Mixed-membership Models
(and an introduction to variational inference)

David M. Blei
Columbia University

November 24, 2015

Introduction

¶ We studied mixture models in detail, models that partition data into a collection of latent groups. We now discuss mixed-membership models, an extension of mixture models to *grouped data*. In grouped data, each "data point" is itself a collection of data; each collection can belong to multiple groups.

¶ Here are the basic ideas:

- 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 groups.
- The mixture proportions vary from group to group

We will see details later. For now, Figure 1 is the graphical model that describes these independence assumptions. This involves the following (generic) generative process,

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

A mixture model is a piece of this graphical model, but there is more to it. Intuitively, mixed-membership models capture 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 the model captures *homogeneity* and *heterogeneity*.

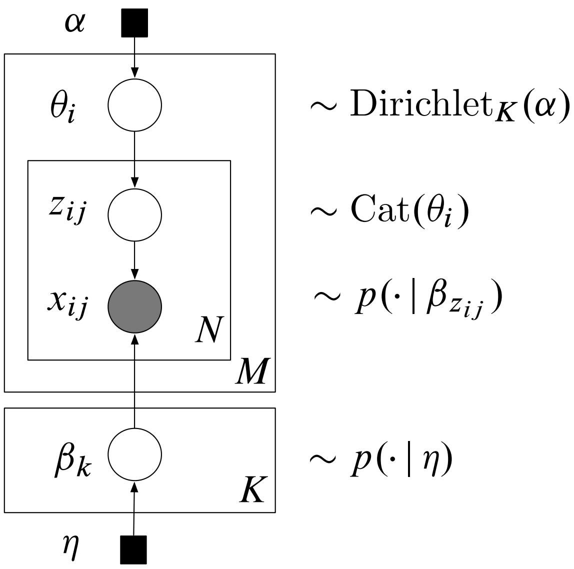

α
θ_i
$\sim \text{Dirichlet}_K(\alpha)$
z_{ij}
$\sim \text{Cat}(\theta_i)$
x_{ij}
$\sim p(\cdot \mid \beta_{z_{ij}})$
N
M
β_k
$\sim p(\cdot \mid \eta)$
K
η

Figure 1: The mixed-membership model.

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

- Observations are individual words.

- Groups are documents, i.e., collections of words.

- Components are distributions over the vocabulary, recurring patterns of observed words.

- Proportions are how much each document reflects each pattern.

- The posterior components look like "topics"—distributions that place their mass on words that exhibit a theme, such as sports or health. The proportions describe how each document exhibits those topics. For example, a document that is half about sports and half about health will place its proportions in those two topics.

- This algorithm has been adapted to all kinds of other data—images, computer code, music data, recommendation data, and 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. You can see why the GM is wrong—networks are 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 in the same 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. (Mixture models of social network data are called *stochastic block models*.)

- 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 nonparametrics (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

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

Notice that this is the normalized parameter vector, a point on the simplex.

¶ We will gain more intuition about the Dirichlet by looking at independent draws. An *exchangeable* Dirichlet is one where each parameter is the same scalar, $\text{Dir}(\alpha, \ldots, \alpha)$. Its expectation is always the uniform distribution.

Figure 2 shows example draws from the exchangeable Dirichlet (on the 10-simplex) with different values of α.

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.

¶ The Dirichlet is conjugate to the multinomial.

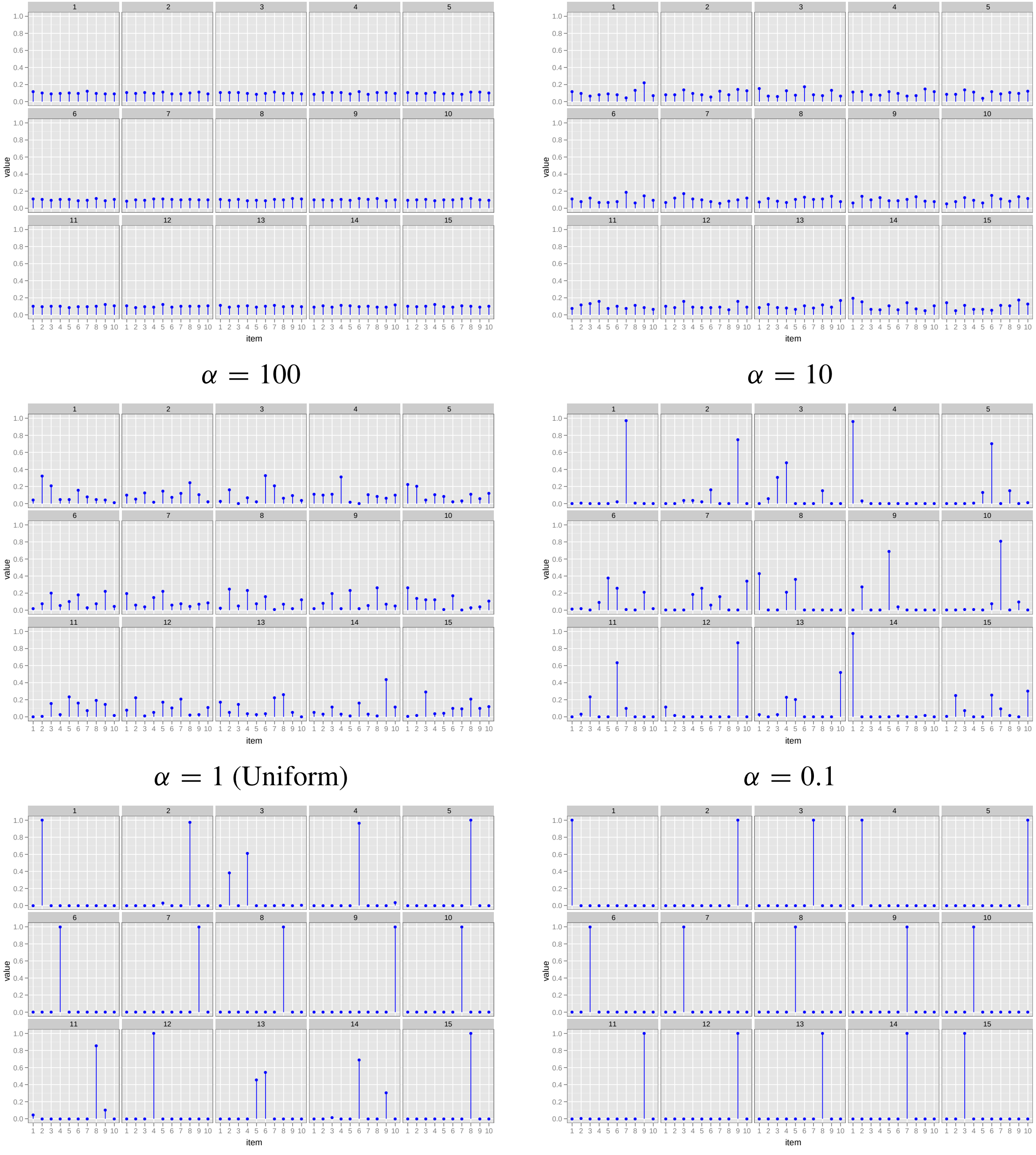
Let z be an indicator vector, i.e., a k-vector that contains a single one. The parameter to z is a point on the simplex θ, denoting the probability of each of the k items. The density function for z is

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

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

¶ Suppose we are in the following model,

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



$\alpha = 100$

$\alpha = 10$

$\alpha = 1$ (Uniform)

$\alpha = 0.1$

$\alpha = 0.01$ (Uniform)

$\alpha = 0.001$

Figure 2: Draws from the (exchangeable) Dirichlet distribution.

Let's compute the posterior distribution of θ,

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

We use the sum $\sum_{i=1}^{n} z_i^j = n_j$; it is the number of times item j appeared in $z_{1:n}$.

Eq. 11 is a Dirichlet distribution with parameter $\hat{\alpha}_j = \alpha_j + n_j$. It is the multivariate analog of our earlier result about the beta distribution.

¶ The expectation of the posterior Dirichlet is interesting,

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

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." (This interpretation is clearer when considering the n_0, x_0 parameterization of this prior; see the notes on exponential families.) The expectation reveals why—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 types of reasons for why we might 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 $w_{d,n} \sim \text{Mult}(\beta_{z_{d,n}})$.

This model is called latent Dirichlet allocation (LDA) (Blei et al., 2003).

¶ [R demo]

¶ 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. (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

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

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".

¶ (Optional): An exchangeable joint distribution is one that is invariant to permutation of its random variables. De Finetti's theorem says that if a collection of random variables are *exchangeable*, then their joint can be written as a "Bayesian model"

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

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

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

In natural language processing, this is called the "bag of words" assumption. Though commonly associated with independence, this assumption is really about exchangeability.

In topic modeling, this is palatable—we can still understand what a document is about (at a high level) even after shuffling its words.

Gibbs sampling in LDA

We derive the basic Gibbs sampler for LDA by calculating the complete conditionals.

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

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

- We used independencies that follow from the graphical model.
- The prior $p(\theta)$ disappears because it doesn't depend on $z_{d,n}$.
- In LDA, the second term is the probability of word $w_{d,n}$ in topic β_k. (We left it general here to enable other likelihoods.)

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

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

- 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 d.
- 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.)

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

- 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.

¶ [ALGORITHM]

¶ A better algorithm is the *collapsed Gibbs sampler*. It integrates out all latent variables except for **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 joint of the assignment and word,

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

Computing this joint gives us the collapsed Gibbs sampler.

We will integrate out the topic proportions θ_d and topic β_j to obtain an integrand independent of the other assignments and words. Given the proportions and topics, the joint distribution of a topic assignment and word is

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

We use this to compute Eq. 23. We short hand $z_{d,n} = k$ to $z_{d,n}$. We integrate out the topic and topic proportions,

$$\begin{aligned} p(z_{d,n} \,|\, \mathbf{z}_{-(d,n)}, \mathbf{w}) &\propto p(z_{d,n}, w_{d,n} \,|\, \mathbf{z}_{-(d,n)}, \mathbf{w}_{-(d,n)}) && (25) \\ &\propto \int_{\beta_k} \int_{\theta_d} p(\theta_d, \beta_k, z_{d,n}, w_{d,n} \,|\, \mathbf{z}_{-(d,n)}, \mathbf{w}_{-(d,n)}) && (26) \\ &= \int_{\beta_k} \int_{\theta_d} p(z_{d,n}, w_{d,n} \,|\, \theta_d, \beta_k) p(\theta_d \,|\, \mathbf{z}_{d,-n}) p(\beta_k \,|\, \mathbf{z}_{-(d,n)}, \mathbf{w}_{-(d,n)}) && (27) \\ &= \int_{\beta_k} \int_{\theta_d} \theta_{d,k} \beta_{k,\mathbf{w}_{d,n}} p(\theta_d \,|\, \mathbf{z}_{d,-n}) p(\beta_k \,|\, \mathbf{z}_{-(d,n)}, \mathbf{w}_{-(d,n)}) && (28) \\ &= \left(\int_{\theta_d} \theta_{d,k} p(\theta_d \,|\, \mathbf{z}_{d,-n}) \right) \left(\int_{\beta_k} \beta_{k,\mathbf{w}_{d,n}} p(\beta_k \,|\, \mathbf{z}_{-(d,n)}, \mathbf{w}_{-(d,n)}) \right). && (29) \end{aligned}$$

Each of these two terms are expectations of posterior Dirichlets.

- In line 2, $\mathbf{z}_{(-d,n)}$ became $\mathbf{z}_{d,-n}$. The proportions θ_d are independent of all assignments $\mathbf{z}_f$ where $f \neq d$.
- The first is like Eq. 20, but using all but $z_{d,n}$ to form counts.
- The second is like Eq. 22, but using all but $w_{d,n}$ to form counts.

The final algorithm is simple

$$p(z_{d,n} = k \,|\, \mathbf{z}_{-(d,n)}, \mathbf{w}) = \left(\frac{\alpha + n_d^k}{k\alpha + n_d} \right) \left(\frac{\eta + m_k^{w_{d,n}}}{v\eta + m_k} \right). \tag{30}$$

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 $w_{d,n}$.

¶ [ALGORITHM]

Mean-field variational inference

¶ LDA is a good testbed for variational inference (Jordan et al., 1999; Wainwright and Jordan, 2008), which is an alternative to MCMC for posterior inference. This was the original algorithm that we derived in Blei et al. (2003). However, we can now derive it in a much simpler way.

¶ Variational inference (VI) is a method of approximate inference. It is an alternative to Gibbs sampling, but is closely related. VI tends to be faster than MCMC, but there is substantially less theory. It is an active area of machine learning research.

We will describe VI in general, and then describe VI for topic models. Consider a general model $p(\mathbf{z}, \mathbf{x})$, where $\mathbf{x}$ are observations and $\mathbf{z}$ are hidden variables. Our goal is to calculate the posterior

$$p(\mathbf{z} \,|\, \mathbf{x}) = \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})}. \tag{31}$$

As we have seen, this is hard because the denominator is hard to compute. Gibbs sampling constructs a Markov chain whose stationary distribution is the posterior. Variational inference takes a different approach.

VI first posits a new distribution over the hidden variables $q(\mathbf{z}; \nu)$, indexed by *variational parameters* ν; note this defines a *family* of distributions over the hidden variables. VI then tries to find the value ν^* which indexes the distribution closest to the exact posterior. (Closeness is measured by Kullback-Leibler divergence.) VI turns the inference problem into an *optimization* problem. Turning computation into optimization is the hallmark of variational algorithms.

Once VI has found ν^*, it uses $q(\mathbf{z}; \nu^*)$ as a proxy for the posterior. The *fitted variational distribution* can be used to explore the data or to form posterior predictive distributions.

¶ An aside: Kullback-Leibler divergence. The KL divergence from $q(\mathbf{z}; \nu)$ to $p(\mathbf{z} \,|\, \mathbf{x})$ is

$$\mathrm{KL}\left(q(\mathbf{z}; \nu) || p(\mathbf{z} \,|\, \mathbf{x})\right) = \mathbb{E}_q\left[\frac{\log q(\mathbf{Z}; \nu)}{\log p(\mathbf{Z} \,|\, \mathbf{x})}\right]. \tag{32}$$

Alternatively,

$$\mathrm{KL}\left(q(\mathbf{z}; \nu) || p(\mathbf{z} \,|\, \mathbf{x})\right) = \mathbb{E}_q\left[\log q(\mathbf{Z}; \nu)\right] - \mathbb{E}_q\left[\log p(\mathbf{Z} \,|\, \mathbf{x})\right]. \tag{33}$$

We gain intuitions about KL by drawing a picture. Consider:

- Mass at $q(\cdot)$; no mass at $p(\cdot \,|\, \mathbf{x})$.
- Mass at $p(\cdot \,|\, \mathbf{x})$; no mass at $q(\cdot)$.
- Mass at both
- Two equal distributions have zero KL.

¶ We return to variational inference. The optimization problem is

$$\nu^* = \arg\min_{\nu} \mathrm{KL}\left(q(\mathbf{z}; \nu) || p(\mathbf{z} \mid \mathbf{x})\right) \tag{34}$$

First, we define the family of distributions. Many VI methods use the *mean-field family*, where each hidden variable is independent and governed by its own parameter. The mean-field distribution is

$$q(\mathbf{z}; \nu) = \prod_{i=1}^{m} q(z_i; \nu_i) \tag{35}$$

At first this looks funny—this is a "model" that contains no data and where nothing is shared between the variables. The idea is that Eq. 35 is a *family* of distributions; its connection to the data, specifically to the posterior, is through the optimization problem in Eq. 34.

The mean-field variational distribution is flexible in that it can capture any configuration of marginal distributions of the latent variables. However it is also limited in that it does not capture any dependencies between them. In general, latent variables are dependent in the posterior distribution.

If the family $q(\cdot; \nu)$ ranged over all distributions of $\mathbf{z}$ then the optimization problem in Eq. 34 would have its optimal at the posterior $p(\mathbf{z} \mid \mathbf{x})$. However, we would not be able to find this optimum—recall that we are doing variational inference because we cannot compute the posterior. The reason we limit the family is to facilitate the optimization.

To see how, expand the objective function,

$$\begin{aligned} \mathrm{KL}\left(q(\mathbf{z}; \nu) || p(\mathbf{z} \mid \mathbf{x})\right) &= \mathbb{E}_q\left[\log q(\mathbf{Z}; \nu)\right] - \mathbb{E}_q\left[\log p(\mathbf{Z} \mid \mathbf{x})\right] & (36) \\ &= \mathbb{E}_q\left[\log q(\mathbf{Z}; \nu)\right] - \mathbb{E}_q\left[\log p(\mathbf{Z}, \mathbf{x})\right] - \log p(\mathbf{x}). & (37) \end{aligned}$$

In variational inference we optimize the first two terms, i.e., the terms that depend on $q(\cdot; \nu)$. Taking these expectations is not possible without simplifying the variational family.

¶ We now describe a general algorithm for mean-field variational inference.

Traditionally, the variational objective is a quantity called the *evidence lower bound* (ELBO). It negates the first two terms from Eq. 37,

$$\begin{aligned} \mathcal{L} &= \mathbb{E}_q\left[\log p(\mathbf{Z}, \mathbf{x})\right] - \mathbb{E}_q\left[\log q(\mathbf{Z}; \nu)\right] & (38) \\ &= \mathbb{E}_q\left[\log p(\mathbf{Z})\right] + \mathbb{E}_q\left[\log p(\mathbf{x} \mid \mathbf{Z})\right] - \mathbb{E}_q\left[\log q(\mathbf{Z}; \nu)\right], & (39) \end{aligned}$$

and our goal is to maximize the ELBO. Maximizing the ELBO is equivalent to minimizing the KL divergence in Eq. 37. (The name ELBO comes from the fact that it is a lower bound on the *evidence*, $\log p(\mathbf{x})$. Many derivations of variational inference use the lower-bound perspective to develop the objective.)

Aside: The ELBO gives alternative intuitions about the variational objective function. The term $\mathbb{E}_q[\log p(\mathbf{x} \mid \mathbf{Z})]$ encourages $q(\cdot; \nu)$ to place its mass on configurations of $\mathbf{z}$ that explain the data. (This is tempered by the probability of the latent variables $\mathbb{E}_q[\log p(\mathbf{Z})]$.)

The last term $-\mathbb{E}_q[\log q(\mathbf{Z}; \nu)]$ is the entropy of the variational distribution. It "regularizes" the objective to prefer variational distributions that spread their mass across many configurations of the latent variables. Without this term, the objective would choose a variational distribution that placed all of its mass on the best configuration.

Note that the entropy decomposes in the mean-field family,

$$-\mathbb{E}_q[\log q(\mathbf{Z}; \nu)] = \sum_{i=1}^{m} \mathbb{E}_q[\log q(Z_i; \nu_i)]. \tag{40}$$

¶ We have now transformed approximate inference into an optimization problem. This opens the door to the (large) world of optimization techniques to help with computation in probabilistic models. For good references see Spall (2003); Boyd and Vandenberghe (2004); Kushner and Yin (1997). But for now we will use one method, coordinate ascent.

Coordinate ascent iteratively optimizes each variational parameter while holding the others fixed. Each step goes uphill in the ELBO; variational inference with coordinate ascent converges to a local optimum.

The coordinate update in mean-field variational inference is

$$q(z_i; \nu_i) \propto \exp\left\{\mathbb{E}_{q_{-i}}[\log p(z_i, \mathbf{Z}_{-i}, \mathbf{x})]\right\}, \tag{41}$$

where $q_{-i}(\mathbf{z}_{-i})$ is the mean-field distribution with the ith factor removed. This update says that the optimal variational factor for z_i is proportional to an exponentiated expected log joint with z_i fixed to its value.

Recall $p(z_i \mid \mathbf{z}_{-i}, \mathbf{x})$ is the complete conditional. (This is the distribution we sample from in the Gibbs sampler.) A trivial consequence of Eq. 41 is that

$$q(z_i; \nu_i) \propto \exp\left\{\mathbb{E}_{q_{-i}}[\log p(z_i \mid \mathbf{Z}_{-i}, \mathbf{x})]\right\} \tag{42}$$

This update reveals a connection between variational inference and Gibbs sampling.

Finally, suppose the complete conditional is in the exponential family,

$$p(z_i \mid \mathbf{z}_{-i}, \mathbf{x}) = \exp\left\{\eta_i(\mathbf{z}_{-i}, \mathbf{x})^\top z_i - a(\eta_i(\mathbf{z}_{-i}, \mathbf{x}))\right\}. \tag{43}$$

(This is the case for most of the models that we will study.) Now place ν_i in the same exponential family, i.e., it is a free parameter that indexes the family with the same dimension and the same log normalizer $a(\cdot)$. The coordinate update is

$$\nu_i = \mathbb{E}_{q_{-i}}[\eta_i(\mathbf{Z}_{-i}, \mathbf{x})] \tag{44}$$

Algorithm 1: Coordinate-ascent mean-field variational inference.

```
Input: A data set $\mathbf{x}$
Output: A variational distribution $q(\mathbf{z}; \nu) = \prod_{i=1}^{m} q(z_i; \nu_i)$
Initialize: Variational factors $q(z_i; \nu_i)$
while the ELBO has not converged do
    for $i \in \{1, \dots, m\}$ do
        Set $q(z_i; \nu_i) \propto \exp\{\mathbb{E}_{-i}[\log p(z_i \mid \mathbf{Z}_{-i}, \mathbf{x})]\}$
    end
    Compute ELBO $= \mathbb{E}[\log p(\mathbf{Z}, \mathbf{x})] + \mathbb{E}[\log q(\mathbf{Z})]$
end
return $q(\mathbf{z}; \nu)$
```

¶ The algorithm is in Algorithm 1.

¶ Note that the coordinate updates involve the complete conditional. Recall, from our lecture on Gibbs sampling, that this involves the *Markov blanket* of of the node z_i, i.e., its children, its parents, and the other parents of its children.

As for the Gibbs sampler, variational inference can also be seen as a message-passing algorithm. The variational parameters live on the nodes in the network; a node passes its "messages" to its neighbor when its neighbor is updating its parameter (Winn and Bishop, 2005).

¶ Let's return to LDA. The mean-field variational family is

$$q(\beta, \boldsymbol{\theta}, \mathbf{z}; \nu) = \prod_{k=1}^{K} q(\beta_k; \lambda_k) \prod_{d=1}^{D} \left(q(\theta_d; \gamma_d) \prod_{n=1}^{N} q(z_n; \varphi_{d,n}) \right). \tag{45}$$

The variational parameters are a V-Dirichlet distribution λ_k for each topic, a K-Dirichlet distribution γ_d for each document's topic proportions, and a K-categorical distribution $\varphi_{d,i}$ for each word in each document.

Consider each update in turn. The update for the variational topic assignment $\varphi_{d,n}$ applies the complete conditional in Eq. 24 in Eq. 42,

$$\varphi_{d,n} \propto \exp\left\{ \mathbb{E}_{\gamma_d}[\log \theta_d] + \mathbb{E}_{\lambda_k}\left[\log \beta_{\cdot, w_{d,n}}\right] \right\}. \tag{46}$$

In this update, $\beta_{\cdot, w_{d,n}}$ is the vector of probabilities of word $w_{d,n}$ under each of the topics. The expectations are

$$\mathbb{E}_{\gamma_d}[\log \theta_{d,k}] = \Psi(\gamma_{d,k}) - \Psi\left(\textstyle\sum_j \gamma_{d,j}\right) \tag{47}$$

$$\mathbb{E}_{\lambda_k}[\log \beta_{k,w}] = \Psi(\lambda_{k,w}) - \Psi\left(\textstyle\sum_v \lambda_{k,v}\right), \tag{48}$$

where $\Psi(\cdot)$ is the digamma function, the first derivative of $\log \Gamma(\cdot)$. (This function is available in most mathematical libraries.) These identities come from the exponential family representation of the Dirichlet in Eq. 2. The sufficient statistic is $\log \theta_k$, and so its expectation is the first derivative of the log normalizer.

Now we turn to the variational Dirichlet parameters. These updates come from the exponential family result in Eq. 44. For the variational topic proportions γ_d, we take the expectation of Eq. 20,

$$\gamma_d = \alpha + \sum_{n=1}^{N} \varphi_{d,n}. \tag{49}$$

(Note that the expectation of an indicator vector is its vector of probabilities.) For each document, this is a "posterior Dirichlet." The second term counts the expected number of times each topic appears in each document.

Similarly, the variational update for the topics comes from Eq. 22,

$$\lambda_k = \eta + \sum_{d=1}^{D} \sum_{n=1}^{N} \varphi_{d,n,k} w_{d,n}. \tag{50}$$

The second term counts the expected number of times each word appears in each topic.

¶ [ALGORITHM]

¶ Final notes:

- VI does not find a global optimum of the KL; it converges to a local optimum; it is sensitive to the starting point.

- We can move beyond the mean-field, finding structured variational distributions that account for posterior dependence in the latent variables.

- We can also move beyond assumptions around the complete conditional. The ADVI algorithm in Stan that Alp presented is an example of this. (There are others, e.g., that connect neural network research to variational methods.)

- There are many open theoretical problems around variational inference.

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.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

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.

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999). Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233.

Kushner, H. and Yin, G. (1997). *Stochastic Approximation Algorithms and Applications*. Springer New York.

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

Spall, J. (2003). *Introduction to stochastic search and optimization: Estimation, simulation, and control*. John Wiley and Sons.

Wainwright, M. and Jordan, M. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305.

Winn, J. and Bishop, C. (2005). Variational message passing. *Journal of Machine Learning Research*, 6:661–694.