1 Discussed models

This section will introduce the reader into the two Topic modeling approaches which will be compared in this Thesis. The aim of both procedures is to assign one or more topics to different documents. Even if the vocabulary and the notation are similar for both approaches, the notation should be resumed at the beginning of the description of the respective model. The basic structural notation of the data consists of the following variables.

A collection of documents is called corpus $D = (\mathbf{w}_1, \ldots, \mathbf{w}_M)$. It consists out of M documents $\mathbf{w} = (w_1, \ldots, w_N)$ which are itself separated in words w_i. These words are vectors of length V. V refers to the length of a vocabulary which holds all the words occurring in the corpus. The vector for a specific word w_i contains all 0 except for index $j \in \{1, ..., V\}$ which represents this very one word in the vocabulary.

This notation may indeed be extended through the addition of indices for documents, but this is not done here or in the standard literature on topic models due to its unnecessary complexity.

1.1 LDA model

Latent Dirichlet Allocation is a Bayesian approach. The idea is based on the representation of exchangeable random variables (acc. to de Finetti) as mixture of distributions. Given that documents $\mathbf{w}$ and words w_i in each document - both considered as random variables in this setting - are exchangeable in such a way, a mixed model such as the LDA model are appropriate.[1]

Let z_j be the topics with $j \in \{1, \ldots, k\}$. In the LDA setting we assume for e every topic z_j there is a term distribution

$$\beta_j \sim Dir(\delta)$$

We further assume each document w has a distribution of topics.

$$\theta \sim Dir(\alpha)$$

Then each word w_i of $\mathbf{w}$ is generated by the following process:

1. Choose $z_i \sim Mult(\theta)$
2. Choose $w_i \sim Mult(\beta_i)$ This distribution will be referred to as $p(w_i|z_i, \beta)$

You can summarize this setup in a plate diagram as figure 1.1.[2]

The task now is to calculate the posterior distribution, which consists of the joint distribution in the numerator and the marginal distribution in the denominator.

[1] Cf. [D. Blei, 2003]

[2] This notation coincides with the notation of [K. Hornik, 2011]

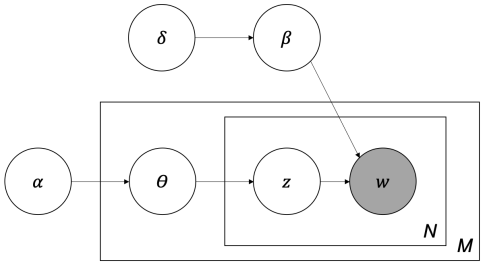

Figure 1: The well-established plate diagram for the standard LDA model extended by the parameter δ. The slightly bigger box represents the generative model of the corporis M documents. The smaller plate represents the iterative generation process of the N words of each document with the aid of the topics. See also "smoothed LDA model" in [D. Blei, 2003] for comparisons.

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)} \tag{1}$$

The joint distribution numerator can be derived straight forward.

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{i=1}^{N} p(w_i|z_i, \beta)\ p(z_i|\theta) \tag{2}$$

One can obtain the marginal distribution of a document w, by integrating out the parameter θ and summing over the topics z_j. Nevertheless, this distribution is intractable.

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{i=1}^{N} \sum_{z_i} p(z_i|\theta) p(w_n|z_i, \beta) \right) d\theta \tag{3}$$

References

[D. Blei, 2003] D. Blei, A. Ng, M. J. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3.

[K. Hornik, 2011] K. Hornik, B. G. (2011). topicmodels: An r package for fitting topic models. *Journal of Statistical Software*, 40(13).