4 Latent Dirichlet Allocation

Topic modelling is a sub-task of natural language processing that tackles the problem which arises from the massive digitisation of information - in the form of books, magazines, journals, sound and images. It works as a tool to organise and search for relevant information in the sea of data. In the context of a text corpus, the main assumption of a topic model is that each word of each document of the corpus has a hidden topic assignment.

> *"The difference between capitalism per se and an economic system merely involving private ownership of material wealth is, rather curiously, provided by the Hindu tale of Ganesha and Kubera. For those who may not be aware, Ganesha, typically portrayed with an elephant head and pot belly, is the lord of wealth in Hinduism, he is traditionally associated with all things related to business and commerce and is the favourite god of all traders and merchants."*

Figure 3: An extract of an article written by Harsh Tiwari from The Pangean. Highlighted showing some topic assignments - labelled by hand and meant for illustrative purposes only.

Take the extract of an article titled *"American Psycho, Capitalism and the Hindu Gods"* shown in Figure 3. It is an article about the parallels between the movie American Psycho, capitalism and a traditional Hindu tale. Some of the words have been highlighted by hand to illustrate the potential[12] hidden topic assignments of the extract. Three topics have been indicated; politics (yelllow), trade (green) and religion (red). Labelling these words by hand isn't a difficult task, however, it is one that is dependent on knowledge and understanding of language. Therefore, these assignments may vary depending on the person who labels it.

Topic modelling has the goal of grouping words of the same topic without the need for pre-labelling data - this is an unsupervised learning task; whereas a supervised learning approach would require labelling the topic assignment of each word of the training data by hand. This means that massive datasets can be easily prepared for topic modelling. This has the caveat that, once words are grouped, the label of the group (topic) needs to be manually assigned; most of the time, this is fairly straightforward. For example, the grouping of the words `Hindu`, `Ganesha`, `Kubera` and `Hinduism` is very easily labelled *religion*; on the other hand, `material, wealth, business, commerce, traders` and `merchants` is labelled *trade*, but could have quite easily been labelled *economics*. The task of grouping these words in a text may seem straightforward, especially given our knowledge and experience using language, however, language is riddled with historical nuances and how we process it is still very unclear. Topic modelling tries to learn these groupings, not through understanding the nuances of language, but through statistical modelling of the textual information.

[12] These topic assignments are done for illustrative purposes and are by no means the output of some machine learning algorithm.

Latent Dirichlet Allocation (LDA) is one such topic model. The overall assumptions of the model is that for a fixed vocabulary and pre-determined number of topics, each topic asserts a different distribution of words over the vocabulary. For example, words like law, legal and legislation are more probable under the topic corresponding to politics than they are under the topic corresponding to economics. This is not to say that it is impossible for the word 'legislation' to be drawn from a topic assignment corresponding to economics, it is just very improbable. Furthermore, the model assumes that within the corpus, each word of each document has some topic assignment which follows the proportion of topics assigned to the document; therefore, if a document is made up of three topics with equal proportions, the topic assignment of the words that make up the document would reflect that.

Given that there are K topics in the corpus which has a vocabulary of size V, the assumptions of the model are best formalised using the generative stochastic process of LDA (Hoffman et al. 2013) outlined in Figure 4.

1. For each topic $k \in \{1, ..., K\}$,
   (a) sample a distribution over the vocabulary $\beta_k \sim Dir_V(\eta)$.
2. For each document $d \in \{1, ..., D\}$
   (a) sample a distribution over topics $\theta_d \sim Dir_K(\alpha)$.
   (b) For each word of the document $n \in \{1, ..., N_d\}$,
      i. sample a topic assignment $z_{d,n} \sim Cat(\theta_d)$, then
      ii. sample a word $w_{d,n} \sim Cat(\beta_{z_{d,n}})$.

Figure 4: The generative process of LDA.

The model is made up of $(K \times V + D \times K + \sum_{d=1}^{D} N_d \times K)$ variables. The interdependence of these variables on one another is fully described using the graphical model in Figure 5. The word and topic proportions, β_k and θ_d, require V and K positive values that sum to one respectively. Therefore, they are each modelled by a Dirichlet distribution (Appendix A.3). The topic and word assignments, $z_{d,n}$ and $w_{d,n}$ then use the corresponding proportions to select from K and V discrete values respectively. It follows that they are each modelled by a categorical distribution (Appendix A.1). Finally, α acts as our initial assumptions of how the topics are distributed in the documents, and η is our assumptions of how the vocabulary is distributed within the topics. These act as prior knowledge, a way of adding external information to the model.

The input of the LDA model is data from a corpus of documents. Before learning, the word distributions for each topic, the topic proportions of each document and the topic assignment of each word of each document is unknown - these are the hidden variables. The only information known is the generated word - the observed variable. This means that the model has $(K \times V + D \times K + \sum_{d=1}^{D} N_d \times K)$ variables that need to be learnt. We can phrase the problem as finding the conditional joint probability distribution of the hidden variables given the observed variables and prior knowledge - this is commonly known as the posterior of the model. Once we have this distribution, an analysis on it would yield the best setting of the hidden variables which generated the data, and how good the setting actually is. The difficult part is the inference of the

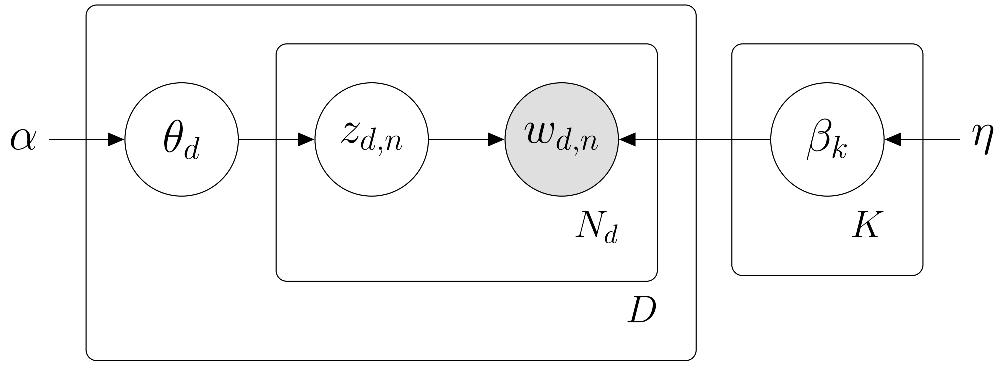

Figure 5: The graphical model representation of LDA.

posterior, $p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}|\mathbf{w}; \alpha, \eta)$. We make a quick note that this notation for the posterior is equivalent to the following fully expanded one:

$$p(\beta_1, ..., \beta_k, \theta_1, ..., \theta_D, z_{1,1}, ..., z_{1,N_1}, ..., z_{D,N_D}|w_{1,1}, ..., w_{1,N_1}, ..., w_{D,N_D}; \alpha, \eta).$$

Using our knowledge of probability (in particular, definition 3.4.1), and the interdependence of the random variables (as seen in Figure 5), the posterior can be rewritten as the following:

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}|\mathbf{w}; \alpha, \eta) = \frac{p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w}; \alpha, \eta)}{p(\mathbf{w}; \alpha, \eta)} = \frac{p(\mathbf{w}|\mathbf{z}, \boldsymbol{\beta})p(\mathbf{z}|\boldsymbol{\theta})p(\boldsymbol{\beta}; \eta)p(\boldsymbol{\theta}; \alpha)}{p(\mathbf{w}; \alpha, \eta)}. \tag{5}$$

The difficulty in finding the posterior comes specifically from the denominator of equation 5 - commonly called the evidence of the model. It is found by marginalising out $\boldsymbol{\beta}, \boldsymbol{\theta}$ and $\mathbf{z}$ from the numerator using the *law of total probability* (Theorem 3.3.2), such that:

$$p(\mathbf{w}; \alpha, \eta) = \int \int \sum_{\mathbf{z}} p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w}; \alpha, \eta) d\boldsymbol{\beta} d\boldsymbol{\theta}. \tag{6}$$

Computing this evidence requires considering - by summing or integrating over - all the different combinations of all the possible values of each of the $(K \times V + D \times K + \sum_{d=1}^{D} N_d \times K)$ hidden variables of the LDA model. This considers (1) all possible word distributions for each topic, (2) all possible topic proportions for each document, and (3) all possible topic assignments for each word for each document. Furthermore, the posterior cannot be analytically determined, due to the lack of a conjugate prior. Therefore, the posterior of the LDA model (equation 6) is computationally intractable. That said, there are methods of approximate inference of the posterior, which we will see in the next chapter.