Mixed Membership Models

David M. Blei
Columbia University

October 31, 2014

Introduction

¶ We have seen mixture models in detail, which partition data into a collection of latent groups. We now discuss mixed-membership models, an extension of mixture models to *grouped data*. Each "data point" is a collection of data, and each collection can belong to multiple groups.

¶ Here are the basic ideas behind mixed-membership modeling:

- Data are grouped, each group x_i is a collection of x_{ij}, were $j \in \{1, \ldots, n_i\}$.
- Each group is modeled with a mixture model.
- The mixture components are shared across all the groups.
- The mixture proportions vary from group to group

We will get into details later. For now, here is the graphical model that describes these independence assumptions:

[graphical model]

This involves the following (generic) generative process,

1. Choose components $\beta_k \sim f(\cdot \mid \eta)$.
2. For each group i:
   (a) Choose proportions $\theta_i \sim \text{Dir}(\alpha)$.
   (b) For each data point j within the group:
      i. Choose a mixture assignment $z_{ij} \sim \text{Cat}(\theta_i)$.
      ii. Choose the data point $x_{ij} \sim g(\cdot \mid \beta_{z_{ij}})$.

You can see a mixture model as a piece of this graphical model. But there is more to it as well. Intuitively this captures that

- Each group of data is built from the same components or—as we will see—from a subset of the same components.

- How each group exhibits those components varies from group to group. Thus, there is both *homogeneity* and *heterogeneity*.

¶ **Text analysis** (Blei et al., 2003)

- Observations are individual words.
- Groups are documents, i.e., collections of words.
- Components are distributions over the vocabulary. These represent recurring patterns of observed words.
- Proportions represent how much each document reflects each pattern.
- The posterior components look like "topics"—like sports or health—and the proportions describe how each document describes those topics.
- This algorithm has been adapted to all kinds of other data—images, computer code, music data, others. More generally, it is a model of high-dimensional discrete data.
- This will be our running example.

¶ **Social network analysis** (Airoldi et al., 2008)

- Somewhat different from the graphical model, but the same ideas apply.
- Observations are single connections between members of a network.
- Groups are the set of connections for each person. Here you can see why the GM is wrong—this is not nested data.
- Components are communities, represented as distributions over which other communities each community tends to link to. In a simplified case, each community only links to others exhibiting that community.
- Proportions represent how much each person reflects a set of communities. You might know several people from your graduate school cohort, others from your neighborhood, others from the chess club, etc.
- Capturing these overlapping communities is not possible with a mixture model of people, where each person is in just one community.
- Conversely, modeling each person individually doesn't tell us anything about the global structure of the network.

¶ **Survey analysis** (Erosheva, 2003)

- Much of social science analyzes carefully designed surveys.

- There might be several social patterns that are present in the survey, but each respondent exhibits different ones.
- (Adjust the graphical model here so that there is no plate around X, but rather individual questions and parameters for each question.)
- The observations are answers to individual questions.
- The groups are the collection of answers by a single respondent.
- Components are collections of likely answers for each question, representing recurring patterns in the survey.
- Proportions represent how much each individual exhibits those patterns.
- A mixture model assumes each respondent only exhibits a single pattern.
- Individual models tell us nothing about the global patterns.

¶ **Population genetics** (Pritchard et al., 2000)

- Observations are the alleles on the human genome, i.e., at a particular site are you an A, G, C, or T?
- Groups are the genotype of individuals—each of our collection of alleles at each of our loci.
- Components are patterns of alleles at each locus. These are "types" of people, or the genotypes of ancestral populations.
- Proportions represent how much each individual exhibits each population.
- Application #1: Understanding population history and differences. For example, in India everyone is part Northern ancestral Indian/Southern ancestral indian and no one is 100% of either. This model gives us a picture of the original genotypes.
- Application #2: "Correcting" for latent population structure when trying to associate genotypes with diseases. For example, prostrate cancer is more likely in African American males than European American males. If we have a big sample of genotypes, an allele that shows up in African American males will look like it is associated with cancer. Correcting for population-level frequencies helps mitigate this confounding effect.
- Application #3: "Chromosome painting." Use the ancestral observations to try to find candidate regions for genome associations. Knowing the AA males get prostrate cancer more than EA males, look for places where a gene is more exhibited than expected (in people with cancer) and less so (in people without

cancer). This is a candidate region. (This was really done successfully for prostrate cancer.)

¶ Compare these assumptions to a single mixture model. A mixture is less heterogeneous—each group can only exhibit one component. (There is still some heterogeneity because different groups come from different parameters.)

Modeling each group with a completely different mixture (proportions and components) is *too* heterogeneous—there is no connection or way to compare groups in terms of the underlying building blocks of the data.

¶ This is an example of a *hierarchical model*, a model where information is shared across groups of data. The sharing happens because we treat parameters as hidden random variables and estimate their posterior distributions.

There are two important characteristics for a successful hierarchical model.

[Use a running example of the graphical model with a few groups.]

One is that information is shared across groups. Here this happens via the unknown mixture components. Consider if they were fixed. The groups of data would be independent.

The other is that within-group data is more similar than across-group data. Suppose the proportions were fixed for each group. Because of the components, there is still sharing across groups. But two data points within the same group are just as similar as two data points across groups. In fact, this is a simple mixture as though the group boundaries were not there. When we involve the proportions as a group-specific random variable, within-group data are more tightly connected than across-group data.

The Dirichlet distribution

¶ The observations x and the components β are tailored to the data at hand. Across mixed membership models, however, the assignments z are discrete and drawn from the proportions θ. Thus, all MMM need to work with a distribution over θ.

The variable θ lives on the *simplex*, the space of positive vectors that sum to one. The exponential prior on the simplex is called the *Dirichlet* distribution. It's important across statistics and machine learning, and particularly important in Bayesian non-parametrics (which we will study later). So, we'll now spend some time studying the Dirichlet.

¶ The parameter to the Dirichlet is a k-vector α, where $\alpha_i > 0$. In its familiar form,

the density of the Dirichlet is

$$p(\theta \mid \alpha) = \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_j\right)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}. \tag{1}$$

The Gamma function a real-valued version of factorial. (For integers, it is factorial.)

You can see that this is in the exponential family because

$$p(\theta \mid \alpha) \propto \exp\left\{ \alpha^\top \log \theta - \sum_j \log \theta_j \right\}. \tag{2}$$

But we'll work with the familiar parameterization for now.

As you may have noticed, the Dirichlet is the multivariate extension of the beta distribution,

$$p(\pi \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha - 1} (1 - \pi)^{\beta - 1} \tag{3}$$

A number between 0 and 1 is a point on the "1-simplex".

¶ The expectation of the Dirichlet is

$$\mathbb{E}[\theta_\ell] = \frac{\alpha_\ell}{\sum_j \alpha_j}. \tag{4}$$

¶ Case #1, $\alpha_j = 1$.

- This is a uniform distribution.
- Every point on the simplex is equally likely.

Case #2, $\alpha_j > 1$.

- This is a "bump."
- It is centered around the expectation.

Case #3, $\alpha_j < 1$.

- This is a *sparse* distribution.
- Some (or many) components will have near zero probability.
- This will be important later, in Bayesian nonparametrics.

(Show pictures here)

¶ The Dirichlet is conjugate to the multinomial.

Let z be an indicator vector, i.e., a k-vector that contains a single one. The parameter to z is a point on the simplex θ, denoting the probability of each of the k items. The

density function for z is

$$p(z \mid \theta) = \prod_{j=1}^{k} \theta_j^{z^j}, \tag{5}$$

which "selects" the right component of θ. (This is a multivariate version of the Bernoulli.)

¶ Suppose we are in the following model,

$$\theta \sim \text{Dir}(\alpha) \tag{6}$$

$$z_i \mid \theta \sim \text{Mult}(\theta) \quad \text{for } i \in \{1, \ldots, n\}. \tag{7}$$

Let's compute the posterior distribution of θ,

$$p(\theta \mid z_{1:n}, \alpha) \propto p(\theta, z_{1:n} \mid \alpha) \tag{8}$$

$$= p(\theta \mid \alpha) \prod_{i=1}^{n} p(z_i \mid \theta) \tag{9}$$

$$= \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_j\right)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} \prod_{i=1}^{n} \prod_{j=1}^{k} \theta_j^{z_i^j} \tag{10}$$

$$\propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1 + \sum_{i=1}^{n} z_i^j}. \tag{11}$$

Note that $\sum_{i=1}^{n} z_i^j = n_j$, i.e., the number of times we saw item j in the variables $z_{1:n}$.

Eq. 11 is a Dirichlet distribution with parameter $\hat{\alpha}_j = \alpha_j + n_j$.

¶ The expectation of the posterior Dirichlet is interesting,

$$\mathbb{E}\left[\theta_\ell \mid z_{1:n}, \alpha\right] = \frac{\alpha_\ell + n_\ell}{n + \sum_{j=1}^{k} \alpha_j} \tag{12}$$

This is a "smoothed" version of the empirical proportions. As n gets large relative to α, the empirical estimate dominates this computation. This is the old story—when we see less data, the prior has more of an effect on the posterior estimate.

When used in this context, α_j can be interpreted as "fake counts." In fact this interpretation is clearer when considering the n_0, x_0 parameterization of this prior. (See our lecture on exponential families.) This expectation reveals why in a different way—it is the MLE as though we saw $n_j + \alpha_j$ items of each type. This is used in language modeling as a "smoother".

Topic models

¶ We will study topic models as a testbed for mixed-membership modeling ideas. But keep in mind the other applications that we mentioned in the beginning of the lecture.

¶ The goal of topic modeling is to analyze massive collections of documents. There are two categories of motivations for why we want to do this:

- Predictive: Search, recommendation, classification, etc.
- Exploratory: Organizing the collection for browsing and understanding.

¶ Our data are documents.

- Each document is a group of words $w_{d,1:n}$.
- Each word $w_{d,i}$ is a value among V words.

The hidden variables are

- Multinomial parameters $\beta_{1:K}$ (compare to Gaussian).
  - Each component is a distribution over the vocabulary.
  - These are called "topics."
- Topic proportions $\theta_{1:D}$.
  - Each is a distribution over the K components.
- Topic assignments $z_{1:D,1:N}$.
  - Each is a multinomial indicator of the k topics.
  - There is one for every word in the corpus.

¶ The basic model has the following generative process. This is an adaptation of the generic mixed-membership generative process.

1. Draw $\beta_k \sim \text{Dir}_V(\eta)$, for $k \in \{1, \ldots, K\}$.
2. For each document d:
   (a) Draw $\theta_d \sim \text{Dir}_K(\alpha)$.
   (b) For each word n in each document,
      i. Draw $z_{d,n} \sim \text{Cat}(\theta_d)$.
      ii. Draw $w_{d,n} \sim \text{Mult}(\beta_{z_{d,n}})$.

¶ Let's contemplate the posterior. Note, this is usually a more productive (and interesting) activity than wondering whether your data really comes from the model. (The usual answer: It doesn't.)

The posterior is proportional to the joint. We have seen in Gibbs sampling that we are doing something that looks like optimizing the joint, getting to configurations of the latent variables that have high enough probability under the prior & explain the data.

In LDA, the log joint is

$$\log p(\cdot) = \sum_{k=1}^{K} \log p(\beta_k) + \sum_{d=1}^{D} \left(\log p(\theta_d) + \sum_{n=1}^{N} \log p(z_{d,n} \mid \theta_d) + \log p(w_{d,n} \mid z_{d,n}, \boldsymbol{\beta}, \theta_d) \right) \quad (13)$$

Substitute in the simple categorical parameterizations,

$$\log p(\cdot) = \sum_{k=1}^{K} \log p(\beta_k) + \sum_{d=1}^{D} \left(\log p(\theta_d) + \sum_{n=1}^{N} \log \theta_{d,z_{d,n}} + \log \beta_{w_{d,n},z_{d,n}} \right)$$

We see that the posterior gets bonuses for choosing topics with high probability in the document (θ_d) and words with high probability in the topic (β_k).

These two latent variables must sum to one. Therefore, the model prefers documents to have peaky topic proportions, i.e., few topics per document, and for topics to have peaky distributions, i.e., few words per topic. But these goals are at odds—putting a document in few topics means that those topics must cover all the words of the document. Putting few words in a topic means that we need many topics to cover the documents.

This intuition is why LDA gives us the kind of sharp co-occurrences.

Again, contrast to a mixture model. Mixtures assert that each document has one topic. That means that the topics must cover all the words that each document contains. They are less peaky and "sharp".

¶ (Do the demo with Jonathan's code.)

¶ Aside: de Finetti theorem says that if a collection of random variables are *exchangeable*, then their joint can be written as a "Bayesian model"

$$p(x_1, x_2, \ldots, x_n) = \int p(\theta) \prod_{i=1}^{n} p(x_i \mid \theta) d\theta \quad (14)$$

In document collections this says that the order of words doesn't matter,

$$p(w_1, w_2, \ldots, w_n \mid \beta) = \int p(\theta) \prod_{i=1}^{n} p(w_i \mid \beta) d\theta \quad (15)$$

This is called the "bag of words" assumption in NLP. Note this is an assumption about exchangeability, rather than independence.

In topic modeling, this is palatable—shuffling the words of a document still tell you what it's about.

Gibbs sampling in LDA

¶ We implement the basic Gibbs sampler by writing down the complete conditionals.

The conditional of the topic (component) assignment $z_{d,n}$ is a multinomial over K elements. Each probability is

$$\begin{aligned} p(z_{d,i} = k \mid \mathbf{z}_{-i}, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{w}) &= p(z_{d,i} = k \mid \theta_d, w_{d,i}, \boldsymbol{\beta}) && (16) \\ &\propto p(\theta_d)\, p(z_{d,i} = k \mid \theta_d)\, p(w_{d,i} \mid \beta_k) && (17) \\ &\propto \theta_{d,k}\, p(w_{d,i} \mid \beta_j) && (18) \end{aligned}$$

- Independence follows from the graphical model.
- The prior $p(\theta)$ disappears because it doesn't depend on z_i.
- In LDA, the second term is the probability of word $w_{d,i}$ in topic β_j. We leave it general here enable other kinds of mixed-membership models.

The conditional of the topic (component) proportions θ_d is a posterior Dirichlet,

$$\begin{aligned} p(\theta_d \mid z, \boldsymbol{\theta}_{-d}, w, \boldsymbol{\beta}) &= p(\theta_d \mid \mathbf{z}_d) && (19) \\ &= \text{Dir}\left(\alpha + \textstyle\sum_{i=1}^{n} z_{d,i}\right). && (20) \end{aligned}$$

- Independence follows from the graphical model.
- The posterior Dirichlet follows from our discussion of the Dirichlet.
- The sum of indicators creates a count vector of the topics in document ℓ.
- This is general for all mixed-membership models.

Finally, the conditional of the topic β_k is a Dirichlet. (For other types of likelihoods, this will be a different posterior.)

$$\begin{aligned} p(\beta_k \mid z, \theta, w, \boldsymbol{\beta}_{-k}) &= p(\beta_j \mid z, w) && (21) \\ &= \text{Dir}\left(\eta + \textstyle\sum_{d=1}^{D} \sum_{n=1}^{N} z_{d,n}^{j} \circ w_{d,n}\right) && (22) \end{aligned}$$

- Independence follows from the graphical model.
- The posterior Dirichlet follows from the discussion of the Dirichlet.
- The double sum counts how many times each word occurs under topic k.

¶ A *collapsed Gibbs sampler* is available, where we integrate out all the latent variables except for $\mathbf{z}$ (Griffiths and Steyvers, 2004).

Each $z_{d,n}$ takes one of K values. It is a simple categorical distribution. The conditional probability of topic assignment k is proportional to the following joint,

$$p(z_{d,n} = k \mid \mathbf{z}_{-(d,n)}, \mathbf{w}) \propto p(z_{d,n}, w_{d,n} \mid \mathbf{z}_{-(d,n)}, \mathbf{w}_{-(d,n)}) \tag{23}$$

We will integrate out the topic proportions θ_d and topic β_j to obtain an integrand independent of the other assignments and words. First, note that the joint distribution of a topic and word within a document is

$$\begin{aligned} p(z_{d,n} = k, w_{d,n} \mid \theta_d, \beta_{1:K}) &= p(z_{d,n} = k \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n} = k) \\ &= \theta_{d,k} \beta_{k,w_{d,n}} \end{aligned} \tag{24}$$

We use this to compute Eq. 23. We integrate out the topic and topic proportions,

$$\begin{aligned} p(z_{d,n} = k \mid \mathbf{z}_{-(d,i)}, w) &\propto \int_{\beta_k} \int_{\theta_d} p(z_{d,n} = k, w_{d,n} \mid \theta_d, \beta_k) p(\theta_d \mid \mathbf{z}_{d,-n}) p(\beta_k \mid \mathbf{z}_{-(d,n)}, \mathbf{w}_{-(d,n)}) \\ &= \int_{\beta_k} \int_{\theta_d} \theta_{d,k} \beta_{k,w_{d,n}} p(\theta_d \mid z_{d,-n}) p(\beta_k \mid z_{-(d,n)}, w_{-(d,n)}) \qquad (25) \\ &= \left(\int_{\theta_d} \theta_{d,k} \, p(\theta_d \mid z_{d,-n}) \right) \left(\int_{\beta_k} \beta_{k,w_{d,n}} p(\beta_k \mid z_{-(d,n)}, w_{-(d,n)}) \right). \qquad (26) \end{aligned}$$

Each of these two terms are expectations of posterior Dirichlets.

- The first is like Eq. 20, but using all but $z_{d,i}$ to form counts.
- The second is like Eq. 22, but using all but $w_{d,i}$ to form counts.

The final algorithm is simple

$$p(z_{d,n} = k \mid \mathbf{z}_{-(d,n)}, \mathbf{w}) = \left(\frac{\alpha + n_d^k}{k\alpha + n_d} \right) \left(\frac{\eta + m_k^{w_{d,n}}}{v\eta + m_k} \right). \tag{27}$$

The counts n_d are per-document counts of topics and the counts m_j are per topic counts of terms. Each is defined excluding $z_{d,n}$ and $w_{d,n}$.

[Give the algorithm]

References

Airoldi, E., Blei, D., Fienberg, S., and Xing, E. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014.

Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Erosheva, E. (2003). Bayesian estimation of the grade of membership model. *Bayesian Statistics*, 7:501–510.

Griffiths, T. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Science*, 101:5228–5235.

Pritchard, J., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959.