Chapter 3

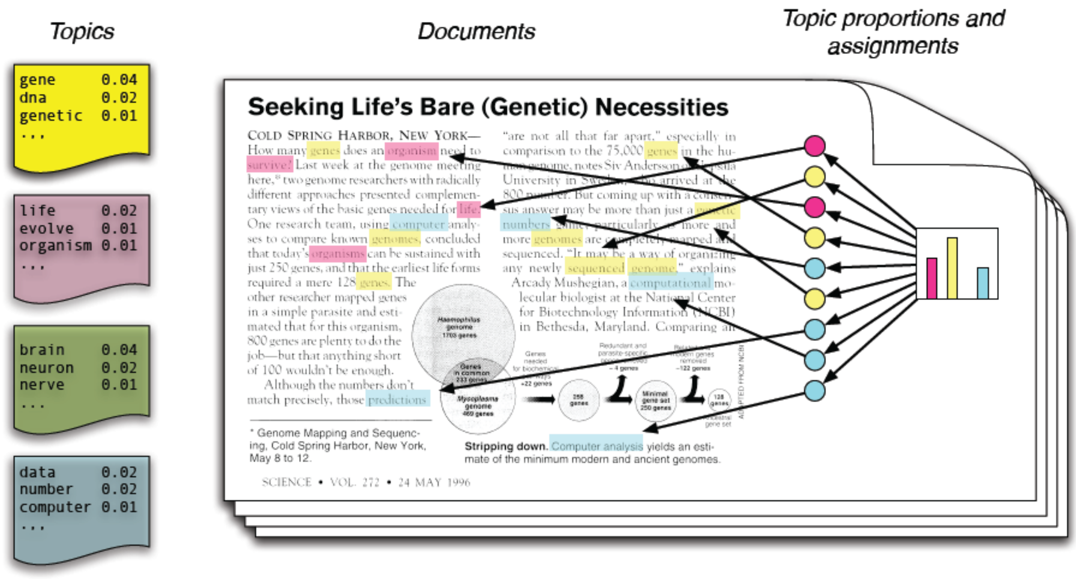
Topic Modeling

Keywords: topic modeling, probabilistic topic modeling, barycentric coordinate system, Dirichlet distribution, EM-algorithm, LDA, ARTM.

3.1 Topic Modeling. What Is It, What Task Does It Resolve

In natural language processing, a **topic modeling** model is a statistical model which aims to highlight topics contained within documents. Its actions is about creating clusters of words targeting the same definition or concept. This process may be used to guess the subject of some documents or benches of texts regarding their content or to classify documents (in order to create some kind of search engines).

As simple used of topic modeling is only based on statistical approaches, it has to be understood that results may present biases according to the prepossessing of data. However, more complete structures and algorithms allow very accurate uses, such as latent Dirichlet allocation (LDA, 3.6).


Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Seeking Life's Bare (Genetic) Necessities
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

3.2 Graphical Models

3.2.1 Introduction to Graphical Models

Graphical models bring together graph theory and probability theory in a powerful formalism for multivariate statistical modeling. In various applied fields including bioinformatics, speech processing, image processing and control theory, statistical models have long been formulated in terms of graphs, and algorithms for computing basic statistical quantities such as likelihoods and score functions have often been expressed in terms of recursions operating on these graphs; examples include phylogenies, pedigrees, hidden Markov models, Markov random fields, and Kalman filters. These ideas can be understood, unified, and generalized within the formalism of **graphical models**. Indeed, graphical models provide a natural tool for formulating variations on these classical architectures, as well as for exploring entirely new families of statistical models. Accordingly, in fields that involve the study of large numbers of interacting variables, graphical models are increasingly in evidence.

We begin with background on graphical models. The key idea is that of factorization: a graphical model consists of a collection of probability distributions that factorize according to the structure of an underlying graph. Here, we are using the terminology ”distribution” loosely; our notation p should be understood as a mass function (density with respect to counting measure) in the discrete case, and a density with respect to Lebesgue measure in the continuous case. There are two main kinds of structured probabilistic models: directed and undirected. Both kinds of graphical models use a graph $\mathcal{G}$ in which each node in the graph corresponds to a random variable, and an edge connecting two random variables means that the probability distribution is able to represent direct interactions between those two random variables.

3.2.2 Directed Models

Directed models use graphs with directed edges, and they represent factorizations into conditional probability distributions, as in the example above. Specifically, a directed model contains one factor for every random variable x_i in the distribution, and that factor consists of the conditional distribution over x_i given the parents of x_i, denoted $Pa_{\mathcal{G}}(\mathrm{x}_i)$:

$$p(\mathbf{x}) = \prod_i p(\mathrm{x}_i \mid Pa_{\mathcal{G}}(\mathrm{x}_i)).$$

See figure 3.1 for an example of a directed graph and the factorization of probability distributions it represents. This graph allows us to quickly see some properties of the distribution. For example, a and c interact directly, but a and e interact only indirectly via c.

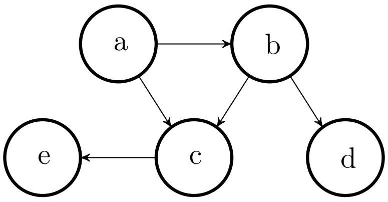


Figure 3.1: A directed graphical model over random variables a, b, c, d and e.

This graph corresponds to probability distributions that can be factored as follows:

$$p(\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e}) = p(\mathrm{a})p(\mathrm{b} \mid \mathrm{a})p(\mathrm{c} \mid \mathrm{a}, \mathrm{b})p(\mathrm{d} \mid \mathrm{b})p(\mathrm{e} \mid \mathrm{c}).$$

3.2.3 Undirected Models

Undirected models use graphs with undirected edges, and they represent factorizations into a set of functions; unlike in the directed case, these functions are usually not probability distributions of any kind. Any set of nodes that are all connected to each other in $\mathcal{G}$ is called a clique. Each clique $\mathcal{C}^{(i)}$ in an undirected model is associated with a factor $\varphi^{(i)}\left(\mathcal{C}^{(i)}\right)$. These factors are just functions, not probability distributions. The output of each factor must be non-negative, but there is no constraint that the factor must sum or integrate to 1 like a probability distribution.

The probability of a configuration of random variables is proportional to the product of all of these factors-assignments that result in larger factor values are more likely. Of course, there is no guarantee that this product will sum to 1. We therefore divide by a normalizing constant Z, defined to be the sum or integral over all states of the product of the φ functions, in order to obtain a normalized probability distribution:

$$p(\mathbf{x}) = \frac{1}{Z}\prod_i \varphi^{(i)}\left(\mathcal{C}^{(i)}\right).$$

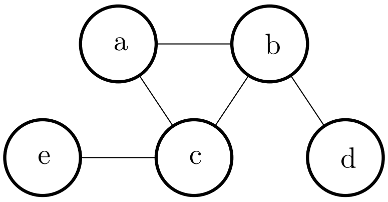


Figure 3.2: An undirected graphical model over random variables a, b, c, d and e.

The graph 3.2 corresponds to probability distributions that can be factored as

$$p(\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e}) = \varphi^{(1)}(\mathrm{a}, \mathrm{b}, \mathrm{c})\varphi^{(2)}(\mathrm{b}, \mathrm{d})\varphi^{(3)}(\mathrm{c}, \mathrm{e}).$$

This graph allows us to quickly see some properties of the distribution. For example, a and c interact directly, but a and e interact only indirectly via c.

3.2.4 Observable and Latent Variables

Before going further, let us define two important statistical concepts.

Observable variables are variables that can be observed and directly measured.

Latent (hidden) variables are variables that are not directly observed but are rather inferred (through a mathematical model) from other variables that are observed (directly measured).

Sometimes latent variables correspond to aspects of physical reality, which could in principle be measured, but may not be for practical reasons. In this situation, the term hidden variables is commonly used (reflecting the fact that the variables are ”really there”, but hidden). Other times, latent variables correspond to abstract concepts, like categories, behavioral or mental states, or data structures. The terms **hypothetical variables** may be used in these situations.

To distinguish the observable variables from the hidden ones, we will use different background colors of graph vertices. Traditionally, vertices corresponding to observable variables are colored grey and vertices corresponding to latent variables are colored white.

Also, usually a number of variables behave in the same way. To simplify our pictures, we will use a rectangular border over the "repeated" variable with a number of such variables on the corner. 3.3

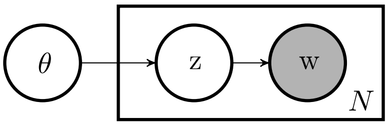


Figure 3.3: In this example θ is an unknown parameter, $\mathrm{z}_1, \dots, \mathrm{z}_N$ are hidden (latent) variables, and $\mathrm{w}_1, \dots, \mathrm{w}_N$ are observable variables.

Given θ, the joint probability for such model can be computes as follows:

$$p(\mathbf{w}, \mathbf{z} \mid \theta) = \prod_{i=1}^{N} p(\mathrm{z}_i \mid \theta) p(\mathrm{w}_i \mid \mathrm{z}_i).$$

3.3 EM-algorithm

In this section we will introduce **EM-algorithm**, a powerful method that is widely used in probabilistic models. In general case, we will use $\mathbf{X}$ to denote the observable variables, $\mathbf{Z}$ to denote the latent variables, and $\boldsymbol{\Theta}$ to denote the model parameters (that we want to estimate).

Assume that we know how to compute a joint distribution over $\mathbf{X}$ and $\mathbf{Z}$ given $\boldsymbol{\Theta}$: $p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta})$.

Let's write a log-likelihood function as an expectation over $\mathbf{Z}$:

$$\log p(\mathbf{X} \mid \boldsymbol{\Theta}) = \int q(\mathbf{Z}) \log p(\mathbf{X} \mid \boldsymbol{\Theta}) d\mathbf{Z}$$

where $q(\mathbf{Z})$ is an arbitrary probability density for $\mathbf{Z}$.

Now that we will transform this formula.

$$\begin{aligned}
\log p(\mathbf{X} \mid \boldsymbol{\Theta}) &= \int q(\mathbf{Z}) \log p(\mathbf{X} \mid \boldsymbol{\Theta}) d\mathbf{Z} = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta})}{p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\Theta})} \frac{q(\mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta})}{q(\mathbf{Z})} d\mathbf{Z} + \int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\Theta})} d\mathbf{Z} \\
&= \mathscr{L}(q, \boldsymbol{\Theta}) + D_{\mathrm{KL}}(q(\mathbf{Z}) \| p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\Theta})) \geq \mathscr{L}(q, \boldsymbol{\Theta}).
\end{aligned}$$

A full log-likelihood function is hard to optimize. Instead of maximizing $\log p(\mathbf{X} \mid \boldsymbol{\Theta})$ by $\boldsymbol{\Theta}$, we are going to maximize $\mathscr{L}(q, \boldsymbol{\Theta})$ by $q, \boldsymbol{\Theta}$. Notice that $\log p(\mathbf{X} \mid \boldsymbol{\Theta})$ does not depend on q so we are not limited to choose q. Thus we can turn the last inequality into equality by putting $q(\mathbf{Z}) = p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\Theta})$ so that $D_{\mathrm{KL}}(q(\mathbf{Z}) \| p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\Theta})) = 0$.

These ideas lead us to an iterative EM-algorithm.

- **E-step** creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters:

$$q(\mathbf{Z})^{(n+1)} = \arg\max_{q'} \mathscr{L}(q', \boldsymbol{\Theta}^{(n)}) = p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\Theta}^{(n)}).$$

- **M-step** computes parameters maximizing the expected log-likelihood found on the E-step:

$$\begin{aligned}
\mathbf{\Theta}^{(n+1)} &= \arg\max_{\mathbf{\Theta}'} \mathscr{L}(q^{(n+1)}, \mathbf{\Theta}') \\
&= \arg\max_{\mathbf{\Theta}'} \int q(\mathbf{Z})^{(n+1)} \log \frac{p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Theta}')}{q(\mathbf{Z})^{(n+1)}} d\mathbf{Z} \\
&= \arg\max_{\mathbf{\Theta}'} \int q(\mathbf{Z})^{(n+1)} \log p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Theta}') d\mathbf{Z} \\
&= \arg\max_{\mathbf{\Theta}'} \mathbb{E}_{\mathbf{Z} \sim q(\mathbf{Z})^{(n+1)}} [\log p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Theta}')].
\end{aligned}$$

In several cases the formulas can be derived analytically.

At that point, all mandatory knowledge have been described which allows us to introduce yet some hidden variable models for topic modeling.

3.4 Probabilistic Generative Topic Model

The joint probability of a document d and a word w can be expressed with conditional probability formula as

$$P(d, w) = P(d)P(w \mid d).$$

Now we will define a model with hidden variables $\mathcal{Z} = \{z_1, \dots, z_T\}$. We assume that each document has its on distribution over (latent) topics, and that each topic has its on distribution over (observable) words. So our goal is to find the matrices $\mathbf{\Phi}$ and $\mathbf{\Theta}$ such that

$$P(w \mid d) = \sum_{z \in \mathcal{Z}} P(w \mid z, d) P(z \mid d) = \sum_{z \in \mathcal{Z}} P(w \mid z) P(z \mid d) = \sum_{z \in \mathcal{Z}} \Phi_{w,z} \Theta_{z,d}.$$

In other words, we want to get a low-rank matrix factorization.

We can write a log-likelihood function for our model:

$$\begin{aligned}
\log P(\mathcal{D}, \mathcal{W} \mid \mathbf{\Phi}, \mathbf{\Theta}) &= \log \prod_{d \in \mathcal{D}} \prod_{w \in \mathcal{W}} P(d, w)^{\#(w,d)} \\
&= \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} \#(w, d) \log P(d, w) \\
&= \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} \#(w, d) \log(P(d) P(w \mid d)) \\
&= \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} \#(w, d) \log \left(P(d) \sum_{z \in \mathcal{Z}} \Phi_{w,z} \Theta_{z,d} \right) \longrightarrow \max_{\mathbf{\Phi}, \mathbf{\Theta}}.
\end{aligned}$$

which is an objective function with restrictions:

$$\sum_{w \in \mathcal{W}} \Phi_{w,z} = 1,\ \Phi_{w,z} \geq 0,\ \sum_{z \in \mathcal{Z}} \Theta_{z,d} = 1,\ \Theta_{z,d} \geq 0.$$

We can also add a non-negative regularizer to the objective function so that it becomes

$$\sum_{d\in\mathcal{D}}\sum_{w\in\mathcal{W}} \#(w,d)\log\left(P(d)\sum_{z\in\mathcal{Z}}\Phi_{w,z}\Theta_{z,d}\right) + R(\mathbf{\Phi},\mathbf{\Theta}) \longrightarrow \max_{\mathbf{\Phi},\mathbf{\Theta}}.$$

If $P(d)$ does not depend on $\mathbf{\Phi}$ or $\mathbf{\Theta}$, then the problem is equivalent to

$$\sum_{d\in\mathcal{D}}\sum_{w\in\mathcal{W}} \#(w,d)\log\sum_{z\in\mathcal{Z}}\Phi_{w,z}\Theta_{z,d} + R(\mathbf{\Phi},\mathbf{\Theta}) \longrightarrow \max_{\mathbf{\Phi},\mathbf{\Theta}}.$$

Let $R(\mathbf{\Phi},\mathbf{\Theta})$ be a continuously differentiable function. Then a point $(\mathbf{\Phi},\mathbf{\Theta})$ of a local extremum can be found as a solution of the equation system:

$$\begin{aligned}
P(z \mid d,w) &\propto \Phi_{w,z}\Theta_{z,d},\\
\Phi_{w,z} &\propto \sum_{d\in\mathcal{D}} \#(w,d)P(z \mid d,w) + \frac{\partial R}{\partial \Phi_{w,z}},\\
\Theta_{z,d} &\propto \sum_{w\in\mathcal{W}} \#(w,d)P(z \mid d,w) + \frac{\partial R}{\partial \Theta_{z,d}}.
\end{aligned}$$

It means that on an E-step we should compute $P(z \mid d,w)$, and on an M-step we should compute $P(w \mid z) = \Phi_{w,z}$ and $P(z \mid d) = \Theta_{z,d}$ according to the formulas obtained.

3.5 PLSA

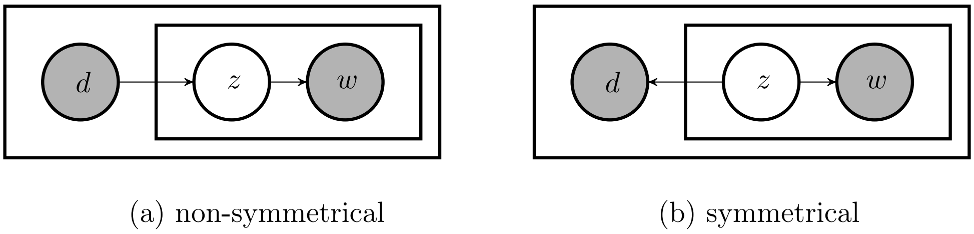
Probabilistic latent semantic analysis (PLSA), proposed by Thomas Hofmann in 1999, was the first technique in history for probabilistic topic modeling. PLSA uses a statistical model which has been called aspect model. The aspect model is a latent variable model for co-occurrence data which associates an unobserved class variable $z \in \mathcal{Z} = \{z_1,\dots,z_T\}$ with each observation. A joint probability model over $\mathcal{D}\times\mathcal{W}$ is defined by the mixture

$$P(d,w) = P(d)P(w \mid d),\; P(w \mid d) = \sum_{z\in\mathcal{Z}} P(w \mid z)P(z \mid d).$$

Like virtually all statistical latent variable models the aspect model introduces a conditional independence assumption, namely that d and w are independent conditioned on the state of the associated latent variable (the corresponding graphical model representation is depicted in Figure 3.4(a). Since the cardinality of z is smaller than the number of documents/words in the collection, z acts as a bottleneck variable in predicting words. It is worth noticing that the model can be equivalently parameterized by (cf. 3.4(b)):

$$P(d,w) = \sum_{z\in\mathcal{Z}} P(z)P(d \mid z)P(w \mid z)$$

which is perfectly symmetric in both entities, documents and words.



(a) non-symmetrical

(b) symmetrical

Figure 3.4: Two versions of the aspect model.

Aspect model uses EM algorithm for maximum likelihood estimation.

- **E-step**. For each $z \in \mathcal{Z}$, $d \in \mathcal{D}$, $w \in \mathcal{W}$ compute

$$P(z \mid d, w) = \frac{P(d, z, w)}{P(d, w)} = \frac{P(z)P(d \mid z)P(w \mid z)}{\sum\limits_{z' \in \mathcal{Z}} P(z')P(d \mid z')P(w \mid z')}.$$

- **M-step** For each $z \in \mathcal{Z}$, $d \in \mathcal{D}$, $w \in \mathcal{W}$ compute

$$\begin{aligned} P(w \mid z) &\propto \sum_{d \in \mathcal{D}} \#(w, d) P(z \mid d, w), \\ P(d \mid z) &\propto \sum_{w \in \mathcal{W}} \#(w, d) P(z \mid d, w), \\ P(z) &\propto \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} \#(w, d) P(z \mid d, w). \end{aligned}$$

To clarify the relation to LSA, let us rewrite the aspect model as parameterized in matrix notation. Hence define matrices by $\hat{\boldsymbol{U}} : \hat{U}_{i,k} = P(w^{(i)} \mid z_k)$, $\hat{\boldsymbol{\Sigma}} : \hat{\Sigma}_{k,k} = P(z_k)$, $\hat{\boldsymbol{V}}^\top : \hat{V}^\top_{k,j} = P(d^{(j)} \mid z_k)$. The joint probability model $\boldsymbol{P}$ can then be written as a matrix product $\boldsymbol{P} = \hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^\top$. Comparing this with SVD, one can make the following observations: (i) outer products between rows of $\hat{\boldsymbol{U}}$ and $\hat{\boldsymbol{V}}^\top$ reflect conditional independence in PLSA, (ii) the K factors correspond to the mixture components in the aspect model, (iii) the mixing proportions in PLSA substitute the singular values. The crucial difference between PLSA and LSA, however, is the objective function utilized to determine the optimal decomposition/approximation. In LSA, this is the L_2 or Frobenius norm, which corresponds to an implicit additive Gaussian noise assumption on (possibly transformed) counts. In contrast, PLSA relies on the likelihood function of multinomial sampling and aims at an explicit maximization of the predictive power of the model.

3.6 LDA

Latent Dirichlet allocation (LDA) is a topic modeling algorithm based on a generative graphical model. It was firstly introduced in 2003 by David Blei, Andrew Ng and Michael I. Jordan in Journal of Machine Learning Research. In this section, we will consider the simplest version of LDA. For more complicated models, SEE SOMETHING. The concept of LDA model is to consider all documents as composed of set of topics or clusters of words and to observe the distribution of them.

More formally, we have a document collection $\mathcal{D}$ and a vocabulary $\mathcal{W}$. Each document $d \in \mathcal{D}$ consists of N_d words $w_{d,1}, \dots, w_{d,N_d}$. The probability distributions over topics $\mathcal{Z}$ are sampled for each document $d \in \mathcal{D}$ from a Dirichlet distribution with parameter $\boldsymbol{\alpha}$. Thus, $\Theta_{:,d} \sim Dir(\boldsymbol{\alpha})$. For each topic $z \in \mathcal{Z}$ we generate a probability distribution over all words $\mathcal{W}$ $\Phi_{:,z}$. After that, for each document d we generate N_d topics $z_{d,1}, \dots, z_{d,N_d}$ from distribution $\Theta_{:,d}$, and for each of these topics $z_{d,k}$ generate per a single word from distribution $\Phi_{:,z_{d,k}}$.

3.6.1 Dirichlet Distribution

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

$$p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

where the parameter $\boldsymbol{\alpha}$ is a K-vector with components $\alpha_k > 0$, 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. These properties facilitate the development of inference and parameter estimation algorithms for LDA.

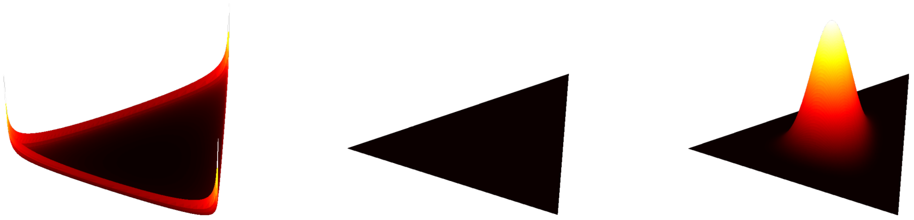


Figure 3.5: Dirichlet distribution plots for $\boldsymbol{\alpha} = (0.75, 0.75, 0.75), (1, 1, 1), (10, 10, 10)$

LDA posits that each word of both the observed and unseen documents is generated by a randomly chosen topic which is drawn from a distribution with a randomly chosen parameter. This parameter is sampled once per document from a smooth distribution on the topic simplex.

3.6.2 LDA Model

Consider the following graphical model.

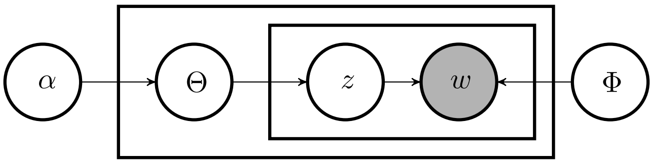


Figure 3.6: A graphical model for LDA.

In this model, $\boldsymbol{\alpha}$ and $\boldsymbol{\Phi}$ are parameters. $\boldsymbol{\Phi}$ is a matrix of size $|\mathcal{W}| \times |\mathcal{Z}|$, $\boldsymbol{\Theta}$ is a matrix of size $|\mathcal{Z}| \times |\mathcal{D}|$, $\boldsymbol{\alpha}$ is a vector of size $|\mathcal{Z}|$. Moreover, $\boldsymbol{\Theta}_{:,d} \sim Dir(\boldsymbol{\alpha})$.

The joint probability of corpora $(\mathcal{D}, \mathcal{W})$ and latent variables $\mathcal{Z}$ and $\boldsymbol{\Theta}$ can be expressed with the formula below.

$$\begin{aligned}
P(\mathcal{D},\mathcal{W},\mathcal{Z},\boldsymbol{\Theta} \mid \boldsymbol{\Phi},\boldsymbol{\alpha}) &= \prod_{d=1}^{|\mathcal{D}|} P(\boldsymbol{\Theta}_{:,d} \mid \boldsymbol{\alpha}) \prod_{n=1}^{N_d} P(z_{d,n} \mid \boldsymbol{\Theta}_{:,d}) P(w_{d,n} \mid z_{d,n}, \boldsymbol{\Phi}) \\
&= \prod_{d=1}^{|\mathcal{D}|} \frac{\Gamma(\sum_{t=1}^{|\mathcal{Z}|} \alpha_t)}{\prod_{t=1}^{|\mathcal{Z}|} \Gamma(\alpha_t)} \prod_{t=1}^{|\mathcal{Z}|} \Theta_{t,d}^{\alpha_t - 1} \prod_{n=1}^{N_d} \Theta_{z_{d,n},d} \Phi_{w_{d,n},z_{d,n}} \\
&= \prod_{d=1}^{|\mathcal{D}|} \frac{\Gamma(\sum_{t=1}^{|\mathcal{Z}|} \alpha_t)}{\prod_{t=1}^{|\mathcal{Z}|} \Gamma(\alpha_t)} \prod_{t=1}^{|\mathcal{Z}|} \Theta_{t,d}^{\alpha_t - 1} \prod_{n=1}^{N_d} \prod_{t=1}^{|\mathcal{Z}|} \Theta_{t,d}^{[z_{d,n}=t]} \Phi_{w_{d,n},t}^{[z_{d,n}=t]}.
\end{aligned}$$

Now we will take the logarithm of it.

$$\log P(\mathcal{D},\mathcal{W},\mathcal{Z},\boldsymbol{\Theta} \mid \boldsymbol{\Phi},\boldsymbol{\alpha}) = \sum_{d=1}^{|\mathcal{D}|} \left[\sum_{t=1}^{|\mathcal{Z}|} (\alpha_t - 1) \log \Theta_{t,d} + \sum_{n=1}^{N_d} \sum_{t=1}^{|\mathcal{Z}|} [z_{d,n} = t] (\log \Theta_{t,d} + \log \Phi_{w_{d,n},t}) \right] + \mathit{const}.$$

After that, we should apply EM algorithm.

- **E-step**.

  $$P(\boldsymbol{\Theta},\mathcal{Z} \mid \mathcal{D},\mathcal{W},\boldsymbol{\Phi},\boldsymbol{\alpha}) \approx q(\boldsymbol{\Theta})q(\mathcal{Z}) = \arg \min_{q(\boldsymbol{\Theta}),q(\mathcal{Z})} D_{\mathrm{KL}}(q(\boldsymbol{\Theta})q(\mathcal{Z}) || P(\boldsymbol{\Theta},\mathcal{Z} \mid \mathcal{D},\mathcal{W},\boldsymbol{\Phi},\boldsymbol{\alpha})).$$

- **M-step**.

  $$\boldsymbol{\Phi} = \arg \max_{\boldsymbol{\Phi}} \mathbb{E}_{\boldsymbol{\Theta} \sim q(\boldsymbol{\Theta}), \mathcal{Z} \sim q(\mathcal{Z})} \log P(\boldsymbol{\Theta},\mathcal{Z} \mid \mathcal{D},\mathcal{W},\boldsymbol{\Phi},\boldsymbol{\alpha}).$$

Here we are using mean field approximation. See ADVANCED BOOK ON BAYESIAN METHODS IF IT'S NOT ENOUGH FOR YOU.

$$\begin{aligned}
\log q(\boldsymbol{\Theta}) &= \mathbb{E}_{\mathcal{Z} \sim q(\mathcal{Z})} \log P(\boldsymbol{\Theta},\mathcal{Z} \mid \mathcal{D},\mathcal{W},\boldsymbol{\Phi},\boldsymbol{\alpha}) + \mathit{const} \\
&= \mathbb{E}_{\mathcal{Z} \sim q(\mathcal{Z})} \log \frac{P(\boldsymbol{\Theta},\mathcal{Z},\mathcal{D},\mathcal{W} \mid \boldsymbol{\Phi},\boldsymbol{\alpha})}{P(\mathcal{D},\mathcal{W} \mid \boldsymbol{\Phi},\boldsymbol{\alpha})} + \mathit{const} \\
&= \mathbb{E}_{\mathcal{Z} \sim q(\mathcal{Z})} \log P(\boldsymbol{\Theta},\mathcal{Z},\mathcal{D},\mathcal{W} \mid \boldsymbol{\Phi},\boldsymbol{\alpha}) + \mathit{const} \\
&= \mathbb{E}_{\mathcal{Z} \sim q(\mathcal{Z})} \sum_{d=1}^{|\mathcal{D}|} \left[\sum_{t=1}^{|\mathcal{Z}|} (\alpha_t - 1) \log \Theta_{t,d} + \sum_{n=1}^{N_d} \sum_{t=1}^{|\mathcal{Z}|} [z_{d,n} = t] (\log \Theta_{t,d} + \log \Phi_{w_{d,n},t}) \right] + \mathit{const} \\
&= \mathbb{E}_{\mathcal{Z} \sim q(\mathcal{Z})} \sum_{d=1}^{|\mathcal{D}|} \left[\sum_{t=1}^{|\mathcal{Z}|} (\alpha_t - 1) \log \Theta_{t,d} + \sum_{n=1}^{N_d} \sum_{t=1}^{|\mathcal{Z}|} [z_{d,n} = t] \log \Theta_{t,d} \right] + \mathit{const} \\
&= \sum_{d=1}^{|\mathcal{D}|} \left[\sum_{t=1}^{|\mathcal{Z}|} (\alpha_t - 1) \log \Theta_{t,d} + \sum_{n=1}^{N_d} \sum_{t=1}^{|\mathcal{Z}|} \mathbb{E}_{\mathcal{Z} \sim q(\mathcal{Z})} [z_{d,n} = t] \log \Theta_{t,d} \right] + \mathit{const} \\
&= \sum_{d=1}^{|\mathcal{D}|} \left[\sum_{t=1}^{|\mathcal{Z}|} (\alpha_t - 1) \log \Theta_{t,d} + \sum_{n=1}^{N_d} \sum_{t=1}^{|\mathcal{Z}|} \gamma_{d,n}^{t} \log \Theta_{t,d} \right] + \mathit{const} \\
&= \sum_{d=1}^{|\mathcal{D}|} \left[\sum_{t=1}^{|\mathcal{Z}|} (\alpha_t - 1) + \sum_{n=1}^{N_d} \gamma_{d,n}^{t} \right] \log \Theta_{t,d} + \mathit{const}.
\end{aligned}$$

Thus,

$$q(\boldsymbol{\Theta}) = const \prod_{d=1}^{|\mathcal{D}|} \prod_{t=1}^{|\mathcal{Z}|} \Theta_{t,d}^{\alpha_t + \sum\limits_{n=1}^{N_d} \gamma_{d,n}^t - 1} = \prod_{d=1}^{|\mathcal{D}|} q(\boldsymbol{\Theta}_{:,d}),\ q(\boldsymbol{\Theta}_{:,d}) \sim Dir(\boldsymbol{\alpha} + \sum_{n=1}^{N_d} \boldsymbol{\gamma}_{d,n}),$$

where

$$\gamma_{d,n}^t = \mathbb{E}_{z_{d,n} \sim q(z_{d,n})}[z_{d,n} = t].$$

What does *const* equal to here?

These formulas allow us to compute $q(\boldsymbol{\Theta})$ on the E-step. Similarly, we compute $q(\mathcal{Z})$.

$$\begin{aligned}
\log q(\mathcal{Z}) &= \mathbb{E}_{\boldsymbol{\Theta} \sim q(\boldsymbol{\Theta})} \log P(\boldsymbol{\Theta}, \mathcal{Z}, \mathcal{D}, \mathcal{W} \mid \boldsymbol{\Phi}, \boldsymbol{\alpha}) + const \\
&= \mathbb{E}_{\boldsymbol{\Theta} \sim q(\boldsymbol{\Theta})} \sum_{d=1}^{|\mathcal{D}|} \sum_{n=1}^{N_d} \sum_{t=1}^{|\mathcal{Z}|} [z_{d,n} = t] \left(\log \Theta_{t,d} + \log \Phi_{w_{d,n},t} \right) + const \\
&= \sum_{d=1}^{|\mathcal{D}|} \sum_{n=1}^{N_d} \sum_{t=1}^{|\mathcal{Z}|} [z_{d,n} = t] \left(\mathbb{E}_{\boldsymbol{\Theta} \sim q(\boldsymbol{\Theta})} \log \Theta_{t,d} + \log \Phi_{w_{d,n},t} \right) + const.
\end{aligned}$$

And we can write

$$q(\mathcal{Z}) = \prod_{d=1}^{|\mathcal{D}|} \prod_{n=1}^{N_d} q(z_{d,n}),\ q(z_{d,n} = t) = \frac{\Phi_{w_{d,n},t} \exp(\mathbb{E}_{\boldsymbol{\Theta} \sim q(\boldsymbol{\Theta})} \log \Theta_{t,d})}{\sum_{t'=1}^{|\mathcal{Z}|} \Phi_{w_{d,n},t'} \exp(\mathbb{E}_{\boldsymbol{\Theta} \sim q(\boldsymbol{\Theta})} \log \Theta_{t',d})} = \gamma_{d,n}^t.$$

E-step is finished.

On the M-step, we need to construct a Lagrange function because of restrictions for all $t \in \{1, \ldots, |\mathcal{Z}|\} \sum\limits_{w \in \mathcal{W}} \Phi_{w,t} = 1$.

$$\begin{aligned}
L &= \mathbb{E}_{\boldsymbol{\Theta} \sim q(\boldsymbol{\Theta}), \mathcal{Z} \sim q(\mathcal{Z})} \sum_{d=1}^{|\mathcal{D}|} \sum_{n=1}^{N_d} \sum_{t=1}^{|\mathcal{Z}|} [z_{d,n} = t] \log \Phi_{w_{d,n},t} + \sum_{t=1}^{|\mathcal{Z}|} \lambda_t \left(\sum_{w=1}^{|\mathcal{W}|} \Phi_{w,t} - 1 \right) \\
&= \sum_{d=1}^{|\mathcal{D}|} \sum_{n=1}^{N_d} \sum_{t=1}^{|\mathcal{Z}|} \gamma_{d,n}^t \log \Phi_{w_{d,n},t} + \sum_{t=1}^{|\mathcal{Z}|} \lambda_t \left(\sum_{w=1}^{|\mathcal{W}|} \Phi_{w,t} - 1 \right).
\end{aligned}$$

$$\begin{aligned}
\frac{\partial L}{\partial \Phi_{\omega,\tau}} &= \sum_{d=1}^{|\mathcal{D}|} \sum_{n=1}^{N_d} \gamma_{d,n}^\tau \frac{1}{\Phi_{\omega,\tau}} [w_{d,n} = \omega] + \lambda_\tau = 0 \Rightarrow \\
\Phi_{\omega,\tau} &= \frac{\sum\limits_{d=1}^{|\mathcal{D}|} \sum\limits_{n=1}^{N_d} \gamma_{d,n}^\tau [w_{d,n} = \omega]}{-\lambda_\tau}.
\end{aligned}$$

Summing over $\omega' \in \mathcal{W}$, we will get

$$-\lambda_\tau = \sum_{\omega' \in \mathcal{W}} \sum_{d=1}^{|\mathcal{D}|} \sum_{n=1}^{N_d} \gamma_{d,n}^{\tau} [w_{d,n} = \omega'].$$

Thus

$$\Phi_{\omega,\tau} = \frac{\sum\limits_{d=1}^{|\mathcal{D}|} \sum\limits_{n=1}^{N_d} \gamma_{d,n}^{\tau} [w_{d,n} = \omega]}{\sum\limits_{\omega' \in \mathcal{W}} \sum\limits_{d=1}^{|\mathcal{D}|} \sum\limits_{n=1}^{N_d} \gamma_{d,n}^{\tau} [w_{d,n} = \omega']}.$$

And that means that we have formulas to update $\mathbf{\Phi}$ on the M-step.

It was the simplest version of LDA. The more complicated model assumes that the rows of $\mathbf{\Phi}$ are sampled from another Dirichlet distribution so that

$$\mathbf{\Phi}_{:,t} \sim Dir(\boldsymbol{\beta}).$$

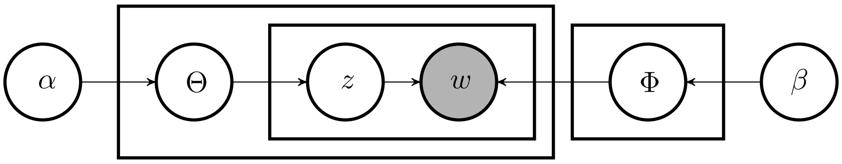


Figure 3.7: More complicated graphical model for LDA.

3.7 BigARTM

TODO

3.8 Seminar

- Practical application of topic models. Clusterization.
- Hands-on tutorial on LDA (gensim). Visualization.
- **Homework:** PLSA and LDA for topic modeling on Hillary Clinton emails.
- Hands-on tutorial on BigARTM.

Bibliography

[1] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine learning research*, 3(Jan):993–1022.