UROP Notes

Michael Diao

Contents

1 LDA

The goal of LDA is to provide a probabilistic model for "documents," which consist of "words" from different latent "topics." With some simplifying assumptions such as **exchangeability** of words (and documents within "corpora"), LDA aims to represent documents as mixtures of topics, which are distributions of words.

> **Remark 1.1** — LDA is motivated by the de Fitnetti representation theorem, which states that any collection of exchangeable random variables has a representation as a mixture distribution (in general, an infinite mixture).

1.1 Setup

We can represent LDA by the graphical model in Figure 1. It has the following hierarchical specification:

- We observe a corpus $\mathcal{D}$ of D documents $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_D$.
- Each of these documents is a vector of N words $\langle w_1, w_2, \ldots, w_N \rangle$.
- The words w_{dn} are generated by latent topics z_{dn}, which are generated according to a distribution θ_d. We have a Dirichlet prior with parameter α on the distributions θ_d.[1] Specifically,

$$\begin{aligned} \theta_d &\sim \text{Dirichlet}(\alpha) \\ z_{dn} \mid \theta_d &\sim \theta_d \\ w_{dn} \mid z_{dn} &\sim \beta_{z_{dn}} \end{aligned}$$

There are some nuances about this model. For instance, do we know the dimensionality K of θ (and thus the number of topics), and if not, how do we do updates? What if N is not fixed?[2] What shape does β take?[3] We can discuss generalizations in more detail, but for now assume that K is known and N is fixed.[4]

The Dirichlet density may be written as follows:

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \left(\prod_{k=1}^{K} \theta_k^{\alpha_k - 1}\right).$$

[1] Why Dirichlet? This is for convenience reasons. It is an exponential family, and it is conjugate to the multinomial distribution (which we are using for topics).
[2] Add a prior on N, e.g. Poisson.
[3] β is a $K \times V$ matrix where V is the size of the dictionary.
[4] N in particular is easy to deal with, since it is independent of α, θ, z so its randomness can easily be ignored.

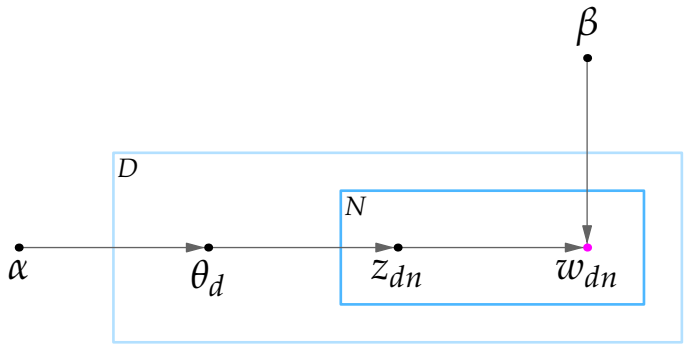

Figure 1: Graphical model representation of LDA.

Given parameters α and β, the joint distribution of $\theta, \mathbf{z}, \mathbf{w}$ ($\mathbf{z}, \mathbf{w}$ are N-dimensional, representing a document) is given by

$$\begin{aligned} p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) &= p(\theta \mid \alpha) \cdot p(\mathbf{z} \mid \theta) \cdot p(\mathbf{w} \mid \mathbf{z}, \beta) \\ &= p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) \cdot p(w_n \mid z_n, \beta). \end{aligned}$$

Then we can retrieve the marginal distribution of $\mathbf{w}$ conditioned on α, β by integrating over θ and summing over $\mathbf{z}$, and thus the probability of a given corpus. We have

$$\begin{aligned} p(\mathbf{w} \mid \alpha, \beta) &= \int_{\Theta} p(\theta \mid \alpha) \sum_{\mathbf{z} \in \mathcal{Z}} p(\mathbf{z} \mid \theta) \cdot p(\mathbf{w} \mid \mathbf{z}, \beta) \, \mathrm{d}\theta \\ &= \int_{\Theta} p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_n \in \mathcal{Z}} p(z_n \mid \theta) \cdot p(w_n \mid z_n, \beta) \right) \mathrm{d}\theta \, . \end{aligned}$$

Therefore, we get

$$\begin{aligned} p(\mathcal{D} \mid \alpha, \beta) &= \prod_{d=1}^{D} p(\mathbf{w}_d \mid \alpha, \beta) \\ &= \prod_{d=1}^{D} \int_{\Theta} p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_{dn} \in \mathcal{Z}} p(z_{dn} \mid \theta_d) \cdot p(w_{dn} \mid z_{dn}, \beta) \right) \mathrm{d}\theta_d \, . \end{aligned}$$

Remark 1.2 — It is important to distinguish LDA from a Dirichlet-multinomial clustering model. The key is that LDA uses an **admixture**, in which a document can be a mixture of topics, whereas a classical clustering model would restrict a document to be associated with a single topic.

There are other ways to interpret LDA. By marginalizing over the latent topic variable z, we can write LDA as a two-level model rather than a three-level one. Indeed, we can

generate a document by choosing $\theta \sim \text{Dir}(\alpha)$ and sampling each word from $p(w \mid \theta, \beta)$, which is given by

$$p(w \mid \theta, \beta) = \sum_{z \in \mathcal{Z}} p(w \mid z, \beta) \cdot p(z \mid \theta)$$
$$\implies p(\mathbf{w} \mid \alpha, \beta) = \int_{\Theta} p(\theta \mid \alpha) \left(\prod_{n=1}^{N} p(w_n \mid \theta, \beta) \right) \mathrm{d}\theta \,.$$

Question 1.3. We have established the motivation and setup behind LDA. Now, how do we perform inference and parameter estimation?

1.2 Inference

In order to use LDA, we must be able to obtain the posterior distribution of the latent variables $\theta, \mathbf{z}$. However, this distribution is intractable because we have

$$p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{p(\mathbf{w} \mid \alpha, \beta)}$$

and computing $p(\mathbf{w} \mid \alpha, \beta)$ is infeasible (NP-hard!) in general.[5] Therefore, we resort to approximation methods to estimate the posterior instead. Prominent ways include Laplace approximation, MCMC, variational inference. For the rest of the section, we will deal with variational inference in particular.

The key idea of variational inference is that we want to approximate the intractable true posterior with a more computationally reasonable distribution. To do this, we consider a simplified graphical model (Figure 2), which removes the edges between $\theta, \mathbf{z}$ and drops the $\mathbf{w}$ nodes. This then gives us a much simpler posterior

$$q(\theta, \mathbf{z} \mid \gamma, \varphi) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \varphi_n).$$

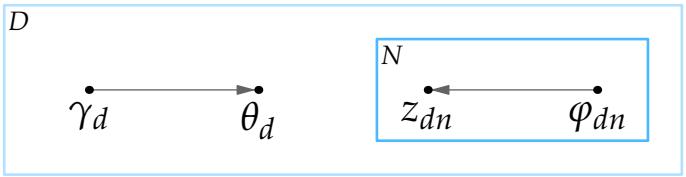

Figure 2: Simplified graphical model with free variational parameters (γ, φ). γ is a Dirichlet parameter, while $\varphi = \langle \varphi_1, \ldots, \varphi_N \rangle$ are multinomial parameters.

We then proceed to do a variant of the EM algorithm as follows:

[5] Proven NP-hard by Sontag and Roy for $\alpha \ll 1$.

1. We initialize our parameters α, β.

2. **E-step**.

   For each document $\mathbf{w}$, we compute

$$(\gamma^*(\mathbf{w}), \varphi^*(\mathbf{w})) = \underset{(\gamma,\varphi)}{\operatorname{argmin}} D\Big(q_{\theta,\mathbf{z}|\gamma,\varphi} \parallel p_{\theta,\mathbf{z}|\mathbf{w},\alpha,\beta}\Big),$$

   which corresponds to an I-projection of $p_{\theta,\mathbf{z}|\mathbf{w},\alpha,\beta}$ onto $\mathcal{Q}$, the family of distributions q. This can be done via an iterative fixed-point method. In particular, we have the update equations

$$\begin{aligned}\varphi_{nk} &\propto \beta_{kw_n} \exp\Big(\mathbb{E}_{q_{\theta|\gamma}}[\log \theta_k \mid \gamma]\Big) \\ \gamma_k &= \alpha_k + \sum_{n=1}^{N} \varphi_{nk}.\end{aligned}$$

   These are obtained by setting the derivative of KL divergence equal to zero, but they have an intuitive interpretation as well. The update for φ_{nk} can be thought of in terms of Bayes' Theorem. Indeed, we have

$$\begin{aligned}\phi_{nz_n} = \mathbb{P}(\mathsf{z}_n = z_n \mid \mathsf{w}_n = w_n) &= p(z_n \mid w_n) \\ &\propto p(w_n \mid z_n) \cdot p(z_n) \\ &\approx \beta_{z_n w_n} \exp\Big(\mathbb{E}_{q_{\theta,\mathbf{z}|\gamma,\varphi}}[\log q(z_n) \mid \gamma, \varphi]\Big) \\ &= \beta_{z_n w_n} \exp\Big(\mathbb{E}_{q_{\theta|\gamma}}[\log \theta_{z_n} \mid \gamma]\Big),\end{aligned}$$

   giving the first equation. On the other hand, the update for γ_k can be thought of as a Dirichlet update with

$$\theta \mid \gamma \sim \text{Dirichlet}\left(\alpha + \sum_{n=1}^{N} \phi_n\right).$$

   This gives us Algorithm 1.4.

   The line $\varphi_{nk}^{t+1} \leftarrow \beta_{kw_n} \exp(\Psi(\gamma_i^t))$ follows from the fact that

$$\begin{aligned}\varphi_{nk} &\propto \beta_{kw_n} \exp\Big(\mathbb{E}_{q_{\theta|\gamma}}[\log \theta_k \mid \gamma]\Big) \\ &= \beta_{kw_n} \exp\left(\Psi(\gamma_k) - \Psi\left(\sum_{k=1}^{K} \gamma_k\right)\right) \\ &\propto \beta_{kw_n} \exp(\Psi(\gamma_k)).\end{aligned}$$

   This completes the E-step.

Algorithm 1.4 Determining $\gamma^*(\mathbf{w}), \varphi^*(\mathbf{w})$.

$\varphi_n \leftarrow$ uniform over $1, \dots K$ for $n = 1, \dots, N$
$\gamma_k \leftarrow \alpha_k + \frac{N}{K}$ for $k = 1, \dots, K$
repeat
  for $n = 1, \dots, N$ **do**
    for $k = 1, \dots, K$ **do**
      $\varphi_{nk}^{t+1} \leftarrow \beta_{kw_n} \exp(\Psi(\gamma_i^t))$
    end for
    normalize φ^{t+1}
  end for
  $\gamma^{t+1} \leftarrow \alpha + \sum_{n=1}^{N} \varphi_n^{t+1}$
until convergence
return φ, γ

3. **M-step**.

   We wish to maximize the log-likelihood of the data with respect to the model parameters α, β, which is

$$\log p(\mathcal{D} \mid \alpha, \beta) = \sum_{d=1}^{D} \log p(\mathbf{w}_d \mid \alpha, \beta) = \sum_{d=1}^{D} \sum_{\mathbf{z}} \int_{\Theta} \log p(\theta, \mathbf{z}, \mathbf{w}_d \mid \alpha, \beta) \, \mathrm{d}\theta \, .$$

   For a single document, we can compute the joint distribution as

$$\begin{aligned} \log p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) &= \log \left(\prod_{d=1}^{D} \left[p(\theta_d \mid \alpha) \prod_{n=1}^{N} p(z_{dn} \mid \theta_d) \cdot p(w_{dn} \mid z_{dn}, \beta) \right] \right) \\ &= \sum_{d} \log p(\theta_d \mid \alpha) + \sum_{d,n} \log p(z_{dn} \mid \theta_d) \cdot p(w_{dn} \mid z_{dn}, \beta). \end{aligned}$$

   Computing this joint probability is at least tractable (if not, then we have a problem), unlike $p(\mathbf{w} \mid \alpha, \beta)$ as we originally described. Now we want to use our variational approximation by incorporating q. We see that by Jensen's inequality,

$$\begin{aligned} \log p(\mathbf{w} \mid \alpha, \beta) &= \sum_{\mathbf{z}} \int_{\Theta} \log p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) \, \mathrm{d}\theta \\ &= \log \mathbb{E}_{q_{\theta, \mathbf{z} \mid \gamma, \varphi}} \left[\frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{q(\theta, \mathbf{z} \mid \gamma, \varphi)} \right] \\ &\geq \mathbb{E}_{q_{\theta, \mathbf{z} \mid \gamma, \varphi}} \left[\log \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{q(\theta, \mathbf{z} \mid \gamma, \varphi)} \right] \\ &= \mathbb{E}_{q_{\theta, \mathbf{z} \mid \gamma, \varphi}} [\log p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)] + h\left(q_{\theta, \mathbf{z} \mid \gamma, \varphi} \right) \\ &\stackrel{\text{def}}{=} \mathcal{L}(\mathbf{w}; \alpha, \beta, \gamma, \varphi). \end{aligned}$$

We call this lower bound the **Evidence Lower Bound** or **ELBO**. The gap between $\mathcal{L}$ and the true log marginal is

$$\log p(\mathbf{w} \mid \alpha, \beta) - \mathcal{L}(\mathbf{w}; \alpha, \beta, \gamma, \varphi) = D\Big(q_{\theta,\mathbf{z}|\gamma,\varphi} \parallel p_{\theta,\mathbf{z}|\mathbf{w}}\Big).$$

The bound is tight iff our variational approximation $q_{\theta,\mathbf{z}|\gamma,\varphi}$ matches the real distribution over a given document $p_{\theta,\mathbf{z}|\mathbf{w}}$.

Accordingly, we update α and β to maximize the ELBO.

4. We repeat the E-step and M-step until convergence.

One issue is that sometimes rare words are given probability zero in the multinomial parameters β. We thus want to "smooth" these parameters by assigning all relevant words a nonzero probability. We can try to amend this problem by placing a Dirichlet prior on the multinomial parameter β. This will give us an intractable posterior on β, so we will need to extend our use of approximation techniques such as VI, for instance by considering the separable density

$$q(\beta, \theta, \mathbf{z} \mid \lambda, \gamma, \varphi) = \prod_{k=1}^{K} \mathrm{Dir}(\beta_k \mid \lambda_k) \prod_{d=1}^{D} q_d(\theta_d, \mathbf{z}_d \mid \gamma_d, \varphi_d).$$

Here q_d is the variational distribution defined for LDA above. This gives us the original update equations as well as a new update for λ.

2 RCTD

Our ultimate goal is to determine the fractional contributions of each cell type to a particular sample. We do this by maximum likelihood, using the following hierarchical model:

$$\begin{aligned} Y_{i,j} \mid \lambda_{i,j} &\sim \text{Poisson}(N_i \lambda_{i,j}) \\ \log \lambda_{i,j} &= \log(\vec{\beta}_i \cdot \vec{\mu}_j) + \alpha_i + \gamma_j + \varepsilon_{i,j}, \end{aligned}$$

where

- $Y_{i,j}$ is the random variable corresponding to the observed expression of gene j at pixel i,
- N_i is the number of transcripts for pixel i,
- $\vec{\beta}_i$ is the K-dimensional row vector of contributions from each cell type (where K is the number of cell types in question) at pixel i,
- $\vec{\mu}_j$ is the K-dimensional column vector of mean expressions of gene j for each cell type,
- α_i is a fixed pixel-specific effect.
- γ_j and $\varepsilon_{i,j}$ are random effects that introduce noise. γ_j in particular is intended to account for platform effects that may over- or underrepresent certain genes. We let these be normally distributed with mean 0 and variance σ_γ, σ_ε respectively.

Therefore, determining the fractional contributions of each cell type reduces to finding the maximum likelihood parameter $\vec{\beta}_i$ for each i.

> **Question 2.1.** Now we have a ton of parameters, potentially thousands. β alone introduces $K \times J$ of them. How do we do any useful estimation here?

We proceed in the following steps:

1. **Supervised estimation of cell type profiles**.

   Using a reference dataset, we estimate the parameters μ_j of expression levels for gene j, giving $\hat{\vec{\mu}}_j$ which will be used in the next steps.

   We can do this by obtaining a (e.g. scRNA-seq) reference annotated with cell types, after which $\vec{\mu}_j$ can be estimated as the empirical average normalized expression of gene j within each cell type.

2. **Gene filtering**.

   Using the estimated expression profiles $\hat{\vec{\mu}}_j$, we filter out genes that are not highly variable across cell types.

We can do this by taking the expression profiles $\hat{\vec{\mu}}_j$ and selecting genes with a minimum average expression and sufficiently high variance.

3. **Platform Effect Normalization**.

With an estimate for $\vec{\mu}_j$, it turns out we now have a way to estimate the platform effects γ_j as well. The idea is that we can consider the average observed expression across pixels

$$\mathsf{M}_j = \frac{1}{I}\sum_{i=1}^{I} \mathsf{Y}_{i,j},$$

whence

$$\begin{aligned}
\log \mathbb{E}_{\mathsf{M}_j|\vec{\lambda}_j}\left[M_j \mid \lambda_{1,j},\ldots,\lambda_{I,j}\right] &= \log\left(\frac{1}{I}\sum_{i=1}^{I} N_i\lambda_{i,j}\right) \\
&= \log\left(\frac{1}{I}\sum_{i=1}^{I} N_i \exp\Big(\log(\vec{\beta}_i\cdot\vec{\mu}_j)+\alpha_i+\gamma_j+\varepsilon_{i,j}\Big)\right) \\
&= \gamma_j + \log\left(\frac{1}{I}\sum_{i=1}^{I} N_i(\vec{\beta}_i\cdot\vec{\mu}_j)\exp\left(\alpha_i+\varepsilon_{i,j}\right)\right) \\
&= \gamma_j + \log\left(\frac{1}{I}\sum_{i=1}^{I}\left(\sum_{k=1}^{K}\beta_{i,k}\cdot\mu_{k,j}\right) N_i \exp\left(\alpha_i+\varepsilon_{i,j}\right)\right) \\
&= \gamma_j + \log\left(\sum_{k=1}^{K}\mu_{k,j}\sum_{i=1}^{I}\frac{N_i}{I}\beta_{i,k}\exp\left(\alpha_i+\varepsilon_{i,j}\right)\right) \\
&= \gamma_j + \log\left(\overline{N}\sum_{k=1}^{K}\mu_{k,j}\left(\frac{1}{I}\sum_{i=1}^{I}\frac{N_i}{\overline{N}}\beta_{i,k}\exp\left(\alpha_i+\varepsilon_{i,j}\right)\right)\right) \\
&= \gamma_j + \log\left(\overline{N}\sum_{k=1}^{K}\mu_{k,j}B_{k,j}\right),
\end{aligned}$$

where

$$\overline{N} \overset{\text{def}}{=} \frac{1}{I}\sum_{i=1}^{I} N_i \quad \text{and} \quad B_{k,j} \overset{\text{def}}{=} \frac{1}{I}\sum_{i=1}^{I}\frac{N_i}{\overline{N}}\beta_{i,k}\exp(\alpha_i+\varepsilon_{i,j}).$$

$\mathsf{B}_{k,j}$ has some desirable properties. In particular, we have

$$\begin{aligned}
\mathbb{E}_{\hat{p}_{\mathsf{B}_{k,j}}}\left[B_{k,j}\right] &= \mathbb{E}_{\hat{p}_{\mathsf{N},\varepsilon_j,\beta_k,\alpha}}\left[\frac{N}{\mathbb{E}_i[N]}\beta_k\exp(\alpha+\varepsilon_j)\right] \\
&= \mathbb{E}_{\hat{p}_{\varepsilon_j,\beta_k,\alpha}}\left[\beta_k\exp(\alpha+\varepsilon_j)\right] \\
&= \overline{\beta}_k\mathbb{E}_{\hat{p}_{\varepsilon_j,\alpha}}\left[\exp(\alpha+\varepsilon_j)\right] \\
&\overset{\text{def}}{=} \overline{\beta}_k\beta_0,
\end{aligned}$$

and as $I \to \infty$, $\text{Var}[B_{k,j}] \to 0$ (by Chebyshev's Inequality). Therefore, as $I \to \infty$ we get $B_{k,j} \approx \overline{\beta}_k \beta_0$ so

$$\log \mathbb{E}_{\mathsf{M}_j | \vec{\lambda}_j} \left[M_j \,\middle|\, \lambda_{1,j}, \ldots, \lambda_{I,j} \right] \approx \gamma_j + \log \beta_0 + \log \left(\overline{N} \sum_{k=1}^{K} \mu_{k,j} \overline{\beta}_k \right).$$

Now we can estimate the platform effects γ_j. Let $W_k \overset{\text{def}}{=} \overline{\beta}_k \beta_0$. The idea is that we can approximate a model for M by

$$\begin{aligned} \mathsf{M}_j \mid \gamma_j &\sim \text{Poisson} \left(I \overline{N} e^{\gamma_j} \sum_{k=1}^{K} \mu_{k,j} W_k \right) \\ \gamma_j &\sim \mathcal{N}(0, \sigma_\gamma^2). \end{aligned}$$

We can then find the MLE estimates of W and σ_γ according to this model and hence obtain $\hat{\gamma}$. This is okay because conditioning on W_k, the distribution of γ_j behaves like a delta function so we can approximate γ_j by the MLE of γ.

4. **Robust Cell Type Decomposition**.

   With $\hat{\mu}_{k,j}$ and $\hat{\gamma}_j$ determined, we then find the MLE estimate for $\alpha_i, \vec{\beta}_i$ and σ_ε in our original model.

We are not yet done after estimating each $\hat{\beta}_{i,k}$. We want to incorporate our prior information that each pixel i is a mixture of a small number of cell types. Hence, we consider models that only consider one or two cell types, and favor "singlet" models $\mathcal{M}$ by minimizing the Akaike Information Criterion

$$\text{AIC}(\mathcal{M}) = \mathcal{L}(\mathcal{M}) + V \cdot p(\mathcal{M}),$$

where $\mathcal{L}$ is the log likelihood, V is a penalty parameter and $p(\cdot)$ is the number of parameters of the model.