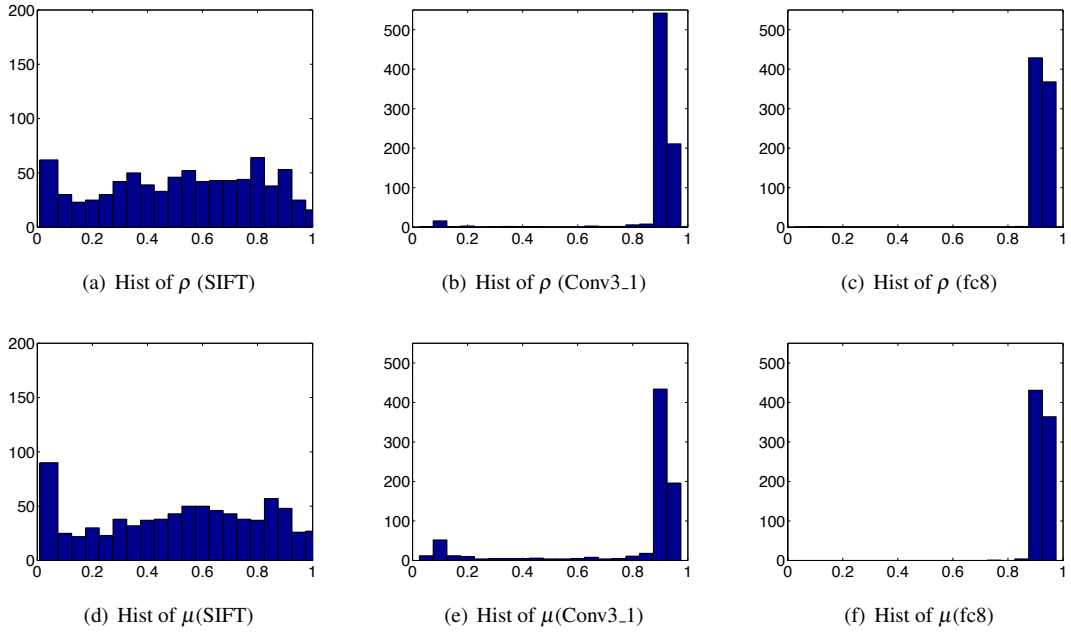
Table 3: Performance comparison of LabelMe image experiments using SIFT.

In Table 1, the classification results are compared using different features, different vocabulary sizes and different number of topics. Comparing the performance with *different features*, the results show that the performance in general increases when the off-the-shelf CNN features are extracted from a deeper layer of the CNN. These results are consistent with the results from this study on deep representations [1]. In general, the deeper the layer is, the more specific information the features represent. Thus, the features extracted from deeper layers benefit the classification tasks more. Comparing the performance with *different vocabulary sizes*, the differences are quite small. We conclude that the performance is in general robust when the vocabulary size is sufficient. Comparing the performance with *different number of topics*, the performance decreases rapidly when the topic size is small, which means that there are not sufficiently many topics to represent the images. However, as long as the number of topics is sufficient, the performance increment is limited or none. This result is consistent with the experimental results from multiple works on different topic models [2, 6, 7], where the performance is low when the number of topics is small and becomes stable when the number of topics is sufficient.

Along with Table 1 in the paper, where performance is compared with different baseline results using features extracted from Conv5_1 layer, Table 2 gives another example using features extracted from fc8 layer. We can see that the advancement of IBTM compared to other methods is consistent with the results presented in the paper. In the end, Table 3 presents the experimental results with SIFT features compared to other methods using the same dataset (only using the images). The results are significantly lower using SIFT features compared to off-the-shelf CNN features, however, IBTM can still improve the performance against other methods using the same type of features. Both the quality of the features and the quality of the models are key to the performance.

Figure 1 shows the histograms of the partition parameters using different features (SIFT, Conv3_1 and fc8) with vocabulary size 1024 and number of topics $T = 15, K = 15, S = 15$. Figure 6 in the paper presents results with Conv5_1 features, which can be used for comparison here. We can see that the better the feature quality, the larger the partition parameters. As expected, with better feature quality, less noise is introduced in the feature representation of the data, hence, the more information can be shared between paired images using IBTM. This group of results further confirms that IBTMl can learn a good representation with different noise levels in the data.

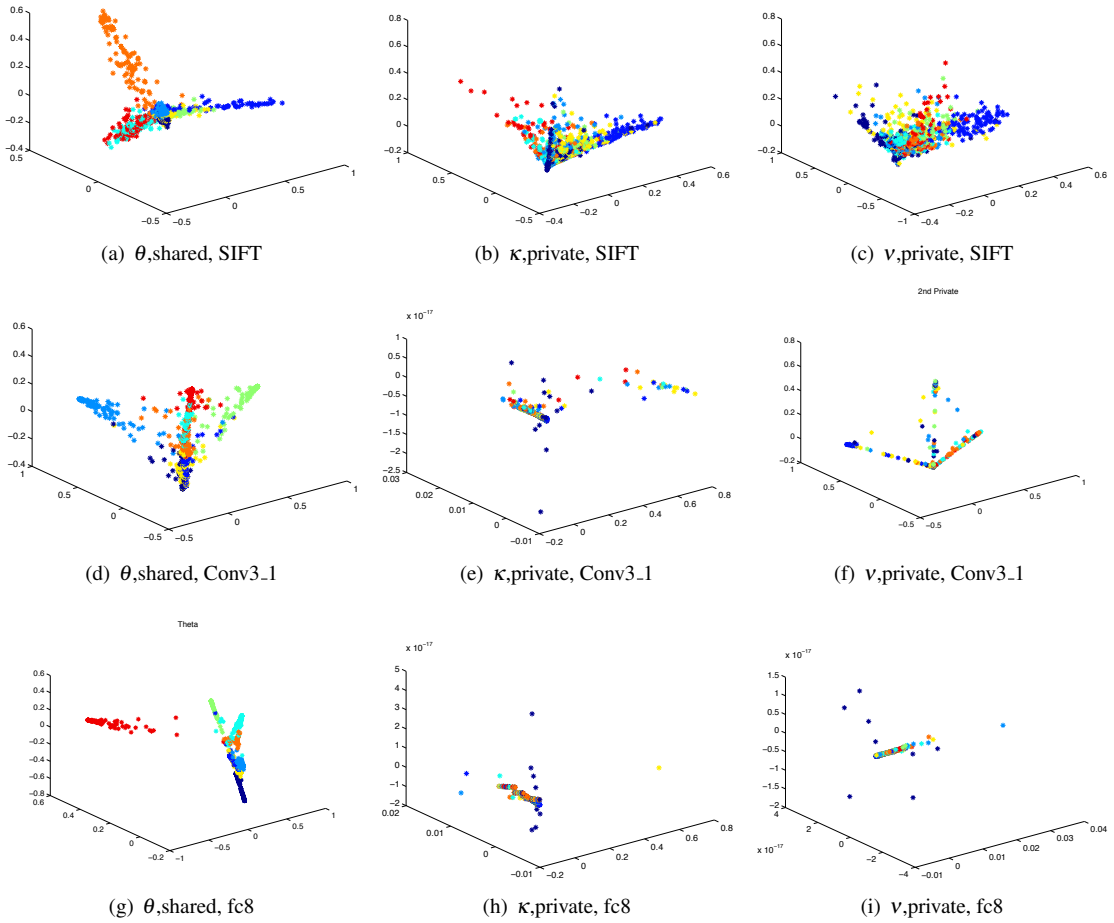
Similarly to Figure 5 in the paper, Figure 2 visualizes the document distributions in different topic representation spaces using different features. The shared topics representations all naturally separate images from different classes while the private ones do not. However, the noisier the features are, the more information is presented in the private topic space with bigger variances and the need of larger dimensionality.



(a) Hist of ρ (SIFT) (b) Hist of ρ (Conv3_1) (c) Hist of ρ (fc8)

(d) Hist of μ(SIFT) (e) Hist of μ(Conv3_1) (f) Hist of μ(fc8)

Figure 1: The histograms of the partition parameters



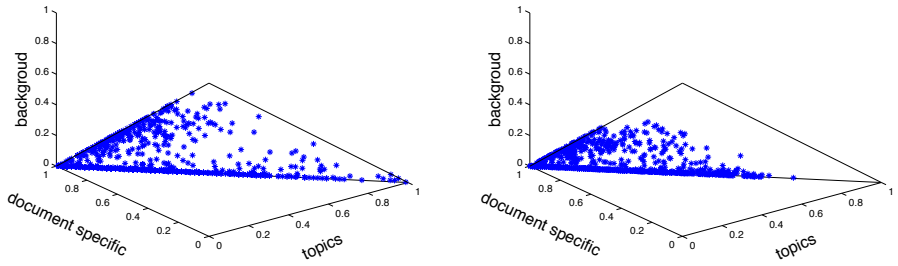
(a) θ,shared, SIFT (b) κ,private, SIFT (c) ν,private, SIFT

(d) θ,shared, Conv3_1 (e) κ,private, Conv3_1 (f) ν,private, Conv3_1

(g) θ,shared, fc8 (h) κ,private, fc8 (i) ν,private, fc8

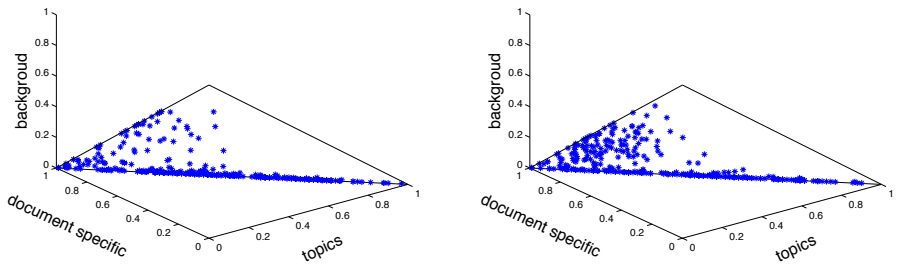
Figure 2: Visualization of the shared topic representation (θ) and private topic representations (κ and ν) with different features for LabelMe experiments using randomly paired images from the same class. The documents of different classes are colored differently and the plots show the first three principal components after applying PCA on the per document topic distributions for all the training data.

Additional Analysis for SWB To understand the performance of SWB [3], the factorization variable λ is plotted for different experiment settings in the paper in Figure 3. In SWB, λ is a three dimensional per document latent variable. $\lambda(1)$ indicates the proportion of words that are sampled from shared topics; $\lambda(2)$ indicates the proportion of words that are sampled from document specific words distribution; and $\lambda(3)$ indicates the proportion of words that are sampled from general background words distribution. SWB has shown promising results for information retrieval since it can match documents both at general topic level (using shared topics) and at specific words level (using document specific words distribution). However, the performance is unsatisfactory for computer vision tasks as shown in the paper. The reason is that the visual words are much more noisy than words in text. All plots in Figure 3 show that the distributions are biased towards document specific information. Comparing the plots in two rows with different dataset, the distributions are more concentrated in document specific information with LabelMe dataset. This shows that most information are modeled as document specific distribution with SWB, which may due to that visual words for image are very noisy. SWB performs better with Leeds butterfly dataset than with LabelMe dataset is probably due to that the visual words are cleaner and more information can be modeled by shared topics.



(a) LabelMe Image-Image

(b) LabelMe Image-Caption



(c) Leeds Image-Image

(d) Leeds Image-Segmenation

Figure 3: Latent variable λ in SWB. λ is document a three diminutional latent variable in SWB which indicates the factorization. The plots shows the distribution of λ on a simplex for those four different experiment settings in the paper using the training data.

2 Variational Inference

In this section, the derivation details are given. The evidence lower bound (ELBO) will be computed first and the update equations will be derived given the ELBO.

2.1 ELBO

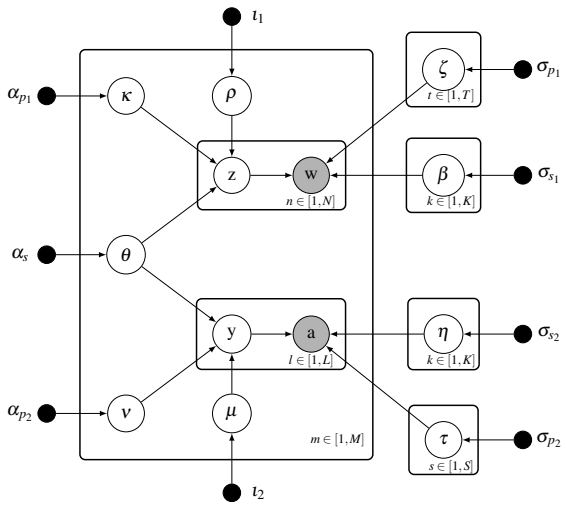


Figure 4: Graphical representation of IBTM

Recalling IBTM in the paper, the graphical representation is shown in Figure 4. The ELBO of IBTM is:

$$
\begin{aligned}
\mathscr{L} &= \mathop{\mathbb{E}}_{q}[\log p(w,a,\mathbb{Z}|\Theta)] - \mathop{\mathbb{E}}_{q}[\log q(\mathbb{Z})] \\
&= \mathop{\mathbb{E}}_{q}[\log p(\kappa,\theta,\nu,\rho,z,w,\mu,y,a,\beta,\eta,\zeta,\tau|\alpha_{p_1},\alpha_s,\alpha_{p_2},\sigma_{p1},\sigma_{p2},\sigma_{s1},\sigma_{s2},\iota_1,\iota_2)] \\
&- \mathop{\mathbb{E}}_{q}[\log q(\kappa,\theta,\nu,\rho,z,\mu,y,\beta,\eta,\zeta,\tau)] \\
&= \sum_{k=1}^{K} \mathop{\mathbb{E}}_{q}[\log p(\beta_k|\sigma_{s1})] + \sum_{k=1}^{K} \mathop{\mathbb{E}}_{q}[\log p(\eta_k|\sigma_{s2})] + \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q}[\log p(\zeta_t|\sigma_{s1})] + \sum_{s=1}^{S} \mathop{\mathbb{E}}_{q}[\log p(\tau_t|\sigma_{s2})] \\
&\sum_{m=1}^{M} \Bigg(\mathop{\mathbb{E}}_{q}[\log p(\kappa_m|\alpha_{p_1})] + \mathop{\mathbb{E}}_{q}[\log p(\theta_m|\alpha_s)] + \mathop{\mathbb{E}}_{q}[\log p(\nu_m|\alpha_{p_2})] + \mathop{\mathbb{E}}_{q}[\log p(\rho_m|\iota_1)] + \mathop{\mathbb{E}}_{q}[\log p(\mu_m|\iota_2)] \\
&+ \sum_{n=1}^{N} \Big(\mathop{\mathbb{E}}_{q}[\log p(z_{mn}|\kappa_m,\theta_m,\rho_m)] + \mathop{\mathbb{E}}_{q}[\log p(w_{mn}|z_{mn},\beta,\zeta)] \Big) \\
&+ \sum_{l=1}^{L} \Big(\mathop{\mathbb{E}}_{q}[\log p(y_{ml}|\nu_m,\theta_m,\mu_m)] + \mathop{\mathbb{E}}_{q}[\log p(a_{ml}|y_{ml},\eta,\tau)] \Big) \Bigg) \\
&- \mathop{\mathbb{E}}_{q}[\log q(\kappa)] - \mathop{\mathbb{E}}_{q}[\log q(\theta)] - \mathop{\mathbb{E}}_{q}[\log q(\nu)] - \mathop{\mathbb{E}}_{q}[\log q(\rho)] - \mathop{\mathbb{E}}_{q}[\log q(z)] \\
&- \mathop{\mathbb{E}}_{q}[\log q(\mu)] - \mathop{\mathbb{E}}_{q}[\log q(y)] - \mathop{\mathbb{E}}_{q}[\log q(\beta)] - \mathop{\mathbb{E}}_{q}[\log q(\eta)] - \mathop{\mathbb{E}}_{q}[\log q(\zeta)] - \mathop{\mathbb{E}}_{q}[\log q(\tau)]
\end{aligned}
\tag{1}
$$

The 1st term:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{q}[\log p(\beta_k|\sigma_{s1})] \\
&= \log\Gamma(V\sigma_{s1}) - V\log\Gamma(\sigma_{s1}) + \sum_{v=1}^{V} (\sigma_{s1}-1)\Big(\Psi(\lambda_{kv}) - \Psi(\sum_{i=1}^{V}\lambda_{ki})\Big)
\end{aligned}
\tag{2}
$$

The 2rd term:

$$\begin{aligned}&\mathop{\mathbb{E}}_{q}[\log p(\eta_t|\sigma_{s2})]\\&=\log\Gamma(W\sigma_{s2})-W\log\Gamma(\sigma_{s2})+\sum_{w=1}^{W}(\sigma_{s2}-1)\Big(\Psi(\upsilon_{tw})-\Psi(\sum_{i=1}^{W}\upsilon_{ti})\Big)\end{aligned}\tag{3}$$

The 3st term:

$$\begin{aligned}&\mathop{\mathbb{E}}_{q}[\log p(\zeta_t|\sigma_{p1})]\\&=\log\Gamma(V\sigma_{p1})-V\log\Gamma(\sigma_{p1})+\sum_{v=1}^{V}(\sigma_{p1}-1)\Big(\Psi(\xi_{tv})-\Psi(\sum_{i=1}^{V}\xi_{ti})\Big)\end{aligned}\tag{4}$$

The 4rd term:

$$\begin{aligned}&\mathop{\mathbb{E}}_{q}[\log p(\tau_s|\sigma_{p2})]\\&=\log\Gamma(W\sigma_{p2})-W\log\Gamma(\sigma_{p2})+\sum_{w=1}^{W}(\sigma_{p2}-1)\Big(\Psi(\mathrm{o}_{sw})-\Psi(\sum_{i=1}^{W}\mathrm{o}_{si})\Big)\end{aligned}\tag{5}$$

The 5th term:

$$\begin{aligned}&\mathop{\mathbb{E}}_{q}[\log p(\kappa_m|\alpha_{p1})]\\&=\log\Gamma(T\alpha_{p1})-T\log\Gamma(\alpha_{p1})+\sum_{t=1}^{T}(\alpha_{p1}-1)\Big(\Psi(\delta_{mt})-\Psi(\sum_{i=1}^{T}\delta_{mi})\Big)\end{aligned}\tag{6}$$

The 6th term:

$$\begin{aligned}&\mathop{\mathbb{E}}_{q}[\log p(\theta_m|\alpha_s)]\\&=\log\Gamma(K\alpha_s)-K\log\Gamma(\alpha_s)+\sum_{k=1}^{K}(\alpha_s-1)\Big(\Psi(\gamma_{mk})-\Psi(\sum_{i=1}^{K}\gamma_{mi})\Big)\end{aligned}\tag{7}$$

The 7th term:

$$\begin{aligned}&\mathop{\mathbb{E}}_{q}[\log p(\nu_m|\alpha_{p2})]\\&=\log\Gamma(S\alpha_{p2})-S\log\Gamma(\alpha_{p2})+\sum_{s=1}^{S}(\alpha_{p2}-1)\Big(\Psi(\varepsilon_{ms})-\Psi(\sum_{i=1}^{S}\varepsilon_{mi})\Big)\end{aligned}\tag{8}$$

The 8th term:

$$\begin{aligned}&\mathop{\mathbb{E}}_{q}[\log p(\rho_m|\iota_1)]\\&=\mathop{\mathbb{E}}_{q}[\log\frac{\Gamma(\iota_{11}+\iota_{12})}{\Gamma(\iota_{11})\Gamma(\iota_{12})}\rho_m^{(\iota_{11}-1)}(1-\rho_m)^{(\iota_{12}-1)}]\\&=\log\Gamma(\iota_{11}+\iota_{12})-\log\Gamma(\iota_{11})-\log\Gamma(\iota_{12})+(\iota_{11}-1)\mathop{\mathbb{E}}_{q}[\log\rho_m]+(\iota_{12}-1)\mathop{\mathbb{E}}_{q}[\log(1-\rho_m)]\\&=\log\Gamma(\iota_{11}+\iota_{12})-\log\Gamma(\iota_{11})-\log\Gamma(\iota_{12})\\&+(\iota_{11}-1)\left(\Psi(r_{m1})-\Psi(r_{m1}+r_{m2})\right)+(\iota_{12}-1)\left(\Psi(r_{m2})-\Psi(r_{m1}+r_{m2})\right)\end{aligned}\tag{9}$$

The 9th term:

$$\begin{aligned}&\mathop{\mathbb{E}}_{q}[\log p(\mu_m|\iota_2)]\\&=\log\Gamma(\iota_{21}+\iota_{22})-\log\Gamma(\iota_{21})-\log\Gamma(\iota_{12})\\&+(\iota_{21}-1)\left(\Psi(u_{m1})-\Psi(u_{m1}+u_{m2})\right)+(\iota_{22}-1)\left(\Psi(u_{m2})-\Psi(u_{m1}+u_{m2})\right)\end{aligned}\tag{10}$$

The 10th term:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{q}[\log p(z_{mn}|\kappa_m,\theta_m,\rho_m)] \\
&=\mathop{\mathbb{E}}_{q}[\log\prod_{k=1}^{K}(\theta_{mk}\rho_m)^{[z_{mn}=k]}\prod_{t=1}^{T}(\kappa_{mt}(1-\rho_m))^{[z_{mn}=K+t]}] \\
&=\sum_{k=1}^{K}\mathop{\mathbb{E}}_{q}[[z_{mn}=k](\log\theta_{mk}+\log\rho_m)]+\sum_{t=1}^{T}\mathop{\mathbb{E}}_{q}[[z_{mn}=K+t](\log\kappa_{mt}+\log(1-\rho_m)] \\
&=\sum_{k=1}^{K}\phi_{mnk}\left(\left(\Psi(\gamma_{mk})-\Psi(\sum_{i=1}^{K}\gamma_{mi})\right)+\left(\Psi(r_{m1})-\Psi(r_{m1}+r_{m2})\right)\right) \\
&+\sum_{t=1}^{T}\phi_{mn(K+t)}\left(\left(\Psi(\delta_{mt})-\Psi(\sum_{i=1}^{T}\delta_{mi})\right)+\left(\Psi(r_{m2})-\Psi(r_{m1}+r_{m2})\right)\right)
\end{aligned}
\tag{11}
$$

Similarly, the 12th term:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{q}[\log p(y_{ml}|\nu_m,\theta_m,\mu_m)] \\
&=\sum_{k=1}^{K}\chi_{mnk}\left(\left(\Psi(\gamma_{mk})-\Psi(\sum_{i=1}^{K}\gamma_{mi})\right)+\left(\Psi(u_{m1})-\Psi(u_{m1}+u_{m2})\right)\right) \\
&+\sum_{s=1}^{S}\chi_{mn(K+t)}\left(\left(\Psi(\varepsilon_{mt})-\Psi(\sum_{i=1}^{T}\varepsilon_{mi})\right)+\left(\Psi(u_{m2})-\Psi(u_{m1}+u_{m2})\right)\right)
\end{aligned}
\tag{12}
$$

The 11th term:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{q}[\log p(w_{mn}|z_{mn},\zeta,\beta)] \\
&=\mathop{\mathbb{E}}_{q}\left[\log\prod_{k=1}^{K}p(w_{mn}|\beta_k)^{[z_{mn}=k]}\prod_{t=1}^{T}p(w_{mn}|\zeta_k)^{[z_{mn}=K+t]}\right] \\
&=\mathop{\mathbb{E}}_{q}\left[\sum_{k=1}^{K}[z_{mn}=k]\log p(w_{mn}|\beta_k)+\sum_{t=1}^{T}[z_{mn}=K+t]\log p(w_{mn}|\zeta_t)\right] \\
&=\sum_{k=1}^{K}\phi_{mnk}\mathop{\mathbb{E}}_{q}[\log p(w_{mn}|\beta_k)]+\sum_{t=1}^{t}\phi_{mn(k+t)}\mathop{\mathbb{E}}_{q}[\log p(w_{mn}|\zeta_t)] \\
&=\sum_{k=1}^{K}\phi_{mnk}\mathop{\mathbb{E}}_{q}\left[\log\prod_{v=1}^{V}\beta_{kv}^{[w_{mn}=v]}\right]+\sum_{t=1}^{t}\phi_{mn(k+t)}\mathop{\mathbb{E}}_{q}\left[\log\prod_{v=1}^{V}\zeta_{tv}^{[w_{mn}=v]}\right] \\
&=\sum_{k=1}^{K}\sum_{v=1}^{V}\phi_{mnk}[w_{mn}=v]\mathop{\mathbb{E}}_{q}[\log\beta_{kv}]+\sum_{t=1}^{T}\sum_{v=1}^{V}\phi_{mn(K+t)}[w_{mn}=v]\mathop{\mathbb{E}}_{q}[\log\zeta_{tv}] \\
&=\sum_{k=1}^{K}\sum_{v=1}^{V}\phi_{mnk}[w_{mn}=v]\left(\Psi(\lambda_{kv})-\Psi(\sum_{i=1}^{V}\lambda_{ki})\right) \\
&+\sum_{t=1}^{T}\sum_{v=1}^{V}\phi_{mn(K+t)}[w_{mn}=v]\left(\Psi(\xi_{tv})-\Psi(\sum_{i=1}^{V}\xi_{ti})\right)
\end{aligned}
\tag{13}
$$

Similarly, the 13th term:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{q}[\log p(a_{ml}|y_{ml},\eta,\tau)] \\
&=\sum_{k=1}^{K}\sum_{w=1}^{W}\chi_{mlk}[a_{ml}=w]\left(\Psi(\upsilon_{kv})-\Psi(\sum_{i=1}^{W}\upsilon_{ki})\right) \\
&+\sum_{s=1}^{S}\sum_{w=1}^{W}\chi_{ml(K+s)}[a_{ml}=w]\left(\Psi(\mathrm{o}_{sw})-\Psi(\sum_{i=1}^{W}\mathrm{o}_{si})\right)
\end{aligned}
\tag{14}
$$

The 14th term:

$$
\begin{aligned}
&-\mathop{\mathbb{E}}_{q}[\log q(\kappa)] \\
&=\sum_{m=1}^{M}\left(-\log\Gamma(\sum_{i=1}^{T}\delta_{mi})+\sum_{t=1}^{T}\log\Gamma(\delta_{mt})-\sum_{t=1}^{T}(\delta_{mt}-1)\left(\Psi(\delta_{mt})-\Psi(\sum_{i=1}^{T}\delta_{mi})\right)\right)
\end{aligned}
\tag{15}
$$

The 15th term:

$$
\begin{aligned}
&-\mathop{\mathbb{E}}_{q}[\log q(\theta)] \\
&=\sum_{m=1}^{M}\left(-\log\Gamma(\sum_{i=1}^{K}\gamma_{mi})+\sum_{k=1}^{K}\log\Gamma(\gamma_{mk})-\sum_{k=1}^{K}(\gamma_{mk}-1)\left(\Psi(\gamma_{mk})-\Psi(\sum_{i=1}^{K}\gamma_{mi})\right)\right)
\end{aligned}
\tag{16}
$$

The 16th term:

$$
\begin{aligned}
&-\mathop{\mathbb{E}}_{q}[\log q(\nu)] \\
&=\sum_{m=1}^{M}\left(-\log\Gamma(\sum_{i=1}^{S}\varepsilon_{mi})+\sum_{s=1}^{S}\log\Gamma(\varepsilon_{ms})-\sum_{s=1}^{S}(\varepsilon_{ms}-1)\left(\Psi(\varepsilon_{ms})-\Psi(\sum_{i=1}^{S}\varepsilon_{mi})\right)\right)
\end{aligned}
\tag{17}
$$

The 17th term:

$$
\begin{aligned}
&-\mathop{\mathbb{E}}_{q}[\log q(\rho)] \\
&=\sum_{m=1}^{M}\Bigg(\log\Gamma(r_{m1})+\log\Gamma(r_{m2})-\log\Gamma(r_{m1}+r_{m2}) \\
&-(r_{m1}-1)\Psi(r_{m1})-(r_{m2}-1)\Psi(r_{m2})+(r_{m1}+r_{m2}-2)\Psi(r_{m1}+r_{m2})\Bigg)
\end{aligned}
\tag{18}
$$

The 18th term:

$$
\begin{aligned}
&-\mathop{\mathbb{E}}_{q}[\log q(z)] \\
&=-\sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{i=1}^{K+T}\phi_{mni}\log\phi_{mni}
\end{aligned}
\tag{19}
$$

The 19th term:

$$
\begin{aligned}
&-\mathop{\mathbb{E}}_{q}[\log q(\mu)] \\
&=\sum_{m=1}^{M}\Bigg(\log\Gamma(u_{m1})+\log\Gamma(u_{m2})-\log\Gamma(u_{m1}+u_{m2}) \\
&-(u_{m1}-1)\Psi(u_{m1})-(u_{m2}-1)\Psi(u_{m2})+(u_{m1}+u_{m2}-2)\Psi(u_{m1}+u_{m2})\Bigg)
\end{aligned}
\tag{20}
$$

The 20th term:

$$
\begin{aligned}
&-\mathop{\mathbb{E}}_{q}[\log q(y)] \\
&=-\sum_{m=1}^{M}\sum_{l=1}^{L}\sum_{i=1}^{K+S}\chi_{mli}\log\chi_{mli}
\end{aligned}
\tag{21}
$$

The 21th term:

$$
\begin{aligned}
&-\mathop{\mathbb{E}}_{q}[\log q(\beta)] \\
&=\sum_{k=1}^{K}\left(-\log\Gamma(\sum_{i=1}^{V}\lambda_{ki})+\sum_{v=1}^{V}\log\Gamma(\lambda_{kv})-\sum_{v=1}^{V}(\lambda_{kv}-1)\left(\Psi(\lambda_{kv})-\Psi(\sum_{i=1}^{V}\lambda_{ki})\right)\right)
\end{aligned}
\tag{22}
$$

The 22th term:

$$
\begin{aligned}
&-\mathop{\mathbb{E}}_{q}[\log q(\eta)] \\
&=\sum_{k=1}^{K}\left(-\log\Gamma(\sum_{i=1}^{W}\upsilon_{ki})+\sum_{w=1}^{W}\log\Gamma(\upsilon_{kw})-\sum_{w=1}^{W}(\upsilon_{kw}-1)\left(\Psi(\upsilon_{kw})-\Psi(\sum_{i=1}^{W}\upsilon_{ki})\right)\right)
\end{aligned}
\tag{23}
$$

The 23th term:

$$
\begin{aligned}
&-\mathop{\mathbb{E}}_{q}[\log q(\zeta)] \\
&=\sum_{t=1}^{T}\left(-\log\Gamma(\sum_{i=1}^{V}\xi_{ti})+\sum_{v=1}^{V}\log\Gamma(\xi_{tv})-\sum_{v=1}^{V}(\xi_{tv}-1)\left(\Psi(\xi_{tv})-\Psi(\sum_{i=1}^{V}\xi_{ti})\right)\right)
\end{aligned}
\tag{24}
$$

The 24th term:

$$
\begin{aligned}
&-\mathop{\mathbb{E}}_{q}[\log q(\tau)] \\
&=\sum_{s=1}^{S}\left(-\log\Gamma(\sum_{i=1}^{W}\mathrm{o}_{si})+\sum_{w=1}^{V}\log\Gamma(\mathrm{o}_{sw})-\sum_{w=1}^{W}(\mathrm{o}_{sw}-1)\left(\Psi(\mathrm{o}_{sw})-\Psi(\sum_{i=1}^{W}\mathrm{o}_{si})\right)\right)
\end{aligned}
\tag{25}
$$

2.2 Compute the update equations

The part of the bound that varies with δ which is the variational parameter for κ, involves terms in 5, 10 and 14 as follows

$$
\begin{aligned}
&\mathscr{L}_{\delta} \\
&=\sum_{m=1}^{M}\left(\sum_{t=1}^{T}(\alpha_{p1}-1)\Big(\Psi(\delta_{mt})-\Psi(\sum_{i=1}^{T}\delta_{mi})\Big)\right. \\
&\left.+\sum_{n=1}^{N}\left(\sum_{t=1}^{T}\phi_{mn(K+t)}\Big(\big(\Psi(\delta_{mt})-\Psi(\sum_{i=1}^{T}\delta_{mi})\big)+\big(\Psi(r_{m2})-\Psi(r_{m1}+r_{m2})\big)\Big)\right)\right) \\
&+\sum_{m=1}^{M}\left(-\log\Gamma(\sum_{i=1}^{T}\delta_{mi})+\sum_{t=1}^{T}\log\Gamma(\delta_{mt})-\sum_{t=1}^{T}(\delta_{mt}-1)\left(\Psi(\delta_{mt})-\Psi(\sum_{i=1}^{T}\delta_{mi})\right)\right) \\
&=\sum_{m=1}^{M}\sum_{t=1}^{T}\left((\alpha_{p1}-1)\Big(\Psi(\delta_{mt})-\Psi(\sum_{i=1}^{T}\delta_{mi})\Big)+\sum_{n=1}^{N}\left(\phi_{mn(K+t)}(\Big(\Psi(\delta_{mt})-\Psi(\sum_{i=1}^{T}\delta_{mi})\Big)\right)\right. \\
&\left.-(\delta_{mt}-1)\left(\Psi(\delta_{mt})-\Psi(\sum_{i=1}^{T}\delta_{mi})\right)\right)+\sum_{m=1}^{M}\left(-\log\Gamma(\sum_{i=1}^{T}\delta_{mi})+\sum_{t=1}^{T}\log\Gamma(\delta_{mt})\right) \\
&=\sum_{m=1}^{M}\sum_{t=1}^{T}\left(\Big(\alpha_{p1}+\sum_{n=1}^{N}\phi_{mn(K+t)}-\delta_{mt}\Big)\Psi(\delta_{mt})-\Big(\alpha_{p1}+\sum_{n=1}^{N}\phi_{mn(K+t)}-\delta_{mt}\Big)\Psi(\sum_{i=1}^{T}\delta_{mi})\right) \\
&+\sum_{m=1}^{M}\left(-\log\Gamma(\sum_{i=1}^{T}\delta_{mi})+\sum_{t=1}^{T}\log\Gamma(\delta_{mt})\right)
\end{aligned}
\tag{26}
$$

$$
\begin{aligned}
&\frac{\partial\mathscr{L}_{\delta}}{\partial\delta_{mt}} \\
&=\Big(\alpha_{p1}+\sum_{n=1}^{N}\phi_{mn(K+t)}-\delta_{mt}\Big)\Psi'(\delta_{mt})-\Big(\alpha_{p1}+\sum_{n=1}^{N}\phi_{mn(K+t)}-\delta_{mt}\Big)\Psi'(\sum_{i=1}^{T}\delta_{mi})
\end{aligned}
\tag{27}
$$

The update equation for δ_{mt} is

$$
\delta_{mt}=\alpha_{p1}+\sum_{n=1}^{N}\phi_{mn(K+t)}. \tag{28}
$$

The part of the bound that varies with γ which is the variational parameter for θ, involves term 6, 10, 12 and 15 as follows

$$
\begin{aligned}
&\mathscr{L}_{\gamma} \\
&= \sum_{m=1}^{M}\left(\sum_{k=1}^{K}(\alpha_s-1)\Big(\Psi(\gamma_{mk})-\Psi(\sum_{i=1}^{K}\gamma_{mi})\Big)\right) \\
&+ \sum_{m=1}^{M}\sum_{n=1}^{N}\left(\sum_{k=1}^{K}\phi_{mnk}\Big(\Psi(\gamma_{mk})-\Psi(\sum_{i=1}^{K}\gamma_{mi})\Big)\right) \\
&+ \sum_{m=1}^{M}\sum_{l=1}^{L}\left(\sum_{k=1}^{K}\chi_{mlk}\Big(\Psi(\gamma_{mk})-\Psi(\sum_{i=1}^{K}\gamma_{mi})\Big)\right) \\
&+ \sum_{m=1}^{M}\left(-\log\Gamma(\sum_{i=1}^{K}\gamma_{mi})+\sum_{k=1}^{K}\log\Gamma(\gamma_{mk})-\sum_{k=1}^{K}(\gamma_{mk}-1)\left(\Psi(\gamma_{mk})-\Psi(\sum_{i=1}^{K}\gamma_{mi})\right)\right) \\
&= \sum_{m=1}^{M}\sum_{k=1}^{K}\Bigg(\Big(\alpha_s+\sum_{n=1}^{N}\phi_{mnk}f_{mn}+\sum_{l=1}^{L}\chi_{mlk}c_{ml}-\gamma_{mk}\Big)\Psi(\gamma_{mk}) \\
&- \Big(\alpha_s+\sum_{n=1}^{N}\phi_{mnk}f_{mn}+\sum_{l=1}^{L}\chi_{mlk}c_{ml}-\gamma_{mk}\Big)\Psi(\sum_{i=1}^{K}\gamma_{mi})\Bigg) \\
&+ \sum_{m=1}^{M}\left(-\log\Gamma(\sum_{i=1}^{K}\gamma_{mi})+\sum_{k=1}^{K}\log\Gamma(\gamma_{mk})\right)
\end{aligned}
\tag{29}
$$

$$
\begin{aligned}
&\frac{\partial\mathscr{L}_{\gamma}}{\partial\gamma_{mk}} \\
&= \Big(\alpha_s+\sum_{n=1}^{N}\phi_{mnk}+\sum_{l=1}^{L}\chi_{mlk}-\gamma_{mk}\Big)\Psi'(\gamma_{mk})-\Big(\alpha_s+\sum_{n=1}^{N}\phi_{mnk}+\sum_{l=1}^{L}\chi_{mlk}-\gamma_{mk}\Big)\Psi'(\sum_{i=1}^{K}\gamma_{mi})
\end{aligned}
\tag{30}
$$

$$
\gamma_{mk} = \alpha_s+\sum_{n=1}^{N}\phi_{mnk}+\sum_{l=1}^{L}\chi_{mlk} \tag{31}
$$

Similarly,

$$
\varepsilon_{ms} = \alpha_{p2}+\sum_{l=1}^{L}\chi_{ml(K+s)} \tag{32}
$$

The part of the bound that varies with ϕ which is the variational parameter for z, involve term 10, 11 and 18 as follows

$$
\begin{aligned}
&\mathscr{L}_\phi \\
&= \sum_{m=1}^{M}\sum_{n=1}^{N}\Bigg(\sum_{k=1}^{K}\phi_{mnk}\Bigg(\Big(\Psi(\gamma_{mk})-\Psi(\sum_{i=1}^{K}\gamma_{mi})\Big)+\Big(\Psi(r_{m1})-\Psi(r_{m1}+r_{m2})\Big)\Bigg) \\
&+\sum_{t=1}^{T}\phi_{mn(K+t)}\Bigg(\Big(\Psi(\delta_{mt})-\Psi(\sum_{i=1}^{T}\delta_{mi})\Big)+\Big(\Psi(r_{m2})-\Psi(r_{m1}+r_{m2})\Big)\Bigg) \\
&+\sum_{k=1}^{K}\sum_{v=1}^{V}\phi_{mnk}[w_{mn}=v]\left(\Psi(\lambda_{kv})-\Psi(\sum_{i=1}^{V}\lambda_{ki})\right) \\
&+\sum_{t=1}^{T}\sum_{v=1}^{V}\phi_{mn(K+t)}[w_{mn}=v]\left(\Psi(\xi_{tv})-\Psi(\sum_{i=1}^{V}\xi_{ti})\right) \\
&-\sum_{i=1}^{K+T}\phi_{mni}\log\phi_{mni}\Bigg) \\
&= \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{k=1}^{K}\Bigg(\phi_{mnk}\Bigg(\Big(\Psi(\gamma_{mk})-\Psi(\sum_{i=1}^{K}\gamma_{mi})\Big)+\Big(\Psi(r_{m1})-\Psi(r_{m1}+r_{m2})\Big)\Bigg) \\
&+\sum_{v=1}^{V}\phi_{mnk}[w_{mn}=v]\left(\Psi(\lambda_{kv})-\Psi(\sum_{i=1}^{V}\lambda_{ki})\right)-\phi_{mni}\log\phi_{mni}\Bigg) \\
&+\sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{t=1}^{T}\Bigg(\phi_{mn(K+t)}\Bigg(\Big(\Psi(\delta_{mt})-\Psi(\sum_{i=1}^{T}\delta_{mi})\Big)+\Big(\Psi(r_{m2})-\Psi(r_{m1}+r_{m2})\Big)\Bigg) \\
&+\sum_{v=1}^{V}\phi_{mn(K+t)}[w_{mn}=v]\left(\Psi(\xi_{tv})-\Psi(\sum_{i=1}^{V}\xi_{ti})\right)-\phi_{mni}\log\phi_{mni}\Bigg)
\end{aligned}
\tag{33}
$$

when $i \le K$

$$
\begin{aligned}
&\frac{\partial\mathscr{L}_\phi}{\partial\phi_{mni}} \\
&= \Big(\Psi(\gamma_{mk})-\Psi(\sum_{i=1}^{K}\gamma_{mi})\Big)+\Big(\Psi(r_{m1})-\Psi(r_{m1}+r_{m2})\Big) \\
&+\sum_{v=1}^{V}[w_{mn}=v]\left(\Psi(\lambda_{iv})-\Psi(\sum_{p=1}^{V}\lambda_{ip})\right)-\log\phi_{mni}-1
\end{aligned}
\tag{34}
$$

$$
\begin{aligned}
&\phi_{mni} \\
&= \exp\Bigg(\Big(\Psi(\gamma_{mk})-\Psi(\sum_{i=1}^{K}\gamma_{mi})\Big)+\Big(\Psi(r_{m1})-\Psi(r_{m1}+r_{m2})\Big) \\
&+\sum_{v=1}^{V}[w_{mn}=v]\left(\Psi(\lambda_{iv})-\Psi(\sum_{p=1}^{V}\lambda_{ip})\right)-1\Bigg)
\end{aligned}
\tag{35}
$$

when $i > K$ (as $i = K+t$)

$$
\begin{aligned}
&\frac{\partial\mathscr{L}_\phi}{\partial\phi_{mni}} \\
&= \Bigg(\Big(\Psi(\delta_{m(i-K)})-\Psi(\sum_{p=1}^{T}\delta_{mp})\Big)+\Big(\Psi(r_{m2})-\Psi(r_{m1}+r_{m2})\Big)\Bigg) \\
&+\sum_{v=1}^{V}[w_{mn}=v]\left(\Psi(\xi_{iv})-\Psi(\sum_{p=1}^{V}\xi_{ip})\right)-\log\phi_{mni}-1
\end{aligned}
\tag{36}
$$

$$
\begin{aligned}
&\phi_{mni} \\
&= \exp\left(\left(\left(\Psi(\delta_{m(i-K)}) - \Psi(\sum_{p=1}^{T}\delta_{mp})\right) + \left(\Psi(r_{m2}) - \Psi(r_{m1}+r_{m2})\right)\right)\right. \\
&\left. + \sum_{v=1}^{V}[w_{mn}=v]\left(\Psi(\xi_{iv}) - \Psi(\sum_{p=1}^{V}\xi_{ip}))\right) - 1\right)
\end{aligned} \tag{37}
$$

The part of the bound that varies with χ which is the variational parameter for y, involves term 12, 13 and 20 as follows

$$
\begin{aligned}
&\mathscr{L}_{\chi} \\
&= \sum_{m=1}^{M}\sum_{l=1}^{L}\left(\sum_{k=1}^{K}\chi_{mnk}\left(\left(\Psi(\gamma_{mk}) - \Psi(\sum_{i=1}^{K}\gamma_{mi})\right) + \left(\Psi(u_{m1}) - \Psi(u_{m1}+u_{m2})\right)\right)\right. \\
&+ \sum_{s=1}^{S}\chi_{mn(K+s)}\left(\left(\Psi(\varepsilon_{mt}) - \Psi(\sum_{i=1}^{T}\varepsilon_{mi})\right) + \left(\Psi(u_{m2}) - \Psi(u_{m1}+u_{m2})\right)\right) \\
&+ \sum_{k=1}^{K}\sum_{w=1}^{W}\chi_{mlk}[a_{ml}=w]\left(\Psi(\upsilon_{kv}) - \Psi(\sum_{i=1}^{W}\upsilon_{ki})\right) \\
&+ \sum_{s=1}^{S}\sum_{w=1}^{W}\chi_{ml(K+s)}[a_{ml}=w]\left(\Psi(\mathrm{o}_{sw}) - \Psi(\sum_{i=1}^{W}\mathrm{o}_{si})\right) \\
&\left. - \sum_{i=1}^{K+S}\chi_{mli}\log\chi_{mli}\right)
\end{aligned} \tag{38}
$$

Similarly, when $i \leq K$

$$
\begin{aligned}
&\chi_{mni} \\
&= \exp\left(\left(\left(\Psi(\gamma_{mi}) - \Psi(\sum_{p=1}^{K}\gamma_{mp})\right) + \left(\Psi(u_{m1}) - \Psi(u_{m1}+u_{m2})\right)\right)\right. \\
&\left. + \sum_{w=1}^{W}[a_{ml}=w]\left(\Psi(\upsilon_{iv}) - \Psi(\sum_{p=1}^{W}\upsilon_{ip})\right) - 1\right)
\end{aligned} \tag{39}
$$

when $i > K$ (as $i = K+s$)

$$
\begin{aligned}
&\chi_{mni} \\
&= \exp\left(\left(\left(\Psi(\varepsilon_{mt}) - \Psi(\sum_{i=1}^{T}\varepsilon_{mi})\right) + \left(\Psi(u_{m2}) - \Psi(u_{m1}+u_{m2})\right)\right)\right. \\
&\left. + \sum_{w=1}^{W}[a_{ml}=w]\left(\Psi(\mathrm{o}_{iv}) - \Psi(\sum_{p=1}^{W}\mathrm{o}_{ip}))\right) - 1\right)
\end{aligned} \tag{40}
$$

The part of the bound that varies with r which is the variational parameter for ρ, involves term 8, 10 and 17

as follows

$$
\begin{aligned}
&\mathscr{L}_r \\
&= \sum_{m=1}^{M} \Bigg((\iota_{11}-1)\left(\Psi(r_{m1})-\Psi(r_{m1}+r_{m2})\right) + (\iota_{12}-1)\left(\Psi(r_{m2})-\Psi(r_{m1}+r_{m2})\right) \Bigg) \\
&+ \sum_{n=1}^{N} \Bigg(\sum_{k=1}^{K} \phi_{mnk} \Big(\Psi(r_{m1}) - \Psi(r_{m1}+r_{m2}) \Big) \\
&+ \sum_{t=1}^{T} \phi_{mn(K+t)} \Big(\Psi(r_{m2}) - \Psi(r_{m1}+r_{m2}) \Big) \Bigg) \\
&+ \log\Gamma(r_{m1}) + \log\Gamma(r_{m2}) - \log\Gamma(r_{m1}+r_{m2}) \\
&- (r_{m1}-1)\Psi(r_{m1}) - (r_{m2}-1)\Psi(r_{m2}) + (r_{m1}+r_{m2}-2)\Psi(r_{m1}+r_{m2}) \Bigg) \qquad (41) \\
&= \sum_{m=1}^{M} \Bigg(\Bigg(\iota_{11} + \sum_{n=1}^{N}\sum_{k=1}^{K} \phi_{mnk} - r_{m1} \Bigg) \Psi(r_{m1}) \\
&+ \Bigg(\iota_{12} + \sum_{n=1}^{N}\sum_{t=1}^{T} \phi_{mn(K+t)} - r_{m2} \Bigg) \Psi(r_{m2}) \\
&- \Bigg(\iota_{11} + \iota_{12} + \sum_{n=1}^{N}\sum_{k=1}^{K} \phi_{mnk} + \sum_{n=1}^{N}\sum_{t=1}^{T} \phi_{mn(K+t)} - r_{m1} - r_{m2} \Bigg) \Psi(r_{m1}+r_{m2}) \\
&+ \log\Gamma(r_{m1}) + \log\Gamma(r_{m2}) - \log\Gamma(r_{m1}+r_{m2}) \Bigg)
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial \mathscr{L}_r}{\partial r_{m1}} \\
&= \Bigg(\iota_{11} + \sum_{n=1}^{N}\sum_{k=1}^{K} \phi_{mnk} - r_{m1} \Bigg) \Psi'(r_{m1}) \qquad (42) \\
&- \Bigg(\iota_{11} + \iota_{12} + \sum_{n=1}^{N}\sum_{k=1}^{K} \phi_{mnk} + \sum_{n=1}^{N}\sum_{t=1}^{T} \phi_{mn(K+t)} - r_{m1} - r_{m2} \Bigg) \Psi'(r_{m1}+r_{m2})
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial \mathscr{L}_r}{\partial r_{m2}} \\
&= \Bigg(\iota_{12} + \sum_{n=1}^{N}\sum_{t=1}^{T} \phi_{mn(K+t)} - r_{m2} \Bigg) \Psi'(r_{m2}) \qquad (43) \\
&- \Bigg(\iota_{11} + \iota_{12} + \sum_{n=1}^{N}\sum_{k=1}^{K} \phi_{mnk} + \sum_{n=1}^{N}\sum_{t=1}^{T} \phi_{mn(K+t)} - r_{m1} - r_{m2} \Bigg) \Psi'(r_{m1}+r_{m2})
\end{aligned}
$$

Setting both to zero, we can get:

$$r_{m1} = \iota_{11} + \sum_{n=1}^{N}\sum_{k=1}^{K} \phi_{mnk} \qquad (44)$$

$$r_{m2} = \iota_{12} + \sum_{n=1}^{N}\sum_{t=1}^{T} \phi_{mn(K+t)} \qquad (45)$$

Similarly, to update u which is the variational parameter for μ is

$$u_{m1} = \iota_{21} + \sum_{l=1}^{L}\sum_{k=1}^{K} \chi_{mns} \qquad (46)$$

$$u_{m2} = \iota_{22} + \sum_{l=1}^{L} \sum_{s=1}^{S} \chi_{mn(K+s)} \tag{47}$$

The derivation for the update of $\xi, \lambda, \upsilon, \mathrm{o}$ follows exactly the same procedure as LDA.

$$\xi_{tv} = \sigma_{p1} + \sum_{m=1}^{M} \sum_{n=1}^{N} \phi_{mn(K+t)} [w_{mn} = v] \tag{48}$$

$$\lambda_{kv} = \sigma_{s1} + \sum_{m=1}^{M} \sum_{n=1}^{N} \phi_{mnk} [w_{mn} = v] \tag{49}$$

$$\upsilon_{kw} = \sigma_{s2} + \sum_{m=1}^{M} \sum_{l=1}^{L} \chi_{mlk} [a_{ml} = w] \tag{50}$$

$$\mathrm{o}_{sw} = \sigma_{p2} + \sum_{m=1}^{M} \sum_{l=1}^{L} \chi_{ml(K+s)} [a_{ml} = w] \tag{51}$$

References

[1] H. Azizpour, A. Razavian, J. Sullivan, A. Maki, and S. Carlsson. From generic to specific deep representations for visual recognition. In *DeepVision workshop, CVPR*, 2015.

[2] D. M. Blei and M. I. Jordan. Modeling annotated data. In *International Conference on Research and Development in Information Retrieval*, pages 127–134. ACM, 2003.

[3] C. Chemudugunta, P. Smyth, and M. Steyvers. Modeling general and specific aspects of documents with a probabilistic topic model. In *NIPS*, volume 19, pages 241–248, 2006.

[4] L. J. Li, H. Su, Y. Lim, and L. Fei-Fei. Objects as attributes for scene classification. In *Trends and Topics in Computer Vision*, pages 57–69. Springer, 2012.

[5] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[6] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, pages 1903–1910. IEEE, 2009.

[7] C. Zhang and H. Kjellstrom. How to Supervise Topic Models. In *ECCV workshop on Graphical Models in Computer Vision*, 2014.