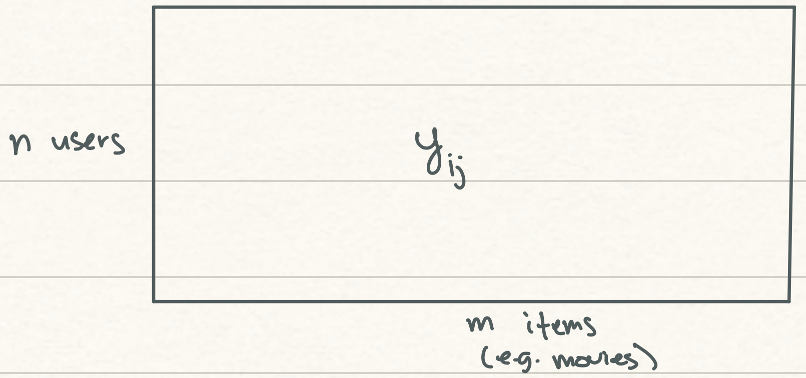
Poisson matrix factorization and auxiliary variables

The "Netflix prize" was a 2009 challenge to build the best movie recommendation system. The winning team won $1M for a solution based on **collaborative filtering and matrix factorization.** Here's the setting. We observe a **very sparse** (users x items) matrix:

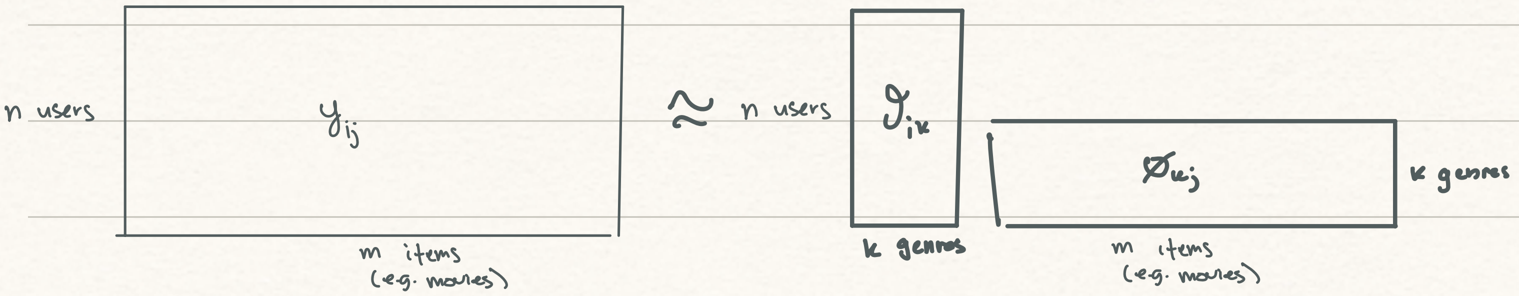


Each element y_ij is the number of minutes the user i spent watching item j.
The goal is to use this data to recommend a movie j to a user i that has never watched it.

For intuition, consider the following three rows:

	Harry Potter movies				Boston movies		
	HP1	...	HP8	...	Good Will Hunting	Departed	Town
Jimmy	1000	...	10000		0	0	0
Sean	2000	...	8000	...	100	0	200
Aaron	2000	...	1000		200	5000	1000

Aaron's row looks very similar to Sean's, except that Sean hasn't seen the Departed. Since Aaron seems to have loved that movie, and Sean is otherwise similar to Aaron, we might recommend Departed to Sean. We can formalize this intuition in terms of matrix factorization. Assume there are **K latent genres** of movies. We can learn what movies are in which genres, and which users like which genres. When a new user likes some films in a certain genre, we can then recommend them the others in that same genre.



Assume the number of minutes i watches j is Poisson distributed:

$$y_{ij} \sim \text{Pois}\left(\sum_k \theta_{ik} \phi_{kj}\right)$$

if the kth summand is large, it means user i has a high rate of watching genre k and that movie j is highly relevant to genre k. We'll place the following priors over the factors:

$$\theta_{ik} \overset{iid}{\sim} \text{Gamma}(a_0, b_0)$$

$$\phi_{kj} \overset{iid}{\sim} \text{Gamma}(a_0, b_0)$$

Although the gamma is the conjugate prior to the Poisson, the Poisson rate is a dot-product of gammas. Unfortunately the complete conditional is not closed-form:

$$p(\theta_{ik} \mid -) = \ ? \qquad \times$$

To deal with we will introduce **auxiliary variables** which will **augment the model**. The Poisson assumption above is equivalent to the following generative process:

$$y_{ij} \overset{ind.}{\sim} \text{Pois}\left(\sum_k \theta_{ik} \phi_{kj}\right) \iff y_{ijk} \overset{ind.}{\sim} \text{Pois}(\theta_{ik} \phi_{kj})$$

$$y_{ij} = \sum_k y_{ijk}$$

where for each (i,j) we have introduced K **latent sub-counts** (sometimes called "**sources**"). We assume the count we observe is the sum of the latent sub-counts, each of which is an independent Poisson random variable. Notice that **if we knew** the latent sub-counts then the model would be conditionally conjugate—the complete conditional would be:

$$p(\theta_{ik} \mid -) = \text{Gamma}\left(a_0 + \sum_j y_{ijk},\ b_0 + \sum_j \phi_{kj}\right) \qquad \checkmark$$

We don't observe the sub-counts, but we can treat them like any other latent variable. For example, if we could easily sample from their complete conditional, we could easily run a Gibbs sampler:

$$p(y_{ij1} \cdots y_{ijK} \mid y_{ij}, \Theta, \Phi) = ?$$

For simplicity, we will drop the (i,j) index. Consider the following generative process, using the bullet notation to denote the sum:

$$y_k \overset{ind.}{\sim} \text{Pois}(\mu_k), \quad y_{\cdot} = \sum_k y_k$$

We can write down the joint distribution, including the sum variable, which is deterministic:

$$p(y_1 \cdots y_K, y_\cdot) = p(y_1 \cdots y_K)\, p(y_\cdot \mid y_1 \cdots y_K)$$

$$= \Big[\prod_k \mathrm{Pois}(y_k; \mu_k)\Big]\, \delta(y_\cdot = \sum_k y_k)$$

$$= \Big[\prod_k \frac{\mu_k^{y_k}}{y_k!} \exp(-\mu_k)\Big]\, \delta(y_\cdot = \sum_k y_k)$$

Now multiply by 1:

$$= \exp(-\mu_\cdot) \Big[\prod_k \frac{\mu_k^{y_k}}{y_k!}\Big]\, \delta(y_\cdot = \sum_k y_k)\, \frac{\mu_\cdot^{y_\cdot}}{\mu_\cdot^{y_\cdot}}\, \frac{y_\cdot!}{y_\cdot!}$$

$$= \underbrace{\frac{\mu_\cdot^{y_\cdot}}{y_\cdot!} \exp(-\mu_\cdot)}_{\shortparallel}\ \underbrace{\frac{y_\cdot!}{\prod_k y_k!} \Big[\prod_k \Big(\frac{\mu_k}{\mu_\cdot}\Big)^{y_k}\Big]\, \delta(y_\cdot = \sum_k y_k)}_{\shortparallel}$$

$$= \mathrm{Pois}(y_\cdot; \mu_\cdot)\ \mathrm{Multinomial}\Big(y_1 \cdots y_K; y_\cdot, \frac{\mu_1}{\mu_\cdot} \cdots \frac{\mu_K}{\mu_\cdot}\Big)$$

$$= p(y_\cdot)\, p(y_1 \cdots y_K \mid y_\cdot)$$

After rearranging terms we find that the joint distribution equals the product of a Poisson PMF for the sum, and then a Multinomial PMF for the sub-counts conditional on the sum. This answers our question about the complete conditional of the sub-counts: it is multinomial. This is the Poisson-thinning property.

Translating this back into our model, the "data augmentation" step is:

$$p(y_{ij1} \cdots y_{ijK} \mid y_{ij}, \Theta, \Phi)$$

$$= \mathrm{Multinomial}\Big(y_{ij}; \frac{\theta_{i1}\phi_{1j}}{\sum_k \theta_{ik}\phi_{kj}} \cdots \frac{\theta_{iK}\phi_{Kj}}{\sum_k \theta_{ik}\phi_{kj}}\Big)$$

You may recognize this step from LDA. In fact, LDA is equivalent to Poisson matrix factorization with Dirichlet priors over the columns of the two parameter matrices, and under the assumption that all y_ij counts are observed (i.e., none are missing).

Data augmentation and auxiliary variables

The more general principle is that if we have a model with latent variables Z and data Y:

$$p(Z, Y), \qquad Z = \{z_1, \dots z_D\}$$

where the complete conditionals are not closed form:

$$p(z_d \mid z_{\backslash d}, Y) = \, ? \quad ✗$$

We can sometimes introduce a new set of **auxiliary variables A**

$$p(Z, Y, A)$$

such that the marginal (and by extension posterior) is the same:

$$p(Z \mid Y) = \mathbb{E}_{A \mid Y}\left[p(Z \mid A, Y)\right]$$

If cleverly designed, this can often lead to the augmented conditionals being nice:

$$p(z_d \mid z_{\backslash d}, Y, A) = \, ✓$$

$$p(A \mid Z, Y) = \, ✓$$

The following Gibbs sampler is then a valid way to approximate P(Z | Y):

$$Z^m \sim p(Z \mid A^{m-1}, Y)$$

$$A^m \sim p(A \mid Z^m, Y)$$

where at the end, we can simply throw away the auxiliary variables:

$$\mathbb{E}_{Z \mid Y}\left[f(Z)\right] = \frac{1}{M} \sum_{m=1}^{M} f(Z^m)$$

The same augmented representations of a model can be used to derive efficient variational inference and other algorithms.

Another example: community detection in networks

Say we observe a network, where the presence or absence of an edge between nodes i and j is given by X_ij:

$$X_{ij} \in \{1, 0\}$$

Say we observe a network, where the presence or absence of an edge between nodes i and j is given by X_ij:

$$p(X_{ij} = 1) = 1 - \exp\left(-\sum_k \theta_{ik} \theta_{jk}\right)$$

Our model for the presence of an edge is as follows, where we assume that each node participates on K latent communities to varying degrees. Assume the following prior for how much i participates in k:

$$\theta_{ik} \overset{iid}{\sim} \text{Gamma}(a_0, b_0)$$

The complete conditional is not in closed form, as the gamma is not conditionally conjugate to the Bernoulli:

$$p(\theta_{ik} \mid -) = ? \qquad \times$$

However we can augment the model with Poisson auxiliary variables. If we sample a Poisson, then the probability it is greater than 0 is equal to 1-exp(-...):

$$p(X_{ij} = 1) = 1 - \exp\left(-\sum_k \theta_{ik} \phi_{kj}\right) \iff \begin{array}{l} y_{ij} \sim \text{Pois}\left(\sum_k \theta_{ik} \phi_{kj}\right) \\ x_{ij} = 1(y_{ij} > 0) \end{array}$$

The complete conditional of this **latent Poisson** is then just a truncated Poisson, if X_ij=1 and 0 otherwise:

$$p(y_{ij} \mid x_{ij} = 1, \Theta, \Phi) = \text{TruncPoisson}_{>0}\left(\sum_k \theta_{ik} \phi_{kj}\right)$$

$$p(y_{ij} \mid x_{ij} = 0, \Theta, \Phi) = \delta_0$$

We can combine this augmentation scheme with the one in the previous section to then get all closed-form complete conditionals.