EECS E6720 Bayesian Models for Machine Learning

Columbia University, Fall 2016

Lecture 12, 12/8/2016

Instructor: John Paisley

Non-negative matrix factorization

- Goal: We have a $M \times N$ data matrix X where $X_{ij} \geq 0$. We want to approximate this with a product of two non-negative matrices,
  - W: a $M \times K$ matrix, $W_{ij} \geq 0$
  - V: a $K \times N$ matrix, $V_{kj} \geq 0$
  - $X_{ij} \approx (WV)_{ij} = \sum_{k=1}^{K} W_{ik}V_{kj}$

- The questions (as usual) are two-fold:
  1. What objective do we use to measure approximation quality?
  2. What algorithm do we run to learn W and V?

- Lee & Seung's NMF paper is a major (non-Bayesian) step in this direction. They propose two objectives:

$$\text{Squared error: } \arg\min_{W,V} \sum_{i,j}(X_{ij} - (WV)_{ij})^2 \tag{1}$$

$$\text{Divergence penalty: } \arg\min_{W,V} \sum_{i,j} X_{ij} \ln \frac{X_{ij}}{(WV)_{ij}} - X_{ij} + (WV)_{ij} \tag{2}$$

- The key contribution is their fast multiplicative update rules for optimizing these objective functions over W and V.

- For example, the algorithm for the Divergence penalty is to update W and V as follows

$$V_{kj} \leftarrow V_{kj} \frac{\sum_{i=1}^{M} W_{ik}X_{ik}/(WV)_{ij}}{\sum_{i=1}^{M} W_{ik}} \tag{3}$$

$$W_{ik} \leftarrow W_{ik} \frac{\sum_{j=1}^{N} V_{kj}X_{ij}/(WV)_{ij}}{\sum_{j=1}^{N} V_{kj}} \tag{4}$$

- The NMF paper shows that these two updates produce new values of W and V that are monotonically decreasing the Divergence penalty. Similar updates are derived for the Squared error penalty.

- This paper is worth studying very closely for its own sake. We discuss it here to make connections with Bayesian methods. For example, what is this Divergence penalty doing? Imagine we had the following model.

- Model: We have a $M \times N$ matrix X of non-negative values. We model this as

$$X_{ij} \sim \text{Poisson}\left(\sum_{k=1}^{K} W_{ik} V_{kj}\right) \tag{5}$$

  where W_{ik} and V_{kj} are non-negative model variables.

- Recall the Poisson distribution: $X \in \{0, 1, 2, \dots\}$ is Poisson distributed with parameter λ if

$$p(X = x|\lambda) = \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}, \qquad \mathbb{E}x = Var(x) = \lambda \tag{6}$$

- Joint likelihood: As usual, write this out first before doing anything else

$$\begin{aligned} p(X|W, V) &= \prod_{i=1}^{M}\prod_{j=1}^{N} \text{Poisson}(X_{ij}|(WV)_{ij}) \\ &= \prod_{i=1}^{M}\prod_{j=1}^{N} \frac{(WV)_{ij}^{X_{ij}}}{X_{ij}!}\mathrm{e}^{-(WV)_{ij}} \end{aligned} \tag{7}$$

- Maximum likelihood: Next, consider maximum likelihood for this model. We want to find

$$\arg\max_{W,V} \ \ln p(X|W, V) = \sum_{i,j} X_{ij} \ln(WV)_{ij} - (WV)_{ij} + \text{constant wrt } W, V \tag{8}$$

- Notice that this objective is simply the negative of the Divergence penalty of NMF. Since the NMF algorithm monotonically increases this objective, this algorithm performs maximum likelihood for the Poisson factorization model described above. See the NMF paper for the original proof. We'll next show that the multiplicative updates are equivalent to the EM algorithm.

- This is another perspective of the NMF algorithm for the Divergence penalty. It's not the one they take in the paper, and it's not required for someone to take it either. However, we'll see that by thinking in terms of probability models, we can introduce prior structure and use other optimization (i.e., inference) algorithms to approximate the posterior distribution of these variables.

- Imagine deriving a gradient algorithm for Equation (8). This would be very difficult. This motivated the usefulness of the multiplicative updates of the original NMF paper.

Maximum likelihood EM for Poisson matrix factorization

- Our goal is to derive an EM algorithm for the Poisson likelihood model described above that will have simple and closed form updates. Therefore, we need to find an appropriate latent variable to add that has the correct marginal distribution. To this end, we digress into a property about sums of Poisson distributed random variables.

- Let $Y^{(k)} \sim \text{Poisson}(\lambda_k)$ independently for $k = 1, \ldots, K$. Then define $X = \sum_{k=1}^{K} Y^{(k)}$. It follows that the marginal distribution of X is $X \sim \text{Poisson}(\sum_{k=1}^{K} \lambda_k)$.

  Proof: A random variable Y with distribution $p(Y)$ is uniquely identified by its moment generating function, $\mathbb{E}_p[e^{tY}]$. For a Poisson(λ) distribution

$$\mathbb{E}_p[e^{tY}] = \sum_{y=0}^{\infty} e^{ty} \frac{\lambda^y}{y!} e^{-\lambda} = e^{-\lambda} e^{\lambda e^t} \sum_{y=0}^{\infty} \frac{(\lambda e^t)^y}{y!} e^{-\lambda e^t} = e^{-\lambda(1-e^t)} \tag{9}$$

  The sum equals 1 because it is over a Poisson(λe^t) distribution. Recognizing that $Y^{(k)}$ are generated independently, the following completes the proof,

$$\mathbb{E}[e^{tX}] = \mathbb{E}[e^{t\sum_{k=1}^{K} Y^{(k)}}] = \prod_{k=1}^{K} \mathbb{E}[e^{tY^{(k)}}] = \prod_{k=1}^{K} e^{-\lambda_k(1-e^t)} = e^{-(\sum_{k=1}^{K} \lambda_k)(1-e^t)} \tag{10}$$

  The proof is complete because we calculated these moment generating functions in the context of $Y^{(k)}$ and found that their sum has exactly the same generating function as a Poisson($\sum_{k=1}^{K} \lambda_k$) random variable. Therefore we can say that $\sum_{k=1}^{K} Y^{(k)}$ has this same distribution, which we will show is the correct marginal distribution.

- Extended model: For each element (i, j) in X, we now use the generative model

$$Y_{ij}^{(k)} \sim \text{Poisson}(W_{ik} V_{kj}), \quad X_{ij} | \vec{Y}_{ij} \sim \mathbb{1}\Big(X_{ij} = \sum_{k=1}^{K} Y_{ij}^{(k)}\Big) \tag{11}$$

  Notice that the "distribution" on X_{ij} puts all of its probability mass on the event $X_{ij} = \sum_{k=1}^{K} Y_{ij}^{(k)}$. Therefore, there's nothing random, but we still can say what $p(X|Y)$ is. Also notice that the marginal of X_{ij}—i.e., the distribution on X_{ij} *not* conditioned on Y (so we integrate Y out)—is $X_{ij} \sim \text{Poisson}(\sum_{k=1}^{K} W_{ik} V_{kj})$ as required.

- Joint likelihood: We now have that the joint likelihood including the extra variables Y is

$$p(X, Y | W, V) = \prod_{i=1}^{M} \prod_{j=1}^{N} p(X_{ij} | \vec{Y}_{ij}) \prod_{k=1}^{K} p(Y_{ij}^{(k)} | W_{ik}, V_{kj}) \tag{12}$$

- EM equation: We use the joint likelihood to set up the EM equation

$$\ln p(X|W, V) = \sum_{Y} q(Y) \ln \frac{p(X, Y|W, V)}{q(Y)} + \sum_{Y} q(Y) \ln \frac{q(Y)}{p(Y|X, W, V)} \tag{13}$$

  The q distribution is on all values $Y_{ij}^{(k)}$ for $i = 1, \ldots, M$, $j = 1, \ldots, N$ and $k = 1, \ldots, K$. Hopefully we can simplify this, otherwise we won't get very far.

- E-Step (part 1): The first part of the E-Step is to set $q(Y) = p(Y|X, W, V)$. We simply use Bayes rule and see what progress we can make.

$$\begin{aligned} p(Y|X,W,V) &\propto p(X|Y,W,V)p(Y|W,V) \\ &\propto p(X|Y)p(Y|W,V) \\ &\propto \prod_{i=1}^{M}\prod_{j=1}^{N} p(X_{ij}|\vec{Y}_{ij})p(\vec{Y}_{ij}|W,V) \end{aligned} \tag{14}$$

We have used the conditional independence defined by the model to write this expression. Notice that for each (i, j) pair, we have a "mini" Bayes rule embedded in this problem. That is

$$p(Y|X,W,V) = \prod_{i=1}^{M}\prod_{j=1}^{N} \frac{p(X_{ij}|\vec{Y}_{ij})p(\vec{Y}_{ij}|W,V)}{p(X_{ij}|W,V)} = \prod_{i=1}^{M}\prod_{j=1}^{N} p(\vec{Y}_{ij}|X_{ij},W,V) \tag{15}$$

Thus we know that $q(Y) = \prod_{i,j} q(\vec{Y}_{ij})$ and $q(\vec{Y}_{ij}) = p(\vec{Y}_{ij}|X_{ij}, W, V)$. We just need to see if we can calculate this for one (i, j) pair.

- In words, this is saying that if someone else generates $Y_{ij}^{(k)} \sim \text{Poisson}(W_{ik}V_{kj})$, calculates $X_{ij} = \sum_{k=1}^{K} Y_{ij}^{(k)}$ and then shows me X_{ij} and W and V, what is my posterior belief about $\vec{Y}_{ij}$?

- Again, we solve Bayes rule for this sub-problem. The result is one of the favorites of probability, here unfortunately derived using bogged-down notation from this problem.

$$\begin{aligned} p(\vec{Y}_{ij}|X_{ij},W,V) &= \frac{p(X_{ij}|\vec{Y}_{ij})p(\vec{Y}_{ij}|W,V)}{p(X_{ij}|W,V)} \\ &= \frac{\mathbb{1}(X_{ij} = \sum_{k=1}^{K} Y_{ij}^{(k)}) \prod_{k=1}^{K} \frac{(W_{ik}V_{kj})^{Y_{ij}^{(k)}}}{Y_{ij}^{(k)}!} e^{-W_{ik}V_{kj}}}{\frac{(WV)_{ij}^{X_{ij}}}{X_{ij}!} e^{-(WV)_{ij}}} \end{aligned} \tag{16}$$

In the numerator, we've multiplied the indicator distribution on X_{ij} with the product of K independent Poisson distributions on $\vec{Y}_{ij}$. In the denominator, we use the fact that the marginal distribution on X_{ij} is Poisson$(\sum_{k=1}^{K} W_{ik}V_{kj})$, which we proved above. We simply write out this distribution here. The final key step is to simplify this,

$$\begin{aligned} p(\vec{Y}_{ij}|X_{ij},W,V) &= \mathbb{1}\Big(X_{ij} = \sum_{k=1}^{K} Y_{ij}^{(k)}\Big) \frac{X_{ij}!}{\prod_{k=1}^{K} Y_{ij}^{(k)}!} \Bigg(\underbrace{\frac{W_{ik}V_{kj}}{\sum_{l=1}^{K} W_{il}V_{lj}}}_{\equiv\, \phi_{ij}(k)}\Bigg)^{Y_{ij}^{(k)}} \\ &= \text{Multinomial}(X_{ij}, \phi_{ij}) \end{aligned} \tag{17}$$

- The conditional posterior on the K Poisson random variables in the vector $\vec{Y}_{ij}$ *given that their sum must equal* X_{ij} is a multinomial distribution with probability distribution equal to the normalization of the K parameters in the Poisson prior on $\vec{Y}_{ij}$. While this wouldn't be obvious *a priori*, it's an intuitively reasonable result and a nice one too. Notice that the indicator in front is superfluous since by definition of the multinomial distribution with these parameters this sum must be true. We can therefore ignore it, but notice that it had to be there to start the calculation.

- E-Step 2: Next we calculate the expectation using this $q(Y)$ distribution.

$$\begin{aligned}\mathcal{L} &= \sum_Y q(Y) \ln p(X, Y|W, V) \\ &= \sum_{i=1}^{M}\sum_{j=1}^{N}\left\{\mathbb{E}_q[\ln p(X_{ij}|\vec{Y}_{ij})] + \sum_{k=1}^{K}\mathbb{E}_q[\ln p(Y_{ij}^{(k)}|W_{ik}, V_{kj})]\right\} \\ &= \sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{k=1}^{K}\mathbb{E}_q[Y_{ij}^{(k)}]\ln(W_{ik}V_{kj}) - W_{ik}V_{kj} + \text{constant}\end{aligned} \tag{18}$$

- In these equalities, we were able to get rid of $\mathbb{E}_q[\ln p(X_{ij}|\vec{Y}_{ij})] = \mathbb{E}_q[\ln \mathbb{1}(X_{ij} = \sum_{k=1}^{K} Y_{ij}^{(k)})]$ because $q(\vec{Y}_{ij})$ is such that $X_{ij} = \sum_{k=1}^{K} Y_{ij}^{(k)}$ with probability equal to one. Therefore the expectation is entirely over $\ln 1 = 0$. (Thankfully! If any of the $q(\vec{Y}_{ij})$ had nonzero probability of $X_{ij} \neq \sum_{k=1}^{K} Y_{ij}^{(k)}$ then $\mathcal{L} = -\infty$ and we couldn't proceed).

- Notice that given $q(\vec{Y}_{ij}) = \text{Multinomial}(X_{ij}, \phi_{ij})$, we can set $\mathbb{E}_q[Y_{ij}^{(k)}] = X_{ij}\phi_{ij}(k)$.

- M-Step: Finally we take derivatives with respect to W_{ik} and V_{kj} and set to zero.

$$\begin{aligned}\nabla_{W_{ik}}\mathcal{L}(W, V) &= 0 = \sum_{j=1}^{N}\frac{X_{ij}\phi_{ij}(k)}{W_{ik}} - \sum_{j=1}^{N} V_{kj} \\ &\Downarrow \\ W_{ik} &= \frac{\sum_{j=1}^{N} X_{ij}\phi_{ij}(k)}{\sum_{j=1}^{N} V_{kj}}\end{aligned} \tag{19}$$

$$\begin{aligned}\nabla_{V_{kj}}\mathcal{L}(W, V) &= 0 = \sum_{i=1}^{M}\frac{X_{ij}\phi_{ij}(k)}{V_{kj}} - \sum_{i=1}^{M} W_{ik} \\ &\Downarrow \\ V_{kj} &= \frac{\sum_{i=1}^{M} X_{ij}\phi_{ij}(k)}{\sum_{i=1}^{M} W_{ik}}\end{aligned} \tag{20}$$

- Now recalling that $\phi_{ij}(k) = W_{ik}V_{kj}/(WV)_{ij}$, we see that the updates are identical to the multiplicative updates for the NMF algorithm using the divergence penalty. Notice that the values of W and V in ϕ are fixed at the values of the previous iteration for all updates in this step. Also notice that these updates do not maximize $\mathcal{L}$ if done only once. However, they do increase $\mathcal{L}$, which is all that is required. We could continue to iterate between updating W and V for a fixed ϕ, or just make one update to each and then update ϕ.

- The EM algorithm also contains a little more information than the multiplicative algorithm for NMF discussed at the beginning of this lecture. (But it's not obvious at first glance if this extra information is useful at all. If it is it would be for computational efficiency.) As written in the original NMF algorithm, it appears that *all* values of W need to be the most recent ones when updating V and vice-versa for updating W. Just looking at Equations (3) and (4) above isn't

enough to say otherwise. However, we know from EM that the functions of W and V in the numerator correspond to ϕ and so we can keep those at the old values and only update the denominators. Therefore, one could iterate the original NMF algorithm several times only updating the sums in the denominators and holding all other values fixed, and separately update W and V in the numerator and multiplied out front every few iterations.

Variational inference for Poisson matrix factorization

- We have motivated a Bayesian approach to this problem. Next, we again define the model, this time along with its priors, and then discuss a variational inference algorithm for approximating the posterior. This algorithm will introduce new challenges that we haven't faced before, and we'll work through a possible solution that can be made more general.

- Model: We have the matrix X and model it with a K-rank non-negative factorization WV such that

$$X_{ij} \sim \text{Poisson}((WV)_{ij}) \tag{21}$$

- Priors: We use gamma priors for W and V as follows:

$$W_{ik} \sim \text{Gamma}(a, b), \qquad V_{kj} \sim \text{Gamma}(c, d) \tag{22}$$

- Posterior: We approximate the intractable posterior $p(W, V|X)$ with $q(W, V)$ using variational inference. We use the factorization

$$q(W, V) \;=\; \left[\prod_{i=1}^{M}\prod_{k=1}^{K} q(W_{ik})\right]\left[\prod_{j=1}^{N}\prod_{k=1}^{K} q(V_{kj})\right] \tag{23}$$

- Problem: We run into another problem this time in that things are *still* intractable. The variational objective function is

$$\begin{aligned}\mathcal{L} &= \mathbb{E}_q[\ln p(X|W,V)] + \mathbb{E}_q[\ln p(W)p(V)] - \mathbb{E}_q[\ln q] \qquad (24)\\ &= \sum_{i,j} X_{ij} \underbrace{\mathbb{E}_q\Big[\ln \sum_{k=1}^{K} W_{ik}V_{kj}\Big]}_{=\,???} - \sum_{k=1}^{K} \mathbb{E}_q[W_{ik}V_{kj}] + \mathbb{E}_q[\ln p(W)p(V)] - \mathbb{E}_q[\ln q] + \text{const.}\end{aligned}$$

- For this model we can't even calculate $\mathcal{L}$, so how do we take derivatives to update variational parameters for each q? The "optimal method" for learning each q has the same problem, since the problematic expectation still needs to be taken.

- Therefore, we don't have an analytic objective function to work with. This is a very common problem and has a few solutions. One trick is to replace the problem function with another function that approximates it (by lower bounding it) and is tractable.

- Since we want to maximize $\mathcal{L}$, we therefore lower bound the problem function.

- Let's work out this problem for this case using more abstract notation.

- Problem function: $\mathbb{E}_q \ln \sum_k Z_k$, $q = \prod_k q(Z_k)$

- The function $\ln$ is concave. Recall that for a concave function of a random variable, $f(Z)$,

$$f(\mathbb{E}Z) \geq \mathbb{E}f(Z) \tag{25}$$

- Notice that the function $\mathbb{E}_q \ln \sum_k Z_k \leq \ln \sum_k \mathbb{E}_q Z_k$. Therefore, we can't lower bound the problem function this way—instead it's an upper bound. The expectation in $\mathbb{E}_q \ln \sum_k Z_k$ is *not* the expectation corresponding to the function f above.

- Instead, we introduce a brand new discrete K-dimensional probability vector $\phi = [\phi(1), \ldots, \phi(K)]$ and write the trivial equality

$$\ln \sum_k Z_k = \ln \sum_k \phi(k) \frac{Z_k}{\phi(k)} \tag{26}$$

- This is a "trick" in that ϕ doesn't necessarily have a modeling purpose. We're just using math to our advantage. (However, we could also motivate this in a way similar to the previous EM algorithm, where we introduced auxiliary variables.)

- We have written $\ln \sum_k Z_k = \ln \mathbb{E}[Z/\phi]$, where this time the expectation is with respect to the distribution ϕ. Since $\ln$ is concave

$$\ln \sum_k Z_k = \ln \mathbb{E}[Z/\phi] \geq \mathbb{E}[\ln Z/\phi] = \sum_{k=1}^{K} \phi(k) \ln \frac{Z_k}{\phi(k)} \tag{27}$$

- Therefore, in the variational objective function

$$\mathbb{E}_q \ln \sum_{k=1}^{K} Z_k \geq \sum_{k=1}^{K} \phi(k) \mathbb{E}_q \ln Z_k - \sum_{k=1}^{K} \phi(k) \ln \phi(k) \tag{28}$$

- Notice that the problem should be fixed now since these are expectations we usually can take. Picking the lower bound as we did is part of the "art" of this technique. There are other lower bounds we could use. Some are no good because they don't result in analytic expectations. Some might be better than this one in that they are tighter—they approximate the original function more closely. This specific lower bound is probably not the only option.

- The next question is how do we set $\phi = [\phi(1), \ldots, \phi(K)]$? We want to set ϕ so that the bound is as good of an approximation as possible.

- Therefore we want to maximize this lower bound over ϕ. Using Lagrange multipliers, we can find that the lower bound is maximized when

$$\phi(k) = \frac{\mathrm{e}^{\mathbb{E}_q \ln Z_k}}{\sum_{\ell=1}^{K} \mathrm{e}^{\mathbb{E}_q \ln Y_\ell}} \tag{29}$$

- At this point, we have two paths we can go down:
  1. Plug this $\phi(k)$ back into the lower-bounded objective function
  2. Keep $\phi(k)$ as an auxiliary variable and use this update for it

- Path 1: Plugging this value of $\phi(k)$ back in and simplifying, we find

$$\mathbb{E}_q \ln \sum_{k=1}^{K} Z_k \;\geq\; \ln \sum_{k=1}^{K} \mathrm{e}^{\mathbb{E}_q \ln Z_k} \tag{30}$$

- This is the tightest possible lower bound of $\mathbb{E}_q \ln \sum_{k=1}^{K} Z_k$ when we limit ourselves to selecting from the family $\sum_{k=1}^{K} \phi(k) \mathbb{E}_q \ln Z_k - \sum_{k=1}^{K} \phi(k) \ln \phi(k)$

- If we go down this path, then the original problem is modified to

$$\mathcal{L} \geq \sum_{i,j} X_{ij} \ln \sum_{k=1}^{K} \mathrm{e}^{\mathbb{E}_q \ln W_{ik} + \mathbb{E}_q \ln V_{kj}} - \sum_{k=1}^{K} \mathbb{E}_q[W_{ik}]\mathbb{E}_q[V_{kj}] + \mathbb{E}_q[\ln p(W)p(V)] - \mathbb{E}_q[\ln q] \tag{31}$$

- The first term is the only place where an approximation is being made. This path has some pros and cons.

- PROS: We have a closed form objective that is the tightest possible approximation given the lower bound we use. (We'll see that Path 2 actually has this same PRO.)

- CONS: We can't use the optimal method to find q. We need to take derivatives and use gradient methods. That is ok, but like the EM story, if we could avoid doing this it would be preferable.

- Path 2: The second option is to keep ϕ as an auxiliary variable that we also optimize over. Since there is a function $\mathbb{E}_q[\ln \sum_k W_{ik} V_{kj}]$ for each (i, j) pair, we introduce a vector ϕ_{ij} for each of these to lower bound it. This is because the best lower bound will be different for each (i, j). Lower bounding them individually rather than using one shared ϕ will result in a much better approximation (it likely wouldn't work well with a single ϕ since the overall approximation will be bad).

- Therefore, we lower bound the variational objective function as

$$\begin{aligned} \mathcal{L} \;\geq\; & \sum_{i,j} \sum_{k=1}^{K} X_{ij} \left[\phi_{ij}(k) \mathbb{E}_q[\ln W_{ik} + \ln V_{kj}] - \phi_{ij}(k) \ln \phi_{ij}(k)\right] \\ & - \sum_{k=1}^{K} \mathbb{E}_q[W_{ik}]\mathbb{E}_q[V_{kj}] + \mathbb{E}_q[\ln p(W)p(V)] - \mathbb{E}_q[\ln q] \end{aligned} \tag{32}$$

- The advantage of this function is that we can now use the optimal method for finding each q. Also, notice that at the point of convergence, none of the parameters change anymore. Therefore, $\phi_{ij}(k)$ will equal its optimal value. Therefore, Path 2 finds a local optimal of the same function

that Path 1 does. It just does it by taking a few more steps. We had a very similar situation with EM and it's worth independently thinking more about the similarities.

- Let $\ln \hat{p}(X, W, V)$ be the log-joint distribution using the lower bound instead of the true objective,

$$\ln \hat{p}(X, W, V) = \sum_{i,j} \sum_{k=1}^{K} X_{ij} \left[\phi_{ij}(k)(\ln W_{ik} + \ln V_{kj}) - \phi_{ij}(k) \ln \phi_{ij}(k) \right] - \sum_{k=1}^{K} W_{ik} V_{kj} + \mathbb{E}_q[\ln p(W)p(V)]$$

- Using the optimal method for finding q with this lower bound, we have the following algorithm.

- Finding $q(W_{ik})$: By the typical route, we have

$$\begin{aligned} q(W_{ik}) &\propto \mathrm{e}^{\mathbb{E}_{-q}[\ln \hat{p}(X,W,V)]} \\ &\propto W_{ik}^{a+\sum_{j=1}^{N} X_{ij}\phi_{ij}(k)-1} \mathrm{e}^{-(b+\sum_{j=1}^{N} \mathbb{E}_q[V_{kj}])W_{ik}} \end{aligned} \tag{33}$$

  Therefore,

$$q(W_{ik}) = \text{Gamma}\left(a + \sum_{j=1}^{N} X_{ij}\phi_{ij}(k) \, , \, b + \sum_{j=1}^{N} \mathbb{E}_q[V_{kj}] \right) \tag{34}$$

- Finding $q(V_{kj})$: By symmetry we can quickly find that

$$q(V_{kj}) = \text{Gamma}\left(c + \sum_{i=1}^{M} X_{ij}\phi_{ij}(k) \, , \, b + \sum_{i=1}^{M} \mathbb{E}_q[W_{ik}] \right) \tag{35}$$

- Optimizing $\phi_{ij}(k)$: After updating each $q(W_{ik})$ and $q(V_{kj})$, set

$$\phi_{ij}(k) \; = \; \frac{\mathrm{e}^{\mathbb{E}_q[\ln W_{ik}] + \mathbb{E}_q[\ln V_{kj}]}}{\sum_{\ell=1}^{K} \mathrm{e}^{\mathbb{E}_q[\ln W_{i\ell}] + \mathbb{E}_q[\ln V_{\ell j}]}} \tag{36}$$

- To assess convergence, we then evaluate the lower bound of the variational objective function, since this is what we are trying to maximize. We hope that the q distributions then are "about as good as" what we would have gotten optimizing $\mathcal{L}$ directly, since the peaks and valleys of the objective function $\mathcal{L}$ should overlap significantly with those of the lower bound we use.

Comparison with EM

- With EM, updating $\phi_{ij}(k)$ didn't involve expectations since we have a point estimate of W and V. Notice that if we remove these expectations above, the update to $\phi_{ij}(k)$ is identical to EM.

- Using the distributions $q(W_{ik})$ and $q(V_{kj})$, compare $\mathbb{E}_q[W_{ik}]$ and $\mathbb{E}_q[V_{kj}]$ with the updates of EM. We can see a close similarity between the two, only now we have a full probability distribution on these terms. Therefore this model can be considered as the Bayesian approach to NMF with a divergence penalty, rather than just being motivated by it.

- This indicates that our lower bound is doing something similar (or perhaps equivalent) to introducing latent random variables $Y_{ij}^{(k)}$ to the model.