Collaborative Topic Modeling for Netflix Prize Problem

Chaofan (Bill) Huang
chuang397@gatech.edu

Tiancheng Ye
tye41@gatech.edu

Yasin Cagatay Gultekin
ygultekin3@gatech.edu

April 22, 2020

Problem Statement

In 2007, Netflix held an open comptition for the best recommender system to predict the users' ratings on films based on their previous ratings on other films. No other outside information about the films and the users can be used to help make the prediction. Let R_{md} denotes the rating on the d-th film by the m-th user, and let $(m, d) \in \Omega$ be the set of rating we observe. One approach is by matrix factorization: let $U = [u_1, \ldots, u_M]^T \in \mathbb{R}^{M \times K}$ denotes the user matrix and $V = [v_1, \ldots, v_D]^T \in \mathbb{R}^{D \times K}$ denotes the item (film) matrix where $K < \min(M, D)$, then the objective is

$$(U^*, V^*) = \arg\min_{U,V} \sum_{(m,d)\in\Omega} (R_{md} - u_m^T v_d)^2$$

The prediction of the m-th user's rating on a film d' where $(m, d') \notin \Omega$ is simply $R_{md'} = u_m^T v_{d'}$. However, there is no easy way to pick the good K, the rank of UV^T, but we want K to be small. The problem can be reformulated into low-rank matrix completion so we do not need to make any assumption about K. The objective becomes

$$\min_X \text{rank}(X) \quad \text{s.t. } X_{md} = R_{md} \; \forall (m, d) \in \Omega$$

The objective function is non-convex, and we can perform convex relaxation to change the objective from minimizing the rank to minimizing the nuclear norm of X, and then solve it efficiently using Singular Value Thresholding (SVT). Though the above approaches shows great result numerically, they still do not provide a easy interpretation about the user and item (film) vector. For example, there are various generes (topics) for the films: action, adverture, comedy, etc. We believe the entries of the item vector should represent the genere information, and the entries of the user vector should represent user's preference on different generes. However, from the user vector and item (film) vector, we cannot easily figure out what genere each entry of the vector represents. Also, if a new item (film) is added to the set, we cannot make any recommendation (prediction on rating) unless one user has rated it. Thus, our goal of this project is to include another dataset containing movie plot information from Wikipedia and to build a probabilistic model that

- use movie plot data to build the item (film) vector so we learn the genres (topics) information of each film (e.g. a film could be 70% comedy and 30% action) precisely from the item (film) vector.
- and make recommendation (predction on rating) for films that no users has ever rated before.

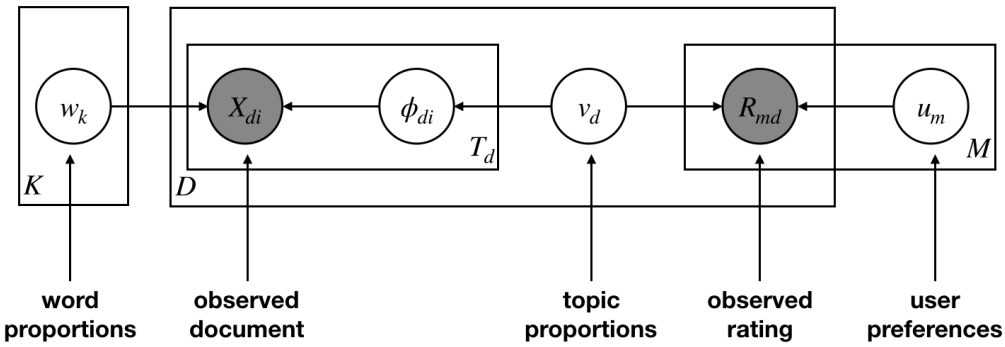
Data

As mentioned in the problem statement, there are two data sources we will use here.

- Netflix Prize Dataset (100,480,507 ratings of 480,189 users on 17,700 films). The dataset is from Kaggle: `https://www.kaggle.com/netflix-inc/netflix-prize-data`
- Wikipedia Movie Plot Dataset (description of 34,886 movie plots from Wikipedia). The dataset is from Kaggle: `https://www.kaggle.com/jrobischon/wikipedia-movie-plots`

Given the constriants on computational resources, we decide to only use a subset of the data: 554 movies released between 1985 and 2005 with 5000 users' in total of 1.5 million ratings (1,2,3,4,5) on those movies. The list of movies includes Independence Day, Pearl Harbor, Anchorman, etc., ranging from various genres. For the 554 movies, we randomly pick 500 movies as the train set, and the other 54 movies with their rating are part of the test set for demonsrating the ability of the model to predict rating for movies with no rating. For the 500 movies in the train set, we randomly keep 75% of the ratings in the train set and put the rest 25% to the test set.

Methodology



In this project, we use Collobrative Topic Modeling, the combination of Matrix Factorization and Topic Modeling, developed by Wang and Blei [3]. In Collaborative Topic Modeling, we have M users' rating on D documents. Let us denote the rating matrix by R where R_{md} is the m-th user's rating on the d-th document. Assuming that the documents $X_1, \ldots, X_D$ are generated from a set of K topics. Each document shares the same set of topics but each with different weight on the K topics. Let v_d be the K-dimensional probability distribution on the topics' weight for the d-th documents. Let u_m be a K-dimensional non-negative vector that represents the m-th user's preference on the K topics. We want R_{md} be closed to $u_m^T v_d$. For each document X_d, it is a T_d-dimensional vector where X_{di} is the i-th word in the document for $i = 1, \ldots, T_d$. Each $X_{di} \in \{1, \ldots, N\}$ where N is the total number of unique words in all D documents. Each topic has different weight on the N words, and let w_k be the N-dimensional probability distribution on the weight of the N words for the k-th topic. Thus, the generative model is

$$
\begin{aligned}
v_{dk} &\sim \text{Gamma}(\alpha, 1) \\
w_k &\sim \text{Dirichlet}([\gamma, \cdots, \gamma]) \\
X_{di} | v_d, w &\sim \sum_{k=1}^{K} \frac{v_{dk}}{\sum_{l=1}^{K} v_{dl}} \text{Categorical}(w_k) \\
u_{mk} &\sim \text{Gamma}(\kappa, \beta) \\
R_{md} | u_m, v_d &\sim \text{Poisson}(u_m^T v_d) \quad \forall (m, d) \in \Omega
\end{aligned}
$$

by noting that for $v_{dk} \sim \text{Gamma}(\alpha, 1)$, then

$$
\left[\begin{array}{ccc} \frac{v_{d1}}{\sum_{l=1}^{K} v_{dl}} & \cdots & \frac{v_{dK}}{\sum_{l=1}^{K} v_{dl}} \end{array} \right] \sim \text{Dirichlet}([\alpha, \cdots, \alpha])
$$

Let $u = \{u_m\}_{m=1}^{M}$, $v = \{v_d\}_{d=1}^{D}$, $w = \{w_k\}_{k=1}^{K}$, and $\theta = \{u, v, w\}$. Let us introduce a latent variable ϕ where ϕ_{di} is a categorical variable that represents which topic the word X_{di} belongs to, where

$$
\phi_{di} \sim \text{Categorical}\left(\frac{v_{d1}}{\sum_{l=1}^{K} v_{dl}}, \cdots, \frac{v_{dK}}{\sum_{l=1}^{K} v_{dl}} \right)
$$

After observing the data $R_{md} = r_{md}$ and $X_{di} = x_{di}$, the joint likelihood including the latent variables is

$$
\begin{aligned}
p(R, X, \theta, \phi) = \prod_{(m,d) \in \Omega} p(r_{md} | u_m, v_d) \times \prod_{m=1}^{M} \prod_{k=1}^{K} p(u_{mk} | \kappa, \beta) \times \prod_{d=1}^{D} \prod_{k=1}^{K} p(v_{dk} | \alpha) \times \\
\prod_{d=1}^{D} \prod_{i=1}^{T_d} p(x_{di} | \phi_{di}, w) \times \prod_{d=1}^{D} \prod_{i=1}^{T_d} p(\phi_{di} | v_d) \times \prod_{k=1}^{K} p(w_k | \gamma)
\end{aligned}
$$

Since Expectation Maximization (EM) only learns the MAP estimator of $\theta = \{u, v, w\}$, we decide to use Variational Inference so we can learn the approximate posterior distribution of θ rather than just a point estimate. The Variational Inference update of the Collaborative Topic Modeling is provided in the Appendix B. However, Wang and Blei also

mentioned that updating the topic weight v_d by the estimate from vanilla Topic Modeling update gave similar performance and saves computation [3]. From our experiment, we even find that using vanilla Topic Modeling to update the topic weight v_d gives results that are more interpretable. This might due to the fact that we have a lot more ratings for each document than the words in each document, causing the imbalance in update step (see update in Appendix for details). Thus, to make the model more flexible, we perform the inference in a two-step approach:

- learn the topic weight v_d and word weight w_k by using Vanilla Topic Modeling.
- learn the user vector u_m given learned topic weight v_d by Poisson Matrix Factorization or Least Sqaure.

We also use Variational Inference to learn $\theta = \{v, w\}$ from the Vanilla Topic Modeling, and the Variational Inference update is also provided in the Appendix A. Another reason that we decide to use Variational Inference is that it can be easily extended to Stochastic Variational Inference by incorporating Natural Gradient Descent and Stochastic Approximation for Conjugate Exponential Family (CEF) models to handle large dataset problem [2], while there is no known closed form update from Stochastic Expectation Maximization for CEF models (we have to rely on stochastic gradient descent, which can be very noisy). From our experiment, we find that Variational Inference gives better result than Stochastic Variational Inference in general where Stochastic Variational Inference suffers from initialization and require careful tuning on batch sizes and learning rate.

Evaluation and Results

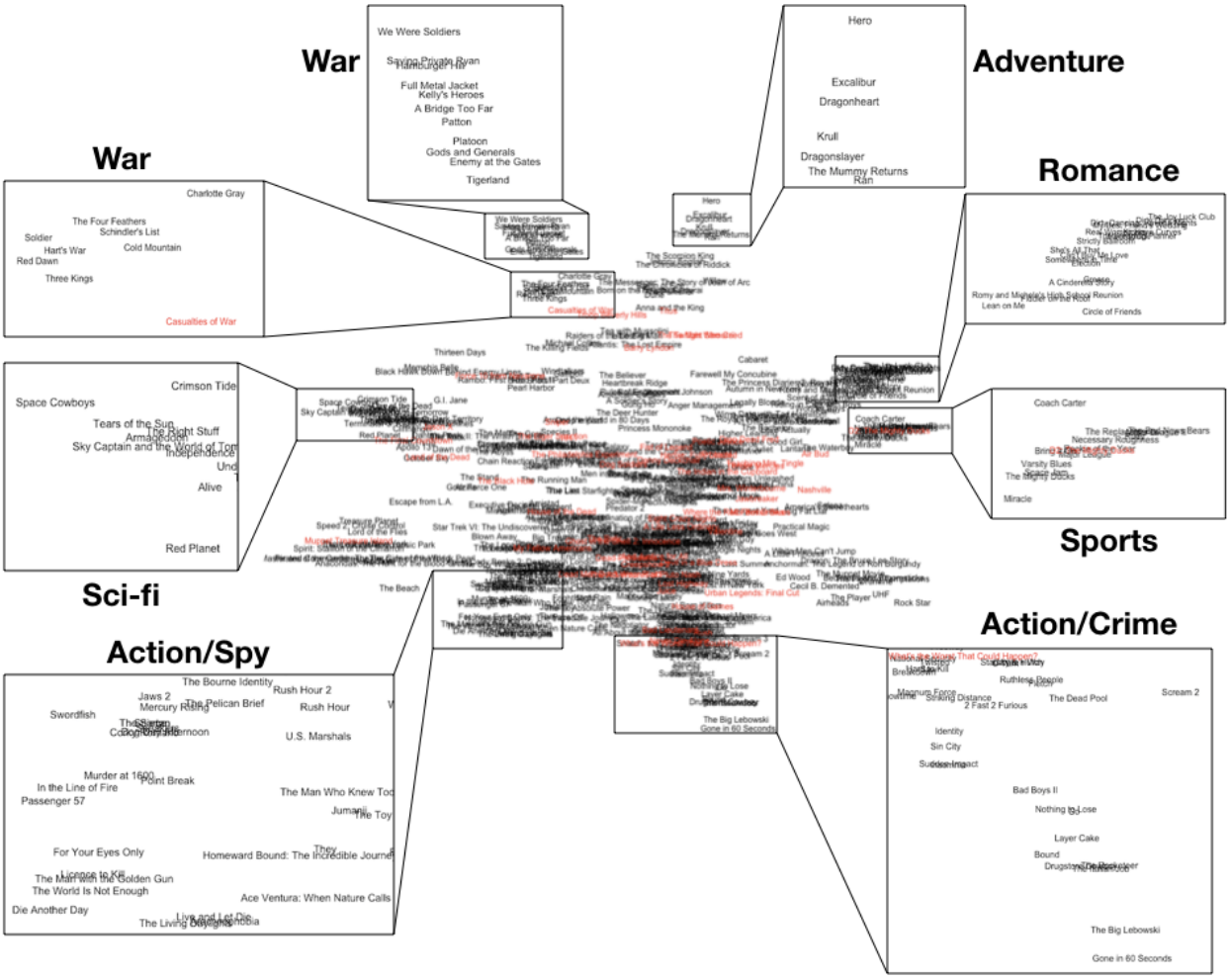
Before we run any experiment, we need to do some natrual langugage processing on the words of the documents such as Lemmatization, deletion of stop words, and etc. We run experiment in $K = 5, 10, 15, 20, 25$. With $K = 25$ (25 topics) gives the result that is most interpretable. In this section, we first present the result from the Vanilla Topic Modeling using posterior mean, and then discuss how to make recommendation.

Topic Modeling Results

Let us first present the posterior topics learned from the 500 movies in the train set. Each topic is a weighted distribution over the words and the following table shows the top 10 frequent words of each topic.

Topic	Top 10 Frequent Words									
1	film	sidney	miss	band	station	studio	star	music	news	spell
2	black	fight	worm	plant	mask	debt	shop	horse	ring	match
3	bond	kill	diamond	escape	agent	spider	jam	steal	destroy	british
4	jesus	truck	christmas	storm	santa	park	death	god	collapse	unit
5	eve	hall	bug	company	skip	clone	world	computer	system	shop
6	alien	earth	superman	planet	happy	ship	nuclear	light	key	infect
7	kill	shoot	police	murder	car	death	gun	escape	killer	house
8	team	game	player	coach	football	ball	field	home	basketball	pitch
9	love	school	home	dance	marry	family	student	party	house	life
10	book	archer	foster	violet	warrior	gun	shoot	gang	judge	knight
11	train	officer	hunt	japanese	colonel	water	worker	general	captain	code
12	wood	shark	enterprise	texas	league	kill	ranch	ranger	river	doctor
13	gang	chinese	escape	china	fight	sing	monk	monkey	governor	royal
14	money	drug	police	car	steal	arrest	rob	kidnap	dealer	escape
15	evil	power	church	robot	angel	death	priest	soul	die	world
16	family	dog	chip	home	house	cat	popeye	animal	toy	voice
17	guard	prison	prisoner	inmate	bishop	escape	cell	rescue	winter	machine
18	mile	race	crew	smoke	court	tiger	shark	battle	forest	divorce
19	ship	boat	island	snake	spirit	crew	captain	rescue	president	dive
20	water	bat	ace	chance	tunnel	creature	attack	shadow	roll	house
21	kill	king	english	army	dragon	castle	sword	attack	lord	battle
22	helicopter	escape	kill	destroy	pilot	mission	team	attack	rocket	rescue
23	bud	judge	firm	court	file	company	office	law	wish	trial
24	soldier	war	kill	american	army	attack	german	nazi	camp	british
25	agent	fbi	bomb	kill	escape	hostage	president	secret	plane	bank

From the above table, we can see that the Topic Modeling is able to identify some important characterstics about the 500 movies in the train set: topic 3 corresponds to series of spy action movies about 007 ("Die Another Day", "The World Is Not Enough", and etc.); topic 6 corresponds to science fiction movies such as "Superman" and "Men in Black"; topic 7 corresponds to action thriller movies such as "Die Hard" and "Sin City"; topic 8 corresponds to sports movies (basball, basketball, football, etc) such as the "Major League" and "Coach Carter"; topic 9 corresponds to romance (love and marriage) movies such as "My Best Friend's Wedding"; topic 14 corresponds to crime thriller movies such as "Gone in 60 Seconds" and "The Italian Job"; topic 21 corresponds to fantasy adventure movies such as Dragonheart and Dragonslayer; topic 22 corresponds to science fiction and adventure movies such as "Red Planet" and "Terminator"; topic 24 corresponds to war movies such as "Saving Private Ryan"; and topic 25 corresponds to action movies related to secret service agency or terrorism such as "Passenger 57" and "Air Force One". Next, let us visualize the movies in the features represented by the topic weights using ISOMAP.



The black text are the 500 movies in the train set, and the red text are the 54 movies that are in the test set where the topic information can be inferred from the learned word weights w_k from the Topic Modeling. From the plot, we can see that we can predict the topic for the movies in the test set accurately as the movie "Casualties of War" is located around the war movies. Moreover, the ISOMAP gives a better images on the connections between the movies. We can see that all action movies are located at the bottom of the plot. The Sci-Fi action movie is on the right side above the action/spy movie. The war movie is on the top left corner. Last, the Romance and Sports movie are located near each other, which is what we expect. In summary, the Topic Modeling does capture important theme/genres information from each movie.

Movie Recommendation

From the Topic Modeling, we obtain the item vectors for all movies. In order to make recommendation, we need to compute the user vectors u_m given the observed rating R_{md} and the item vector v_d. We propose two methods here.

- **Poisson Matrix Factorization.** Same idea in the Collaborative Topic Modeling discussed in the Methodology Section. The objective is to solve

$$\arg\max_{u_1,\ldots,u_M} \sum_{(m,d)\in\Omega} \log p(r_{m,d}|u_m) + \sum_{m=1}^{M}\sum_{k=1}^{K} \log p(u_{mk}|\kappa,\beta)$$

  The prior is $u_{mk} \sim \text{Gamma}(\kappa,\beta)$ and the likelihood is $R_{md} \sim \text{Poisson}(u_m^T v_d)$ where v_d is the posterior mean from the Topic Modeling. We can solve the above problem using Expectation Maximization (EM), and the update for u_m is the same EM update of u_m in Collaborative Topic Modeling provided in the Appendix B.

- **Ridge Least Square.** Same idea in the Least Square approach in matrix completition. The objective is to solve

$$\arg\min_{u_m} \sum_{(m,d)\in\Omega} (R_{md} - u_m^T v_d)^2 + \sum_{m=1}^{M} \lambda_m ||u_m||_2^2$$

  where v_d is the posterior mean from the Topic Modeling. The Ridge Penalty term here is to avoid multi-collinearity and singular matrix issue. We are allowing the Ridge penalty to be different for each u_m, and we use Leave One Out Cross Validation to pick the best λ_m for each u_m.

The two methods above yield the user vectors u_m. To make prediction on the m^*-th user's rating on a film d' where $(m^*, d') \notin \Omega$ is simply $u_{m^*}^T v_{d'}$. As a comparison, we also look into the **Item-based Recommendation** since we have the item vector to compute for similarities. Define the similarity metrics between movie d and d' to be

$$\text{sim}(d, d') = \exp\{-\lambda d(v_d, v_{d'})^2\}$$

where $d(v_d, v_{d'})$ is the l_2 norm of the topic weights of the two movies. To make prediction on the m^*-th user's rating on a film d' where $(m^*, d') \notin \Omega$, we need to compute

$$\frac{\sum_{(m^*,d)\in\Omega} \text{sim}(d, d') R_{m^* d}}{\sum_{(m^*,d)\in\Omega} \text{sim}(d, d')}$$

We can see that the above is in the form of Kernel Regression, and we use Leave One Out Cross Validation to pick the best λ for the similarity metric. Here is the comparison between the 3 methods on predicting the rating in the test. There are two metrics for evaluating the performance: mean square error and Poisson log likelihood (not including the log factorial term, which is a constant).

	Poisson log likelihood			Mean Square Error		
Methods	Movie-Train	Movie-Test	Overall	Movie-Train	Movie-Test	Overall
Poisson Matrix Factorization	0.780	0.477	0.750	0.953	0.906	0.948
Ridge Least Square	0.769	0.472	0.740	0.992	0.935	0.986
Item-Based Recommendation	0.790	0.471	0.759	0.884	0.941	0.890

where the "Movie-Train" is the 500 movies that we use in Topic Modeling training, and the "Movie-Test" is the rest 54 movies in the test set where the topic weights are inferred using the learned word weights vector w_k from Topic Modeling result. From the above results, we can see that Item-Based Recommendation's performance is better than both Poisson Matrix Factorization and Ridge Least Square. The reason is that both Poisson Matrix Factorization and Ridge Least Square approach only capture linear relationship between the rating and the item vector, while the Item-Based Recommendation (Kernel Regression) can also capture nonlinear relationship. Moreover, from the result on the "Movie-test", we can see that the model can also make fairly good prediction on movies that does not have any rating. This is particular useful when a new item (movie) is added to the recommender system since we can now make recommendation for this new item to the users right away.

Conclusion

In summary, we build a probabilistic movie recommender system that

- learn movie genres (topic) information from movie plot data using Topic Modeling

- and propose three different approaches to predict rating for recommendation purpose

The advantage of our model is that it can precisely identify the genres (topics) information for each movie and have the ability to make recommendation for items that no users has ever rated before, where the traditional matrix completion approach cannot achieve.

References

[1] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[2] Matthew D Hoffman et al. "Stochastic variational inference". In: *The Journal of Machine Learning Research* 14.1 (2013), pp. 1303–1347.

[3] Chong Wang and David M Blei. "Collaborative topic modeling for recommending scientific articles". In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2011, pp. 448–456.

Appendix A: Topic Modeling Inference

The generative model is

$$\begin{aligned} v_d &\sim \text{Dirichlet}([\alpha, \cdots, \alpha]) \\ w_k &\sim \text{Dirichlet}([\gamma, \cdots, \gamma]) \\ X_{di}|v_d, w &\sim \sum_{k=1}^{K} v_{dk}\text{Categorical}(w_k) \end{aligned}$$

Let $v = \{v_d\}_{d=1}^{D}$, $w = \{w_k\}_{k=1}^{K}$, and $\theta = \{v, w\}$. Let us introduce a latent variable ϕ where ϕ_{di} is a categorical variable that represents which topic the word X_{di} belongs to, where $\phi_{di} \sim \text{Categorical}(v_d)$. After observing the data $X_{di} = x_{di}$, the joint likelihood including the latent variables is

$$p(X, \theta, \phi) = \prod_{d=1}^{D}\prod_{i=1}^{T_d} p(x_{di}|\phi_{di}, w) \times \prod_{d=1}^{D}\prod_{i=1}^{T_d} p(\phi_{di}|v_d) \times \prod_{d=1}^{D}\prod_{k=1}^{K} p(v_{dk}|\alpha) \times \prod_{k=1}^{K} p(w_k|\gamma)$$

The log joint likelihood function is

$$\begin{aligned} \ln p(X, \theta, \phi) = \text{const.} + \sum_{d=1}^{D}\sum_{i=1}^{T_d}\sum_{k=1}^{K}\sum_{n=1}^{N}\Big(\mathbb{1}(x_{di}=n)\mathbb{1}(\phi_{di}=k)\ln w_{kn}\Big) + \sum_{d=1}^{D}\sum_{i=1}^{T_d}\sum_{k=1}^{K}\Big(\mathbb{1}(\phi_{di}=k)\ln v_{dk}\Big) + \\ \sum_{d=1}^{D}\sum_{k=1}^{K}\Big((\alpha-1)\ln v_{dk}\Big) + \sum_{k=1}^{K}\sum_{n=1}^{N}\Big((\gamma-1)\ln w_{kn}\Big) \end{aligned}$$

Expectation Maximization

E-step. Compute the expected "complete data" log likelihood on $q_t(\phi) = p(\phi|R, X, \theta_{t-1})$.

$$\begin{aligned} Q(\theta|\theta_{t-1}) = \text{const.} + \sum_{d=1}^{D}\sum_{i=1}^{T_d}\sum_{k=1}^{K}\sum_{n=1}^{N}\Big(\mathbb{1}(x_{di}=n)q_{dik}\ln w_{kn}\Big) + \sum_{d=1}^{D}\sum_{i=1}^{T_d}\sum_{k=1}^{K}\Big(q_{dik}\ln v_{dk}\Big) + \\ \sum_{d=1}^{D}\sum_{k=1}^{K}\Big((\alpha-1)\ln v_{dk}\Big) + \sum_{k=1}^{K}\sum_{n=1}^{N}\Big((\gamma-1)\ln w_{kn}\Big) \end{aligned}$$

where

$$q_{dik} = \mathbb{E}_{p(\phi|R,X,\theta_{t-1})}[\mathbb{1}(\phi_{di}=k)] = \frac{v_{dk}\prod_{n=1}^{N} w_{kn}^{\mathbb{1}(x_{di}=n)}}{\sum_{l=1}^{K} v_{dl}\prod_{n=1}^{N} w_{ln}^{\mathbb{1}(x_{di}=n)}}$$

M-step. Solve for $\theta_t = \arg\max_\theta Q(\theta|\theta_{t-1})$.

- Update for v_{dk}. Given the constriant $\sum_{k=1}^{K} v_{dk} = 1$, let us use Lagrange multiplier,

$$L(\lambda, v_{d1}, \ldots, v_{dK}) = \text{const.} + \sum_{i=1}^{T_d}\sum_{k=1}^{K}\Big(q_{dik}\ln v_{dk}\Big) + \sum_{k=1}^{K}\Big((\alpha-1)\ln v_{dk}\Big) + \lambda\Big(1 - \sum_{k=1}^{K} v_{dk}\Big)$$

  Thus,

$$v_{dk} = \frac{\sum_{i=1}^{T_d} q_{dik} + (\alpha-1)}{\sum_{k=1}^{K}\sum_{i=1}^{T_d} q_{dik} + n(\alpha-1)} = \frac{\sum_{i=1}^{T_d} q_{dik} + (\alpha-1)}{T_d + n(\alpha-1)}$$

- Update for w_{kn}. Given the constriant $\sum_{n=1}^{N} w_{kn} = 1$, let us use Lagrange multiplier,

$$L(\lambda, w_{k1}, \ldots, w_{kN}) = \text{const.} + \sum_{d=1}^{D}\sum_{i=1}^{T_d}\sum_{n=1}^{N}\Big(\mathbb{1}(x_{di}=n) q_{dik}\ln w_{kn}\Big) + \sum_{n=1}^{N}\Big((\gamma-1)\ln w_{kn}\Big) + \lambda\Big(1 - \sum_{n=1}^{N} w_{kn}\Big)$$

  Thus,

$$w_{kn} = \frac{\sum_{d=1}^{D}\sum_{i=1}^{T_d}\mathbb{1}(x_{di}=n) q_{dik} + (\gamma-1)}{\sum_{d=1}^{D}\sum_{i=1}^{T_d} q_{dik} + n(\gamma-1)}$$

Variational Inference

In Variational Inference, we use some distribution $q(\theta)$ to approximate the posterior distribution $p(\theta|X)$ that minimizes the Kullback-Leibler (KL) divergence between $q(\theta)$ and $p(\theta|X)$. For this problem, consider $\theta = \{v, w, \phi\}$, which the latent variable are included. In Variational Inference framework, we solve the following optimization problem,

$$\arg\max_{q} \mathcal{L}(q) = \mathbb{E}_q[\ln p(X,\theta)] - \mathbb{E}_q[\ln q(\theta)]$$

where $\mathcal{L}(q)$ is the evidence lower bound (ELOB) and $\ln p(X) = \mathcal{L}(q) + KL(q(\theta)||p(\theta|X))$, so maximizing ELOB is equivalent to minimizing the KL divergence between $q(\theta)$ and the posterior $p(\theta|X)$. Using mean-field assumption, we model all parameters in θ to be independent, which

$$q(\theta) = \prod_{d=1}^{D} q(v_d) \times \prod_{k=1}^{K} q(w_k) \times \prod_{d=1}^{D}\prod_{i=1}^{T_d} q(\phi_{di})$$

where

$$\begin{aligned} q(v_d) &\sim \text{Dirichlet}([o_{d1}, \cdots, o_{dK}]) \\ q(w_k) &\sim \text{Dirichlet}([e_{k1}, \cdots, e_{kN}]) \\ q(\phi_{di}) &\sim \text{Categorical}([z_{di1}, \ldots, z_{diK}]) \end{aligned}$$

Thus, it follows that

$$\begin{aligned} \mathbb{E}_q[\ln p(X,\theta)] = \text{const.} &+ \sum_{d=1}^{D}\sum_{i=1}^{T_d}\sum_{k=1}^{K}\sum_{n=1}^{N}\Big(\mathbb{1}(x_{di}=n)\mathbb{E}_q[\mathbb{1}(\phi_{di}=k)]\mathbb{E}_q[\ln w_{kn}]\Big) + \\ &\sum_{d=1}^{D}\sum_{i=1}^{T_d}\sum_{k=1}^{K}\Big(\mathbb{E}_q[\mathbb{1}(\phi_{di}=k)]\mathbb{E}_q[\ln v_{dk}]\Big) + \sum_{d=1}^{D}\sum_{k=1}^{K}\Big((\alpha-1)\mathbb{E}_q[\ln v_{dk}]\Big) + \sum_{k=1}^{K}\sum_{n=1}^{N}\Big((\gamma-1)\mathbb{E}_q[\ln w_{kn}]\Big) \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}_q[\ln q(\theta)] = \text{const.} &+ \sum_{d=1}^{D}\Big(\ln\Gamma\Big(\sum_{k=1}^{K} o_{dk}\Big) - \sum_{k=1}^{K}\ln\Gamma(o_{dk}) + \sum_{k=1}^{K}(o_{dk}-1)\mathbb{E}_q[\ln v_{dk}]\Big) + \\ &\sum_{k=1}^{K}\Big(\ln\Gamma\Big(\sum_{n=1}^{N} e_{kn}\Big) - \sum_{n=1}^{N}\ln\Gamma(e_{kn}) + \sum_{n=1}^{N}(e_{kn}-1)\mathbb{E}_q[\ln w_{kn}]\Big) + \\ &\sum_{d=1}^{D}\sum_{i=1}^{T_d}\sum_{k=1}^{K}\Big(\mathbb{E}_q[\mathbb{1}(\phi_{di}=k)]\ln z_{dik}\Big) \end{aligned}$$

Our goal is to find $q^*(\theta)$ such that the lower bound of the ELOB is maximized, which

$$q^*(\theta) = \arg\max_{q} \tilde{\mathcal{L}}(q) = \mathbb{E}_q[\ln \tilde{p}(X,\theta)] - \mathbb{E}_q[\ln q(\theta)]$$

Here is the coordinate ascent update.

- $v_d \sim \text{Dirichlet}([o_{d1}, \cdots, o_{dK}])$,

$$\begin{aligned}\ln q^*(v_d) &= \sum_{i=1}^{T_d}\sum_{k=1}^{K}\Big(\mathbb{E}_q[\mathbb{1}(\phi_{di}=k)]\ln v_{dk}\Big) + \sum_{k=1}^{K}\Big((\alpha-1)\ln v_{dk}\Big) + \text{ const.}\\ &= \sum_{k=1}^{K}\Big(\alpha + \sum_{i=1}^{T_d}\mathbb{E}_q[\mathbb{1}(\phi_{di}=k)] - 1\Big)\ln v_{dk} + \text{ const.}\end{aligned}$$

  Thus, $o_{dk} = \alpha + \sum_{i=1}^{T_d}\mathbb{E}_q[\mathbb{1}(\phi_{di}=k)]$.

- $w_k \sim \text{Dirichlet}([e_{k1}, \ldots, e_{kN}])$,

$$\begin{aligned}\ln q^*(w_k) &= \sum_{d=1}^{D}\sum_{i=1}^{T_d}\sum_{n=1}^{N}\Big(\mathbb{1}(x_{di}=n)\mathbb{E}_q[\mathbb{1}(\phi_{di}=k)]\ln w_{kn}\Big) + \sum_{n=1}^{N}\Big((\gamma-1)\ln w_{kn}\Big) + \text{ const.}\\ &= \sum_{n=1}^{N}\Big(\gamma + \sum_{d=1}^{D}\sum_{i=1}^{T_d}\mathbb{1}(x_{di}=n)\mathbb{E}_q[\mathbb{1}(\phi_{di}=k)] - 1\Big)\ln w_{kn} + \text{ const.}\end{aligned}$$

  Thus, $e_{kn} = \gamma + \sum_{d=1}^{D}\sum_{i=1}^{T_d}\mathbb{1}(x_{di}=n)\mathbb{E}_q[\mathbb{1}(\phi_{di}=k)]$.

- $\phi_{di} \sim \text{Categorical}([z_{di1}, \ldots, z_{diK}])$,

$$\begin{aligned}\ln q^*(\phi_{di}) &= \sum_{k=1}^{K}\sum_{n=1}^{N}\Big(\mathbb{1}(x_{di}=n)\mathbb{1}(\phi_{di}=k)\mathbb{E}_q[\ln w_{kn}]\Big) + \sum_{k=1}^{K}\mathbb{1}(\phi_{di}=k)]\mathbb{E}_q[\ln v_{dk}] + \text{ const.}\\ &= \sum_{k=1}^{K}\Big(\sum_{n=1}^{N}\mathbb{1}(x_{di}=n)\mathbb{E}_q[\ln w_{kn}] + \mathbb{E}_q[\ln v_{dk}]\Big)\mathbb{1}(\phi_{di}=k) + \text{ const.}\end{aligned}$$

  Thus, $z_{dik} = \sum_{n=1}^{N}\mathbb{1}(x_{di}=n)\mathbb{E}_q[\ln w_{kn}] + \mathbb{E}_q[\ln v_{dk}]$.

where

$$\begin{aligned}\mathbb{E}_q[\ln v_{dk}] &= \psi(o_{dk}) - \psi\Big(\sum_{k=1}^{K} o_{dk}\Big)\\ \mathbb{E}_q[\ln w_{kn}] &= \psi(e_{kn}) - \psi\Big(\sum_{n=1}^{N} e_{kn}\Big)\\ \mathbb{E}_q[\mathbb{1}(\phi_{di}=k)] &= z_{dik}\end{aligned}$$

where $\psi(x) = \frac{d\log\Gamma(x)}{dx}$ is the diagamma function.

Appendix B: Collaborative Topic Modeling Inference

The log joint likelihood function is

$$\begin{aligned}\ln p(R, X, \theta, \phi) = \text{const.} &+ \sum_{m=1}^{M}\sum_{d=1}^{D}\Big(r_{md}\ln\sum_{k=1}^{K}u_{mk}v_{dk} - \sum_{k=1}^{K}u_{mk}v_{dk}\Big) +\\ &\sum_{m=1}^{M}\sum_{k=1}^{K}\Big((\kappa-1)\ln u_{mk} - \beta u_{mk}\Big) + \sum_{d=1}^{D}\sum_{k=1}^{K}\Big((\alpha-1)\ln v_{dk} - v_{dk}\Big) +\\ &\sum_{d=1}^{D}\sum_{i=1}^{T_d}\sum_{k=1}^{K}\sum_{n=1}^{N}\Big(\mathbb{1}(x_{di}=n)\mathbb{1}(\phi_{di}=k)\ln w_{kn}\Big) +\\ &\sum_{d=1}^{D}\sum_{i=1}^{T_d}\sum_{k=1}^{K}\mathbb{1}(\phi_{di}=k)\Big(\ln v_{dk} - \ln\sum_{l=1}^{K}v_{dl}\Big) + \sum_{k=1}^{K}\sum_{n=1}^{N}\Big((\gamma-1)\ln w_{kn}\Big)\end{aligned}$$

Expectation Maximization

We use Expectation Maximization to obtain the MAP for θ. By looking at the log joint likelihood function, there are two terms with summation inside the log function, causing trouble of getting the closed form update in the M-step. Thus, we replace the two problematic terms by their lower bound.

- $\ln \sum_{k=1}^{K} u_{mk} v_{dk}$. Let us introduce a K-dimensional vector ρ_{md} such that $\sum_{k=1}^{K} \rho_{mdk} = 1$ and $\rho_{mdk} > 0$. Since $\ln(\cdot)$ is concave and by Jensen's inequality,

$$\ln \sum_{k=1}^{K} u_{mk} v_{dk} = \ln \sum_{k=1}^{K} \rho_{mdk} \frac{u_{mk} v_{dk}}{\rho_{mdk}} \geq \sum_{k=1}^{K} \rho_{mdk} \ln \frac{u_{mk} v_{dk}}{\rho_{mdk}} = \sum_{k=1}^{K} \rho_{mdk} (\ln u_{mk} + \ln v_{dk}) - \sum_{k=1}^{K} \rho_{mdk} \ln \rho_{mdk}$$

  where the tightest bound is obtained when $\rho_{mdk} = (u_{mk} v_{dk}) / (\sum_{l=1}^{K} u_{ml} v_{dl})$.

- $-\ln \sum_{l=1}^{K} v_{dl}$. Since $-\ln(\cdot)$ is convex and by the first order Taylor expansion on $\xi_d > 0$,

$$-\ln \sum_{l=1}^{K} v_{dl} \geq -\ln \xi_d - \frac{\sum_{l=1}^{K} v_{dl} - \xi_d}{\xi_d}$$

  where the tightest bound is obtained when $\xi_d = \sum_{l=1}^{K} v_{dl}$.

Replacing the two terms by their lower bounds yields the following lower bound of the log joint likelihood function that we want to maximize,

$$\begin{aligned}
\ln \tilde{p}(R, X, \theta, \phi) = \text{const.} + & \sum_{m=1}^{M} \sum_{d=1}^{D} \sum_{k=1}^{K} \Big(r_{md} \rho_{mdk} (\ln u_{mk} + \ln v_{dk}) - r_{md} \rho_{mdk} \ln \rho_{mdk} - u_{mk} v_{dk} \Big) + \\
& \sum_{m=1}^{M} \sum_{k=1}^{K} \Big((\kappa - 1) \ln u_{mk} - \beta u_{mk} \Big) + \sum_{d=1}^{D} \sum_{k=1}^{K} \Big((\alpha - 1) \ln v_{dk} - v_{dk} \Big) + \\
& \sum_{d=1}^{D} \sum_{i=1}^{T_d} \sum_{k=1}^{K} \sum_{n=1}^{N} \Big(\mathbb{1}(x_{di} = n) \mathbb{1}(\phi_{di} = k) \ln w_{kn} \Big) + \\
& \sum_{d=1}^{D} \sum_{i=1}^{T_d} \sum_{k=1}^{K} \mathbb{1}(\phi_{di} = k) \Big(\ln v_{dk} - \ln \xi_d - \frac{\sum_{l=1}^{K} v_{dl} - \xi_d}{\xi_d} \Big) + \sum_{k=1}^{K} \sum_{n=1}^{N} \Big((\gamma - 1) \ln w_{kn} \Big)
\end{aligned}$$

E-step. Compute the expected "complete data" log likelihood on $q_t(\phi) = p(\phi | R, X, \theta_{t-1})$.

$$\begin{aligned}
Q(\theta | \theta_{t-1}) = \text{const.} + & \sum_{m=1}^{M} \sum_{d=1}^{D} \sum_{k=1}^{K} \Big(r_{md} \rho_{mdk} (\ln u_{mk} + \ln v_{dk}) - r_{md} \rho_{mdk} \ln \rho_{mdk} - u_{mk} v_{dk} \Big) + \\
& \sum_{m=1}^{M} \sum_{k=1}^{K} \Big((\kappa - 1) \ln u_{mk} - \beta u_{mk} \Big) + \sum_{d=1}^{D} \sum_{k=1}^{K} \Big((\alpha - 1) \ln v_{dk} - v_{dk} \Big) + \\
& \sum_{d=1}^{D} \sum_{i=1}^{T_d} \sum_{k=1}^{K} \sum_{n=1}^{N} \Big(\mathbb{1}(x_{di} = n) q_{dik} \ln w_{kn} \Big) + \\
& \sum_{d=1}^{D} \sum_{i=1}^{T_d} \sum_{k=1}^{K} q_{dik} \ln v_{dk} - \sum_{d=1}^{D} T_d \Big(\ln \xi_d + \frac{\sum_{k=1}^{K} v_{dk} - \xi_d}{\xi_d} \Big) + \sum_{k=1}^{K} \sum_{n=1}^{N} \Big((\gamma - 1) \ln w_{kn} \Big)
\end{aligned}$$

where

$$q_{dik} = \mathbb{E}_{p(\phi | R, X, \theta_{t-1})} [\mathbb{1}(\phi_{di} = k)] = \frac{v_{dk} \prod_{n=1}^{N} w_{kn}^{\mathbb{1}(x_{di} = n)}}{\sum_{l=1}^{K} v_{dl} \prod_{n=1}^{N} w_{ln}^{\mathbb{1}(x_{di} = n)}}$$

M-step. Solve for $\theta_t = \arg\max_{\theta} Q(\theta | \theta_{t-1})$.

- Update for u_{mk}. Take the derivative of $Q(\theta|\theta_{t-1})$ with respect to u_{mk},

$$\frac{\partial Q}{\partial u_{mk}} = \sum_{d=1}^{D}\left(\frac{r_{md}\rho_{mdk}}{u_{mk}} - v_{dk}\right) + \frac{\kappa - 1}{u_{mk}} - \beta = 0$$

Thus,

$$u_{mk} = \frac{\sum_{d=1}^{D} r_{md}\rho_{mdk} + \kappa - 1}{\sum_{d=1}^{D} v_{dk} + \beta}$$

- Update for v_{dk}. Take the derivative of $Q(\theta|\theta_{t-1})$ with respect to v_{dk},

$$\frac{\partial Q}{\partial v_{dk}} = \sum_{m=1}^{M}\left(\frac{r_{md}\rho_{mdk}}{v_{dk}} - u_{mk}\right) + \frac{\alpha - 1}{v_{dk}} - 1 + \sum_{i=1}^{T_d}\frac{q_{dik}}{v_{dk}} - \frac{T_d}{\xi_d} = 0$$

Thus,

$$v_{dk} = \frac{\sum_{m=1}^{M} r_{md}\rho_{mdk} + \sum_{i=1}^{T_d} q_{dik} + \alpha - 1}{\sum_{m=1}^{M} u_{mk} + T_d/\xi_d + 1}$$

- Update for w_{kn}. Given the constriant $\sum_{n=1}^{N} w_{kn} = 1$, let us use Lagrange multiplier,

$$L(\lambda, w_{k1}, \ldots, w_{kN}) = \text{const.} + \sum_{d=1}^{D}\sum_{i=1}^{T_d}\sum_{n=1}^{N}\left(\mathbb{1}(x_{di} = n) q_{dik} \ln w_{kn}\right) + \sum_{n=1}^{N}\left((\gamma - 1)\ln w_{kn}\right) + \lambda\left(1 - \sum_{n=1}^{N} w_{kn}\right)$$

Thus,

$$w_{kn} = \frac{\sum_{d=1}^{D}\sum_{i=1}^{T_d}\mathbb{1}(x_{di} = n) q_{dik} + (\gamma - 1)}{\sum_{d=1}^{D}\sum_{i=1}^{T_d} q_{dik} + n(\gamma - 1)}$$

- Update for ρ_{mdk}. To obtain the tightest lower bound, $\rho_{mdk} = (u_{mk}v_{dk})/(\sum_{l=1}^{K} u_{ml}v_{dl})$
- Update for ξ_d. To obtain the tightest lower bound, $\xi_d = \sum_{k=1}^{K} v_{dk}$

Variational Inference

In Variational Inference, we use some distribution $q(\theta)$ to approximate the posterior distribution $p(\theta|R, X)$ that minimizes the Kullback-Leibler (KL) divergence between $q(\theta)$ and $p(\theta|R, X)$. For this problem, consider $\theta = \{u, v, w, \phi\}$, which the latent variable are included. In Variational Inference framework, we solve the following optimization problem,

$$\arg\max_{q} \mathcal{L}(q) = \mathbb{E}_q[\ln p(R, X, \theta)] - \mathbb{E}_q[\ln q(\theta)]$$

where $\mathcal{L}(q)$ is the evidence lower bound (ELOB) and $\ln p(X) = \mathcal{L}(q) + KL(q(\theta)||p(\theta|R, X))$, so maximizing ELOB is equivalent to minimizing the KL divergence between $q(\theta)$ and the posterior $p(\theta|R, X)$. Using mean-field assumption, we model all parameters in θ to be independent, which

$$q(\theta) = \prod_{m=1}^{M}\prod_{k=1}^{K} q(u_{mk}) \times \prod_{d=1}^{D}\prod_{k=1}^{K} q(v_{dk}) \times \prod_{k=1}^{K} q(w_k) \times \prod_{d=1}^{D}\prod_{i=1}^{T_d} q(\phi_{di})$$

where

$$\begin{aligned} q(u_{mk}) &\sim \text{Gamma}(a_{mk}, b_{mk}) \\ q(v_{dk}) &\sim \text{Gamma}(c_{dk}, o_{dk}) \\ q(w_k) &\sim \text{Dirichlet}([e_{k1}, \cdots, e_{kN}]) \\ q(\phi_{di}) &\sim \text{Categorical}([z_{di1}, \ldots, z_{diK}]) \end{aligned}$$

Thus, it follows that

$$
\begin{aligned}
&\mathbb{E}_q[\ln p(R, X, \theta)] \\
&= \text{const.} + \sum_{m=1}^{M}\sum_{d=1}^{D}\left(r_{md}\mathbb{E}_q\left[\ln\sum_{k=1}^{K}u_{mk}v_{dk}\right] - \sum_{k=1}^{K}\mathbb{E}_q[u_{mk}]\mathbb{E}_q[v_{dk}]\right) + \\
&\sum_{m=1}^{M}\sum_{k=1}^{K}\left((\kappa-1)\mathbb{E}_q[\ln u_{mk}] - \beta\mathbb{E}_q[u_{mk}]\right) + \sum_{d=1}^{D}\sum_{k=1}^{K}\left((\alpha-1)\mathbb{E}_q[\ln v_{dk}] - \mathbb{E}_q[v_{dk}]\right) + \\
&\sum_{d=1}^{D}\sum_{i=1}^{T_d}\sum_{k=1}^{K}\sum_{n=1}^{N}\left(\mathbb{1}(x_{di}=n)\mathbb{E}_q[\mathbb{1}(\phi_{di}=k)]\mathbb{E}_q[\ln w_{kn}]\right) + \\
&\sum_{d=1}^{D}\sum_{i=1}^{T_d}\sum_{k=1}^{K}\mathbb{E}_q[\mathbb{1}(\phi_{di}=k)]\left(\mathbb{E}_q[\ln v_{dk}] - \mathbb{E}_q\left[\ln\sum_{l=1}^{K}v_{dl}\right]\right) + \sum_{k=1}^{K}\sum_{n=1}^{N}\left((\gamma-1)\mathbb{E}_q[\ln w_{kn}]\right) \\
&\geq \text{const.} + \sum_{m=1}^{M}\sum_{d=1}^{D}\sum_{k=1}^{K}\left(r_{md}\rho_{mdk}(\mathbb{E}_q[\ln u_{mk}] + \mathbb{E}_q[\ln v_{dk}]) - r_{md}\rho_{mdk}\ln\rho_{mdk} - \mathbb{E}_q[u_{mk}]\mathbb{E}_q[v_{dk}]\right) + \\
&\sum_{m=1}^{M}\sum_{k=1}^{K}\left((\kappa-1)\mathbb{E}_q[\ln u_{mk}] - \beta\mathbb{E}_q[u_{mk}]\right) + \sum_{d=1}^{D}\sum_{k=1}^{K}\left((\alpha-1)\mathbb{E}_q[\ln v_{dk}] - \mathbb{E}_q[v_{dk}]\right) + \\
&\sum_{d=1}^{D}\sum_{i=1}^{T_d}\sum_{k=1}^{K}\sum_{n=1}^{N}\left(\mathbb{1}(x_{di}=n)\mathbb{E}_q[\mathbb{1}(\phi_{di}=k)]\mathbb{E}_q[\ln w_{kn}]\right) + \\
&\sum_{d=1}^{D}\sum_{i=1}^{T_d}\sum_{k=1}^{K}\left(\mathbb{E}_q[\mathbb{1}(\phi_{di}=k)]\mathbb{E}_q[\ln v_{dk}]\right) - \sum_{d=1}^{D}T_d\left(\ln\xi_d + \frac{\sum_{k=1}^{K}\mathbb{E}_q[v_{dk}] - \xi_d}{\xi_d}\right) + \sum_{k=1}^{K}\sum_{n=1}^{N}\left((\gamma-1)\mathbb{E}_q[\ln w_{kn}]\right) \\
&= \mathbb{E}_q[\ln\tilde{p}(R, X, \theta)]
\end{aligned}
$$

where we replace $\mathbb{E}_q[\ln\sum_{k=1}^{K}u_{mk}v_{dk}]$ and $-\mathbb{E}_q[\ln\sum_{l=1}^{K}v_{dl}]$ by their lower bound using the idea that we use in Expectation Maximization. On the other hand, we also have

$$
\begin{aligned}
\mathbb{E}_q[\ln q(\theta)] = \text{const.} &+ \sum_{m=1}^{M}\sum_{k=1}^{K}\left(a_{mk}\ln b_{mk} - \ln\Gamma(a_{mk}) + (a_{mk}-1)\mathbb{E}_q[\ln u_{mk}] - b_{mk}\mathbb{E}_q[u_{mk}]\right) + \\
&\sum_{d=1}^{D}\sum_{k=1}^{K}\left(c_{dk}\ln o_{dk} - \ln\Gamma(o_{dk}) + (c_{dk}-1)\mathbb{E}_q[\ln v_{dk}] - o_{dk}\mathbb{E}_q[v_{dk}]\right) + \\
&\sum_{k=1}^{K}\left(\ln\Gamma\left(\sum_{n=1}^{N}e_{kn}\right) - \sum_{n=1}^{N}\ln\Gamma(e_{kn}) + \sum_{n=1}^{N}(e_{kn}-1)\mathbb{E}_q[\ln w_{kn}]\right) + \\
&\sum_{d=1}^{D}\sum_{i=1}^{T_d}\sum_{k=1}^{K}\left(\mathbb{E}_q[\mathbb{1}(\phi_{di}=k)]\ln z_{dik}\right)
\end{aligned}
$$

Our goal is to find $q^*(\theta)$ such that the lower bound of the ELOB is maximized, which

$$
q^*(\theta) = \arg\max_q \tilde{\mathcal{L}}(q) = \mathbb{E}_q[\ln\tilde{p}(R, X, \theta)] - \mathbb{E}_q[\ln q(\theta)]
$$

Here is the coordinate ascent update.

- $u_{mk} \sim \text{Gamma}(a_{mk}, b_{mk})$,

$$
\begin{aligned}
\ln q^*(u_{mk}) &= \sum_{d=1}^{D}r_{md}\rho_{mdk}\ln u_{mk} - \sum_{d=1}^{D}\mathbb{E}_q[v_{dk}]u_{mk} + (\kappa-1)\ln u_{mk} - \beta u_{mk} + \text{ const.} \\
&= \left(\kappa + \sum_{d=1}^{D}r_{md}\rho_{mdk} - 1\right)\ln u_{mk} - \left(\beta + \sum_{d=1}^{D}\mathbb{E}_q[v_{dk}]\right)u_{mk} + \text{ const.}
\end{aligned}
$$

  Thus, $a_{mk} = \kappa + \sum_{d=1}^{D}r_{md}\rho_{mdk}$ and $b_{mk} = \beta + \sum_{d=1}^{D}\mathbb{E}_q[v_{dk}]$.

- $v_{dk} \sim \text{Gamma}(c_{dk}, o_{dk})$,

$$\begin{aligned}\ln q^*(v_{dk}) = &\sum_{m=1}^{M} r_{md}\rho_{mdk} \ln v_{dk} - \sum_{m=1}^{M} \mathbb{E}_q[u_{mk}]v_{dk} + (\alpha - 1)\ln v_{dk} - v_{dk} + \\ &\sum_{i=1}^{T_d} \mathbb{E}_q[\mathbb{1}(\phi_{di} = k)] \ln v_{dk} - \frac{T_d}{\xi_d} v_{dk} + \text{ const.} \\ = &\left(\alpha + \sum_{m=1}^{M} r_{md}\rho_{mdk} + \sum_{i=1}^{T_d} \mathbb{E}_q[\mathbb{1}(\phi_{di} = k)] - 1\right) \ln v_{dk} - \\ &\left(1 + \sum_{m=1}^{M} \mathbb{E}_q[u_{mk}] + T_d/\xi_d\right) v_{dk} + \text{ const.}\end{aligned}$$

  Thus, $c_{dk} = \alpha + \sum_{m=1}^{M} r_{md}\rho_{mdk} + \sum_{i=1}^{T_d} \mathbb{E}_q[\mathbb{1}(\phi_{di} = k)]$ and $o_{dk} = 1 + \sum_{m=1}^{M} \mathbb{E}_q[u_{mk}] + T_d/\xi_d$.

- $w_k \sim \text{Dirichlet}([e_{k1}, \ldots, e_{kN}])$,

$$\begin{aligned}\ln q^*(w_k) &= \sum_{d=1}^{D}\sum_{i=1}^{T_d}\sum_{n=1}^{N} \left(\mathbb{1}(x_{di} = n)\mathbb{E}_q[\mathbb{1}(\phi_{di} = k)] \ln w_{kn}\right) + \sum_{n=1}^{N} \left((\gamma - 1)\ln w_{kn}\right) + \text{ const.} \\ &= \sum_{n=1}^{N} \left(\gamma + \sum_{d=1}^{D}\sum_{i=1}^{T_d} \mathbb{1}(x_{di} = n)\mathbb{E}_q[\mathbb{1}(\phi_{di} = k)] - 1\right) \ln w_{kn} + \text{ const.}\end{aligned}$$

  Thus, $e_{kn} = \gamma + \sum_{d=1}^{D}\sum_{i=1}^{T_d} \mathbb{1}(x_{di} = n)\mathbb{E}_q[\mathbb{1}(\phi_{di} = k)]$.

- $\phi_{di} \sim \text{Categorical}([z_{di1}, \ldots, z_{diK}])$,

$$\begin{aligned}\ln q^*(\phi_{di}) &= \sum_{k=1}^{K}\sum_{n=1}^{N} \left(\mathbb{1}(x_{di} = n)\mathbb{1}(\phi_{di} = k)\mathbb{E}_q[\ln w_{kn}]\right) + \sum_{k=1}^{K} \mathbb{1}(\phi_{di} = k)]\mathbb{E}_q[\ln v_{dk}] + \text{ const.} \\ &= \sum_{k=1}^{K} \left(\sum_{n=1}^{N} \mathbb{1}(x_{di} = n)\mathbb{E}_q[\ln w_{kn}] + \mathbb{E}_q[\ln v_{dk}]\right)\mathbb{1}(\phi_{di} = k) + \text{ const.}\end{aligned}$$

  Thus, $z_{dik} = \sum_{n=1}^{N} \mathbb{1}(x_{di} = n)\mathbb{E}_q[\ln w_{kn}] + \mathbb{E}_q[\ln v_{dk}]$.

- $\rho_{mdk} = \exp\{\mathbb{E}_q[\ln u_{mk}] + \mathbb{E}_q[\ln v_{dk}]\}/(\sum_{l=1}^{K} \exp\{\mathbb{E}_q[\ln u_{ml}] + \mathbb{E}_q[\ln v_{dl}]\})$ to obtain the tightest lower bound.

- $\xi_d = \sum_{k=1}^{K} \mathbb{E}_q[v_{dk}]$ to obtain the tightest lower bound.

where

$$\begin{aligned}\mathbb{E}_q[u_{mk}] &= a_{mk}/b_{mk} \\ \mathbb{E}_q[\ln u_{mk}] &= \psi(a_{mk}) - \ln b_{mk} \\ \mathbb{E}_q[v_{dk}] &= c_{dk}/o_{dk} \\ \mathbb{E}_q[\ln v_{dk}] &= \psi(c_{dk}) - \ln o_{dk} \\ \mathbb{E}_q[\mathbb{1}(\phi_{di} = k)] &= z_{dik} \\ \mathbb{E}_q[\ln w_{kn}] &= \psi(e_{kn}) - \psi\left(\sum_{n=1}^{N} e_{kn}\right)\end{aligned}$$

where $\psi(x) = \frac{d \log \Gamma(x)}{dx}$ is the diagamma function.