Mixed Membership Models

David M. Blei
Princeton University

October 10, 2013

1 Introduction

- We have reviewed and seen mixture models in detail. And we've seen hierarchical models—particularly those that capture *nested* structure in the data.

- We will now combine these ideas to form *mixed membership models*, which is a powerful modeling methodology.

- The basic ideas are
  - Data are grouped.
  - Each group is modeled with a mixture.
  - The mixture components are shared across all the groups.
  - The mixture proportions are vary from group to group.

- We'll get to the details later, but for now here is the graphical model that describes the independence assumptions. (Draw generic graphical model.)

- 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 *homogoneity* and *heterogeneity*.

- Example #1: 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 fitted 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. More generally, it models high-dimensional discrete data.
  - This will be our running example.

- Example #2: 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. For example, 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.

- Example #3: Survey analysis (Erosheva, 2004)
  - 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.

- Example #4: 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 exampple, African American males are more likely to get prostrate cancer than European American males. If we have a big sample of genotypes, an allele that shows up in AA 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 AA 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 that a single mixture model is less heterogenous—each group can only exhibit one component. (Though, there is heteorgeneity in the sense that different groups can come from different components.)

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

2 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.
  - It's particularly important in Bayesian nonparametrics.
  - 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

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

- 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.

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

- The Dirichlet is the multivariate extension of the beta distribution,

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

- The expectation of the Dirichlet is

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

- Interesting case #1, $\alpha_j = 1$.
  - This is a uniform distribution.
  - Every point on the simplex is equally likely.

- Interesting case #2, $\alpha_j > 1$.
  - This is a "bump."
  - It is centered around the expectation.

- Interesting 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 a multinomial indicator, i.e., a k-vector that contains a single one. (In general, a multinomial vector contains counts, but we'll look at multinomial indicators for now.)

- 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.

- Suppose we are in the following model,

$$\begin{aligned} \theta &\sim \text{Dir}(\alpha) && (6)\\ z_i \mid \theta &\sim \text{Mult}(\theta) \quad \text{for } i \in \{1,\ldots,n\}. && (7) \end{aligned}$$

- Let's compute the posterior distribution of θ,

$$\begin{aligned} p(\theta \mid z_{1:n}, \alpha) &\propto p(\theta, z_{1:n} \mid \alpha) && (8)\\ &= p(\theta \mid \alpha) \prod_{i=1}^{n} p(z_i \mid \theta) && (9)\\ &= \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_j\right)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} \prod_{i=1}^{n} \prod_{j=1}^{k} \theta_j^{z_i^j} && (10)\\ &\propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1 + \sum_{i=1}^{n} z_i^j}. && (11) \end{aligned}$$

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

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

- The expectation of the posterior Dirichlet is interesting,

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

  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 interepreted as "fake counts." This expectation reveals why—this is the MLE as though we saw $n_j + \alpha_j$ items.

- This is used in language modeling as a "smoother." (We will learn about more complicated smoothing—and corresponding distributions—later on in the course.)

3 Probabilistic topic models

- We will study topic models as a testbed for mixed-membership modeling ideas.

- The goal is to model massive collections of documents. There are two motivations.
  - Predictive tasks: Search, collaborative filtering, classification, etc.
  - Exploratory analysis: Organizing the collection for browsing and understanding.

- (Draw the graphical model.)

- Our data are documents
  - Each document is a group of words $w_{d,1:n}$.
  - Each word $w_{d,i}$ is a multinomial 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.
    * Note there is one for every word in the corpus.
    * Billions of hidden variables.

- The basic model has the following generative process.

  1. Draw $\beta_j \sim \text{Dir}_v(\eta)$, for $j \in \{1, \ldots, k\}$.
  2. For each document ℓ, for $\ell \in \{1, \ldots, d\}$
     (a) Draw $\theta_\ell \sim \text{Dir}_k(\alpha)$.
     (b) For each word in each document,
        i. Draw $z_{\ell,i} \sim \text{Mult}(\theta_\ell)$.
        ii. Draw $w_{\ell,i} \sim \text{Mult}(\beta_{z_{\ell,i}})$.

- (Show the intuitions in a couple of slides.)

- (Show this working with Jonathan's code.)

- Think again about heterogeneity.
  - In a *mixtures of multinomials* model, each document exhibits "one topic."
  - Intuitively, we expect text to be more heterogenous.

- But this also (kind of) explains why it "works."
  - A mixture model says all documents exhibit one topic.
  - Fitting a mixture model, therefore, puts words together with high probability that tend to co-occur within a document.
  - In LDA, the Dirichlet distribution can allow for multiple topics at once.
  - But, there are fewer topics than words, so words still need to co-occur. But they can co-occur with other patterns of co-occurrence.
  - Furthermore, the sparsity of the Dirichlet distribution keeps the number of topics per document small. (In practice $\alpha << 1$ both when fixed and when fitted.)

- The bag of words assumption and de Finetti
  - De Finetti's theorem says that if a collection of random variables are *exchangeable*, then their joint can be written as a "Bayesian model"

$$p(x_1, x_2, \ldots, x_n) = \int p(\theta) \prod_{i=1}^{n} p(x_i \mid \theta) d\theta \tag{13}$$

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

$$p(w_1, w_2, \ldots, w_n \mid \beta) = \int p(\theta) \prod_{i=1}^{n} p(w_i \mid \beta) d\theta \tag{14}$$

  - This is called the "bag of words" assumption in NLP. Note this is an assumption about exhchangeability, rather than independence.
  - I think this makes it more palettable—shuffling the words of a document still tell you what it's about.

- LDA is closely related to factor analysis
  - In factor analysis

$$x_i \sim \mathcal{N}(\mu, \Sigma) \tag{15}$$
$$y_i \sim \mathcal{N}(x_i^\top \beta, \text{diag}(\lambda)) \tag{16}$$

  - In LDA

$$\theta_d \sim \text{Dir}(\alpha) \tag{17}$$
$$w_d \sim \text{Mult}(\theta_d^\top \beta, n) \tag{18}$$

  - For this reason, sometimes its called "multinomial PCA."
  - This also connects to latent semantic indexing (Deerwester et al., 1990), which is a singular value decomposition of a document/term matrix.

4 Gibbs sampling in LDA

- We'll discuss two kinds of inference—Gibbs sampling and variational inference. Both have advantages and disadvantages.

- First, let's review MCMC methods and Gibbs sampling in general.

- The idea behind MCMC is to define a Markov chain on the hidden variables of the model, such that the stationary distribution of the Markov chain is the posterior distribution that we are trying to estimate. See Neal (1993) for an excellent discussion.

- When the conditionals of each hidden variable are available then Gibbs sampling is a particularly straightforward MCMC algorithm.

- A Gibbs sampling algorithm repeatedly samples from $p(z_i \mid z_{-i}, x)$. This defines an appropriate chain.
  - An iteration is a complete pass through the hidden variables.
  - After enough iterations have been run (the "burn-in"), collecting these full scans at a lag theoretically collects samples from the posterior.
  - This revolutionized Bayesian statistics. See Gelfand and Smith (1990).

- Notice that these are the same conditionals that we used in defining coordinate-ascent mean-field variational inference.

- In a mixed-membership model like LDA, we can write down the conditionals. The conditional of the topic (component) assignment $z_{d,i}$ is a multinomial over k elements. Each probability is

$$\begin{aligned} p(z_{d,i} = j \mid z_{-i}, \theta, \beta, w) &= p(z_{d,i} = j \mid \theta, w_{d,i}, \beta) && (19) \\ &\propto p(\theta) p(z_{d,i} = j \mid \theta) p(w_{d,i} \mid \beta_j) && (20) \\ &\propto \theta_j p(w_{d,i} \mid \beta_j) && (21) \end{aligned}$$

  - Independence follows from the graphical model.
  - The term $p(\theta)$ disappears because it doesn't depend on z_i.
  - In LDA, the second term is just the probability of word $w_{d,i}$ in topic β_j.
  - Note that this term is a general conditional for any mixed-membership model.

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

$$\begin{aligned} p(\theta_\ell \mid z, \theta_{-\ell}, w, \beta) &= p(\theta_\ell \mid z_\ell) && (22) \\ &= \text{Dir}\left(\alpha + \textstyle\sum_{i=1}^{n} z_{\ell,i}\right). && (23) \end{aligned}$$

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

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

$$\begin{aligned} p(\beta_j \mid z, \theta, w, \beta_{-j}) &= p(\beta_j \mid z, w) && (24) \\ &= \text{Dir}\left(\eta + \textstyle\sum_{\ell=1}^{d} \sum_{i=1}^{n} z_{\ell,i}^{j} \circ w_{\ell,i}\right) && (25) \end{aligned}$$

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

- In LDA, a *collapsed Gibbs sampler* is available, where we integrate out all the latent variables except for z.

- Each $z_{\ell,i}$ takes one of k values. It is drawn from a multinomial.
  - First we note that it is proportional to the following joint,

$$p(z_{\ell,i} = j \mid z_{-(\ell,i)}, w) \propto p(z_{\ell,i}, w_{\ell,i} \mid z_{-(\ell,i)}, w_{-(\ell,i)}) \tag{26}$$

  - We integrate out the topic proportions θ_ℓ and topic β_j to obtain an integrand independent of the other assignments and words,

$$\begin{aligned} p(z_{\ell,i} = j \mid z_{-(\ell,i)}, w) &\propto \int_{\beta_j} \int_{\theta_\ell} p(z_{\ell,i} = j \mid \theta_\ell) p(w_{\ell,i} \mid z_{\ell,i} = j, \beta_j) p(\theta_\ell \mid z_{\ell,-i}) p(\beta_j \mid z_{-(\ell,i)}, w_{-(\ell,i)}) && (27) \\ &= \int_{\beta_j} \int_{\theta_\ell} \theta_j \beta_{j,w_{\ell,i}} p(\theta_\ell \mid z_{\ell,-i}) p(\beta_j \mid z_{-(\ell,i)}, w_{-(\ell,i)}) && (28) \\ &= \left(\int_{\theta_\ell} \theta_{\ell,j} p(\theta_\ell \mid z_{\ell,-i}) \right) \left(\int_{\beta_j} \beta_{j,w_{\ell,i}} p(\beta_j \mid z_{-(\ell,i)}, w_{-(\ell,i)}) \right) && (29) \end{aligned}$$

- Each of these two terms are expectations of posterior Dirichlets.
  - The first is almost like Equation 23, but using all but $z_{\ell,i}$ to form counts.
  - The second is almost like Equation 25, but using all but $w_{\ell,i}$ to form counts.

- The final algorithm is simple

$$p(z_{\ell,i} = j \mid z_{-(\ell,i)}, w) = \left(\frac{\alpha + n_\ell^j}{k\alpha + n_\ell} \right) \left(\frac{\eta + m_j^{w_{\ell,i}}}{v\eta + m_j} \right). \tag{30}$$

  The counts n_ℓ are per-document counts of topics and the counts m_j are per topic counts of terms. Each is defined excluding $z_{\ell,i}$ and $w_{\ell,i}$.

5 Mean-field variational inference for LDA

- (Put a picture of the graphical model and corpus on the board.)

- Recall that in mean-field variational inference, we minimize the KL divergence between a factored distribution q and the posterior.

- In LDA, the factored distribution is

$$q(\beta_{1:k}, \theta_{1:d}, w_{1:d}) = \prod_{j=1}^{k} q(\beta_j \mid \lambda_j) \prod_{\ell=1}^{d} \left(q(\theta \mid \gamma) \prod_{i=1}^{n} q(z_{\ell,i} \mid \phi_{\ell,i}) \right). \tag{31}$$

  - For each topic, λ_j is a posterior Dirichlet over terms. This gives us the word lists.
  - For each document, γ_ℓ is a posterior Dirichlet over topics. This gives the topic proportions—a description of how each document exhibits the topics.
  - For each word in each document, $\phi_{\ell,i}$ is a multinomial over topics. This tells us which topic that particular word is likely to have come from.

- We optimize these parameters with coordinate ascent, computing each update either by looking at the log conditional or the log joint. (Recall the variational notes.)

- (Put a graphical model of the variational distribution on the board.)

- First, let's look at the Dirichlet as an exponential family.

$$\begin{aligned} p(\theta \mid \alpha) &\propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} \\ &= \exp\{(\alpha - 1)^\top \log \theta\} \end{aligned}$$

- This means that the natural parameter is $\alpha - 1$ and the sufficient statistic is $\log \theta$.

- Let's update the variational Dirichlet for the topic proportions.

  - The conditional distribution of the topic proportions is

$$\theta_\ell \mid z_\ell \sim \text{Dir}(\alpha + \textstyle\sum_{i=1}^{n} z_{\ell,i}) \tag{32}$$

  - This has expected natural parameter

$$\text{E}[\alpha + \textstyle\sum_{i=1}^{n} z_{\ell,i} - 1] = \alpha + \textstyle\sum_{i=1}^{n} \phi_{\ell,i} - 1]. \tag{33}$$

  - Which means that the update for the topic proportions is

$$\gamma_\ell^* = \alpha + \textstyle\sum_{i=1}^{n} \phi_{\ell,i}, \tag{34}$$

    which simply moves it into the usual parameterization.

- The same logic applies to the variational topic Dirichlet,

$$\lambda^* = \eta + \sum_{\ell=1}^{d} \sum_{i=1}^{n} w_{\ell,i} \phi_{\ell,i}. \tag{35}$$

- For the multinomial, we use the log of the conditional again,

$$p(z_{\ell,i} = j) \propto \exp\{\mathrm{E}[\log \theta_\ell^j + \log \beta_j^{w_{\ell,i}}]. \tag{36}$$

- Needed fact. If $\theta \sim \mathrm{Dir}(\alpha)$ then $\mathrm{E}[\log \theta] = \Psi(\alpha) - \Psi(\sum_{j=1}^{k} \alpha_j)$ where Ψ is the first derivative of the $\log\Gamma$ function.

- This means that,

$$\phi_{\ell,i}^j \propto \exp\{\Psi(\gamma_{\ell,j}) + \Psi(\lambda_{j,w_{\ell,i}}) - \Psi(\textstyle\sum \lambda_{j,\cdot})\}, \tag{37}$$

  where we used that the second term of $\mathrm{E}[\log \theta]$ doesn't depend on j.

- We now have the final algorithm.

  1. Initialize each λ_j randomly.
  2. Repeat until the ELBO converges:
     (a) For each document ℓ:
        i. Initialize $\phi_{\ell,i} = 1/k$.
        ii. Repeat until the local ELBO converges:
           A. Update γ_ℓ from Equation 34.
           B. For each word, update $\phi_{\ell,i}$ from Equation 37.
     (b) For each topic j, update λ_j from Equation 36.

- Note how much easier this was to derive than in Blei et al. (2003).

- Also note that this is inefficient. We analyze every document with random topics before getting anywhere. We'll fix this later with stochastic optimization.

- (Show the inference examples from my slides.)

- (Show slides of extensions to simple LDA.)