6.882: Bayesian Modeling and Inference **Lecture Date: February 08, 2012**

Latent Dirichlet Allocation

Lecturer: Tamara Broderick *Scribe: Ajinkya Ghorpade and Rahul Singh*

Outline

Last Time:

- Bayes theorem and Bayesian Methods
- Benefits and Challenges of Bayes
- de Finetti's Theorem
- Posterior Approximation

Today:

- Latent Dirichlet Allocation (Blei et al. (2003))
  - Model
  - Posterior Approximation

In the previous lecture, we outlined the Bayesian framework. A model consists of a prior over latent parameters and a likelihood of observations given parameters. Given the assumption of infinite exchangeability, de Finetti's theorem motivates this approach. In today's lecture, we apply Bayesian methods to the problem of modelling corpora. Exchangeability at the level of words within documents corresponds to the 'bag-of-words' assumption. We will study models of increasing complexity leading to latent Dirichlet allocation (LDA). For LDA, we will approximate the posterior over latent parameters using mean field variational Bayes (MFVB).

1 Model

1.1 Notation

We introduce the following notation:

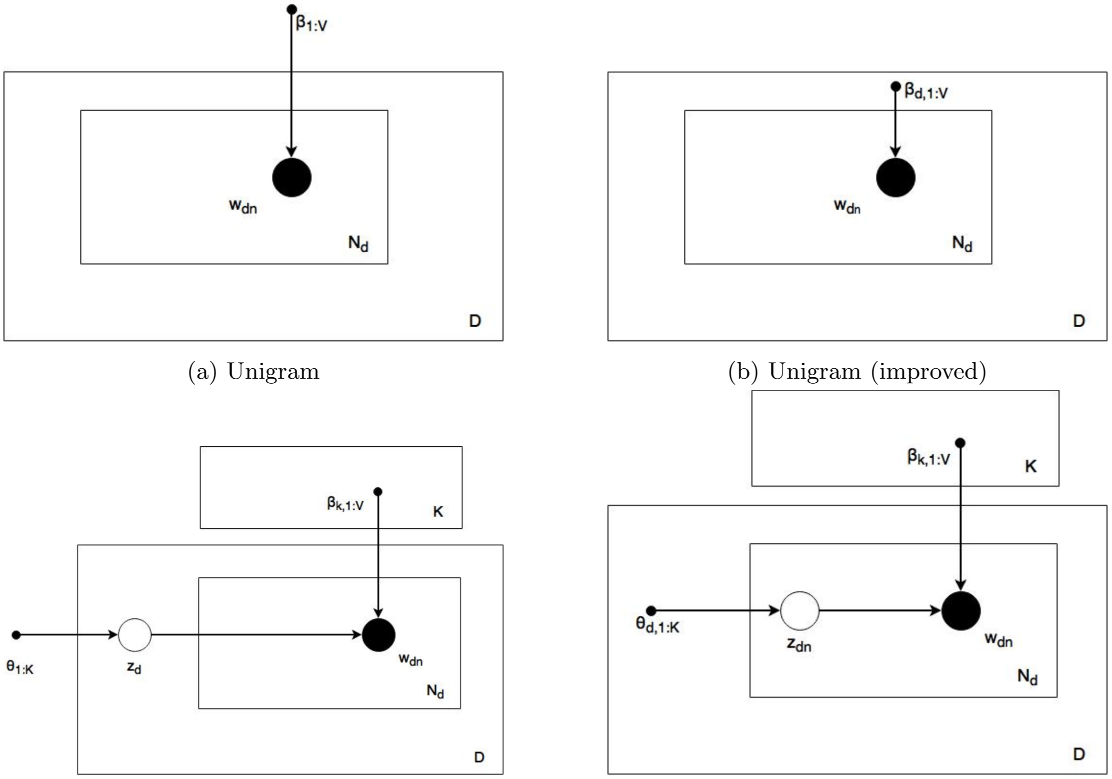
- V - # words in vocabulary
- D - # documents
- N_d - # words in document d
- w_{dn} - the n^{th} word in document d. $w_{dn} \in \{1, \ldots, V\} =: [V]$.

1.2 Graphical Models Review

A graphical model is a diagrammatic representation of the probabilistic relationships among random variables. Nodes represent random variables while edges represent statistical dependence. In this lecture, we use directed edges to represent conditional dependence. Plate notation represents repeated sub-graphs. Graphical model semantics are summarized below:

- ● - Observed variable (observation)
- ○ - Latent variable (parameter)
- • - Constant (hyperparameter)
- □ - Repeat the sub-graph inside the plate
- ⟶ - The direction of the conditional dependence between variables

For further reading on directed graphical models, we recommend Bishop (2006) or (Murphy, 2012, Chapter 8).



(a) Unigram

(b) Unigram (improved)

(c) Mixture of Unigrams

(d) pLSI

Figure 1: Graphical Models for Corpora

1.3 Unigram Model

We begin with a simple model of the data generating process behind the corpora. In the unigram model, we assume that each word is drawn i.i.d. from a categorical distribution i.e., $w_{dn} \sim Cat(\beta_{1:V})$. Under this model, every document has the same distribution over the vocabulary. Figure 1a shows the graphical representation of the unigram model.

But in general we do not expect documents all be the same. And indeed to discover topics, we expect that we will have to observe these differences. We can define a new model, similar to the unigram model, such that each document has its own distribution over the vocabulary, i.e. $w_{dn} \sim Cat(\beta_{d,1:V})$. This unigram model is shown in Figure 1b. Trying to learn the $\beta_{d,1:V}$ would lead to massive overfitting. Also this model does not capture our intuition that words tend to covary across documents; "phylum" and "genus" are more likely to occur together. "Supernova" and "redshift" and more likely to occur together.

1.4 Mixture of Unigrams

We address the drawbacks of the unigram model by defining topics $\beta_{k,1:V}$ separate from the documents and giving it a prior as a form of regularization to avoid overfitting. (See Nigam et al 2000.) Figure 1c shows the corresponding graphical model. For each document, we let z_d represent the topic to which this document is assigned. And we let $\theta_{1:K}$ be the global distribution over topics.

$$z_d \sim Cat(\theta_{1:K})$$

$$w_{dn} \sim Cat(\beta_{z_d,1:K})$$

That is, each document has one topic, and words in that document are drawn from that topic's distribution over the vocabulary. This model is similar to a clustering model, where we choose K to be the number of topics and then assign each document to one of these K topics.

Often the assumption that a document has a single topic is too stringent. For example, a news article about the space program may contain words drawn from a mixture of topics related to technology, astronomy, government, etc.

1.5 (Almost) Probabilistic Latent Semantic Indexing (pLSI)

Next, we consider (a slight change to) the probabilistic latent semantic indexing (pLSI) model, which allows documents to have distributions over topics. Now we imagine each document has its own topic distribution $\theta_{d,1:K}$. For each word,

$$z_{dn} \sim Cat(\theta_{d,1:K})$$

$$w_{dn} \sim Cat(\beta_{z_{dn},1:V})$$

In this model, each word has a latent topic indicator. Suppose we are focusing on document d. N_d latent topics indicators are chosen from that document's distributions over topics. For each topic indicator, we draw a word from the corresponding topic's distribution over the vocabulary.

Since each document has its own distribution over topics determined by an idiosyncratic constant, trying to learn the $\theta_{d,1:K}$ would lead to serious overfitting. One solution will be to treat the $\theta_{d,1:K}$ as parameters and put a prior on them.

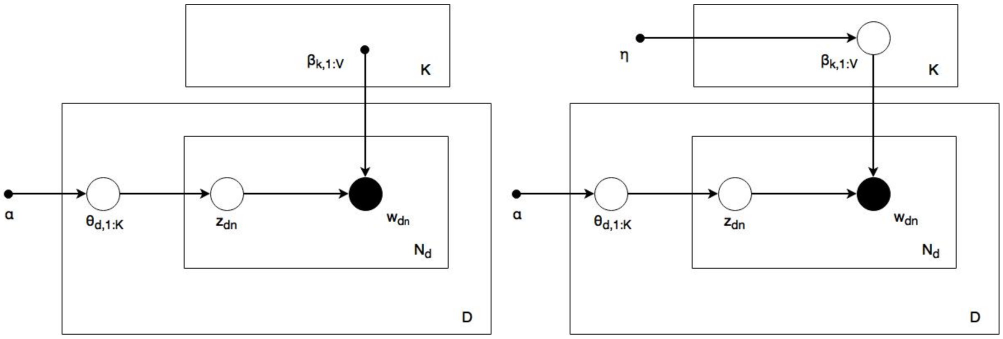
1.6 Latent Dirichlet Allocation (LDA)

LDA was proposed by Blei et al. (2003). The same model was proposed by Pritchard et al. (2000) for modelling population structure using genotype data. LDA addresses the overfitting problem of pLSI by incorporating a prior—that is, by assuming that each document's distribution over topics is a latent parameter drawn i.i.d. from a distribution parametrized by a shared constant, i.e. $\theta_{d,1:K} \sim Dir(\alpha)$. Here we overload the Dir notation slightly; we imagine α is a positive real value and suppose that the input to the K-parameter Dirichlet distribution is actually α repeated K times. The Dirichlet distribution is conjugate to the categorical distribution, which makes it a natural choice (and necessary choice for the basic MFVB-CA algorithm). Figure 2a displays this innovation.

Similarly, one can consider each topic's distribution over the vocabulary to be a latent parameter drawn i.i.d. from a distribution parametrized by a shared constant, i.e. $\beta_{k,1:V} \sim Dir(\eta)$. Again, η is positive real-valued. Figure 2b displays this further innovation, which we hereafter call LDA.

Now that we have defined the constants and parameters in LDA, we suppress subscripts that are not restrictive, e.g. we write $\beta_{k,1:V}$ as β_k. To summarize the generative model in this lighter notation:

```
1. for $k = 1 : K$
    (a) Draw $\beta_k \sim Dir(\eta)$
2. for $d = 1 : D$
    (a) Draw $\theta_d \sim Dir(\alpha)$
    (b) for $n = 1 : N_d$
        i. Draw $z_{dn} \sim Cat(\theta_d)$
        ii. Draw $w_{dn} \sim Cat(\beta_{z_{dn}})$
```



(a) LDA

(b) LDA (improved)

Figure 2: Graphical Models of LDA

2 Posterior Approximation

As Bayesians, we are interested in the posterior over parameters $p(\theta, \beta, z|w)$ derived from the generative model. One reason why the LDA posterior is intractable is the high dimensionality of the parameters. Recall that

- $dim(\beta) = KV$ i.e. (# topics) × (size vocab) $\sim 100 \times 10{,}000$
- $dim(\theta) = DK$ i.e. (# documents) × (# topics)

- $dim(z) = \sum_{d=1}^{D} N_d$ i.e. total # words across documents

This calls for an approximate inference method, and we choose mean field variational Bayes (MFVB).

2.1 Mean Field Variational Bayes

In the following, we will let x refer to data and θ refer to the concatenation of model parameters. Bayes rule is

$$p(\theta|x) \propto p(x|\theta)p(\theta)$$

Variational Bayes (VB) approximates the posterior $p(\theta|x)$ with the distribution $q^*(\theta)$, contrained to class Q, that is closest to $p(\theta|x)$ in the sense of Kullback-Leibler divergence.

$$q^* = \text{argmin}_{q \in Q} KL[q(\cdot)||p(\cdot|x)]$$

where

$$\begin{aligned} KL[q(\cdot)||p(\cdot|x)] &:= \int q(\theta) ln\left(\frac{q(\theta)}{p(\theta|x)}\right) d\theta \\ &= \int q(\theta) ln\left(\frac{p(x)q(\theta)}{p(x,\theta)}\right) d\theta \\ &= ln(p(x)) + \int q(\theta) ln\left(\frac{q(\theta)}{p(x,\theta)}\right) d\theta \\ &= ln(p(x)) - \int q(\theta) ln\left(\frac{p(x,\theta)}{q(\theta)}\right) d\theta \\ &= \ln(\text{evidence}) - \text{ELBO}(q) \end{aligned}$$

The last line uses the definitions

$$\begin{aligned} \text{evidence} &:= p(x) \\ \text{ELBO}(q) &:= \int q(\theta) ln\left(\frac{p(x,\theta)}{q(\theta)}\right) d\theta \end{aligned}$$

The abbreviation ELBO stands for evidence lower bound. By the derivation above,

$$\begin{aligned} \ln(\text{evidence}) &= \text{ELBO}(q) + KL[q(\cdot)||p(\cdot|x)] \\ \ln(\text{evidence}) &\geq \text{ELBO}(q) \end{aligned}$$

since $KL[\cdot||\cdot] \geq 0$. (An aside: $KL[\cdot||\cdot] \geq 0$ due to Jensen's inequality. One may also use Jensen's inequality directly to show that evidence $\geq$ ELBO(q).)

An alternative interpretation of VB, then, is constrained maximization of the variational lower bound on the evidence: the ELBO.

$$q^* = \text{argmax}_{q \in Q} \text{ELBO}(q) \tag{1}$$

MFVB is a type of VB in which Q is specified to factor over components θ_j that partition θ. Note that θ_j may be non-singleton.

$$Q = \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\}$$

The factorization of q is an approximation that permits us to simplify ELBO(q). In particular, it will allow us to reformulate (1) as coordinate ascent in coordinates $q_j := q_j(\theta_j)$.

$$\begin{aligned}
\text{ELBO}(q) &:= \int q(\theta) ln\bigg(\frac{p(x,\theta)}{q(\theta)}\bigg) d\theta \\
&= \int \bigg[\prod_{i=1}^{J} q_i\bigg]\bigg[ln(p(x,\theta)) - \sum_{k=1}^{J} ln(q_k)\bigg] d\theta \\
&\simeq \int q_j \int \bigg[\prod_{i \neq j} q_i\bigg] ln(p(x,\theta)) d\theta_{i\neq j} d\theta_j - \int q_j ln(q_j) d\theta_j \\
&\simeq \int q_j ln(\tilde{p}(\theta_j)) d\theta_j - \int q_j ln(q_j) d\theta_j \\
&= \int q_j ln\bigg(\frac{\tilde{p}(\theta_j)}{q_j}\bigg) d\theta_j \\
&= -KL[q_j || \tilde{p}]
\end{aligned}$$

where $\simeq$ means up to an additive constant in q_j. Note that in the fourth line, we define $\tilde{p}(\theta_j)$ as the distribution that satisfies

$$ln(\tilde{p}(\theta_j)) \simeq \int \bigg[\prod_{i\neq j} q_i\bigg] ln(p(x,\theta)) d\theta_{i\neq j} = \mathbb{E}_{q_{i\neq j}} ln(p(x,\theta))$$

Hence ascent in coordinate q_j amounts to finding q_j^* where

$$\begin{aligned}
q_j^* &= \text{argmax}_{q_j} \text{ELBO}(q) \\
&= \text{argmin}_{q_j} KL[q_j || \tilde{p}] \\
&= \tilde{p}(\theta_j) \\
&\propto exp\bigg(\mathbb{E}_{q_{i\neq j}} ln(p(x,\theta))\bigg) \qquad (2)
\end{aligned}$$

MFVB consists of iteratively performing update (2) over $j = 1, ..., J$.

2.2 MFVB for LDA

We now approximate the exact posterior $p(\theta, z, \beta | w)$ with $q^*(\theta, z, \beta)$. LDA has the following log joint distribution, suppressing dependence on constants α and η.

$$\begin{aligned}
ln(p(\theta,z,\beta,w)) &= \sum_{d,n} ln(p(w_{dn}|z_{dn},\beta)) + \sum_{d,n} ln(p(z_{dn}|\theta_d)) + \sum_{d} ln(p(\theta_d)) + \sum_{k} ln(p(\beta_k)) \\
&\simeq \sum_{d,n}\sum_{k} 1\{z_{dn} = k\} ln(\beta_{kw_{dn}}) \\
&\quad + \sum_{d,n}\sum_{k} 1\{z_{dn} = k\} ln(\theta_{dk}) \\
&\quad + \sum_{d,k} (\alpha - 1) ln(\theta_{dk}) \\
&\quad + \sum_{k,v} (\eta - 1) ln(\beta_{kv})
\end{aligned}$$

where $\simeq$ means up to an additive constant in any parameter. In the first line, we factor the joint distribution according to the child-parent relationships in the directed acyclic graph (DAG). In the second line, we plug in the particular conditional distribution for each child-parent relationship according to the generative model.

We make the MFVB approximation that q factors as

$$q(\theta, z, \beta) = \prod_{d,n} q(z_{dn}) \prod_{d} q(\theta_d) \prod_{k} q(\beta_k) \tag{3}$$

Each factor in the RHS of (3) can be considered a particular q_j. Why not factor the θ and β terms further? Remember that the components of θ_d must sum to one as they form a distribution. Similarly for the β_k. Thus our last step is to tailor the update equation (2) for each one. The approximate posterior components turn out to be categorical and Dirichlet distributions without any further assumptions!

1. z_{dn}

$$\begin{aligned} ln(q^*(z_{dn})) &= \mathbb{E}_{-z_{dn}} ln(p(\theta, z, \beta, w)) \\ &\simeq \sum_k 1\{z_{dn} = k\} \mathbb{E}_{q^*}\Big[ln(\beta_{kw_{dn}}) + ln(\theta_{dk}) \Big] \end{aligned}$$

where $\simeq$ means up to an additive constant in z_{dn}. Therefore

$$\begin{aligned} \phi_{dnk} &:= q^*(z_{dn} = k) \\ &\propto exp\Bigg(\mathbb{E}_{q^*}\Big[ln(\beta_{kw_{dn}}) + ln(\theta_{dk}) \Big] \Bigg) \end{aligned}$$

2. θ_d

$$\begin{aligned} ln(q^*(\theta_d)) &= \mathbb{E}_{-\theta_d} ln(p(\theta, z, \beta, w)) \\ &\simeq \sum_{k,n} [\mathbb{E}_{q^*} 1\{z_{dn} = k\}] ln(\theta_{dk}) + \sum_k (\alpha - 1) ln(\theta_{dk}) \\ &= \sum_{k,n} \phi_{dnk} ln(\theta_{dk}) + \sum_k (\alpha - 1) ln(\theta_{dk}) \\ &= \sum_k \Bigg(\alpha - 1 + \sum_n \phi_{dnk} \Bigg) ln(\theta_{dk}) \end{aligned}$$

where $\simeq$ means up to an additive constant in θ_d. Therefore

$$\begin{aligned} q^*(\theta_d) &= Dir(\theta_d | \gamma_d) \\ \gamma_{dk} &= \alpha + \sum_n \phi_{dnk} \end{aligned}$$

3. β_k

$$\begin{aligned} q^*(\beta_k) &= Dir(\beta_k | \lambda_k) \\ \lambda_{kv} &= \eta + \sum_{d,n} 1\{w_{dn} = v\} \phi_{dnk} \end{aligned}$$

by an argument analogous to that for θ_d

References

Bishop, C. M. (2006). Graphical models. *Pattern recognition and machine learning*, 4:359–422.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.