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 and is often associated with the class of hierarchical models [A. Gelman, 2014]. 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 is appropriate.[1]

The following notation is used in conjunction with the LDA model. Let z_j be the topics with $j \in \{1, \ldots, k\}$. In the LDA setting we assume for 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 $p(w_i|z_i, \beta)$

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

You can summarize this setup in a plate diagram as shown in figure 1. The notation above, which is also used within the diagram, coincides with the notation of [K. Hornik, 2011].

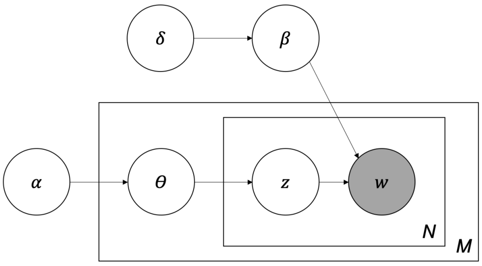

Figure 1: The well-established plate diagram for the standard LDA model extended by the parameter δ. The slightly bigger box represents the generative model of the corporis M documents. The smaller plate represents the iterative generation process of the N words of each document with the aid of the topics.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

In order to estimate the model parameters, the first task is to calculate the posterior distribution, which consists of the joint distribution in the numerator and the marginal distribution in the denominator.

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

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 $\mathbf{w}$, by integrating out the parameter θ and summing over the topics z_j. Nevertheless, this expression is intractable.

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

According to the current state of research, there are two feasible approaches to calculating the posterior density.[2] One approach is to simulate the posterior density by iteratively sampling - the so-called Gibbs sampling. The

[2] Cf. eg. [Blei, 2012]

second approach is a deterministic method, a modified version of the well-known EM algorithm [AP Dempster, 1977]: the Variational EM algorithm (VEM algorithm) [Wainwright and Jordan, 2008]. In the following two sections the both approaches are roughly outlined to give the reader some insight into the Bayesian inference underlying the algorithms.

1.1.1 Variational EM algorithm

In the VEM algorithm for the LDA model is a mean field approach which varies the steps E and M of the EM algorithm in a way such that this algorithm becomes solvable. Note that the main problem of calculating the marginal distribution is, to derive the conditional probability of some hidden variables given some observed values (”evidence”). The variation of the EM algorithms consists mainly in approximating the directly intractable E step. Rewriting the log of the border density of $\mathbf{w}$ as follows in (4), results in the fact that the marginal density can be estimated downwards with the aid of Jensen’s inequality.

$$\begin{aligned}
\log p(\mathbf{w}|\alpha,\beta) &= \log \int \sum_z p(\theta,\mathbf{z},\mathbf{w}|\alpha,\beta)d\theta && (4)\\
&= \log \int \sum_z \frac{p(\theta,\mathbf{z},\mathbf{w}|\alpha,\beta)q(\theta,\mathbf{z})}{q(\theta,\mathbf{z})}d\theta && (5)\\
&\geq \int \sum_z q(\theta,\mathbf{z}) \log p(\theta,\mathbf{z},\mathbf{w}|\alpha,\beta)d\theta - \int \sum_z q(\theta,\mathbf{z}) \log q(\theta,\mathbf{z})d\theta && (6)\\
&= \mathbb{E}_q[\log p(\theta,\mathbf{z},\mathbf{w}|\alpha,\beta)] - \mathbb{E}_q[\log q(\theta,\mathbf{z})] && (7)
\end{aligned}$$

Here $q(\theta,\mathbf{z})$ is an arbitrary distribution which can be called the variational distribution.

$$q(\theta,\mathbf{z}) \widehat{=} q(\theta,\mathbf{z}|\gamma,\phi) = q(\theta|\gamma)\prod_{i=1}^{N} q(z_i|\phi_i) \tag{8}$$

The right hand side $L(\gamma,\phi,\alpha,\beta) := \mathbb{E}_q[\log p(\theta,\mathbf{z},\mathbf{w}|\alpha,\beta)] - \mathbb{E}_q[\log q(\theta,\mathbf{z})]$ be called ”lower bound”. It can be shown that $\log p(\mathbf{w}|\alpha,\beta) - L(\gamma,\phi,\alpha,\beta)$ is the Kullbak Leibler divergence (D_{KL}) of the true posterior and the variational distribution. From equations (4)-(7) follows that:

$$\log p(\mathbf{w}|\alpha,\beta) = D_{KL}(q(\theta,\mathbf{z}|\gamma,\phi)||p(\theta,\mathbf{z},\mathbf{w}|\alpha,\beta)) + L(\gamma,\phi,\alpha,\beta) \tag{9}$$

Since the marginal is fixed, we conclude, that minimizing the KL-divergence is equivalent to maximizing the lower bound.[3]

$$
\begin{aligned}
(\gamma^*, \phi^*) &= \underset{\gamma,\phi}{\operatorname{argmin}}\, D_{KL}(q(\theta, \mathbf{z}|\gamma, \phi)||p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)) && (10) \\
&= \underset{\gamma,\phi}{\operatorname{argmax}}\, L(\gamma, \phi, \alpha, \beta) && (11)
\end{aligned}
$$

The variation of the EM algorithm thus is to use the variational distribution $q(\theta, \mathbf{z}|\gamma^*(\mathbf{w}), \phi^*(\mathbf{w}))$ instead the posterior distribution $p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)$. Now the two steps of the VEM algorithm are:

(1) **E step** Optimize the variational parameters θ and ϕ for every document in the corpus. This can be done analytically by deriving the derivatives of the KL divergence. And set them to zero.

(2) **M Step** Maximize the lower bound using the optimized parameter of the E step with respect to α and β.

1.1.2 Gibbs sampling

References

[A. Gelman, 2014] A. Gelman, J. Carlin, H. S. D. D. A. V. D. R. (2014). *Bayesian Data Analysis.* Chapman & Hall/CRC.

[AP Dempster, 1977] AP Dempster, NM Laird, D. R. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Rojal Statistical Society.*

[Blei, 2012] Blei, D. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77.

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

[et al., 1999] et al., M. J. (1999). An introduction to variational methods for graphical models. *Kluwer Academic Publishers - Machine Learning*, 37:183–233.

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

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

[3] Cf. [et al., 1999] and [Wainwright and Jordan, 2008], for details of the derivation of the lower bound see [D. Blei, 2003]