

We wish to find a probabilistic model of a corpus that not only assigns high probability to members of the corpus, but also assigns high probability to other “similar” documents.

### 3. Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.<sup>1</sup>

LDA assumes the following generative process for each document  $\mathbf{w}$  in a corpus  $D$ :

1. Choose  $N \sim \text{Poisson}(\xi)$ .
2. Choose  $\theta \sim \text{Dir}(\alpha)$ .
3. For each of the  $N$  words  $w_n$ :
  - (a) Choose a topic  $z_n \sim \text{Multinomial}(\theta)$ .
  - (b) Choose a word  $w_n$  from  $p(w_n | z_n, \beta)$ , a multinomial probability conditioned on the topic  $z_n$ .

Several simplifying assumptions are made in this basic model, some of which we remove in subsequent sections. First, the dimensionality  $k$  of the Dirichlet distribution (and thus the dimensionality of the topic variable  $z$ ) is assumed known and fixed. Second, the word probabilities are parameterized by a  $k \times V$  matrix  $\beta$  where  $\beta_{ij} = p(w^j = 1 | z^i = 1)$ , which for now we treat as a fixed quantity that is to be estimated. Finally, the Poisson assumption is not critical to anything that follows and more realistic document length distributions can be used as needed. Furthermore, note that  $N$  is independent of all the other data generating variables ( $\theta$  and  $\mathbf{z}$ ). It is thus an ancillary variable and we will generally ignore its randomness in the subsequent development.

A  $k$ -dimensional Dirichlet random variable  $\theta$  can take values in the  $(k-1)$ -simplex (a  $k$ -vector  $\theta$  lies in the  $(k-1)$ -simplex if  $\theta_i \geq 0$ ,  $\sum_{i=1}^k \theta_i = 1$ ), and has the following probability density on this simplex:

$$p(\theta | \alpha) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \dots \theta_k^{\alpha_k-1}, \quad (1)$$

where the parameter  $\alpha$  is a  $k$ -vector with components  $\alpha_i > 0$ , and where  $\Gamma(x)$  is the Gamma function. The Dirichlet is a convenient distribution on the simplex — it is in the exponential family, has finite dimensional sufficient statistics, and is conjugate to the multinomial distribution. In Section 5, these properties will facilitate the development of inference and parameter estimation algorithms for LDA.

Given the parameters  $\alpha$  and  $\beta$ , the joint distribution of a topic mixture  $\theta$ , a set of  $N$  topics  $\mathbf{z}$ , and a set of  $N$  words  $\mathbf{w}$  is given by:

$$p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^N p(z_n | \theta) p(w_n | z_n, \beta), \quad (2)$$

---

1. We refer to the latent multinomial variables in the LDA model as topics, so as to exploit text-oriented intuitions, but we make no epistemological claims regarding these latent variables beyond their utility in representing probability distributions on sets of words.