Stockholms universitet

Trend Analysis Using Probabilistic Topic Modeling

Evaluating ESG Trends in Earnings Call Data

Anton Stråhle

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2021:8
Matematisk statistik
Juni 2021

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Master Thesis **2021:8**
http://www.math.su.se

Trend Analysis Using Probabilistic Topic Modeling

Evaluating ESG Trends in Earnings Call Data

Anton Stråhle*

June 2021

Abstract

This thesis examines how topic modeling, specifically probabilistic topic modeling, can be used in order to gauge trends in textual data. We introduce two probabilistic topic models, LDA and DMM, and present two types of approximative inference for both models as well as methods of hyperparameter estimation. Furthermore we also introduce methods of guiding the models in the direction of uncovering topics relating to the trends that we are interested in.

The two models are applied to different corpora consisting of questions from the Q & A sessions of financial earnings calls. We focus on how ESG related trends can be examined using these methods, hence we guide the models towards uncovering such topics. We find that DMM models are not applicable to the data examined as the assumptions made are not fulfilled to such an extent as initially thought. LDA models on the other hand, when applied to specific subsets of the data, yield quite promising results.

When further nudging the LDA models in the direction of uncovering specific ESG topics we find that we are able to uncover topics where we have some prior knowledge of the topics existence within the corpus. The ability to uncover topics in such a manner indicates that the framework established in the thesis is usable in practice but that it requires some prior knowledge regarding the topics one wants to uncover.

Using the trained models we also construct two metrics that can be used to gauge the trends of these topics when they are evaluated on a test corpus that spans a longer period of time. We also discuss the drawbacks of the framework introduced as well as some ways in which it can be improved.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: anton.t.strahle@gmail.com. Supervisor: Taras Bodnar.

Acknowledgments

I would like to thank my supervisor Prof. Taras Bodnar at the Department of Mathematics at Stockholm University for his invaluable assistance throughout the entire writing process of this thesis. Furthermore I would also like to thank my external supervisor Andreas E. Johansson at SEB Investment Management for giving me the opportunity to write this thesis and for contributing with valuable insight and ideas. I would also like to thank the people who took their time to proofread this thesis which helped improve it immensely. Lastly, I would like to extend a special thanks to all friends and family that have encouraged and supported me throughout my studies.

Terminology

corporate governance	The corporate governance aspect of ESG includes the responsibilities of the management of the company when it comes to transparency, corruption and more [1].
corpus	A corpus, $\mathbf{w}$, is a collection of M documents such that $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_1, \ldots, \mathbf{w}_M)$.
document	A document, $\mathbf{w}_i$, is a sequence of N words such that $\mathbf{w}_i = (w_{i,1}, w_{i,2}, \ldots, w_{i,N_i})$ where $w_{i,j}$ is the jth word in the ith document.
document-term matrix	A document-term matrix, X is a $M \times V$ matrix where element i, v indicates the number of times the vth word in a vocabulary W appears in document i.
environmental	The environmental aspect of ESG includes the emission of greenhouse gases, waste management and more [1].
ESG	ESG is an umbrella term that includes three central factors, environmental, social and corporate governance. Basing investments on these factors is usually referred to as ESG, or sustainable, investing [1].
social	The social aspect of ESG includes human rights, labor standards, workplace safety and more [1].
topic	A topic, z, is an integer in $\{1, ..., K\}$ where K is some fixed number of unique topics. By definition, in the case of LDA, $z_{i,j}$ is the topic attached to the jth word of document i. For DMM z_i is the topic attached to the ith document.
vocabulary	A vocabulary, W, is defined as the set of unique words in a corpus. These unique words are indexed by the integers $\{1, ..., V\}$.
word	A word, $w_{i,j}$, is an integer in $\{1, ..., V\}$ where V is the length of a vocabulary W. By definition $w_{i,j}$ is the jth word of document i.

Contents

Chapter 1

Introduction

In the last decade both retail and institutional investors have begun to further shift towards sustainable investment strategies. This shift can be seen in the form of both divestment in companies which are deemed to be unsustainable and further investment in companies that are deemed to be sustainable [1]. In 2019 Morgan Stanley reported that the percentage of retail investors interested in sustainable investing had increased to 85% from 71% in 2015 [2]. Consequently the assets under management (AUM) in sustainable mutual funds have seen their steepest increase ever in the last few years, indicating that sustainable investing, which was once a niche, has now begun to move into the mainstream [1].

Sustainable investing can conveniently be represented by the umbrella term ESG which is made up of three central factors, those being environmental, social and corporate governance [1]. These factors can then decomposed into sub-factors that focus on more specific themes, such as climate change or community impact which are environmental and social sub-factors respectively. Although most retail and institutional investors have some perception of which of these factors they consider to be of the largest, or smallest, importance it is difficult to determine how the market as a whole perceives the relative importance of each factor.

In this thesis we propose a framework for quantifying how much focus the market, or rather a proxy of the market, places on each of these factors and sub-factors. In order to determine how much emphasis is put on each of the factors we propose a method which utilizes textual data from the questions and answers (Q & A) session of earnings calls to determine the frequency of questions regarding, and hence the focus on, these different factors. Since the Q & A data consist of transcripts from actual earnings calls, that are of course unlabeled, the frequency of questions relating to specific subjects cannot be determined right away. Given the unlabeled nature of the data our method of choice is to use unsupervised learning in order to evaluate the topics present in current, but also historical and future, questions.

Given a clustering model which performs well enough to be able to broadly determine that a question is related to a specific ESG sub-factor, or at least a factor, we can analyze how the frequency of ESG related questions have varied in the past. Beyond analyzing historical trends we can also use our models to determine the topics of new questions and as such allow for the identification of new trends.

Chapter 2

Probabilistic Topic Modeling

In order to obtain the frequency of questions relating to specific ESG factors we first need to be able to determine which factors or sub-factors that specific questions touch upon. In order to evaluate the questions we have chosen to use probabilistic topic modeling in which the aim is to discover the latent groups, or topics, that exist in a corpus and make the documents within it similar. As such from a topic modeling point of view the aim of the thesis is to uncover topics relating to these ESG factors, or even sub-factors, that we are interested in and to use these discovered topics to examine their frequency in past and future questions to be able to determine their prevalence over time.

The general idea behind topic modeling is that we can assume that there are some latent semantic structures in a corpus that determines the similarity between the documents in it. The aim is then to uncover these semantic structures, otherwise known as topics, in order to determine the similarity between documents. Although it is extremely useful to be able to determine the topics of documents in a corpus, one of the major drawbacks of traditional probabilistic models such as Probabilistic Latent Semantic Indexing (pLSI) as introduced in *Hoffman (1999)*, and as is briefly mentioned in Section 2.2, is that there are no possibilities of determining the topics of new documents [3][4]. This issue occurs as the model is not a proper generative model, meaning that the probabilities of topics given a document can not be determined [3]. Luckily there also exists several newer, and improved, models that have been developed to address this shortcoming and it is models of this type which we will focus on in this thesis as the aim is to be able to evaluate both documents inside, but also outside, of the training corpus [3].

The main model which we examine is Latent Dirichlet Allocation (LDA) which is further discussed in Section 2.3. LDA is perhaps the most popular probabilistic topic model given that it neatly addresses the issues of previous models by being an actual generative model, meaning that it can be used to evaluate new documents [3]. Even though LDA has many advantages, a possible issue it faces is that it models a document as a mixture of topics [3][5]. This is intuitive for longer documents that allow for the existence of several topics, such as chapters in a book or news articles. It could however be considered an unreasonable or problematic assumption in the case of shorter documents, such as tweets, or questions as in our case, that rarely touch upon more than a single topic. This is discussed further in Section 2.4. As such we will also examine so called Dirichlet Multinomial Mixture (DMM) models which assume that every document consists of a single topic, rather than a mixture of topics, which could arguably be more appropriate given the type of data which we are examining in this thesis [5]. Dirichlet Multinomial Mixture models are discussed in Section 2.5.

Although both LDA and DMM can be used to discover unique and reasonable topics the methods are still unsupervised, meaning that we might have a hard time discovering rarer topics that are discussed at a lower frequency. This issue occurs as the aim is to maximize the probability over a corpus, which more often than not leads to

the discovery of shallow and at times uninformative topics [6]. Essentially it becomes beneficial from a probabilistic point of view to ignore rare topics in order to better model more frequent, usually superficial topics and words [6]. We have chosen to address this issue by using a semi-supervised approach that utilizes a small set of seed words to encourage the discovery of topics related to these words. This does of course mean that some prior knowledge of both the seeded topics, and their existence in the specific corpus, is required in order to maximize the chances of discovering the sought after topics. This is further discussed in Section 2.7.

Another issue which conventional probabilistic topic modeling techniques face in practice is scalability. As will be discussed for both LDA and DMM exact inference is not possible, meaning that we have to resort to approximative inference methods. The issue that arises is that although these techniques work well they sadly do not scale as well when dealing with large corpora. Not only does a large corpus imply a large number of documents but it almost always lead to a larger vocabulary, a vocabulary being all unique words in a corpus. Beyond that a larger corpus will most likely be more diverse than a smaller one which will require a larger number of topics to properly describe the similarities of the documents within it. Hence by increasing the size of a corpus we essentially increase the number of calculations required on three different fronts which is of course problematic. This is worth keeping in mind as although objectively better results could likely be obtained by increasing the size of the corpus, the size of the number of topics, this is not always possible due to the aforementioned computational limitations. Some methods have been proposed to address this issue, such as one introduced in *Cong et al. (2019)* where the authors represent texts as word embeddings and then cluster these embeddings in an efficient manner [7]. From these clusters the relative importance of the words in the cluster is then obtained [7]. As most of these approaches are not as probabilistic in nature they are not explored further in this thesis.

2.1 Introduction to Probabilistic Topic Modeling

The aim of probabilistic topic modeling is, as previously mentioned, to discover the underlying semantic structures in a set of documents. This set of documents is usually referred to as a corpus, which we denote by $\mathbf{w}$, containing documents $\mathbf{w}_1, ..., \mathbf{w}_M$ such that $\mathbf{w} = \{\mathbf{w}_1, ..., \mathbf{w}_M\}$. Each document $\mathbf{w}_i$ is a sequence of N_i terms, or words, from a vocabulary $W = \{1, ..., V\}$ such that $\mathbf{w}_i = \{w_{i,1}, ..., w_{i,N_i}\}$.

Most topic models treat documents as a *bag-of-words*, meaning that the sequencing of terms within a document is ignored and that the documents are modeled by their term frequencies. The concept of treating documents as *bag-of-words* is further detailed in Section 3.1. The term frequencies are then represented by the $M \times V$ matrix, often referred to as the document-term matrix, or DTM for short, X in which element i, v represents the number of times the vth word in a vocabulary W occurs in the ith document $\mathbf{w}_i$. By reducing the problem to modeling the frequencies instead of trying to incorporate the semantics of the documents we reduce the complexity of the problem, making analysis more feasible.

2.2 Probabilistic Latent Semantic Indexing

In this section we will briefly touch upon Probabilistic Latent Semantic Indexing (pLSI), also known as Probabilistic Latent Semantic Analysis (pLSA), which was initially introduced in *Hoffman (1999)* [4][8]. Probabilistic Latent Semantic Indexing, like most probabilistic topic models, models the document-term frequencies as mentioned in the previous section.

The idea behind pLSI, and all topic models in some sense, is that we associate some unknown, or latent, grouping variable $z = 1, ..., K$ with each observation, which in the case of pLSI is the individuals terms v [4]. The generative process for a document-term pair (d, v), with $d \in \mathbf{w}$ and $v \in W$, in pLSI is described as follows

1. Select a document d with probability $P(d)$
2. Pick a latent class z with probability $P(z|d)$
3. Generate a word v with probability $P(v|z)$

The generative process can then be explained using plate notation as can be seen below.

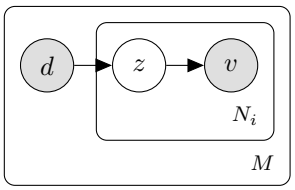


Figure 2.1: Plate representation of pLSI

From this generative process we have that the joint probability of a document word pair (d, v) is

$$P(d, v) = P(d)P(v|d) \tag{2.1}$$

$$= P(d) \underbrace{\sum_{k=1}^{K} P(v|z = k)P(z = k|d)}_{=P(v|d)} \tag{2.2}$$

where it is assumed that conditioned on the latent topic z words v are generated independently of the underlying document d [4]. Here $P(z|d)$ would be the document-topic probabilities whilst $P(v|z)$ would be the topic-word probabilities. Going forward from here $P(d), P(z|d)$ and $P(v|z)$ are inferred by maximizing the log-likelihood function

$$L = \sum_{d=1}^{M} \sum_{v=1}^{V} X_{d,v} \log \left(P(d, v) \right). \tag{2.3}$$

Where X is the DTM with element $X_{d,v}$ being the number of times the vth word in a vocabulary W appears in document d. Given the presence of the latent variables the maximum likelihood estimation is performed by using the EM algorithm [4].

Although pLSI possesses many of the qualities necessary for our scope one of its main issues is that although it is a generative model of the documents in the corpus $\mathbf{w}$ it is not a generative model for new documents, meaning that it cannot be used to determine the topic probabilities of new documents [3].

2.3 Latent Dirichlet Allocation

Latent Dirichlet Allocation, henceforth referred to as LDA, is a generative statistical model of a corpus which represents every document in the corpus as a mixture of latent topics and every latent topic as a mixture of words [3]. Like pLSI, LDA also operates on the *bag-of-words* assumption. The main difference between LDA and pLSI is that whilst pLSI assumes unknown categorical parameters for both the document-topic distribution and the topic-word distributions, LDA on the other hand endows both of these distributions with Dirichlet priors. These Dirichlet priors are conjugate to the categorical document-topic and topic-word distributions as will prove quite useful. Having these priors distinguishes LDA from the previously discussed pLSI as it is a proper generative model, both for documents in the training corpus but also for new documents.

2.3.1 Dirichlet Prior

As mentioned above the Dirichlet distribution is conjugate to the categorical distribution. Hence given our Dirichlet priors for the parameters of our categorical distributions the resulting posteriors are also Dirichlet distributed. The conjugacy also turns out to be quite neat as it simplifies a lot of the derivations relating to the approximative inference algorithms examined as will become evident later on.

We can also choose the hyperparameters of these priors to be sparse as this leads to documents being a mixture of few, highly probable, topics and topics being a mixture of few, highly probable words. This sparsity allows for quite interpretable models as each topic uncovered will be determined by a small selection of words that can usually, given that the model is applicable to the data in question, be interpreted quite easily allowing for the labeling of the topic [3].

2.3.2 Model Definition

In the original paper by Blei et al. the authors propose two different variations of LDA [3]. The first variation, as described in Appendix A.8, treats the word distributions, conditioned on a topic k, as categorical with parameter $\boldsymbol{\phi}_k$. The second variation, which is the one we will focus on in this thesis, is very similar but endows the categorical parameters $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_K)$ with Dirichlet distributions as well.

The advantage of having priors on the topic-word distributions is that it allows for the possibility of generating words which are not present in the training corpora. Without the priors words not present in the training corpora will be assigned probability zero which is problematic when we want to determine the topics of documents not in the training corpora that might include these words. Furthermore through the inclusion of Dirichlet priors on the topic-word distributions we can alter the hyperparameter $\boldsymbol{\beta}$, making it preferably sparse, in order to have topics be distinguished by a small selection of unique words as mentioned in the previous section [3].

The generative process of the second variation introduced in *Blei et al. (2003)* is hence as follows [3].

1. Choose $\boldsymbol{\phi}_k \sim Dir(\boldsymbol{\beta})$ for $k = 1, ..., K$
2. For each document $i = 1, ..., M$
   (a) Choose a document length $N_i \sim Po(\xi)$
   (b) Choose a topic distribution $\boldsymbol{\theta}_i \sim Dir(\boldsymbol{\alpha})$
   (c) For each word $j = 1, ..., N_i$
      i. Choose a topic $z_{i,j} \sim Cat(\boldsymbol{\theta}_i)$
      ii. Choose a word $w_{i,j} \sim Cat(\boldsymbol{\phi}_{z_{i,j}})$

This process can be shown in plate representation as below.

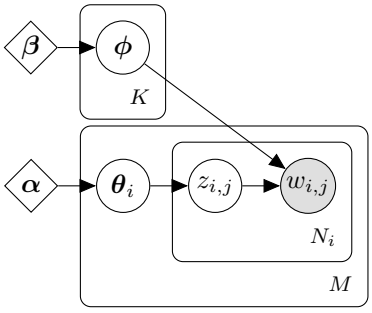


Figure 2.2: Plate representation of LDA

Probabilistically we have that

$$\boldsymbol{\theta}_i \overset{iid}{\sim} Dir(\boldsymbol{\alpha}) \qquad i = 1, ..., M, \tag{2.4}$$

$$\boldsymbol{\phi}_k \overset{iid}{\sim} Dir(\boldsymbol{\beta}) \qquad k = 1, ..., K, \tag{2.5}$$

$$z_{i,j}|\boldsymbol{\theta}_i \overset{indep}{\sim} Cat(\boldsymbol{\theta}_i) \qquad i = 1, ..., M \quad j = 1, ..., N_i, \tag{2.6}$$

$$w_{i,j}|z_{i,j}, \boldsymbol{\Phi} \overset{indep}{\sim} Cat(\boldsymbol{\phi}_{z_{i,j}}) \qquad i = 1, ..., M \quad j = 1, ..., N_i. \tag{2.7}$$

Above we have categorical distributions which are equivalent to multinomial distributions where $n = 1$. We note that the distribution of N_i, the number of words in a document, is independent of everything else in the process, meaning that it can be seen as an ancillary variable which we can treat as a deterministic quantity going forward. As mentioned previously we also have $\boldsymbol{\beta}$ which is the hyperparameter for the Dirichlet priors of the topic-word distributions as well as $\boldsymbol{\alpha}$ which is the hyperparameter for the Dirichlet priors of the document-topic distributions. We also have that the document-topic and topic-word probabilities, i.e. $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ are independent of one another.

Combining the entirety of the probabilistic model we can obtain the full probability of generating a specific corpus $\mathbf{w}$ as can be seen below.

$$P(\mathbf{w}, \mathbf{z}, \boldsymbol{\Theta}, \boldsymbol{\Phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{k=1}^{K} P(\boldsymbol{\phi}_k; \boldsymbol{\beta}) \prod_{i=1}^{M} P(\boldsymbol{\theta}_i; \boldsymbol{\alpha}) \prod_{j=1}^{N_i} P(w_{i,j}|\boldsymbol{\phi}_{z_{i,j}}) P(z_{i,j}|\boldsymbol{\theta}_i). \tag{2.8}$$

Given that we only observe a corpus $\mathbf{w}$ the main inferential problem in this variation of LDA is to determine the following [3]

$$P(\boldsymbol{\Theta}, \boldsymbol{\Phi}, \mathbf{z}|\mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{P(\mathbf{w}, \mathbf{z}, \boldsymbol{\Theta}, \boldsymbol{\Phi}; \boldsymbol{\alpha}, \boldsymbol{\beta})}{P(\mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})}. \tag{2.9}$$

The denominator in equation (2.9) above is intractable to compute and hence we have to resort to approximate inference techniques in order to properly utilize LDA [3]. Luckily there are several choices of approximate inference methods that have been developed for LDA. In the original paper the authors introduce a Variational Bayesian approach to inference whilst in another paper by Griffiths et al. the method of choice is Collapsed Gibbs Sampling [9][3]. In the following sections we will introduce both methods of approximate inference.

2.3.2.1 Variational Bayesian Inference

In the original paper by Blei et al. the authors propose a Variational Bayesian approach to the inference problem in LDA [3]. As the denominator in equation (2.9) is intractable to compute we cannot obtain an analytical expression of the posterior. An alternative to this is to approximate the posterior $P(\boldsymbol{\theta}, \boldsymbol{\Theta}, \mathbf{z}|\mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})$ with an adjustable lower bound indexed by some set of variational parameters. The variational parameters are then chosen by minimizing the difference between the lower bound and the true posterior [3]. The choice of the variational distribution family is chosen to be separable on the random variables $\boldsymbol{\Theta}, \boldsymbol{\Phi}, \mathbf{z}$ and in accordance with *Attias (2000)* as can be seen in Figure 2.3 below [3][10].

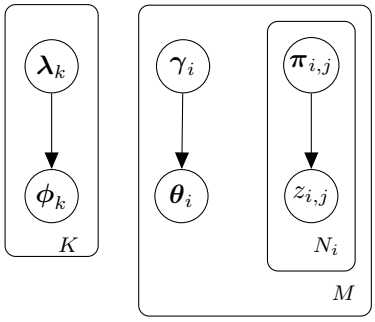


Figure 2.3: Plate representation of the reduced second version of LDA

We use so called mean field approximation where the family of variational distribution is assumed to be fully factorized [10][11]. This yields the variational distribution q with free variational parameters $\boldsymbol{\gamma}, \boldsymbol{\pi}$ and $\boldsymbol{\lambda}$.

$$q(\mathbf{z}, \boldsymbol{\Theta}, \boldsymbol{\Phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}) = \prod_{k=1}^{K} q(\boldsymbol{\phi}_k; \boldsymbol{\lambda}_k) \prod_{i=1}^{M} q(\boldsymbol{\theta}_i; \boldsymbol{\gamma}_i) \prod_{j=1}^{N_i} q(z_{i,j}; \boldsymbol{\pi}_{i,j}). \tag{2.10}$$

The optimal lower bound is determined by the optimal values of the variational parameters $\boldsymbol{\gamma}, \boldsymbol{\pi}$ and $\boldsymbol{\lambda}$ and as such we need to derive their update equations [3]. We obtain these by minimizing the Kullback-Leibler Divergence between the variational distribution q and the true posteriors p, $D(q||p)$, w.r.t. these parameters.

$$D(q||p) = \int \int \sum_{\mathbf{z}} q(\mathbf{z}, \boldsymbol{\Theta}, \boldsymbol{\Phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}) \log \left(\frac{q(\mathbf{z}, \boldsymbol{\Theta}, \boldsymbol{\Phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda})}{p(\mathbf{z}, \boldsymbol{\Theta}, \boldsymbol{\Phi} | \mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})} \right) d\boldsymbol{\Theta} d\boldsymbol{\Phi}. \tag{2.11}$$

This minimization is described in detail in Appendix A.2 and yields the following update equations.

$$\boldsymbol{\gamma}_{i,k} = \alpha_k + \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,j,k}, \tag{2.12}$$

$$\boldsymbol{\pi}_{i,j,k} \propto \exp \left\{ \Psi \left(\boldsymbol{\gamma}_{i,k} \right) + \Psi(\boldsymbol{\lambda}_{k,w_{i,j}}) - \Psi \left(\sum_{v=1}^{V} \boldsymbol{\lambda}_{k,v} \right) \right\}, \tag{2.13}$$

$$\boldsymbol{\lambda}_{k,v} = \beta_v + \sum_{i=1}^{M} \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,j,k} \mathbf{1}_{\{w_{i,j}=v\}}. \tag{2.14}$$

Above we have that Ψ is the digamma function, i.e. $\Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$. We update the variational parameters as per the update equations above until convergence. Upon convergence we have a set of variational parameters that define our lower bound of the posterior. It is also possible to further estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ using a Variational Expectation Maximization (VEM) approach as is discussed in Section 2.3.3.1 [3][11].

In Algorithm 1 below we detail how the Variational Bayesian Inference approach as discussed in this section is used to obtain optimal values for the variational parameters through iteration until convergence. Upon convergence we have an approximation of the posterior which is indexed by the converged variational parameters.

Algorithm 1: Variational Bayesian Inference for LDA

input : A corpus $\mathbf{w} = (\mathbf{w}_1, ..., \mathbf{w}_M)$ and hyperparameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$
output: Optimal values of $\boldsymbol{\gamma}$, $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$
initialize $\boldsymbol{\pi}_{i,j,k}^{(0)}$ as $\frac{1}{K}$.
initialize $\boldsymbol{\gamma}_{i,k}^{(0)}$ as $\alpha_k + \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,j,k}^{(0)}$.
initialize $\boldsymbol{\lambda}_{v,k}^{(0)}$ as $\beta_v + \sum_{i=1}^{M} \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,j,k}^{(0)} \mathbf{1}_{\{w_{i,j}=v\}}$
set $t = 1$
while *not converged* **do**
  for *topic* $k = 1, ..., K$ **do**
    for *document* $i = 1, ..., M$ **do**
      update
      $\boldsymbol{\gamma}_{i,k}^{(t)} = \alpha_k + \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,j,k}^{(t-1)}$
      for *word* $j = 1, ..., N_i$ **do**
        update
        $\boldsymbol{\pi}_{i,j,k}^{(t)} = \exp\left\{\Psi\left(\boldsymbol{\gamma}_{i,k}^{(t-1)}\right) + \Psi\left(\boldsymbol{\lambda}_{k,w_{i,j}}^{(t-1)}\right) - \Psi\left(\sum_{v=1}^{V} \boldsymbol{\lambda}_{k,v}^{(t-1)}\right)\right\}$
      end
    end
    for *word in vocabulary* $v = 1, ...V$ **do**
      update
      $\boldsymbol{\lambda}_{k,v}^{(t)} = \beta_v + \sum_{i=1}^{M} \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,j,k}^{(t-1)} \mathbf{1}_{\{w_{i,j}=v\}}$
    end
    normalize all $\boldsymbol{\pi}_{i,j,k}^{(t)}$ by dividing by $\sum_{l=1}^{K} \boldsymbol{\pi}_{i,j,l}^{(t)}$
  end
  $t = t + 1$
end

As $\boldsymbol{\gamma}_{i,k}$ is approximately the kth Dirichlet topic prior α_k plus the expected number of words generated by the kth topic, we can estimate $\boldsymbol{\theta}_{i,k}$ as [3]

$$\hat{\boldsymbol{\theta}}_{i,k} = \frac{\boldsymbol{\gamma}_{i,k}}{\sum_{l=1}^{K} \boldsymbol{\gamma}_{i,l}}. \tag{2.15}$$

The same argument can be made for $\boldsymbol{\lambda}_{k,v}$ and the estimation of $\boldsymbol{\phi}_{k,v}$

$$\hat{\boldsymbol{\phi}}_{k,v} = \frac{\boldsymbol{\lambda}_{k,v}}{\sum_{l=1}^{K} \boldsymbol{\lambda}_{l,v}}. \tag{2.16}$$

Lastly we also have the topic probabilities $\boldsymbol{\pi}_{i,j,k}$ which represent the final topic probabilities for all terms in the corpus.

2.3.2.2 Collapsed Gibbs Sampler

Another option for the inference problem is to obtain an approximation of the sought after posterior distribution by using a Collapsed Gibbs Sampler. This approach is discussed in a paper by Griffiths et al. and is based on the regular Gibbs Sampler as introduced in *Geman et al. (1984)* [12][9]. The idea of the Collapsed Gibbs Sampler within the scope of LDA is to use the fact that the Dirichlet priors are conjugate priors to the categorical topic and word distributions which allows us to integrate out both $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$. The integration w.r.t. $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ is often referred to as the collapsing of the Dirichlet priors, hence the name. Once the priors have been collapsed we determine the posterior distribution of a specific topic, e.g. that of the bth word in document a, $z_{a,b}$, conditioned on all other topics $\mathbf{z}_{\neg a,b}$. The full derivations can be found in Appendix A.1 and the posterior distribution obtained is as follows

$$P(z_{a,b} = \kappa | \mathbf{w}, \mathbf{z}_{\neg a,b}; \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \frac{\alpha_\kappa + n_{a,\bullet}^{(\kappa)\neg a,b}}{\sum_{k=1}^{K} \alpha_k + n_{a,\bullet}^{(k)\neg a,b}} \times \frac{\beta_{w_{a,b}} + n_{\bullet,w_{a,b}}^{(\kappa)\neg a,b}}{\sum_{v=1}^{V} \beta_v + n_{\bullet,v}^{(\kappa)\neg a,b}} \tag{2.17}$$

where $n_{i,v}^{(k)}$ is the number of times the vth word in a vocabulary W, appears in document i and is assigned topic k. We can formally define $n_{i,v}^{(k)}$ as

$$n_{i,v}^{(k)} = \sum_{j=1}^{N_i} \mathbf{1}_{\{w_{i,j}=v \cap z_{i,j}=k\}} \tag{2.18}$$

$$= \sum_{j=1}^{N_i} \mathbf{1}_{\{w_{i,j}=v\}} \mathbf{1}_{\{z_{i,j}=k\}}. \tag{2.19}$$

We also let the $\bullet$ indicate the summation over the corresponding index, such that

$$n_{\bullet,v}^{(k)} = \sum_{i=1}^{M} \sum_{j=1}^{N_i} \mathbf{1}_{\{w_{i,j}=v\}} \mathbf{1}_{\{z_{i,j}=k\}} \tag{2.20}$$

and let $\neg$ indicate the exclusion of specific components such that

$$n_{a,\bullet}^{(k)\neg a,b} = \sum_{j \neq b} \mathbf{1}_{\{z_{a,j}=k\}}. \tag{2.21}$$

The left fraction in equation (2.17) can intuitively be seen as topic-document probability whilst the right fraction can be seen as the topic-word probability.

As in regular Gibbs Sampling we then sample from the above posterior for each document-word pair (a, b) and update the assigned topics iteratively. This is done until the posterior converges as is detailed in Algorithm 2 below.

Algorithm 2: Collapsed Gibbs Sampling for LDA

input : A corpus $\mathbf{w} = (\mathbf{w}_1, ..., \mathbf{w}_M)$

output: Topic assignments $\mathbf{z}$ for all words $w_{i,j}$ and counts $n_{i,v}^{(k)}$ for all documents, words and topics.

initialize $z_{i,j}^{(0)}$ by selecting a topic $1, ..., K$ with equal probabilities $\frac{1}{K}$.

calculate the initial counts $n_{i,v}^{(k)}$.

$t = 1$

while *not converged* **do**

  for *document* $a = 1, ..., M$ **do**

    for *word* $b = 1, ..., N_i$ **do**

      sample a new topic $z_{a,b}^{(t)}$ according to

      $P(z_{a,b}^{(t)} | \mathbf{w}, \mathbf{z}_{\{(x,y):(x<a) \text{ or } (x=a,y<b)\}}^{(t)}, \mathbf{z}_{\{(x,y):(x>a) \text{ or } (x=a,y>b)\}}^{(t-1)}; \boldsymbol{\alpha}, \boldsymbol{\beta})$

      update $n_{a,w_{a,b}}^{\left(z_{a,b}^{(t)}\right)}$ += 1 and $n_{a,w_{a,b}}^{\left(z_{a,b}^{(t-1)}\right)}$ −= 1

    end

  end

  $t = t + 1$

end

Given the final topic assignments $\mathbf{z}$ and the final counts $n_{i,v}^{(k)}$ we can then estimate the document-topic and topic-word probabilities as follows, using the conjugacy of the Dirichlet and categorical distributions [9]. This gives us an estimate of $\boldsymbol{\theta}_{i,k}$ as follows

$$\hat{\boldsymbol{\theta}}_{i,k} = \frac{\alpha_k + n_{i,\bullet}^{(k)}}{\sum_{l=1}^{K} \alpha_k + n_{i,\bullet}^{(l)}}. \tag{2.22}$$

The same argument can be made for $\boldsymbol{\phi}_{k,v}$ giving us

$$\hat{\boldsymbol{\phi}}_{k,v} = \frac{\beta_{k,v} + n_{\bullet,v}^{(k)}}{\sum_{u=1}^{V} \beta_{k,u} + n_{\bullet,u}^{(k)}}. \tag{2.23}$$

2.3.3 Choice of Hyperparameter

When using either of the approximative inference algorithms discussed in the previous sections we need to specify the hyperparameters of our Dirichlet priors, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, as well as the number of topics K. These can of course be set to fixed values as is done in a lot of the literature [13][9]. We can however also determine these hyperparameters using a variety of different methods.

2.3.3.1 Empirical Bayes

A more statistical approach to the inference of the hyperparameters for the Dirichlet priors, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, is an Empirical Bayes approach which is described in *Blei et al. (2003)* [3].

We use a Variational EM algorithm where the variational E-step corresponds to minimizing the KL-divergence between the true posterior and the variational distribution w.r.t. the variational parameters $\boldsymbol{\gamma}$, $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$. This is equivalent to maximizing $L(\boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}; \boldsymbol{\alpha}, \boldsymbol{\beta})$ w.r.t. the variational parameters as is described in Section 2.3.2.1 and detailed in Appendix A.2. Once the approximate posterior has been found, indexed by some optimal values of the variational parameters, we perform the variational M-step which is the maximization of $L(\boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}; \boldsymbol{\alpha}, \boldsymbol{\beta})$ w.r.t. the hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ where we fix the variational parameters to their optimal values obtained in the E-step.

In a similar fashion as for the update equations for the Variational Bayesian Inference algorithm we can find optimal values for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that minimize the KL-divergence between the approximate and true posteriors, i.e. they maximize $L(\boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}; \boldsymbol{\alpha}, \boldsymbol{\beta})$ as described in Section 2.3.2.1. These derivations are detailed in Appendix A.5. It turns out that no analytical solution exists and we instead have to resort to numerical means of maximizing the above expression. In *Blei et al. (2003)* the proposed method is a Newton-Raphson approach for which the derivations are detailed in Appendix A.5 [3].

The Newton-Raphson approach is based upon the following derivatives which are used for the maximization of $L(\boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}; \boldsymbol{\alpha}, \boldsymbol{\beta})$. We have that

$$\frac{\partial L}{\partial \alpha_k} = M\left(\Psi\left(\sum_{l=1}^{K}\alpha_l\right) - \Psi\left(\alpha_k\right)\right) + \sum_{i=1}^{M}\left(\Psi\left(\boldsymbol{\gamma}_{i,k}\right) - \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_{i,l}\right)\right), \quad (2.24)$$

$$\frac{\partial L}{\partial \alpha_k \alpha_h} = M\left(\delta_{k,h}\Psi'\left(\alpha_k\right) - \Psi'\left(\sum_{l=1}^{K}\alpha_l\right)\right), \quad (2.25)$$

where $\delta_{k,h}$ is the Kronecker delta. In an almost identical manner we have that

$$\frac{\partial L}{\partial \beta_v} = K\left(\Psi\left(\sum_{u=1}^{V}\beta_u\right) - \Psi\left(\beta_v\right)\right) + \sum_{k=1}^{K}\left(\Psi\left(\boldsymbol{\lambda}_{k,v}\right) - \Psi\left(\sum_{u=1}^{V}\boldsymbol{\lambda}_{k,u}\right)\right), \quad (2.26)$$

$$\frac{\partial L}{\partial \beta_v \beta_s} = K\left(\delta_{v,s}\Psi'\left(\beta_v\right) - \Psi'\left(\sum_{u=1}^{V}\beta_u\right)\right). \quad (2.27)$$

With estimates of the variational parameters we can then use Newton-Raphson as described previously to update $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. This process is then repeated until convergence which finally yields a final set of values for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ (whilst also returning optimal values for the variational parameters $\boldsymbol{\gamma}, \boldsymbol{\pi}$ and $\boldsymbol{\lambda}$ from the last Variational E-step). This procedure can be neatly summarized in Algorithm 3 below.

Algorithm 3: Variational EM

input : A corpus $\mathbf{w} = (\mathbf{w}_1, ..., \mathbf{w}_M)$ and an initial set of hyperparameters $\boldsymbol{\alpha}^{(0)}$, $\boldsymbol{\beta}^{(0)}$

output: Estimates for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$

set $t = 1$

while *not converged* **do**

Variational E-Step:

$$(\boldsymbol{\gamma}^{(t)}, \boldsymbol{\pi}^{(t)}, \boldsymbol{\lambda}^{(t)}) = \underset{\boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}}{\operatorname{argmax}} \, L(\boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}; \boldsymbol{\alpha}^{(t-1)}, \boldsymbol{\beta}^{(t-1)})$$

Variational M-Step:

$$(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}) = \underset{\boldsymbol{\alpha}, \boldsymbol{\beta}}{\operatorname{argmax}} \, L(\boldsymbol{\gamma}^{(t)}, \boldsymbol{\pi}^{(t)}, \boldsymbol{\lambda}^{(t)}; \boldsymbol{\alpha}, \boldsymbol{\beta})$$

$t = t + 1$

end

2.3.3.2 Model Comparison Using Coherence

The Empirical Bayes approach discussed in the previous section provides a good foundation for inferring the hyperparameters of the Dirichlet priors. However, it does not provide a way to determine the optimal number of topics K. A straightforward way to determine K is to compare the performance obtained using a different number of topics. In order to do this however we require a way to compare the performance of different models.

One of the most frequent metrics used to compare probabilistic topic models is perplexity which can be seen as a variation of the likelihood on a test corpus, i.e. a set of documents not used in the training of the model [3][14]. In topic modeling in general the use of perplexity to determine model performance has faced some criticism as the metric correlates negatively with topic interpretability [15]. Given that the main scope of this thesis is to organize a corpus into interpretable topics it seems questionable to optimize our hyperparameters using a measure that does not correspond to improved interpretability [16]. As such we have chosen to use topic coherence as the basis of our model comparison.

There exist several different coherence metrics that all measure the fraction of in-topic co-occurrences for the terms with the highest probability in the conditional categorical distribution, i.e. the terms v in topic k with the largest values $\boldsymbol{\phi}_{k,v}$. This approach is suitable for LDA as the most probable words in each topic are usually distinct and directly associated with the topic itself, at least as long as proper pre-processing has been conducted, as can be seen in the word clouds in Chapter 4.

2.3.3.2.1 C_V Coherence

A quite nuanced approach to determining the coherence of a topic model is introduced in *Röder et al. (2015)* [17]. As with other coherence measures we examine the top n words in a specific topic and define this set of n words for topic k as $W^{(k)}$. The measure, referred to as the C_V measure, is made up of several steps as detailed below[18].

(i).

For each topic k and each word $w \in W^{(k)}$ we let $S_l^{(k)}$ be the pair $(w, W^{(k)})$. These pairs are then used to measure the extent to which $W^{(k)}$ supports w.

(ii).

Estimate the probabilities of words $P\left(W_l^{(k)}\right)$ and pairs of words $P\left(W_l^{(k)}, W_h^{(k)}\right)$

through boolean document calculations. These probabilities can be expressed as

$$P\left(W_l^{(k)}\right) = \frac{\sum_{i=1}^{M} \mathbf{1}_{\left\{W_l^{(k)} \in \mathbf{w}_i\right\}}}{M}, \tag{2.28}$$

$$P\left(W_l^{(k)}, W_h^{(k)}\right) = \frac{\sum_{i=1}^{M} \mathbf{1}_{\left\{W_l^{(k)} \in \mathbf{w}_i \cap W_h^{(k)} \in \mathbf{w}_i\right\}}}{M}. \tag{2.29}$$

The probabilities above are not identical to those proposed in *Röder et al. (2015)*. The original C_V measure uses sliding window boolean document calculations instead of the regular approach detailed above [17]. In the sliding window approach we have a window length s, and instead of examining the actual documents $D_{\text{test}} = (\mathbf{w}_1, ..., \mathbf{w}_M)$, we examine the virtual documents $\mathbf{w}'_1 = (w_{1,1}, ..., w_{1,s})$, $\mathbf{w}'_2 = (w_{1,2}, ..., w_{1,s+1})$ and so forth where we move the window one word at the time [18]. This approach attempts to capture the proximity of words within documents and not only their co-occurrences. Although the idea of incorporating a sliding window might prove fruitful in longer documents where different parts may differ greatly, it seems unlikely that it would be of any use in the corpus we are examining. As the documents in our corpus are short, if not very short, each document will most likely consist of a single window, making the virtual documents $\mathbf{w}'_i$ identical to the actual documents $\mathbf{w}_i$. It could also be argued that the sliding window could be seen as redundant as every document should consist of a single question, making the need for incorporating the proximity somewhat unnecessary unless the length of the sliding window is small.

(iii).
For every pair $S_l^{(k)}$ from **(i).** we determine $\rho_{S_l^{(k)}}$ which indicates the similarity between $W_l^{(k)}$ and $W^{(k)}$. We define the normalized pointwise mutual information (NPMI) as follows [18]

$$\text{NPMI}\left(W_l^{(k)}, W_h^{(k)}\right) = \left(\frac{\log\left(\frac{P\left(W_l^{(k)}, W_h^{(k)} + \varepsilon\right)}{P\left(W_l^{(k)}\right) P\left(W_h^{(k)}\right)}\right)}{-\log\left(P\left(W_l^{(k)}, W_h^{(k)} + \varepsilon\right)\right)} \right) \tag{2.30}$$

where ε is some small number to prevent taking the logarithm of 0. Using the definition of NPMI above we calculate $\rho_{S_l^{(k)}}$ as follows

$$u_l = \left\{ \text{NPMI}\left(W_l^{(k)}, W_h^{(k)}\right)^{\zeta} \right\}_{h=1,...,n}, \tag{2.31}$$

$$v = \left\{ \sum_{l=1}^{n} \text{NPMI}\left(W_l^{(k)}, W_h^{(k)}\right)^{\zeta} \right\}_{h=1,...,n}, \tag{2.32}$$

$$\rho_{S_l^{(k)}}(u_l, v) = \frac{\sum_{h=1}^{n} u_{l,h} \cdot v_h}{\|u_l\|_2 \cdot \|v\|_2}. \tag{2.33}$$

By altering ζ we can choose to place further, or less, emphasis on higher NPMI values. Also note that $\rho_{S_l^{(k)}}(u_l, v)$ is the cosine similarity between u_l and v.

(iv).
The measures for all pairs can then be aggregated using the arithmetic mean into either metrics for each topic or for the model as a whole [18]. That is

$$\rho_k = \frac{\sum_{l=1}^{n} \rho_{S_l^{(k)}}(u_l, v)}{n}, \tag{2.34}$$

$$\rho = \frac{\sum_{k=1}^{K} \rho_k}{K}, \tag{2.35}$$

where ρ corresponds to the final metric that we use to compare our models as we are usually interested in the model as a whole and not specific topics. We set $\zeta = 1$, $\varepsilon = 10^{-12}$ and $n = 5$ as is suggested in previous literature [18][17][19]. We also set the length of the sliding window in such a way that the set of virtual documents $\mathbf{w}'$ is equivalent to the set of actual documents, i.e. the corpus. Here it should be noted that $\rho, \rho_k, \rho_{S_l^{(k)}} \in [0, 1]$.

2.4 Issues with LDA on Short Texts

One of the issues with LDA, in specific cases, is that it assumes that a document is a mixture of multiple topics. Although this assumption might hold for certain types of documents such as news articles, chapters in a book or Wikipedia articles it could be considered dubious for shorter types of documents, such as tweets or as in our case, shorter questions, where a specific document will contain few words and rarely the same word more than once or twice, especially after the data have been pre-processed.

The specific issues which arise is that we have a very sparse document-term matrix which makes inference problematic both mathematically as it is difficult to efficiently capture the co-occurrences when the documents are short. It is also computationally difficult as corpora consisting of short texts generally include numerous documents in which each is assumed to have a unique mixture of topics. There are several different models that deal with this issue in different ways, such as Biterm Topic Models (BTM) and Dirichlet Multinomial Mixture (DMM) models [5][20]. In this thesis we have decided to examine, beside regular LDA, a variant of Dirichlet Multinomial Mixture models as will be described in the following section.

2.5 Dirichlet Multinomial Mixture Models

DMM models bear very close resemblance to LDA models. The main difference between the two models is that whilst LDA assumes that a document is a mixture of topics, DMM assumes that a document is short and hence only contains a single topic [5][3].

2.5.1 Model Definition

Similar for LDA we can describe the generative process of a corpus in a DMM model as follows [5]

1. Choose $\phi_k \sim Dir(\boldsymbol{\beta})$ for $k = 1, ..., K$
2. Choose a topic distribution $\boldsymbol{\theta} \sim Dir(\boldsymbol{\alpha})$
3. For each document $i = 1, ..., M$
   (a) Choose a document length $N_i \sim Po(\xi)$
   (b) Choose a topic $z_i \sim Cat(\boldsymbol{\theta})$
   (c) For each word $j = 1, ..., N_i$
      i. Choose a word $w_{i,j} \sim Cat(\boldsymbol{\phi}_{z_i})$

The generative process of a corpus as detailed above can then be described in plate representation as can be seen in Figure 2.4 below.

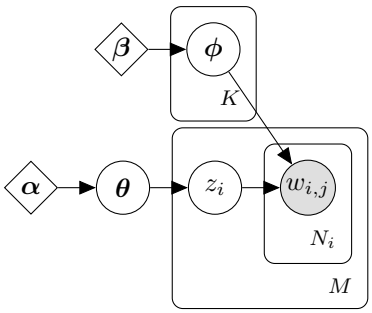


Figure 2.4: Plate representation of a DMM

We can also present the model using a more probabilistic formulation

$$\boldsymbol{\theta} \overset{iid}{\sim} Dir(\boldsymbol{\alpha}), \tag{2.36}$$

$$\boldsymbol{\phi}_k \overset{iid}{\sim} Dir(\boldsymbol{\beta}) \qquad k = 1, ..., K, \tag{2.37}$$

$$z_i|\boldsymbol{\theta} \overset{indep}{\sim} Cat(\boldsymbol{\theta}) \qquad i = 1, ..., M, \tag{2.38}$$

$$w_{i,j}|z_i, \boldsymbol{\Phi} \overset{iid}{\sim} Cat(\boldsymbol{\phi}_{z_i}) \qquad i = 1, ..., M \quad j = 1, ..., N_i. \tag{2.39}$$

Furthermore since we have independent and identically distributed categorical distributions for words in the same document, as they all depend on the same latent variable z_i, we can express the distribution of the elements in the document-term matrix X, where X_i is the word frequencies in the ith document, as follows

$$X_i|z_i, \boldsymbol{\Phi} \sim MN(N_i, \boldsymbol{\phi}_{z_i}). \tag{2.40}$$

We can then obtain the joint probability of (assuming that we have fixed document lengths, $N_1, ..., N_M$) the entire model as follows

$$P(\mathbf{w}, \mathbf{z}, \boldsymbol{\Theta}, \boldsymbol{\Phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = P(\boldsymbol{\Theta}; \boldsymbol{\alpha}) \prod_{k=1}^{K} P(\boldsymbol{\phi}_k; \boldsymbol{\beta}) \prod_{i=1}^{M} P(z_i|\boldsymbol{\theta}) \prod_{j=1}^{N_i} P(w_{i,j}|\boldsymbol{\phi}_{z_i}). \tag{2.41}$$

Like for LDA, DMM suffers from the same issue of intractability as described in relation to equation (2.9) [5]. Due to this we once again have to resort to approximate inference. Like with LDA there exist several different applicable inference methods for DMM. In the original paper by Yin et al. the authors specifically propose Collapsed Gibbs Sampling, labeling their method GSDMM (Gibbs Sampling Dirichlet Multinomial Mixtures) where the results bear close resemblance to those for LDA introduced by Griffiths et al. [5][9]. Although the idea of using Collapsed Gibbs Sampling for inference is the method mentioned in the paper by Yin et al. we can also use a Variational Bayesian Inference, as is done for LDA [3].

2.5.1.1 Collapsed Gibbs Sampler

As for LDA the goal of the Collapsed Gibbs Sampler for DMM is to collapse the Dirichlet priors of the multinomial document-topic and topic-word distributions. In a very similar fashion to LDA we derive the posterior probability of the topic of document a, z_a, conditioned on the topics of all other documents, $\mathbf{z}_{\neg a}$. The posterior is as follows and the full derivations can be found in Appendix A.3,

$$\begin{aligned} P(z_a = \kappa|\mathbf{z}_{\neg a}, \mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto & \frac{\left(\alpha_\kappa + m^{(\kappa)\neg a}\right)}{\sum_{k=1}^{K} \alpha_k + m^{(k)}} \\ & \times \frac{\prod_{v \in \mathbf{w}_a} \prod_{n=1}^{m_{a,v}^{(\bullet)}} \left(\beta_v + m_{\bullet,v}^{(\kappa)\neg a} - m_{a,v}^{(\bullet)} + n - 1\right)}{\prod_{n=1}^{N_a} \left(\sum_{v=1}^{V} \left(\beta_v + m_{\bullet,v}^{(\kappa)\neg a}\right) - N_a + n - 1\right)} \end{aligned} \tag{2.42}$$

where $m^{(k)}$ is the number of documents assigned topic k and $m_{i,v}^{(k)}$ is the number of times the vth word in a vocabulary W appears in document i when document i has topic k, i.e. $z_i = k$. Formally these quantities can be defined as

$$m^{(k)} = \sum_{i=1}^{M} \mathbf{1}_{\{z_i=k\}}, \tag{2.43}$$

$$m_{i,v}^{(k)} = \mathbf{1}_{\{z_i=k\}} \sum_{j=1}^{N_i} \mathbf{1}_{\{w_{i,j}=v\}}. \tag{2.44}$$

We also use the same summation, $\bullet$, and exclusion, $\neg$, notation as for LDA. We once again note that the upper fraction in equation (2.42) can be seen as the document-topic probability and that the lower fraction can be seen as the topic-word probability. As for LDA we then sample from the above posterior for each document a and update the assigned topics iteratively. This is done until the posterior converges as is detailed in Algorithm 4 below.

Algorithm 4: Collapsed Gibbs Sampling for DMM

input : A corpus $\mathbf{w} = (\mathbf{w}_1, ..., \mathbf{w}_M)$

output: Topic assignments $\mathbf{z}$ for all documents, counts $m^{(k)}$ and $m_{\bullet,v}^{(k)}$ for all documents, terms, and topics.

initialize $z_{i,j}^{(0)}$ by selecting a topic $1, ..., K$ with equal probabilities $\frac{1}{K}$.

calculate the initial counts $m_{i,v}^{(k)}$ in the corpus $\mathbf{w}$.

set $t = 1$

while *not converged* **do**

  for *document* $a = 1, ..., M$ **do**

    sample a new topic $z_a^{(t)}$ according to $P(z_a^{(t)} | \mathbf{w}, \mathbf{z}_{\{x:x<a\}}^{(t)}, \mathbf{z}_{\{x:x>a\}}^{(t-1)}, \boldsymbol{\alpha}, \boldsymbol{\beta})$

    update $m^{(z_a^{(t)})}$ += 1 and $m^{(z_a^{(t-1)})}$ −= 1

    for *word in vocabulary* $v = 1, ..., V$ **do**

      update $m_{\bullet,v}^{(z_a^{(t)})}$ += $m_{a,v}^{(\bullet)}$ and $m_{\bullet,v}^{(z_a^{(t-1)})}$ −= $m_{a,v}^{(\bullet)}$

    end

  end

  $t = t + 1$

end

Note that we do not need to update $m_{a,v}^{(\bullet)}$ in the algorithm as this is the number of times the vth word in the vocabulary appears in document a, regardless of the topic, which is fixed.

Once the posterior converges, giving us the final set of topics $\mathbf{z}$, as well as the counts $m^{(k)}$ and $m_{\bullet,v}^{(k)}$, we can, as for LDA, obtain the following estimates due to the conjugacy between the categorical and Dirichlet distributions [5]

$$\hat{\boldsymbol{\theta}}_k = \frac{\alpha_k + m^{(k)}}{\sum_{k=1}^{K} \alpha_k + m^{(k)}}, \tag{2.45}$$

$$\hat{\boldsymbol{\phi}}_{k,v} = \frac{\beta_v + m_{\bullet,v}^{(k)}}{\sum_{v=1}^{V} \beta_v + m_{\bullet,v}^{(k)}}. \tag{2.46}$$

2.5.1.2 Variational Bayesian Inference

Like for LDA we can also use Variational Bayesian methods to make inference for DMM. In order to do this we reduce our model to that of a variational family of distributions which is parameterized by $\boldsymbol{\gamma}$, $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$. The reduced model is very similar to the one which was introduced for LDA and is also chosen in such a way that we have separability between $\mathbf{z}$, $\boldsymbol{\Phi}$ and $\boldsymbol{\theta}$, as is suggested for Bayesian networks

with latent variables [10]. We also note that since we only have one $\boldsymbol{\theta}$ we no longer have the vector of document-topic distributions $\boldsymbol{\Theta}$, as we did for LDA.

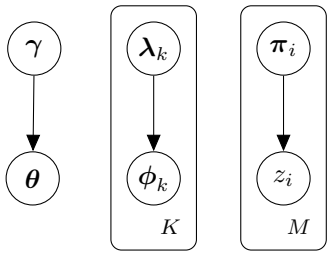


Figure 2.5: Plate representation of the reduced DMM model

The mean field approximation which assumes separability yields the following, fully factorized, family of variational distributions [10]

$$q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\Phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}) = P(\boldsymbol{\theta}; \boldsymbol{\gamma}) \prod_{k=1}^{K} P(\boldsymbol{\phi}_k; \boldsymbol{\lambda}_k) \prod_{i=1}^{M} P(z_i; \boldsymbol{\pi}_i). \tag{2.47}$$

To find the update equations for the variational parameters we once again minimize the Kullback-Leibler Divergence defined as follows

$$D(q||p) = \int \int \sum_{\mathbf{z}} q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\Phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}) \log \left(\frac{q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\Phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda})}{p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\Phi} | \mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})} \right) d\boldsymbol{\theta} d\boldsymbol{\phi}. \tag{2.48}$$

The derivations of this minimization are included in Appendix A.4 and yields the following update equations

$$\boldsymbol{\gamma}_k = \alpha_k + \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,k}, \tag{2.49}$$

$$\boldsymbol{\pi}_{i,k} \propto \exp \left\{ \Psi\left(\boldsymbol{\gamma}_k\right) + \sum_{j=1}^{N_i} \Psi(\boldsymbol{\lambda}_{k,w_{i,j}}) - \Psi\left(\sum_{u=1}^{V} \boldsymbol{\lambda}_{k,u} \right) \right\}, \tag{2.50}$$

$$\boldsymbol{\lambda}_{k,v} = \beta_v + \sum_{i=1}^{M} \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,k} \mathbf{1}_{\{w_{i,j}=v\}}. \tag{2.51}$$

We update the variational parameters as per the update equations above until convergence as described in Algorithm 5 below.

Algorithm 5: Variational Bayes for DMM

input : A corpus $\mathbf{w} = (\mathbf{w}_1, ..., \mathbf{w}_M)$
output: Converged estimates of $\boldsymbol{\gamma}$, $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$
initialize $\boldsymbol{\pi}_{i,k}^{(0)}$ as $\frac{1}{K}$.
initialize $\boldsymbol{\gamma}_k^{(0)}$ as $\alpha_k + \sum_{i=1}^{M} \boldsymbol{\pi}_{i,k}^{(0)}$.
initialize $\boldsymbol{\lambda}_{k,v}^{(0)}$ as $\beta_v + \sum_{i=1}^{M} \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,k}^{(0)} \mathbf{1}_{\{w_{i,j}=v\}}$
set $t = 1$
while *not converged* **do**
  for *topic* $k = 1, ..., K$ **do**
    for *document* $i = 1, ..., M$ **do**
      update
      $\boldsymbol{\gamma}_k^{(t)} = \alpha_k + \sum_{i=1}^{M} \boldsymbol{\pi}_{i,k}^{(t-1)}$
      update
      $\boldsymbol{\pi}_{i,k}^{(t)} = \exp\left\{\Psi\left(\boldsymbol{\gamma}_k^{(t-1)}\right) + \sum_{j=1}^{N_i} \Psi(\boldsymbol{\lambda}_{k,w_{i,j}}^{(t-1)}) - \Psi\left(\sum_{u=1}^{V} \boldsymbol{\lambda}_{k,u}^{(t-1)}\right)\right\}$
    end
    for *word in vocabulary* $v = 1, ...V$ **do**
      update
      $\boldsymbol{\lambda}_{k,v}^{(t)} = \beta_v + \sum_{i=1}^{M} \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,k}^{(t-1)} \mathbf{1}_{\{w_{i,j}=v\}}$
    end
    normalize all $\boldsymbol{\pi}_{i,k}^{(t)}$ by dividing by $\sum_{l=1}^{K} \boldsymbol{\pi}_{i,l}^{(t)}$
  end
  $t = t + 1$
end

As $\boldsymbol{\gamma}_k$ is approximately the kth Dirichlet topic prior α_k plus the expected number of documents generated from the kth topic, we can approximate $\boldsymbol{\theta}_k$ as follows due to the conugacy of the Dirichlet and multinomial distributions

$$\hat{\boldsymbol{\theta}}_k = \frac{\boldsymbol{\gamma}_k}{\sum_{l=1}^{K} \boldsymbol{\gamma}_l}. \tag{2.52}$$

The same argument can be made for $\boldsymbol{\lambda}_{k,v}$ and the approximation of $\boldsymbol{\phi}_{k,v}$

$$\hat{\boldsymbol{\phi}}_{k,v} = \frac{\boldsymbol{\lambda}_{k,v}}{\sum_{l=1}^{K} \boldsymbol{\lambda}_{l,v}}. \tag{2.53}$$

Given the approximate posterior determined by the variational parameters we can also use a variational EM algorithm to find maximum likelihood estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ which is briefly discussed in Section 2.5.3.2. [3].

2.5.2 Movie Group Process

A key property of DMM models, where Gibbs Sampling is the choice of inference (i.e. GSDMM models as introduced by Yin et al.), is that topics, or clusters, are removed during the training process [5]. The idea is then that we supply the models with a maximum number of possible topics K, instead of a fixed number of topics as for LDA models. In *Yin et al. (2014)* the authors propose a so called Movie Group Process which serves as an analogy to help explain how their model functions [5]. In short a Movie Group Process describes how a group of students in a class, with a list of movies they like, can be seated around K tables such that the students at each table have similar movies on their lists to discuss whilst also making the number of students at each table is as large as possible. In the case of GSDMM models the students are documents, the movies are words, and the tables are the topics.

Initially students are allocated randomly across the K tables. In each iteration students are asked to choose a table according to the following rules

1. Choose a table with more students
2. Choose a table whose students have similar movies on their list

Per this analogy, which is equivalent to GSDMM, we have that some tables will be completely empty once the students stop changing tables if their movie lists are similar enough and there is a large enough number of initial tables [5]. This leads to the final clusters being large as per the first rule whilst also being composed of similar documents as per the second rule.

2.5.3 Choice of Hyperparameters

Like for LDA when using either of the approximative inference algorithms discussed in the previous sections we need to specify the hyperparameters, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. These can of course be set to fixed values as is done in the initial literature [5].

2.5.3.1 Model Evaluation

When using the GSDMM approach where the number of topics K is inferred automatically as described in Section 2.5.2 whilst balancing the similarity within the clusters and simultaneously maximizing the size of the clusters [5]. Due to this property it is not necessary to evaluate the model for a vast number of K as the total number of uncovered topics, i.e. tables filled at the end of a movie group process, converges for all the data sets examined in *Yin et al. (2014)* [5]. Although an evaluation metric might not be needed for determining K one is still needed if we want to compare models with different Dirichlet hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

In the case of DMM models in general coherence, as introduced in Section 2.3.3.2, is not a suitable evaluation metric. As each document is made up of a single topic most topics will be diluted by general words that occur frequently across all topics. Although most of these are stripped from the vocabulary when pre-processing those that still remain, do more often than not, end up as the most frequently occurring words in multiple topics. Hence when using standard coherence measures that focus on the defining, i.e. the most frequently occurring, words of the topics these superficial terms are more often than not included, and given their occurrences across several topics, leads to a decrease in overall coherence even though the topics might be more distinct.

In order to determine appropriate values for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ we could instead use a wide variety of metrics to determine the model performance as the number of clusters are inferred automatically (note that the inference of the number of cluster depends on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as discussed in [5]). In this thesis we have however chosen to limit the choices of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to the default values of $\alpha_k = 0.1$ for all k and $\beta_v = 0.1$ for all v as is done in the original paper by Yin et al. [5].

2.5.3.2 Empirical Bayes

Given the similarities between LDA and DMM models we can use an almost identical Variational EM approach to estimating the hyperparameters as introduced for LDA in Section 2.3.3.1. The only thing that differs in Algorithm 3 is the slight alterations to the functions maximized in both the Variational E-step and the Variational M-step.

2.6 Determining the Topics of New Documents

Although we can infer the document-topic probabilities $\boldsymbol{\Theta}$ for documents in the training corpus doing so for new documents can be somewhat tricky. A straightforward, but computationally expensive, approach is to use folding-in approach, altered to suit LDA and DMM respectively, as heuristically introduced in *Hoffman (1999)* [4]. The idea of folding-in can be applied to both types of inference algorithms discussed

in this thesis and the idea is to incorporate the new documents in the main corpus and run the appropriate inference algorithm until convergence once more. For a Collapsed Gibbs Sampling approach we keep the topics of all terms for LDA and documents for DMM in the original training corpus unchanged, though they might change when iterating depending on the new documents supplied. As the topics in the initial corpus the have already converged in the previous training of the model the convergence of the new topics, which are initialized randomly as per Algorithm 2 and 4, should not require a significant number of iterations unless the new documents supplied are vastly different or in great quantity. For a Variational Bayesian approach we would keep $\boldsymbol{\lambda}$ as is whilst initializing components for the new documents in $\boldsymbol{\gamma}$ and $\boldsymbol{\pi}$ according to the initialization step in Algorithm 1 and 5. We then run the algorithm until convergence which, as for the Collapsed Gibbs Sampler, should not take too long. This process becomes computationally expensive as we iterate over the entire training corpus once more. Although computationally expensive the process can be sped up by freezing all other topics than the ones we wish to evaluate, hence we iterate only over the new documents until they converge.

A more simplistic approach would be to extend the idea of documents being composed of a single topic to LDA. Through this extension we can determine the probabilities of documents belonging to specific topics, under the assumption of equal probabilities for all topics, as follows

$$P(z_i = k|\mathbf{w}_i, \boldsymbol{\Theta}, \boldsymbol{\Phi}) \propto P(\mathbf{w}_i|z_i = k, \boldsymbol{\Phi}) = \prod_{j=1}^{N_i} \boldsymbol{\phi}_{k,w_{i,j}}. \tag{2.54}$$

For DMM models we can approximate the updated document-topic distribution obtained by folding-in new documents with the distribution obtained in the initial training, i.e. $\boldsymbol{\theta}$, in order to obtain the following simplified evaluation method

$$P(z_i = k|\mathbf{w}_i, \boldsymbol{\theta}, \boldsymbol{\Phi}) \propto P(z_i = k|\boldsymbol{\theta})P(\mathbf{w}_i|z_i = k, \boldsymbol{\Phi}) = \boldsymbol{\theta}_k \prod_{j=1}^{N_i} \boldsymbol{\phi}_{k,w_{i,j}}. \tag{2.55}$$

2.7 Incorporating Prior Knowledge

In most implementations of LDA and DMM it is assumed that the document-topic and topic-word priors are symmetric, i.e. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the vector of all ones multiplied by some scalars α and β. This assumption places an equal probability on all topics but also on all words. Having an equal probability of all topics seems to be a fair assumption as we rarely know the distribution of topics within a corpus. However the symmetry of the topic-word distributions could be questioned, especially if we have some prior knowledge of the topics that we wish to uncover or that we believe to exist in the corpus.

In this thesis we implement, and extend, the idea of, *Weak Supervision with Minimal Prior Knowledge* as introduced in *Lu et al. (2010)* where we change the symmetric prior $Dir(\boldsymbol{\beta})$ into the asymmetric prior $Dir(\boldsymbol{\beta}+\mathbf{c})$ [14]. Here $\mathbf{c}$ is the vector of length V that indicates the additional weight to be put on words, $1, ..., V$. If we have no prior knowledge at all we have that $\mathbf{c} = \mathbf{0}$. Note that $\mathbf{c}$ can be viewed as a vector of pseudo-counts if we examine the estimates of $\boldsymbol{\phi}_{k,v}$ through the lens of the Collapsed Gibbs Sampling inference approach in both models.

$$\hat{\boldsymbol{\phi}}_{k,v}^{\text{LDA}} = \frac{\beta_v + \mathbf{c}_v + n_{\bullet,v}^{(k)}}{\sum_{u=1}^{V} \beta_u + \mathbf{c}_u + n_{\bullet,u}^{(k)}}, \tag{2.56}$$

$$\hat{\boldsymbol{\phi}}_{k,v}^{\text{DMM}} = \frac{\beta_v + \mathbf{c}_v + m_{\bullet,v}^{(k)}}{\sum_{u=1}^{V} \beta_u + \mathbf{c}_u + m_{\bullet,u}^{(k)}}. \tag{2.57}$$

Instead of viewing it as altering the Dirichlet prior we can see the addition of $\mathbf{c}$ as inflating the counts n in LDA or m in DMM [14]. From this it is obvious that the weights are dependent on the number of words in the corpus and have to be adjusted if the size of training corpus is altered.

This approach does however only utilize the fact we are aware of the existence of specific words, henceforth referred to as seed words, in a corpus. Although this approach will guide the model to uncover topics relating to these seed words we can further extend it by associating certain seed words with one another [14]. In the model proposed in *Lu et al. (2010)* we have that

$$\boldsymbol{\phi}_k \sim Dir(\boldsymbol{\beta} + \mathbf{c}) \quad \text{for all } k = 1, ..., K. \tag{2.58}$$

In order to associate seed words with one another we can define $\mathbf{c}_k$, denoting the additional weights for words $1, ..., V$ in topic k which leads to unique priors for all topic-word distributions.

$$\boldsymbol{\phi}_k \sim Dir(\boldsymbol{\beta} + \mathbf{c}_k) \quad \text{for all } k = 1, ..., K. \tag{2.59}$$

This extended approach further facilitates the possibility to extract specific topics if one has some prior knowledge of specific seed words that the topic should contain. An example relating to the data set we are examining in this thesis could be the want to discover a topic relating to Covid-19. In this case we could use our prior knowledge and add extra weight to for example *"covid", "vaccine"* and *"quarantine"* to some specific topic k.

The choice of the non-zero values in $\mathbf{c}_k$ depend on the how certain we are of the topics existence and specifically the existence of the seed words within that topic. Small values can be seen to represent a weak assumption of the existence of a certain topic containing the set of seed words and is essentially ignored (by ignored we mean that the counts n or m dominate the pseudo-counts $\mathbf{c}$) if there is no support for the set of seed words in the data. Larger values translate to a strong assumption of the existence of a topic containing the set of seed words and might be more appropriate if the topic we are looking for is rare, but its existence is almost certain. A possible issue with large values is that they might dominate the probabilities and as such they have to be applied with care.

Chapter 3

Data

Our data set consists of earnings call transcripts collected from 2008 up until the start of 2021. In all there are several million questions in the data set. We have chosen to focus our empirical study on the Q & A sessions, specifically the questions, in these transcripts as they contain questions primarily asked by institutional investors. As mentioned previously we can then choose to see the questions asked by these institutional investors as a proxy of what the market itself is concerned with.

In our transcripts we treat each question as a separate document, hence also assuming that questions are asked independently of one another which should be true in general (excluding the rare occasion of follow-up questions and other deviations). It turns out that including all available questions in our training corpus is problematic for several reasons as we will detail below.

As each question is asked to a specific company we can group these companies in several ways in order to create partitions of the complete data set. These partitions can be based on the industry of operation or some other features of these companies. For each of these partitions we have some set of companies that should have somewhat similar business practices if the partitions are distinct enough. For this set of companies we (generally) have quarterly earnings call that have been transcribed for over the greater part of a decade. For each earnings call we then finally have a set of questions that were asked during the Q & A session, which constitute our documents, $\mathbf{w}_i$, that make up our corpus, $\mathbf{w}$.

One of the main issues which arises is that there is an inherent difference in the questions asked in these different partitions. This is of course quite intuitive as one would assume that the questions would differ between pharmaceutical companies and companies operating in the energy sector. More closely related to our problem one would also assume that the ESG related questions would differ between the different industries, for example the questions related to ESG for pharmaceuticals companies will most likely be different compared to ESG questions companies in the energy sector. There are of course inherent differences between specific companies within the same partitions as well though these could be assumed to be less prominent in general than the differences between the partitions. In Section 4.1 we examine the issues in practice when modeling using non-partitioned data.

Furthermore given that the data are observed over a longer period of time it is not unreasonable to assume that there are some temporal differences in the data. Specifically in relation to our examination of ESG related topics we can see that the ESG trend was essentially non-existent before 2016 when simply examining the frequency of questions including *esg* in Figure 3.1 below (this is somewhat dubious as its only the umbrella term ESG which is new a not necessarily the aspects that it encompasses).

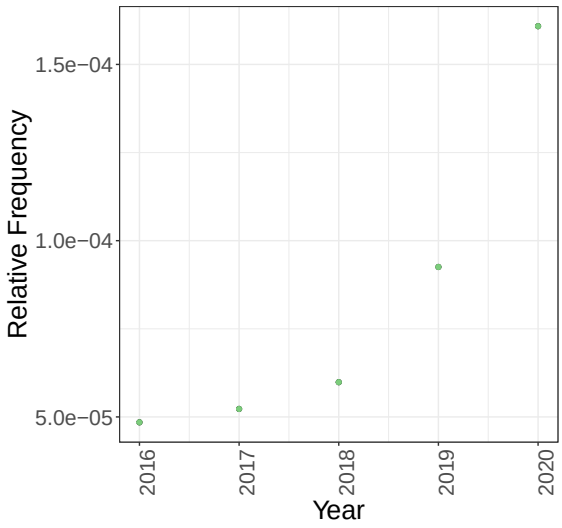


Figure 3.1: ESG Over Time

In all given the inherent differences and the size of our data set it is not unreasonable to examine some partition of the companies, e.g. sectors or industries, by themselves. It is also not unreasonable to examine them for some shorter period of time rather than throughout the entire span of the data set as ESG topics will be much more difficult to uncover if they are non-existent in a majority of the documents used. Hence when training our models we focus on documents from more recent years whilst then evaluating the rest of the documents using the trained model. This approach allows us to uncover topics that can intuitively be linked to certain ESG aspects and once those topics are verified we can evaluate the remainder of the data set in order to determine the prominence of those topics in a test corpus that spans some longer period of time.

3.1 Pre-processing

In order to properly utilize our models to their fullest extent we need to process the data in a variety of ways before passing it to our models. As the data set consists of transcripts of individuals speaking there are occurrences where words are considered indiscernible, these words are marked as being indiscernible in the transcripts and are as such removed before the rest of the pre-processing begins as to not inflate the corpus with occurrences of the word *indiscernible*.

For both LDA and DMM we have chosen a fairly straightforward pre-processing approach, which can surely be improved. To begin each document, or question, is tokenized, meaning that they are reduced to a list of lower case words, called tokens. In order to reduce the feature space further we have also chosen to exclude all numbers, symbols and punctuation, leaving us only with words. Using this approach we have that the following question

What caused the lack of growth during the last quarter?

is expressed as the following collection of tokens

{*what, caused, the, lack, of, growth, during, the, last, quarter*}

The list of tokens that represent each document are then further stripped of so called stop words. Stop words are extremely common words that would appear to be of little value in topic modeling in general as they should occur equally frequently in all topics [21]. In our case we reduce our vocabulary W by filtering out the list of SMART stop words as presented in *Lewis et al. (2004)*[22]. In the case of the aforementioned question the removal of these stop words would leave us with the following list of tokens

$$\{lack,\ growth,\ quarter\}$$

This final list of tokens would then represent the original question when modeling the corpus in which it is contained. Turning this list into a *bag-of-words*, which we then ultimately present in the form of a row in the document-term matrix, would yield a vector of length V with ones at the index of *lack*, *growth* and *quarter* and zeroes at all other positions.

3.1.1 Lemmatization and Stemming

After the documents have been tokenized and the stop words have been removed we lemmatize or stem the remaining tokens. Documents are, for grammatical reasons, going to include different forms of a word, such as *study*, *studying* and *studies*. In a model using *bag-of-words* these three words can be assumed to have approximately the same meaning and as such it is beneficial to treat them as the same token, both for computational and modeling reasons. This can be achieved by reducing them all to the same base form via some algorithm. Generally speaking there are two main methods for achieving this reduction, stemming and lemmatization. The difference between stemming in this case is that whilst lemmatization uses a lexicon to find the lemma of the three words, which is *study* in all three cases, stemming removes common prefixes and suffixes from words entirely to create reduced tokens, leading to the reduced *study*, *study* and *studi* [21]. Although stemming is in general a much simpler, and less time consuming, approach as it only requires a list of common prefixes and suffixes, whilst lemmatization requires and entire lexicon, it is more coarse and hence prone to the issue seen above where not all words are reduced to the same base form. As a result we have chosen to use lemmatization as the method of further reducing the feature space.

3.1.2 Token Frequencies

Token frequencies in corpora are in general very right-skewed with few high frequency words and an abundance of low frequency words, at least that is the case in our data. Low frequency words are in general difficult to model as it takes a sizable number of word occurrences in order to model the co-occurrences with other words into distinct topics. As such since low frequency words provide little to no value from a modeling point of view we can further reduce the size of our vocabulary W in order to improve computational efficiency. Below we have the word frequency distribution for documents dated during December of 2020 in the Earnings Call Q & A data set.

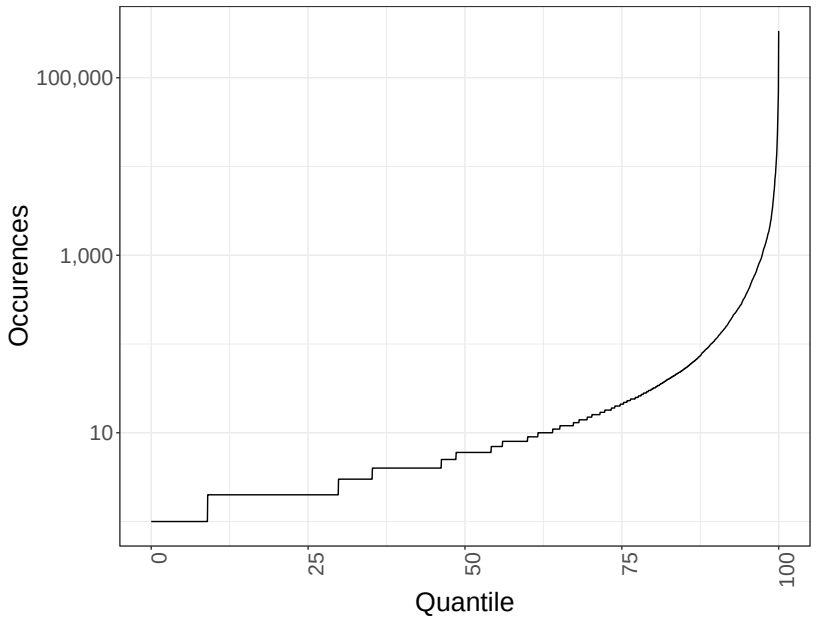


Figure 3.2: Word Occurrences

From Figure 3.2 above we see that a majority of the words in the corpus appear fewer than ten times. Furthermore we also see that there is a small subset of words that have drastically higher frequencies than the rest. After examination it turns out that these words are more or less superficial and can in some sense be perceived as corpus specific stop words, i.e. words appearing often yet containing no actual information in regards to topic distinction. As such we have chosen to focus on the words with counts between the 75th to 99.9th percentile.

3.1.3 Document Length

Another aspect which might cause problems in the modeling of our reduced documents is their length. As we at times trim documents quite drastically as seen with the hypothetical question in Section 3.1.1 (which is before the reduction based on token frequencies) we might encounter some issues when training the model. As such we have chosen to exclude documents that could be considered to be too short as the modeling of co-occurrences is problematic for these documents, specifically in the case of LDA, which as discussed might have some issues in the case of extremely short documents. As such we have chosen to filter out all documents that, after the previous pre-processing, contain fewer than 5 tokens.

Chapter 4

Empirical Illustrations

In this chapter we discuss the empirical findings from the models introduced in Chapter 2 when evaluated on our earnings call transcript data described in Chapter 3. The results presented are based on implementations of the models using `topicmodels` in R and `gensim` in Python for LDA [23][24]. For DMM we use both `gsdmm` as well as our own implementation of the model, based on Algorithm 4, in Python [25].

In this chapter we also present multiple word clouds that are based on the most frequent words in each topic, determined by the estimated topic-word probabilities, $\mathbf{\Phi}$, where the size of the word in a cloud depends on the probability of that word within the topic. Hence the largest words will have the highest probability within the topic.

4.1 Standard LDA and DMM

In order to highlight the need for the partitioning of the data into specific categories, as mentioned in the previous chapter, we begin by examining the non-partitioned data set. For this example we use a small fraction of the data, consisting of approximately 100 000 questions, as the aim is to highlight the need for the partitioning mentioned earlier and not necessarily to produce the best model possible. The questions were all asked during earnings calls dated during December of 2020.

4.1.1 LDA

For this example we use the Collapsed Gibbs Sampling inference approach discussed earlier and the default symmetric hyperparameters as introduced in the `textmineR` package, setting $\alpha = 0.1$ [13]. In the original paper by Griffiths et al., and in the `topicmodles` package, the suggested value for β is 0.1, this did however not coincide with the estimated hyperparameters obtained using the Empirical Bayes approach introduced in Section 2.3.3.1. The values obtained in this way were significantly smaller than 0.1, hence we have chosen to use 0.01 as our symmetric prior in accordance with the `textmineR` package as well [13][23]. Furthermore we use the C_V coherence metric as introduced in Section 2.3.3.2.1 in order to evaluate the performance for different numbers of topics K.

This example yields the following C_V coherence, with $\zeta = 1$, $\varepsilon = 10^{-12}$ and $n = 5$, for different values of K. The length of the sliding window has also been set to the maximum document length in the corpus such that the set of virtual documents is equivalent to the corpus as we believe that the length of the documents in our case makes the sliding window redundant.

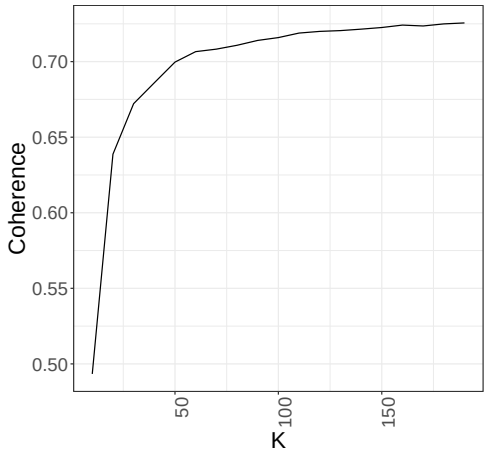


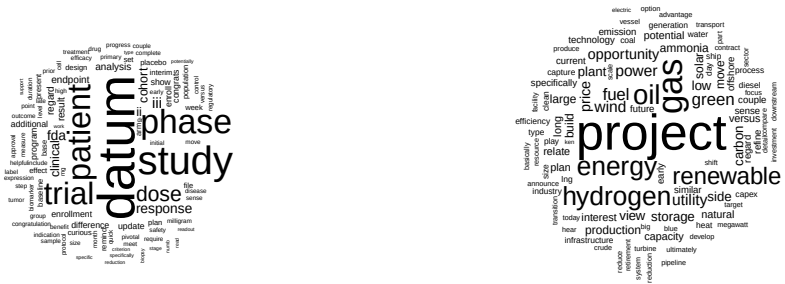
Figure 4.1: Coherence

In Figure 4.1 above we note that the in-sample coherence seems to stagnate at around $K = 50$, hence using something similar to elbow method, we choose that as the number of topics when proceeding with this example. Here it should be noted that although an increased number of cluster seems to have a fairly limited contribution w.r.t. topic coherence the new topics discovered when increasing K further can still be of interest, i.e. when attempting to uncover rarer topics. As this is just an illustrative example we choose the lowest K that still provides adequate coherence.

Here it should importantly be noted that this is the C_V coherence as evaluated on the training corpus, hence we are more than likely partially overfitting the model, meaning that the coherence on a corpus of held-out documents could be poor. Within the scope of this thesis however this is not really as much of a problem as one would initially believe as the aim is more related to the discovery of specific topics. Therefore we are somewhat fine with uncovering niche topics that might simply exist in the training corpus as long as the in-topic coherence, ρ_k, on the corpus of held-out documents for the sought after topics is good. This likely overfitting is of course something we will have to keep in mind going forward as some methods that we examine depend on the general coherence of the model. Figure 4.1 above can hence be seen as an indicator of the fact that there are further distinct topics to uncover in the training corpus, even though these might not exist in general.

Although this model seems to have distinct topics as is suggested by the coherence score they are of little use within the scope of this thesis. The topics that we uncover when examining the data set as a whole are broader and do not contribute in a meaningful way as they capture much larger aspects of the corpus than we would like. Furthermore in order to obtain more detailed topics that would be of value within the aim of this thesis we would have to use an incredibly large number of topics, which would make the computations infeasible, especially if we were to increase the number of documents as well as would be required for such a model.

Many of the broader topics that we uncover with this model encompass entire industries or sectors which is not surprising as a significant amount of questions will include some parts that are specific to either the company the question was asked in relation to or to the sector or industry of that company. Below are word clouds for two of the topics that could quite directly be attributed to specific industries or sectors.



(a) Pharmaceuticals **(b)** Energy

Figure 4.2: Industry/Sector Related Topics

Figure 4.2 a) includes a variety of words, such as *study* and *fda* (U.S Food and Drug Administration), that would indicate that the topic is fairly focused on the pharmaceutical industry. In Figure 4.2 b) we have quite clear indications, such as *gas* and even *energy*, that the topic is focused on the energy sector. In Figure 4.2 b) we also get a glimpse of words that actually relate to the scope of the thesis, such as *hydrogen*, *carbon* and *renewable* that can quite clearly be related to the environmental aspect of ESG within the energy sector.

Beside the industry related topics uncovered above we also obtain a variety of specific topics relating to the type of documents we are examining. Given that our corpus is made up of questions asked during earnings calls we obtain a plethora of topics relating to different aspects of finance. Furthermore we also obtain topics relating to the specific corpus as well. As our data consist of questions asked during the Q & A session of earnings calls we obtain topics in which this is reflected, as can be seen in the figure below.

Figure 4.3: Earnings Call

In the Figure 4.3 above we have that the word cloud includes words, such as *conference* and *presentation* that can be attributed to earnings calls in general. These types of topics do in general not provide any meaningful insight, at least not within the scope of this thesis, but are a by-product of the modeling and are necessary to avoid the dilution of other, perhaps more informative, topics with more general tokens.

Lastly, given that this example is using a small, but recent, data set we also obtain topics that could be considered time specific. By time specific we mean topics that would generally only appear in a fraction of the documents that were dated during a specific time period.

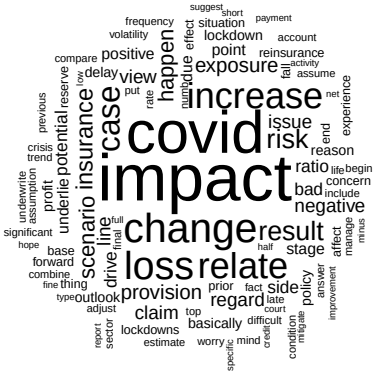


Figure 4.4: Covid-19

In Figure 4.4 above we have obtained a topic that centers around the Covid-19 pandemic including words such as *case* and *covid*. Given that the pandemic is temporary the topic should become highly infrequent once the pandemic is over, and once questions regarding the pandemic are fewer.

4.1.2 DMM

For this initial example we use GSDMM, i.e. a DMM model using the Collapsed Gibbs Sampler as the choice of inference, to examine the same data set as in the example for LDA. We use symmetric hyperparameters, setting $\alpha = 0.1$ and $\beta = 0.1$ as suggested in the initial paper by Yin et al. [5]. In the case of DMM models we set the maximum number of topics K, instead of a fixed number of topics as for LDA. We examine a large number of topics, that sadly all encounter the same issues.

When using the above DMM models the number of populated topics, i.e. occupied tables in the Movie Group Process, never decreased below the maximum number of topics as we would expect. The highest value of K that we examined was 1000. In the final example we obtained a fitted model with 1000 different topics which is not feasible to work with given the scope of this thesis. We could of course attempt to train a model with an even larger number of topics but this would also yield a model with at least 1000 topics populated and the same issues would arise once more. Examining all these topics becomes unfeasible, specifically when keeping in mind that topics are also diluted by words that appear frequently in the dictionary, even after pre-processing, as was discussed in relation to the usage of coherence as an evaluation metric for DMM models in Section 2.5.3.1.

This result indicates that DMM models cannot efficiently be employed on our data set as a whole, at least within the scope of this thesis, as the potential number of hidden topics is far to vast. DMM models also suffer from the fact that it is difficult to properly inspect the topics uncovered due to the aforementioned dilution caused by general words. As discussed in the previous section when examining LDA models we were essentially able to filter these general words into specific topics as can be seen in Figure 4.3 where both word clouds consist of many such general and corpus specific words. This could perhaps also suggest that the aforementioned assumption of a one-to-one relation between topics and documents could be dubious in the case of our corpus, or at least in the case of our pre-processed version of the corpus, as a substantial amount of topics uncovered by the GSDMM model includes general terms that could otherwise be grouped into a separate topic in an LDA model.

4.2 Partitioned Data

As is evident from the LDA example in the previous section an inherent cause of separation in the data is the groups to which the company the question was asked to belongs. As such in order to be able to extract more meaningful topics in regards to the scope of this thesis, without making computations infeasible due to the large number of topics needed as observed in the DMM example, we separate the data set by these industries as mentioned in Chapter 3. The main question that arises from this is the choice of level at which we separate the data set. The companies that we examine can all be separated according to the Global Industry Classification Standard (GICS) as expressed in Figure 4.5 below [26].

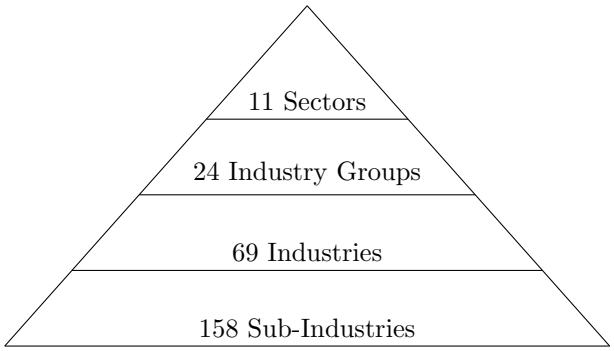


Figure 4.5: MSCI GICS Levels

We can partition the data set according to any of these levels with the tradeoff being fewer, but larger and less specific, partitions at the top and more numerous, but smaller and more distinct, partitions at the bottom. Splitting the data set too thin by using for example sub-industries might lead to a lack of data in certain partitions whilst using sectors might lead to similar issues as when examining the non-partitioned data, i.e. the uncovered topics representing the lower levels.

In order to attempt to solve the issues that arise with both extremes mentioned earlier we have chosen to examine the 69 industries as we believe that they should be distinct enough and that data availability should not be a problem for any of the industries. For the sake of the length of this thesis we will not be able to examine all 69 industries, instead we focus our efforts on a small subset where we believe that the methods discussed in this thesis would be most applicable as this would best highlight how the framework introduced could be used in practice.

In order to illustrate how the methods can be used we examine two specific industries, Metals & Mining and Oil, Gas & Consumable Fuels as we believe that the environmental aspect of ESG is likely to be discussed quite a lot in both. As such it seems reasonable to assume that we should be able to uncover at least an environmentally related topic in both of these industries.

As we now have a small subset of questions relating to a specific industry it becomes computationally feasible to work with questions from a wider time span than when we previously examined the data set as a whole. We also pre-process the data in a similar fashion as described in Section 3.1.

4.2.1 Metals & Mining

The Metals & Mining industry, which is categorized under the Materials sector, encompasses the production of metals such as aluminum, the mining of precious metals such as gold or silver, as well as several other sub-industries [26]. In all the training corpus consists of approximately 125000, after pre-processing, documents dated between 2018 and 2021. The vocabulary includes about 4400 unique tokens and the corpus includes 2.1 million tokens in total.

We begin by examining an LDA model, still using the default symmetric hyperparameters $\alpha = 0.1$ and $\beta = 0.01$ and once again using the C_V coherence for a wide array of K in order to determine the optimal number of topics. This yields the following results.

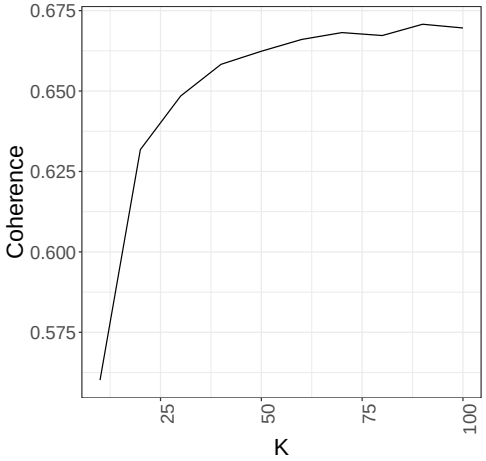


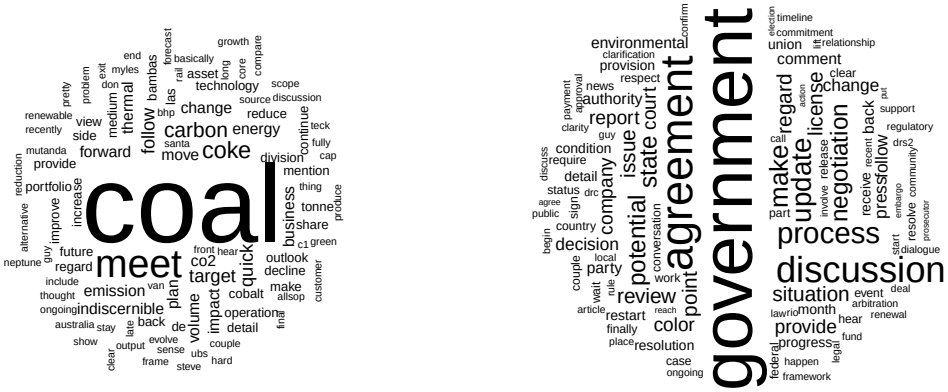
Figure 4.6: Coherence: Metals & Mining

As before we choose K at a level where the coherence starts to stagnate, hence we proceed with $K = 50$ when examining the specific topics obtained. When examining this reduced data set we discover a variety of topics relating to the mining of specific ores such as iron, or the production of others such as aluminum. As for the non-partitioned corpus we obtain topics that seem to include more earnings call and Q & A specific terms as well as general corpus specific terms. Two examples can be seen in the word clouds below.

(a) Mining **(b)** Growth

Figure 4.7: General Topics

In Figure 4.7 a) we have what seems to be a somewhat general mining topic that does not take the specific ore or material mined as these are covered in separate topics. Figure 4.7 b) on the other hand seems to concern growth and expansion in general. We also uncover, as expected, a topic somewhat relating to environmental sustainability as well as a topic that touches upon corporate governance



(a) Environmental **(b)** Corporate Governance

Figure 4.8: ESG Topics

In Figure 4.8 a) we have a topic which quite clearly seems to center around *coal* whilst also including more environmental aspects such as *co2*, *carbon* and *emission* which seems to further the notion that environmental questions are discussed to some degree within the industry. In Figure 4.8 b) the topic centers around government and includes a variety of sub-topics, i.e. topics that can be grouped under a single topic, that seemingly could be related to dealings with said government. This occurrence of sub-topics is one of the reasons as to why increasing the number of topics K might be fruitful in the case of wanting to uncover rare subjects, as these will be separated into their own, still distinct, topics. In the case of Figure 4.8 a) we might also be able to discover a more pure environmental topic if the total number of topics is larger. This is further discussed in Section 4.3 where seed words are applied in order to nudge the models in the direction of discovering these rare topics.

4.2.2 Oil, Gas & Consumable Fuels

The Oil, Gas & Consumable Fuels is categorized under the Energy sector and includes the production and storage of oil, gas and coal as well as several other sub-industries [26]. The training corpus consists of about 175000 documents, after pre-processing, dated between 2018 to 2020. The vocabulary includes about 5300 unique tokens and the corpus includes 2.7 million tokens in total.

We begin by examining an LDA model, as in the example for Metals & Mining with the default, symmetric, hyperparameters $\alpha = 0.1$ and $\beta = 0.01$.

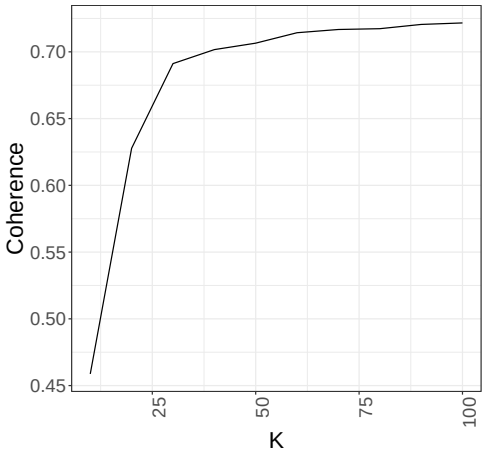
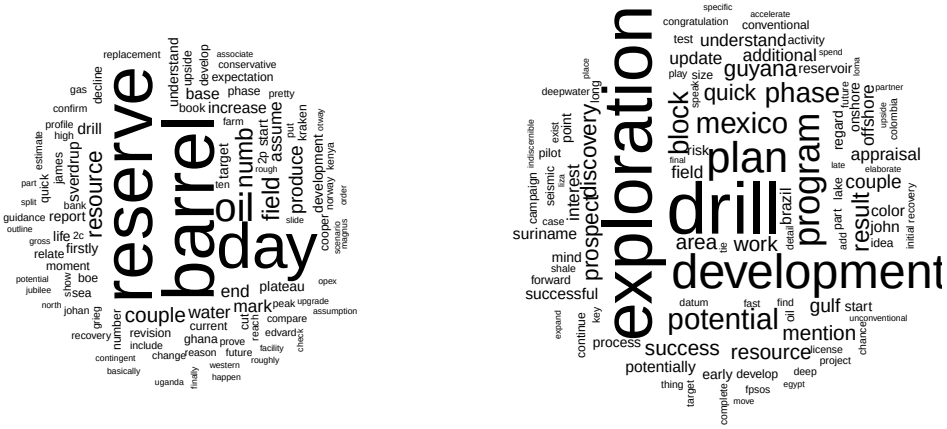


Figure 4.9: Coherence: Oil, Gas & Consumable Fuels

Once again we choose K at a level where the coherence starts to stagnate, hence we proceed with $K = 50$ when examining the specific topics obtained. As when examining the data set as a whole we still end up with topics that could be associated with lower levels of the GICS hierarchy as well as further general topics associated with different parts of the industry as a whole.



(a) Oil Price **(b)** Development

Figure 4.10: General Topics

In Figure 4.10 a) we have a word cloud which one would likely associate with questions regarding oil, and specifically the price of oil. In Figure 4.10 b) we have a more general topic that seems to focus on the development process, including both exploration, prospecting and drilling.

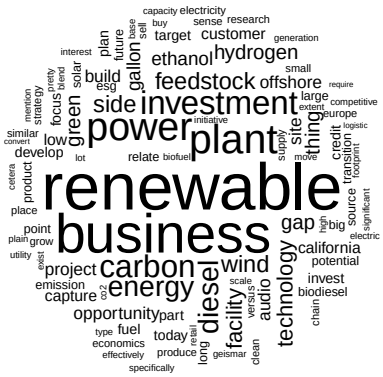


Figure 4.11: Environmental

In Figure 4.11 we also uncover a topic which we could associate with the environmental aspect of ESG. In the case of Oil, Gas & Consumable Fuels it seems as if the main focus relates to the transition towards more renewable alternatives for energy.

As was discussed in the previous section in regards to applying DMM models we still encounter the same issues when examining the reduced data sets, i.e. the number of topics required are far too large. When examining the topics produced by the LDA models it also seems as if that the assumption of documents being a mixture of topics is quite reasonable. In the case of Metals & Mining for example we have topics relating to mining as well as topics relating to specific ores, hence a question regarding iron mining would likely be a mixture including these two topics. As such it seems as if the assumption of questions being made up of a single topic is not as reasonable as initially thought. Although the questions are in general very short it is evident from previously examined topics that they contain a lot of overlapping, such as Figure 4.3 b) in which the topic includes general Q & A specific terms and Figure 4.3 a) that includes general earnings call terms. As such, although they posses some nice properties, it seems as if DMM models are not as applicable to this corpus as initially thought. Due to these findings the proceeding sections will only focus on LDA models.

4.3 Seeded LDA

As we discovered in the previous section we are able to uncover several different interesting topics relating to the two industries we have examined. Unfortunately none of the topics we uncovered, besides the environmental topics, for either of the two industries could directly be linked to the ESG themes that we aim to uncover in this thesis. This is of course quite subjective as we only obtain the contents of the uncovered topics, in the form of topic-word probabilities, whilst labeling the topics ourselves. To the best of our ability, and given our limited prior knowledge, we believe that none, except the environmental topic in Oil, Gas & Consumable Fuels could be seen as sufficient.

As discussed in Section 2.7 we can use seed words in order to attempt to nudge our models in the direction of discovering topic related to ESG. As mentioned previously this requires some prior knowledge of not only the industry we examine but also the types of questions typically asked in relation to ESG within the industry. As our prior knowledge in regards to the industries we are examining is fairly limited we have to keep in mind that the impacts we observe when seeding our topics depend entirely on these somewhat uninformed choices.

It should however be noted that we do not require an excessive amount of seed words as, per the nature of LDA, words that are similar and that co-occur with the initial seed words will be prominent in the model regardless if they are given extra weight or not.

4.3.1 Choice of Weight

As mentioned in Section 2.7 the seeding alters the topic-word hyperparameters $\boldsymbol{\beta}$ such that additional weight is put on specific words, in specific topics, such that

$$\boldsymbol{\phi}_k \sim Dir(\boldsymbol{\beta} + \mathbf{c}_k) \tag{4.1}$$

where $\mathbf{c}_k$ is the vector of length V where the non-zero elements represent the additional weight for those specific words. As mentioned previously this additional weight can be seen as adding pseudo-counts to the specified words in the specific topic. Going forward we have chosen, quite arbitrarily but also with the total size of the corpus in consideration, the number of added pseudo-counts to be 500, which is quite a significant amount in comparison to the sizes of the corpora. essentially makes sure that if such a topic that we want to uncover exists, it will likely be found if there are no other, vastly more frequent topics, that override it. In the case where we want to uncover rarer topics it is not unreasonable to further increase the number of topics K as to not have more frequent topics overriding those that we are attempting to seed. As noted by the Figure 4.1, 4.6 and 4.9 thus far increasing the total number of topics does not seem to negatively impact the coherence, as one would assume since more topics should imply more diversity between the top terms of each topic. As mentioned w.r.t. those figures we do have to keep in mind that increasing K likely leads to overfitting.

4.3.2 Environmental Topics

As both our example industries, Metals & Mining and Oil, Gas & Consumable Fuels, have somewhat similar profiles when it comes to the environmental aspect of ESG we can seed them using the same seed words. As our current models do not incorporate bigrams, we use the two seed words *carbon* and *emission* (the lemmatized version of *emissions*) as this seems to be discussed quite frequently in both sample industries. For this example we also use $K = 100$ in an attempt to create more nuanced topics that only include terms that clearly relate to environmental questions. As discussed earlier this should not impact the training coherence negatively but instead only provide us with a larger set of distinct topics. Increasing K does of course lead to some overfitting as discussed previously as well, hence we make sure to examine the held-out coherence for the models at large in Table 4.1 below. The held-out coherence is calculated using a test corpus for both industries with documents dated between 2015 and 2021, excluding any documents used in the training of the models. In the case of Metals & Mining the test corpus includes about 250000 documents and for Oil, Gas & Consumable Fuels the test corpus includes approximately 400000 documents.

Industry	In-Sample Coherence	Held-Out Coherence
Metals & Mining	0.67	0.66
Oil, Gas & Consumable Fuels	0.69	0.68

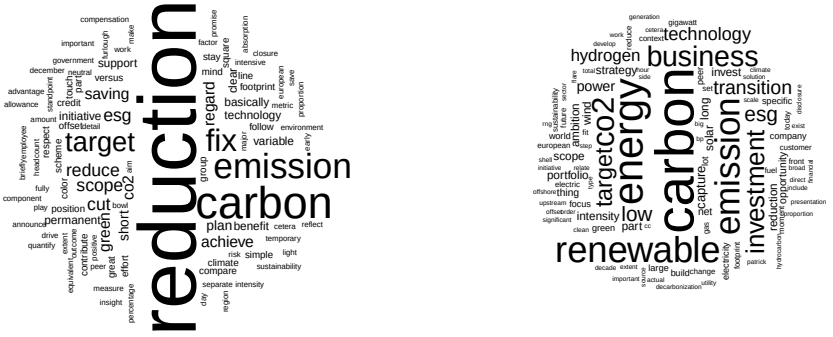
Table 4.1: Seeded LDA Coherence (Model Average)

As can be seen above with $K = 100$ we do not seem to overfit as much as initially thought as the difference between the in-sample and held-out coherence is negligible. When examining the seeded environmental topics specifically in Table 4.2 we see that the in-topic coherence of these on the held-out set are both adequate, meaning that we should be able to utilize them within the scope of gauging trends.

Industry	In-Sample Coherence	Held-Out Coherence
Metals & Mining	0.76	0.71
Oil, Gas & Consumable Fuels	0.82	0.81

Table 4.2: Seeded LDA In-Topic Coherence (Environmental Topics)

With this in mind it does seem that these models should be applicable, not only in order to evaluate the documents in the initial corpus during training, but also new documents supplied later on. The seeded models yield the following environmental topics



(a) Metals & Mining **(b)** Oil, Gas & Consumable Fuels

Figure 4.12: Seeded Environmental Topics

As mentioned previously, and if we allow ourselves to speculate, it seems as if the main environmental questions within Metals & Mining relate to emission targets, either set by the company themselves or by some other body or organization as well as the general reduction of emissions. In Oil, Gas & Consumable Fuels on the other hand there seems to be less discussion in regards to emission targets and more concerning renewables and the transition towards more sustainable alternatives.

These are examples of topics that we could then use to measure the frequency of their occurrence over time, hence being able to crudely determine the change in focus w.r.t. the environmental aspect of ESG over time.

4.3.3 Corporate Governance and Social Topics

A logical extension would be to seed topics for the corporate governance and social aspects of ESG. Unlike the environmental topics seeded in the previous section our prior knowledge regarding corporate governance and social aspects within the two example industries is quite limited. Furthermore it also seems as if corporate governance and social aspect is not nearly as frequently discussed in either of the industries as the environmental aspect, meaning that seeding coherent topics might require a larger number of total topics K as well as a good selection of seed words. Given that the scope of the thesis is to highlight the methods of how these topics can be obtained and then utilized we have as such chosen to not examine these topics individually. We do however make an attempt to seed a corporate governance topic in Section 4.4.1 where we want to illustrate how the trend in different topics can be compared.

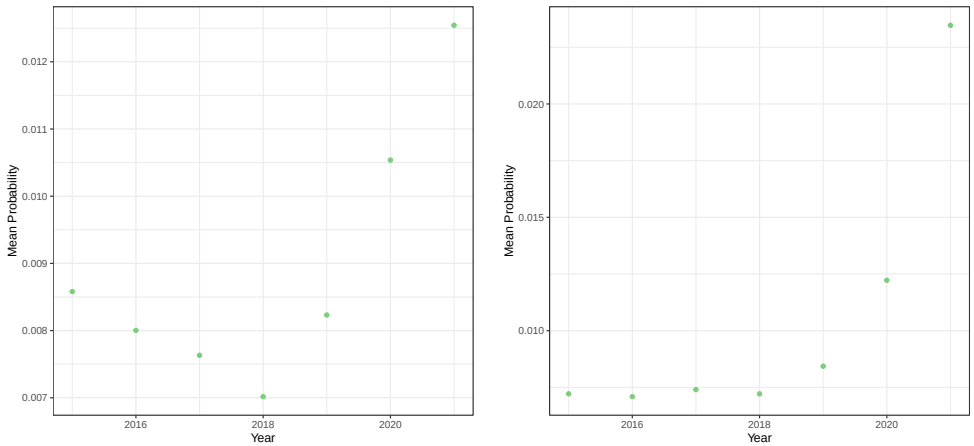
4.4 Extracting Trends

In this section we will discuss how these seeded topics can be used to determine trends in ESG related questions.

The main idea is to utilize the seeded topics found in order to determine the frequency at which different ESG questions are asked over time. The simplest, and most computationally efficient, way to achieve this is to train the model on a smaller data set where we can verify that the respective topics behave as we would like them to. Whilst doing this however we have to keep in mind that it assumes that the contents of the topics remain unchanged over time. Using the trained model we can then, as discussed in Section 2.6, evaluate new documents, either by folding-in or using the assumption of documents being made up of a single topic and utilizing the estimated topic-word probabilities $\mathbf{\Phi}$. Using folding-in with a large number of documents may alter the topics quite significantly, which is not really what we are after as we aim to evaluate the frequency of the same topic over time. In order to not alter the topics of the already examined corpus we can instead choose to freeze those topics whilst folding-in the new documents. This would imply that we randomly initialize topics for the new documents which we want to evaluate and then iterate those until convergence without touching the documents in the initial corpus. Upon convergence we can easily obtain estimates for the document-topic probabilities of the new documents.

As mentioned previously, w.r.t. the sparsity of the document-topic prior we have that documents are in general made up of a few selections of topics, hence it is not enough to simply focus on the topic with the highest probability as this neglects other potential topics of importance. Another issue that arises from this approach is that some documents express low probabilities for all topics with no specific topic that is vastly more probable than the rest. Instead of determining which specific topics document touch upon we can instead choose to examine the probabilities directly.

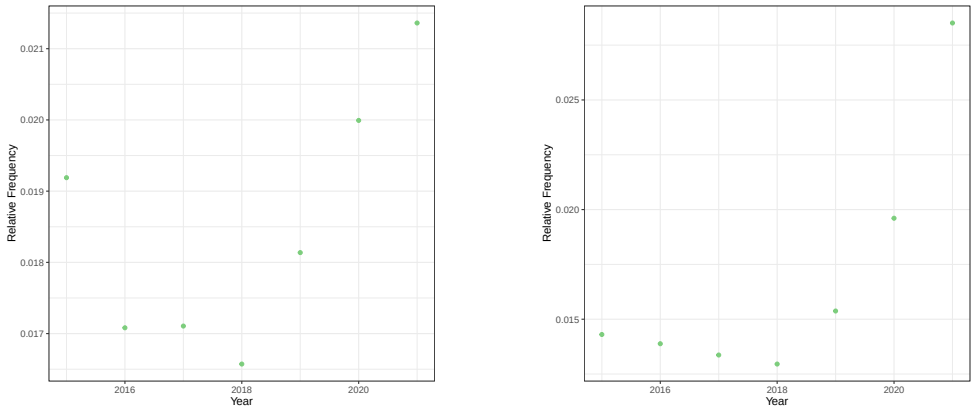
In *Bickel (2019)* the author proposes a simple and straightforward approach to evaluating trends using the aforementioned probabilities [27]. We begin by determining the document-topic distribution of the documents that we aim to use in our trend analysis. As all questions have a time stamp indicating when they were asked we can then group the documents by some time span, e.g. years or fiscal quarters, and determine the mean, or some other appropriate measure, of the topic probabilities in each period [27]. Using this approach we take the average of the probabilities of the environmental topics in each industry for each year. This yields the following historical trends for the environmental aspect in Metals & Mining and Oil, Gas & Consumable Fuels.



(a) Metals & Mining **(b)** Oil, Gas & Consumable Fuels

Figure 4.13: Trend Evaluation (Bickel)

Another approach as introduced by *Hwang et al. (2018)* is, like for the previously discussed coherence metrics, to examine the defining words all topics [28]. With these top n (the authors use $n = 10$) words we then simply calculate the relative frequency, as is shown for *esg* in Figure 3.1, of these top terms across the entire corpus, creating a final trend score which is equal to the sum of these relative frequencies [28]. This sum could also be replaced by a weighted sum using the normalized topic-word probabilities $\mathbf{\Phi}$ for the top words. Below are the relative term frequencies for the top 10 words, as in the word clouds in Figure 4.12 for Metals & Mining and Oil, Gas & Consumable Fuels.



(a) Metals & Mining **(b)** Oil, Gas & Consumable Fuels

Figure 4.14: Trend Evaluation (Hwang)

Given the fact that terms can be included in multiple topics it does seem more reasonable to somehow weight the term frequencies in the method introduced by Hwang et al., as the term *target* can refer to both emission targets and financial targets. Hence it seems unreasonable to attribute all occurrences of *target* and other similar words with multiple meanings to a single topic. It should however be noted that the method used by Hwang et al. is much easier to interpret than the one introduced by Bickel. The method as introduced by Bickel does suffer from the fact that it is dependent on the other topics which is not an issue for the method by Hwang et al. [27][28]. Lastly we have to keep in mind that the seeding of the topics, using the approach discussed in Section 2.7, inflate the topic-word probabilities $\boldsymbol{\phi}_{k,v}$ as pseudo-counts are added for the seed words. It could also be worth considering deflating these specific words to get more accurate results.

To summarize the previous figures it becomes quite clear that there is a growing focus on environmental aspects in both industries, which one could perhaps suspect given the nature of those industries.

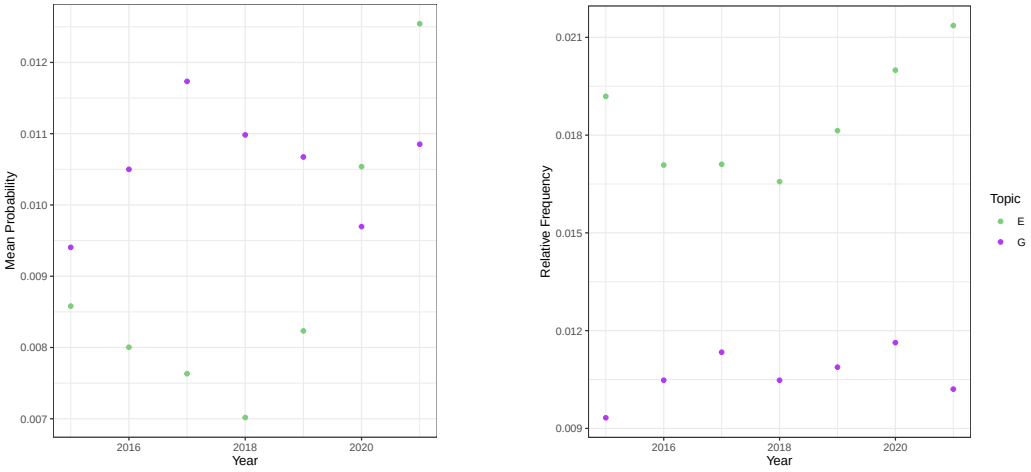
4.4.1 Comparing Topics

Lastly we can compare the values obtained from either of the aforementioned methods in order to effectively weight different topics against each other, hence creating a weighting scheme in which the weights correspond to how much relative focus is being put on the corresponding topics in the Q & A sessions of earnings calls. Although we have not seeded corporate governance nor social topics for either of the industries we can at least include an illustrative example of what this process would look like. The main idea would be to seed topics relating to the ESG aspects that we would like to compare and to then determine the frequencies of those topics over time, using either of the methods discussed in the previous section or some other applicable method. In the example below we have an environmental topic as well as what could be considered to be a corporate governance topic that has been seeded in the Metals & Mining data set using $K = 100$ whilst adding 500 pseudo-counts as before to the chosen seed words. This yields the two following topics.

(a) Environmental **(b)** Corporate Governance

Figure 4.15: ESG Topics

The environmental aspect of ESG within Metals & Mining can once again be seeded in to a topic as is seen in Figure 4.15 a). In Figure 4.15 b) we have an attempt at seeding a corporate governance topic which seems to focus on regulations and obtaining permits for operations. As we are not experts when it comes Metals & Mining nor ESG we acknowledge that these topics might not be as representative as they perhaps could be with proper seeding. As such this should only be seen as an illustrative example and the results obtained should be taken with a grain of salt. We can use these two topics in order to examine the trends of both, as in the previous section, whilst also determining how focus has shifted between the two over time. Below we illustrate how both topics have varied over time using both methods discussed in the previous section.

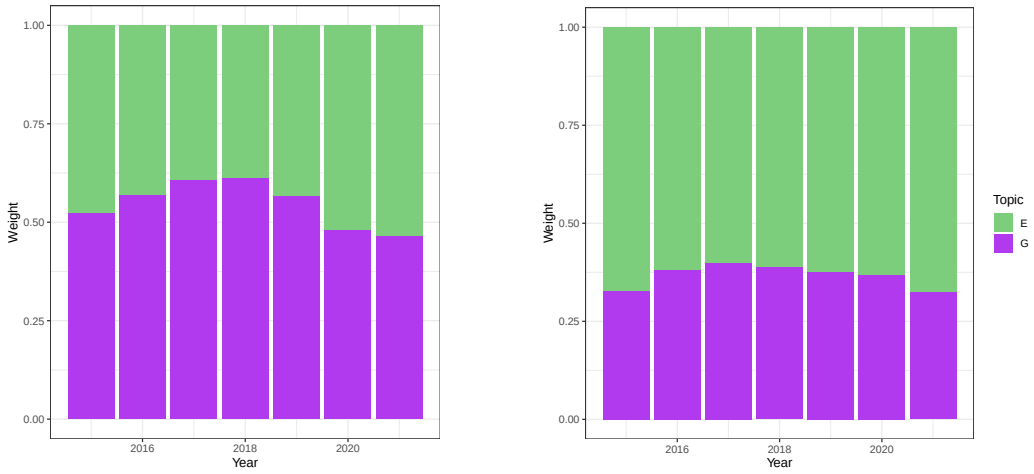


(a) Bickel **(b)** Hwang et al.

Figure 4.16: Environmental (E) & Corporate Governance (G) Trends

In Figure 4.16 we note that there are some quite significant differences between the two methods, primarily in the corporate governance topic. This likely stems from the fact that the approach by Hwang et al. only uses the defining features of topics without taking their frequency across all topics into account. The leads to the full inclusion of terms in the frequency counts that might appear across multiple topics which is problematic as we could assume that words such as *process* in the corporate governance topic might appear in other topics as well.

Lastly these trends can be used to generate weights, which is just the trend score divided by the sum of the trend scores of all topics examined. These weights then indicate how much focus is put on the different topics, relative to one another, at different points in time. For the environmental and corporate governance topics we obtain the following weights using both methods.



(a) Bickel **(b)** Hwang et al.

Figure 4.17: Environmental (E) & Corporate Governance (G) Weights

As discussed previously there is quite a difference between the two methods which of course impacts the weights as well, as can be seen in Figure 4.17. Although the actual weights differ we note that the general trend of an increase, followed by a decrease, in the focus on corporate governance seems to exist in both weighting schemes.

The seeded corporate governance topic in this example is, as previously mentioned, quite questionable but this example illustrates how the methods discussed in this thesis could be used, given sufficient prior knowledge and support in the data. This approach can of course also be extended to include social topics and non-ESG topics altogether.

One thing to keep in mind when using this framework is that the topics discovered, which are then used for evaluation, are assumed to be time independent. This means that the topics found should not change drastically in content over time, e.g. the topics represented in Figure 4.15 are assumed to not vary over time. If this assumption holds it becomes feasible to train the models on a smaller data set whilst then evaluating it on a larger one which might encompass different periods of time. A way to examine if this assumption holds is to examine the coherence on a held-out data set that spans a longer period of time, or several sets that encompass different time periods than the one used in training. It should however be noted that this only verifies that the defining words co-occur in the specific period and not that the topic remains unchanged over time. We might for example have other words that co-occur with the defining words as well as a re-ordering of the defining words in different time periods. In this case it could be of interest to actually use perplexity given that it is actually a likelihood based approach which takes the above issues into account.

In the method introduced by Bickel we have to ensure that the model generalizes well for all topics w.r.t. coherence, as the probability of a single topic does depend on the probabilities of all others. In the approach introduced by Hwang et al. we are not actually concerned with the others topics as we are only interested in the unnormalized frequencies of the defining words in the topics that we are examining.

Topic	In-Sample Coherence	Held-Out Coherence
Environmental	0.62	0.60
Corporate Governance	0.74	0.75
Model Average	0.69	0.66

Table 4.3: Seeded LDA In-Topic Coherence

From Table 4.3 above we note, as previously, that the model at large seems to generalize well whilst the specific in topic coherences do not seem to decrease too drastically. As a result the trends uncovered in Figure 4.16 and the weights in 4.17 should be usable in further projects, assuming that the seeding and the topic content could be considered to be up to standard.

Chapter 5

Discussion

In this thesis we examine and discuss how one can apply probabilistic topic models, and specifically seeded probabilistic topic models, in order to uncover specific topics of interest within a corpus. In order to achieve this we introduce two key topic models, LDA and DMM, which we use to attempt to uncover the latent topics sought after. As both models require approximative inference methods we also introduce two different varieties of these for both model types. Furthermore we also introduce methods for estimating the hyperparameters required in our models if the need for specific tuning should be required.

One issue with LDA models, although they prove to work quite well on our data, is that they might struggle when the documents in a corpus are shorter, as questions can sometimes be [5]. As such we also examine DMM models that are created specifically to deal with this issue as they assume that documents are made up of a single topic instead of a mixture of topics as in LDA [3][5]. When applying both models to our financial corpus it does however become quite evident that the assumption of a single topic per document is quite unreasonable. The issue being that although most question are quite short they still include multiple different, quite separable, aspects. To briefly summarize we might have some terms relating to earnings calls, Q & A sessions, the specific industry, or finance in general as can be seen, and as is discussed, in Section 4.1. These possibilities and the mixing of different topics in documents leads to DMM models requiring an unfeasible, as well as uninterpretable, number of topics, which severely restricts the usage of the models on our data set. In general it seems as if the topics in the corpus needs to be more distinct with less overlap between the actual topics in order for DMM models to work well.

As the scope of the thesis is to uncover ESG related topics in our financial earnings call data set we further extend LDA to allow for the inclusion of seed words to guide the discovery of the latent topics. By supplying our models with seed words we are in some cases able to uncover specific, quite nuanced, topics relating to different aspects of ESG. In this thesis we only examine the application of our methods on two different industries, Metals & Mining as well as Oil, Gas & Consumable Fuels. In these example industries, where we have some reasonable prior knowledge regarding the existence of environmental topics we manage to seed such topics using LDA models. These models can then be used in order to quantify the frequency of questions regarding specific topics in a new corpus. If the documents in the corpus are assigned timestamps it also becomes feasible to evaluate the frequency of these topics over time. We are able to evaluate the frequency of environmental questions in both industries over time, as can be seen in Figure 4.13 and 4.14. Furthermore we also compare an environmental topic with a corporate governance topic in the example of Metals & Mining in order to gauge how the focus between the two aspects of ESG has shifted over time within that industry, as can be seen in Figure 4.16 and 4.17.

As initially stated, if we view the content of the questions asked during the Q & A sessions as a proxy for what the market is concerned with the trend in these topics

could also be used as an indicator for trends in the market. In this thesis we have focused on ESG but the methods discussed could of course be used to find other topics and determine other trends as well.

Furthermore we have to keep in mind that LDA, like most conventional clustering methods, will always find clusters, or in our case topics, regardless of whether they exist or not. By this we mean that k-means for example will always yield k clusters regardless of whether there is some logical partition in the data. Due to this if no reasonable partition of the data can be found the models will still return the best possible fit, which might produce indiscernible topics. As such it is important to examine the topics obtained in order to verify that they make sense and are not simply made up of seemingly randomized terms. This can also be verified by examining the in-sample coherence which determines how distinct the topics uncovered are.

Lastly it should be reiterated that the method of seeding as discussed in this thesis often requires extensive knowledge of the topics being seeded as well as the existence of those topics within the corpus at large. As our prior knowledge of ESG, and the existence of ESG related topics within the earnings call data set, is quite limited the results presented in this thesis should be taken with a grain of salt. With extensive knowledge of the topics being seeded the framework proposed in this thesis should be useful in quantifying certain markets trends, given that there is support for the trends in data. Furthermore we also have to keep in mind the assumption made when tuning a model and then using it for evaluation. When examining the trends of specific topics over a longer period of time this assumes that the content of that topic does not vary drastically over the examined time period. As mentioned previously in relation to our environmental topics this would imply that question relating to the environmental aspect of ESG are assumed to not change drastically over the course of the examined time period, which in our case is 2015-2021.

5.1 Possible Improvements

The framework presented in this thesis for trend analysis using probabilistic topic modeling could be further improved and extended in multiple ways. As mentioned previously we introduce two inference methods but only use one of them, the Collapsed Gibbs Sampler, in the illustrative example. Although both are approximative methods of inference and should yield similar results it would nonetheless be interesting to compare models where the inference method differs. In relation to this it would also be interesting to examine if the construction of the variational distribution by Attias is actually reasonable, which has not been discussed or touched upon in this thesis. We also extend the Variational Bayesian Inference approach with the use of Variational EM as can be used to infer the hyperparameters for the Dirichlet priors. Although this method was applied initially and proved to yield hyperparameters quite similar to the default values used later on it would of course be interesting, albeit a bit time consuming, to put more attention of the optimization of the hyperparameters.

It would of course also be of interest to examine the GSDMM models further in order to examine how they would perform on even more granular partitions of the data sets such as sub-industries where the total number of topics required might not be as significant. We do however speculate that the results would be fairly similar as the existence of certain aspects of the data should exist regardless of the level we partition the data at.

As discussed in the introduction of the thesis one of the main issues with these models is their scalability meaning that although we might have large quantities of data it cannot be properly utilized due to computational constraints. One of the proposed solutions would be, as mentioned previously, to apply the framework in *Cong et al. (2019)* in order to create separable topics in an efficient manner [7]. If such a framework could efficiently retain the information when shrinking the feature space it would not only allow for the modeling of large amounts of data but also for the usage

of a large number of topics. We could of course also extend this framework by using seed words, or seed embeddings, as well as some appropriate, albeit less efficient, clustering algorithm such as k-means++ where we center the clusters around the specific seed embeddings.

We will also briefly mention some quite feasible and easy to implement changes that could be made to the current framework in order to improve it. Most improvements, beyond those discussed in relation to using other inference methods and optimizing the hyperparameters, to be made lies within the pre-processing step.

We currently use lemmatization and some other, less nuanced, methods for trimming the size of the vocabulary. Stemming, as was touched upon briefly, could for example also be applied, although this would perhaps make the topics a bit more difficult to interpret. The filtering on token frequencies could of course also have been examined further. In its current state the framework only uses unigrams, i.e. single words, which brings with it quite a lot of issues. As mentioned earlier we can have the same word appearing in multiple different topics as it might have multiple meanings, as words do. A way to attempt to combat this is to use bigrams, or even n-grams. Bigrams are two words following one another whilst n-grams is a sequence of n words following one another. By extending the framework to not only use unigrams we can retain some of the meaning behind the words, even when using a *bag-of-words* approach. The use of bigrams does for example solve the aforementioned issue where we might have the bigram *emission target* instead of the unigrams *emission* and *target* where the second token might be confused with for example a target relating to financial growth. By using n-grams we can of course further extend this to include even more complicated sequences. The application of bigrams would for example also allow for the seeding of environmental topics in both industries examined using the seed token *carbon emission* (the lemmatized version of *carbon emissions*) instead of the two seed words *carbon* and *emission*. The drawback of using bigrams, and specifically n-grams, is that it further increases the size of the vocabulary which is not exactly what we desire as scalability is already an issue.

Bibliography

[1] "MSCI What is ESG?." https://www.msci.com/our-solutions/esg-investing/what-is-esg.

[2] "Morgan Stanley Institute For Sustainable Investment Sustainable Signals: Individual Investor Interest Driven by Impact, Conviction and Choice." https://www.morganstanley.com/content/dam/msdotcom/infographics/sustainable-investing.

[3] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 05 2003.

[4] T. Hofmann, "Probabilistic Latent Semantic Indexing," p. 50–57, 1999.

[5] J. Yin and J. Wang, "A Dirichlet multinomial mixture model-based approach for short text clustering," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 08 2014.

[6] J. Jagarlamudi, R. Udupa, and H. Daumé III, "Incorporating Lexical Priors into Topic Models," 01 2012.

[7] L. Cong, T. Liang, and X. Zhang, "Textual Factors: A Scalable, Interpretable, and Data-driven Approach to Analyzing Unstructured Information," 2019.

[8] T. Hoffman, "Learning the Similarity of Documents: An Information-Geometric Approach to Document Retrieval and Categorization," vol. 12, 03 2000.

[9] T. Griffiths and M. Steyvers, "Finding Scientific Topics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101 Suppl 1, pp. 5228–35, 04 2004.

[10] H. Attias, "A Variational Bayesian Framework for Graphical Models," *Adv. Neural Inf. Process. Syst*, vol. 12, 09 2000.

[11] C. Geigle, "Inference methods for Latent Dirichlet Allocation," 2016.

[12] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, pp. 721–741, 11 1984.

[13] T. Jones, *textmineR: Functions for Text Mining and Topic Modeling*, 2019. R package version 3.0.4.

[14] B. Lu, M. Ott, C. Cardie, and B. Tsou, "Multi-aspect Sentiment Analysis with Topic Models," *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 81–88, 12 2011.

[15] J. Chang, J. Boyd-Graber, C. Wang, S. Gerrish, and D. M. Blei, "Reading Tea Leaves: How Humans Interpret Topic Models," 2009.

[16] D. Blei, "Probabilistic Topic Models," *Communications of the ACM*, vol. 55, 08 2011.

[17] M. Röder, A. Both, and A. Hinneburg, "Exploring the Space of Topic Coherence Measures," *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 2015.

[18] S. Syed and M. Spruit, "Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation," *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 165–174, 2017.

[19] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, ""Exploring Topic Coherence over Many Models and Many Topics"," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, (Jeju Island, Korea), pp. 952–961, Association for Computational Linguistics, July 2012.

[20] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," pp. 1445–1456, 05 2013.

[21] R. Larson, "Introduction to Information Retrieval," *JASIST*, vol. 61, pp. 852–853, 04 2010.

[22] D. Lewis, Y. Yang, T. Russell-Rose, and F. Li, "RCV1: A New Benchmark Collection for Text Categorization Research," *Journal of Machine Learning Research*, vol. 5, pp. 361–397, 04 2004.

[23] B. Grün and K. Hornik, "topicmodels: An R Package for Fitting Topic Models," *Journal of Statistical Software*, vol. 40, no. 13, pp. 1–30, 2011.

[24] R. Rehurek and P. Sojka, "Gensim–python framework for vector space modelling," *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, vol. 3, no. 2, 2011.

[25] R. Walker, "GSDMM: Short text clustering," 2017.

[26] "MSCI Global Industry Classification Standard (GICS®) Methodology." `https://www.msci.com/gics`.

[27] M. Bickel, "Reflecting trends in the academic landscape of sustainable energy using probabilistic topic modeling," *Energy, Sustainability and Society*, vol. 9, 12 2019.

[28] M.-H. Hwang, S. Ha, M. In, and K. Lee, "A Method of Trend Analysis using Latent Dirichlet Allocation," *International Journal of Control and Automation*, vol. 11, pp. 173–182, 05 2018.

[29] M. Hoffman, D. Blei, and F. Bach, "Online learning for Latent Dirichlet Allocation," *Advances in Neural Information Processing Systems*, vol. 23, pp. 856–864, 11 2010.

[30] R. Sundberg, *Statistical Modelling by Exponential Families*. Institute of Mathematical Statistics Textbooks, Cambridge University Press, 2019.

Appendices

Appendix A

Derivation of Theoretical Results

In this chapter we present the derivations of all theoretical results.

A.1 Derivation of a Collapsed Gibbs Sampler for LDA

Note that these derivations are partially inspired by [11] and [9].

We begin by collapsing the conjugate Dirichlet priors.

$$P(\mathbf{w}, \mathbf{z}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \int\int P(\mathbf{w}, \mathbf{z}, \boldsymbol{\Theta}, \boldsymbol{\Phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) d\boldsymbol{\Theta} d\boldsymbol{\Phi} \tag{A.1}$$

$$= \int\int \prod_{k=1}^{K} P(\boldsymbol{\phi}_k; \boldsymbol{\beta}) \prod_{i=1}^{M} P(\boldsymbol{\theta}_i; \boldsymbol{\alpha}) \prod_{j=1}^{N_i} P(w_{i,j}|\boldsymbol{\phi}_{z_{i,j}}) P(z_{i,j}|\boldsymbol{\theta}_i) d\boldsymbol{\Theta} d\boldsymbol{\Phi} \tag{A.2}$$

$$= \underbrace{\int \prod_{k=1}^{K} P(\boldsymbol{\phi}_k; \boldsymbol{\beta}) \prod_{i=1}^{M} \prod_{j=1}^{N_i} P(w_{i,j}|\boldsymbol{\phi}_{z_{i,j}}) d\boldsymbol{\Phi}}_{(1)} \times \underbrace{\prod_{i=1}^{M} \int P(\boldsymbol{\theta}_i; \boldsymbol{\alpha}) \prod_{j=1}^{N_i} P(z_{i,j}|\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i}_{(2)}. \tag{A.3}$$

We now examine the two separate integrals.
(1)

$$(1) = \int \prod_{k=1}^{K} P(\boldsymbol{\phi}_k; \boldsymbol{\beta}) \prod_{i=1}^{M} \prod_{j=1}^{N_i} P(w_{i,j}|\boldsymbol{\phi}_{z_{i,j}}) d\boldsymbol{\Phi} \tag{A.4}$$

$$= \int \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \prod_{v=1}^{V} \boldsymbol{\phi}_{k,v}^{\beta_v - 1} \prod_{i=1}^{M} \prod_{j=1}^{N_i} P(w_{i,j}|\boldsymbol{\phi}_{z_{i,j}}) d\boldsymbol{\Phi}. \tag{A.5}$$

We let $n_{i,v}^{(k)}$ denote the number of times the vth word in the vocabulary, assigned to

topic k, appears in the ith document. We formally define $n_{i,v}^{(k)}$ as

$$n_{i,v}^{(k)} = \sum_{j=1}^{N_i} \mathbf{1}_{\{w_{i,j}=v \cap z_{i,j}=k\}} \tag{A.6}$$

$$= \sum_{j=1}^{N_i} \mathbf{1}_{\{w_{i,j}=v\}} \mathbf{1}_{\{z_{i,j}=k\}}. \tag{A.7}$$

Furthermore we also adopt the $\bullet$ notation to indicate the sum over a specific index, hence

$$n_{\bullet,v}^{(k)} = \sum_{i=1}^{M} n_{i,v}^{(k)} \tag{A.8}$$

$$= \sum_{i=1}^{M} \sum_{j=1}^{N_i} \mathbf{1}_{\{w_{i,j}=v\}} \mathbf{1}_{\{z_{i,j}=k\}}. \tag{A.9}$$

Using the above we can rewrite $P(w_{i,j}|\boldsymbol{\phi}_{z_{i,j}})$ as

$$\prod_{i=1}^{M}\prod_{j=1}^{N_i} P(w_{i,j}|\boldsymbol{\phi}_{z_{i,j}}) = \prod_{i=1}^{M}\prod_{j=1}^{N_i}\prod_{v=1}^{V} \boldsymbol{\phi}_{z_{i,j},v}^{\mathbf{1}_{\{w_{i,j}=v\}}} \tag{A.10}$$

$$= \prod_{v=1}^{V}\prod_{k=1}^{K} \boldsymbol{\phi}_{k,v}^{\sum_{i=1}^{M}\sum_{j=1}^{N_i}\mathbf{1}_{\{w_{i,j}=v\cap z_{i,j}=k\}}} \tag{A.11}$$

$$= \prod_{v=1}^{V}\prod_{k=1}^{K} \boldsymbol{\phi}_{k,v}^{n_{\bullet,v}^{(k)}}. \tag{A.12}$$

Continuing where we left off we obtain

$$(1) = \prod_{k=1}^{K} \int \frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v\right)} \prod_{j=1}^{V} \boldsymbol{\phi}_{k,v}^{\beta_v-1} \prod_{v=1}^{V} \boldsymbol{\phi}_{k,v}^{n_{\bullet,v}^{(k)}} d\boldsymbol{\phi}_k \tag{A.13}$$

$$= \prod_{k=1}^{K} \int \frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v\right)} \boldsymbol{\phi}_{k}^{n_{\bullet,v}^{(k)}+\beta_v-1} d\boldsymbol{\phi}_k \tag{A.14}$$

$$= \prod_{k=1}^{K} \frac{\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma(\beta_v)}}{\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v+n_{\bullet,v}^{(k)}\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v+n_{\bullet,v}^{(k)}\right)}} \underbrace{\int \frac{\Gamma\left(\sum_{v=1}^{V}\beta_v+n_{\bullet,v}^{(k)}\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v+n_{\bullet,v}^{(k)}\right)} \boldsymbol{\phi}_{k}^{n_{\bullet,v}^{(k)}+\beta_v-1} d\boldsymbol{\phi}_k}_{=1} \tag{A.15}$$

$$= \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v\right)} \frac{\prod_{v=1}^{V}\Gamma\left(\beta_v+n_{\bullet,v}^{(k)}\right)}{\Gamma\left(\sum_{v=1}^{V}\beta_v+n_{\bullet,v}^{(k)}\right)}. \tag{A.16}$$

We obtain the 1 on the right hand side in (A.15) as we integrate the density function of a $Dir\left(\beta_1+n_{\bullet,v}^{(1)},...,\beta_V+n_{\bullet,v}^{(K)}\right)$ over its entire support.

(2)

$$(2) = \prod_{i=1}^{M} \int P(\boldsymbol{\theta}_i;\boldsymbol{\alpha}) \prod_{j=1}^{N_i} P(z_{i,j}|\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \tag{A.17}$$

$$= \prod_{i=1}^{M} \int \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)} \prod_{k=1}^{K} \theta_{i,k}^{\alpha_k-1} \prod_{j=1}^{N_i} P(z_{i,j}|\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i. \tag{A.18}$$

We once again use the notation of $n_{i,v}^{(k)}$. As such we can rewrite $P(z_{i,j}|\boldsymbol{\theta}_i)$ as follows

$$\prod_{j=1}^{N_i} P(z_{i,j}|\boldsymbol{\theta}_i) = \prod_{j=1}^{N_i}\prod_{k=1}^{K} \theta_{i,k}^{\mathbf{1}_{\{z_{i,j}=k\}}} = \prod_{k=1}^{K} \theta_{i,k}^{\sum_{j=1}^{N_i}\mathbf{1}_{\{z_{i,j}=k\}}} = \prod_{k=1}^{K} \theta_{i,k}^{n_{i,\bullet}^{(k)}}. \tag{A.19}$$

Where $n_{i,\bullet}^{(k)}$ is the number of words (any words, hence the bullet) in document i assigned to topic k. Continuing from where we left off we get

$$(1) = \prod_{i=1}^{M} \int \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)} \prod_{k=1}^{K} \theta_{i,k}^{\alpha_k-1} \prod_{k=1}^{K} \theta_{i,k}^{n_{i,\bullet}^{(k)}} d\boldsymbol{\theta}_i \tag{A.20}$$

$$= \prod_{i=1}^{M} \int \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)} \prod_{k=1}^{K} \theta_{i,k}^{n_{i,\bullet}^{(k)}+\alpha_k-1} d\boldsymbol{\theta}_i \tag{A.21}$$

$$= \prod_{i=1}^{M} \frac{\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}}{\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k+n_{i,\bullet}^{(k)}\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k+n_{i,\bullet}^{(k)}\right)}} \underbrace{\int \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k+n_{i,\bullet}^{(k)}\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k+n_{i,\bullet}^{(k)}\right)} \prod_{k=1}^{K} \theta_{i,k}^{n_{i,\bullet}^{(k)}+\alpha_k-1} d\boldsymbol{\theta}_i}_{=1} \tag{A.22}$$

$$= \prod_{i=1}^{M} \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)} \frac{\prod_{k=1}^{K}\Gamma\left(\alpha_k+n_{i,\bullet}^{(k)}\right)}{\Gamma\left(\sum_{k=1}^{K}\alpha_k+n_{i,\bullet}^{(k)}\right)}. \tag{A.23}$$

We obtain the 1 on the right hand side in (A.22) as we integrate the density function of a $Dir\left(\alpha_1+n_{i,\bullet}^{(1)},...,\alpha_K+n_{i,\bullet}^{(K)}\right)$ over its entire support.

We now return to the initial integral and combine integrals (1) and (2).

$$P(\mathbf{w},\mathbf{z};\boldsymbol{\alpha},\boldsymbol{\beta}) = (1)\times(2) \tag{A.24}$$

$$\begin{aligned} &= \prod_{i=1}^{M} \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)} \frac{\prod_{k=1}^{K}\Gamma\left(\alpha_k+n_{i,\bullet}^{(k)}\right)}{\Gamma\left(\sum_{k=1}^{K}\alpha_k+n_{i,\bullet}^{(k)}\right)} \\ &\times \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v\right)} \frac{\prod_{v=1}^{V}\Gamma\left(\beta_v+n_{\bullet,v}^{(k)}\right)}{\Gamma\left(\sum_{v=1}^{V}\beta_v+n_{\bullet,v}^{(k)}\right)}. \end{aligned} \tag{A.25}$$

What remains is to compute the posterior of $z_{a,b}$, $P(z_{a,b}|\mathbf{w},\mathbf{z}_{\neg a,b};\boldsymbol{\alpha},\boldsymbol{\beta})$, where $\mathbf{z}_{\neg a,b}$ includes all latent variables $z_{i,j}$ except $z_{a,b}$.

$$P(z_{a,b}|\mathbf{w},\mathbf{z}_{\neg a,b};\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{P(z_{a,b},\mathbf{z}_{\neg a,b},\mathbf{w};\boldsymbol{\alpha},\boldsymbol{\beta})}{P(\mathbf{z}_{\neg a,b},\mathbf{w};\boldsymbol{\alpha},\boldsymbol{\beta})} \tag{A.26}$$

$$\propto P(\mathbf{z},\mathbf{w};\boldsymbol{\alpha},\boldsymbol{\beta}). \tag{A.27}$$

Using the proportionality above we obtain

$$\begin{aligned} P(z_{a,b}=\kappa|\mathbf{w},\mathbf{z}_{\neg a,b};\boldsymbol{\alpha},\boldsymbol{\beta}) &\propto \prod_{i=1}^{M} \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)} \frac{\prod_{k=1}^{K}\Gamma\left(\alpha_k+n_{i,\bullet}^{(k)}\right)}{\Gamma\left(\sum_{k=1}^{K}\alpha_k+n_{i,\bullet}^{(k)}\right)} \\ &\times \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v\right)} \frac{\prod_{v=1}^{V}\Gamma\left(\beta_v+n_{\bullet,v}^{(k)}\right)}{\Gamma\left(\sum_{v=1}^{V}\beta_v+n_{\bullet,v}^{(k)}\right)} \end{aligned} \tag{A.28}$$

$$\propto \frac{\prod_{k=1}^{K}\Gamma\left(\alpha_k+n_{a,\bullet}^{(k)}\right)}{\Gamma\left(\sum_{k=1}^{K}\alpha_k+n_{a,\bullet}^{(k)}\right)} \prod_{k=1}^{K} \frac{\Gamma\left(\beta_{w_{a,b}}+n_{\bullet,w_{a,b}}^{(k)}\right)}{\Gamma\left(\sum_{v=1}^{V}\beta_v+n_{\bullet,v}^{(k)}\right)}. \tag{A.29}$$

We now let $n_{i,j}^{(k)\neg a,b}$ be the corresponding counts without the inclusion of a, b. Formally we define these altered counts as follows

$$n_{i,j}^{(k)\neg a,b} = \sum_{i,j \neq a,b} \mathbf{1}_{\{w_{i,j}=v\}} \mathbf{1}_{\{z_{i,j}=k\}}. \tag{A.30}$$

Continuing from where we left off we obtain the following

$$\begin{aligned} P(z_{a,b} = \kappa | \mathbf{w}, \mathbf{z}_{\neg a,b}; \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto & \frac{\prod_{k \neq \kappa} \Gamma\left(\alpha_k + n_{a,\bullet}^{(k)\neg a,b}\right) \times \Gamma\left(1 + \alpha_\kappa + n_{a,\bullet}^{(\kappa)\neg a,b}\right)}{\Gamma\left(1 + \sum_{k=1}^K \alpha_k + n_{a,\bullet}^{(k)\neg a,b}\right)} \\ & \times \prod_{k \neq \kappa} \frac{\Gamma\left(\beta_{w_{a,b}} + n_{\bullet,w_{a,b}}^{(k)\neg a,b}\right)}{\Gamma\left(\sum_{v=1}^V \beta_v + n_{\bullet,v}^{(k)\neg a,b}\right)} \\ & \times \frac{\Gamma\left(1 + \beta_{w_{a,b}} + n_{\bullet,w_{a,b}}^{(\kappa)\neg a,b}\right)}{\Gamma\left(1 + \sum_{v=1}^V \beta_v + n_{\bullet,v}^{(\kappa)\neg a,b}\right)}. \end{aligned} \tag{A.31}$$

Above we use that $n_{a,\bullet}^{(\kappa)\neg a,b} + 1 = n_{a,\bullet}^{(\kappa)}$ and that $n_{\bullet,w_{a,b}}^{(\kappa)\neg a,b} + 1 = n_{\bullet,w_{a,b}}^{(\kappa)}$ since if $z_{a,b} = \kappa$ the mentioned counts differ by one. We now use that the Gamma function possess the following property, $\Gamma(x) = x\Gamma(x-1)$ to obtain

$$\begin{aligned} P(z_{a,b} = \kappa | \mathbf{w}, \mathbf{z}_{\neg a,b}; \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto & \frac{\prod_{k \neq \kappa} \Gamma\left(\alpha_k + n_{a,\bullet}^{(k)\neg a,b}\right)}{\Gamma\left(\sum_{k=1}^K \alpha_k + n_{a,\bullet}^{(k)\neg a,b}\right)} \\ & \times \frac{\Gamma\left(\alpha_\kappa + n_{a,\bullet}^{(\kappa)\neg a,b}\right)\left(\alpha_\kappa + n_{a,\bullet}^{(\kappa)\neg a,b}\right)}{\sum_{k=1}^K \alpha_k + n_{a,\bullet}^{(k)\neg a,b}} \\ & \times \prod_{k \neq \kappa} \frac{\Gamma\left(\beta_{w_{a,b}} + n_{\bullet,w_{a,b}}^{(k)\neg a,b}\right)}{\Gamma\left(\sum_{v=1}^V \beta_v + n_{\bullet,v}^{(k)\neg a,b}\right)} \\ & \times \frac{\Gamma\left(\beta_{w_{a,b}} + n_{\bullet,w_{a,b}}^{(\kappa)\neg a,b}\right)\left(\beta_{w_{a,b}} + n_{\bullet,w_{a,b}}^{(\kappa)\neg a,b}\right)}{\Gamma\left(\sum_{v=1}^V \beta_v + n_{\bullet,v}^{(\kappa)\neg a,b}\right)\left(\sum_{v=1}^V \beta_v + n_{\bullet,v}^{(\kappa)\neg a,b}\right)} \\ = & \underbrace{\frac{\prod_{k=1}^K \Gamma\left(\alpha_k + n_{a,\bullet}^{(k)\neg a,b}\right)}{\Gamma\left(\sum_{k=1}^K \alpha_k + n_{a,\bullet}^{(k)\neg a,b}\right)}}_{(1)} \frac{\left(\alpha_\kappa + n_{a,\bullet}^{(\kappa)\neg a,b}\right)}{\sum_{k=1}^K \alpha_k + n_{a,\bullet}^{(k)\neg a,b}} \\ & \times \underbrace{\prod_{k=1}^K \frac{\Gamma\left(\beta_{w_{a,b}} + n_{\bullet,w_{a,b}}^{(k)\neg a,b}\right)}{\Gamma\left(\sum_{v=1}^V \beta_v + n_{\bullet,v}^{(k)\neg a,b}\right)}}_{(2)} \frac{\left(\beta_{w_{a,b}} + n_{\bullet,w_{a,b}}^{(\kappa)\neg a,b}\right)}{\left(\sum_{v=1}^V \beta_v + n_{\bullet,v}^{(\kappa)\neg a,b}\right)}. \end{aligned} \tag{A.32}$$

(A.33)

Lastly we drop (1) and (2) (as they are the same for all κ as a, b is excluded) to obtain the final posterior. Note that we could also drop the denominator of the left fraction below as it is the same for all choices of κ. However by doing this some of the intuitions regarding the two fractions would be lost and hence we have chosen to keep it for clarity.

$$P(z_{a,b} = \kappa | \mathbf{w}, \mathbf{z}_{\neg a,b}; \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \frac{\alpha_\kappa + n_{a,\bullet}^{(\kappa)\neg a,b}}{\sum_{k=1}^K \alpha_k + n_{a,\bullet}^{(k)}} \times \frac{\beta_{w_{a,b}} + n_{\bullet,w_{a,b}}^{(\kappa)\neg a,b}}{\sum_{v=1}^V \beta_v + n_{\bullet,v}^{(\kappa)\neg a,b}}. \tag{A.34}$$

The result can intuitively be seen as the left fraction being the probability of topics among documents and the second being the probability of words among a topic.

A.2 Variational Bayesian Inference for LDA

Recall that

$$P(\mathbf{w}, \mathbf{z}, \boldsymbol{\Theta}, \boldsymbol{\Phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{k=1}^{K} P(\boldsymbol{\phi}_k; \boldsymbol{\beta}) \prod_{i=1}^{M} P(\mathbf{z}_i|\boldsymbol{\theta}_i) P(\boldsymbol{\theta}_i; \boldsymbol{\alpha}) \prod_{j=1}^{N_i} P(w_{i,j}|\boldsymbol{\phi}_{\mathbf{z}_i}). \quad \text{(A.35)}$$

These derivations are based on [11], [3] and [29].

We examine the smooth definition of LDA in [3]. In this case we let the variational distribution $q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi})$ be defined as per the reduction in Figure 2.3.

$$q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}) = \prod_{k=1}^{K} q(\boldsymbol{\phi}_k|\boldsymbol{\lambda}_k) \prod_{i=1}^{M} q(\mathbf{z}_i, \boldsymbol{\theta}_i|\boldsymbol{\gamma}_i, \boldsymbol{\pi}_i) \quad \text{(A.36)}$$

$$= \prod_{k=1}^{K} q(\boldsymbol{\phi}_k|\boldsymbol{\lambda}_k) \prod_{i=1}^{M} q(\boldsymbol{\theta}_i|\boldsymbol{\gamma}_i) \prod_{j=1}^{N_i} q(z_{i,j}|\boldsymbol{\pi}_{i,j,k}). \quad \text{(A.37)}$$

We then find the update equations for $\boldsymbol{\gamma}$, $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$ by minimizing the Kullback-Leibler Divergence between the variational distribution q and the true posteriors p w.r.t. these parameters.

$$D(q||p) = \int\int \sum_{\mathbf{z}} q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}) \log\left(\frac{q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda})}{p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})}\right) d\boldsymbol{\theta} d\boldsymbol{\phi} \quad \text{(A.38)}$$

$$= \int\int \sum_{\mathbf{z}} q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}) \log\left(\frac{q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda})}{p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})} p(\mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})\right) d\boldsymbol{\theta} d\boldsymbol{\phi} \quad \text{(A.39)}$$

$$= \int\int \sum_{\mathbf{z}} q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}) \log\left(\frac{q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda})}{p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})}\right) d\boldsymbol{\theta} d\boldsymbol{\phi}$$
$$+ \int\int \sum_{\mathbf{z}} q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}) \log\left(p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})\right) d\boldsymbol{\theta} d\boldsymbol{\phi}. \quad \text{(A.40)}$$

In equation (A.40) we note that $\log(p(\mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta}))$ does not depend on neither $\mathbf{z}$, $\boldsymbol{\theta}$ nor $\boldsymbol{\phi}$. As such we obtain

$$D(q||p) = \int\int \sum_{\mathbf{z}} q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}) \log\left(\frac{q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda})}{p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})}\right) d\boldsymbol{\theta} d\boldsymbol{\phi}$$
$$+ \log\left(p(\mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})\right) \underbrace{\int\int \sum_{\mathbf{z}} q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}) d\boldsymbol{\theta} d\boldsymbol{\phi}}_{=1} \quad \text{(A.41)}$$

$$= \int\int \sum_{\mathbf{z}} q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}) \log\left(\frac{q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda})}{p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})}\right) d\boldsymbol{\theta} d\boldsymbol{\phi}$$
$$+ \log\left(p(\mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})\right). \quad \text{(A.42)}$$

As $\log\left(p(\mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})\right)$ does not depend on q we simply have to minimize the first part of equation (A.42).

$$\int\int \sum_{\mathbf{z}} q(\mathbf{z},\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{\gamma},\boldsymbol{\pi},\boldsymbol{\lambda}) \log\left(\frac{q(\mathbf{z},\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{\gamma},\boldsymbol{\pi},\boldsymbol{\lambda})}{p(\mathbf{z},\boldsymbol{\theta},\boldsymbol{\phi},\mathbf{w};\boldsymbol{\alpha},\boldsymbol{\beta})}\right) d\boldsymbol{\theta} d\boldsymbol{\phi} \tag{A.43}$$

$$\begin{aligned} &= \int\int \sum_{\mathbf{z}} q(\mathbf{z},\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{\gamma},\boldsymbol{\pi},\boldsymbol{\lambda}) \log\left(q(\mathbf{z},\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{\gamma},\boldsymbol{\pi},\boldsymbol{\lambda})\right) d\boldsymbol{\theta} d\boldsymbol{\phi} \\ &- \int\int \sum_{\mathbf{z}} q(\mathbf{z},\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{\gamma},\boldsymbol{\pi},\boldsymbol{\lambda}) \log\left(p(\mathbf{z},\boldsymbol{\theta},\boldsymbol{\phi},\mathbf{w};\boldsymbol{\alpha},\boldsymbol{\beta})\right) d\boldsymbol{\theta} d\boldsymbol{\phi} \end{aligned} \tag{A.44}$$

$$= E_q\left[\log\left(q(\mathbf{z},\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{\gamma},\boldsymbol{\pi},\boldsymbol{\lambda})\right)\right] - E_q\left[\log\left(p(\mathbf{z},\boldsymbol{\theta},\boldsymbol{\phi},\mathbf{w};\boldsymbol{\alpha},\boldsymbol{\beta})\right)\right]. \tag{A.45}$$

We can then define L as follows.

$$L(\boldsymbol{\gamma},\boldsymbol{\pi},\boldsymbol{\lambda};\boldsymbol{\alpha},\boldsymbol{\beta}) = E_q\left[\log\left(p(\mathbf{z},\boldsymbol{\theta},\boldsymbol{\phi},\mathbf{w};\boldsymbol{\alpha},\boldsymbol{\beta})\right)\right] - E_q\left[\log\left(q(\mathbf{z},\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{\gamma},\boldsymbol{\pi},\boldsymbol{\lambda})\right)\right] \tag{A.46}$$

$$\begin{aligned} &= E_q\left[\log(p(\boldsymbol{\theta};\boldsymbol{\alpha})p(\mathbf{z}|\boldsymbol{\theta})p(\boldsymbol{\phi};\boldsymbol{\beta})p(\mathbf{w}|\mathbf{z},\boldsymbol{\phi}))\right] \\ &- E_q\left[\log(q(\boldsymbol{\phi};\boldsymbol{\lambda})q(\boldsymbol{\theta};\boldsymbol{\gamma})q(\mathbf{z};\boldsymbol{\pi}))\right] \end{aligned} \tag{A.47}$$

$$\begin{aligned} &= \underbrace{E_q[\log(p(\boldsymbol{\theta};\boldsymbol{\alpha}))]}_{(1)} + \underbrace{E_q[\log(p(\mathbf{z}|\boldsymbol{\theta}))]}_{(2)} \\ &+ \underbrace{E_q[\log(p(\boldsymbol{\phi};\boldsymbol{\beta}))]}_{(3)} + \underbrace{E_q\left[p(\mathbf{w}|\mathbf{z},\boldsymbol{\phi})\right]}_{(4)} \\ &- \underbrace{E_q[\log(q(\boldsymbol{\phi};\boldsymbol{\lambda}))]}_{(5)} - \underbrace{E_q[\log(q(\boldsymbol{\theta};\boldsymbol{\gamma}))]}_{(6)} - \underbrace{E_q[\log(q(\mathbf{z};\boldsymbol{\pi}))]}_{(7)}. \end{aligned} \tag{A.48}$$

We can now evaluate the seven expectations above individually.
(1)

$$E_q[\log(p(\boldsymbol{\theta};\boldsymbol{\alpha}))] = E_q\left[\log\left(\prod_{i=1}^{M} \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)} \prod_{k=1}^{K} \boldsymbol{\theta}_{i,k}^{\alpha_k - 1}\right)\right] \tag{A.49}$$

$$= \sum_{i=1}^{M} \log\left(\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)}\right) + \sum_{k=1}^{K}(\alpha_k - 1)E_q[\log(\boldsymbol{\theta}_{i,k})]. \tag{A.50}$$

In the reduced model $\boldsymbol{\theta}_i \sim Dir(\boldsymbol{\gamma}_i)$ which is in the exponential family of distributions as is shown in Appendix A.6. Using the fact that $\boldsymbol{\theta}_i$ is of an exponential family we have that the following is true by one of the properties of the exponential family as described in Appendix A.7

$$E_q[\log(\boldsymbol{\theta}_{i,k})] = \Psi\left(\boldsymbol{\gamma}_{i,k}\right) - \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_{i,l}\right) \tag{A.51}$$

where Ψ is the digamma function, i.e. $\frac{\Gamma'}{\Gamma}$. Now returning to where we left off we get

$$E_q[\log(p(\boldsymbol{\theta}_i;\boldsymbol{\alpha}))] = \sum_{i=1}^{M} \log\left(\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)}\right) + \sum_{k=1}^{K}(\alpha_k - 1)\left(\Psi\left(\boldsymbol{\gamma}_{i,k}\right) - \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_{i,l}\right)\right). \tag{A.52}$$

(2)

$$E_q[\log(p(\mathbf{z}|\boldsymbol{\theta}))] = E_q\left[\log\left(\prod_{i=1}^{M}\prod_{j=1}^{N_i}\prod_{k=1}^{K}\boldsymbol{\theta}_{i,k}^{\mathbf{1}_{\{z_{i,j}=k\}}}\right)\right] \tag{A.53}$$

$$= E_q\left[\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{K}\mathbf{1}_{\{z_{i,j}=k\}}\log\left(\boldsymbol{\theta}_{i,k}\right)\right] \tag{A.54}$$

$$= \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{K}E_q[\mathbf{1}_{\{z_{i,j}=k\}}]E_q\left[\log\left(\boldsymbol{\theta}_{i,k}\right)\right] \tag{A.55}$$

$$= \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{K}\boldsymbol{\pi}_{i,j,k}\left(\Psi\left(\boldsymbol{\gamma}_{i,k}\right) - \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_{i,l}\right)\right). \tag{A.56}$$

(3)

$$E_q\left[\log\left(p(\boldsymbol{\phi};\boldsymbol{\beta})\right)\right] = E_q\left[\log\left(\prod_{k=1}^{K}\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v\right)}\prod_{v=1}^{V}\boldsymbol{\phi}_{k,v}^{\beta_v-1}\right)\right] \tag{A.57}$$

$$= \sum_{k=1}^{K}\log\left(\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v\right)}\right) + \sum_{v=1}^{V}(\beta_k-1)E_q\left[\log\left(\boldsymbol{\phi}_{k,v}\right)\right]. \tag{A.58}$$

Like in **(1)** we use that $\boldsymbol{\phi}_k \sim Dir(\boldsymbol{\lambda}_k)$ to obtain

$$E_q\left[\log\left(p(\boldsymbol{\phi}|\boldsymbol{\beta})\right)\right] = \sum_{k=1}^{K}\log\left(\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v\right)}\right) + \sum_{v=1}^{V}(\beta_v-1)\left(\Psi(\boldsymbol{\lambda}_{k,v}) - \Psi\left(\sum_{u=1}^{V}\boldsymbol{\lambda}_{k,u}\right)\right). \tag{A.59}$$

(4)

$$E_q[\log(p(\mathbf{w}|\mathbf{z},\boldsymbol{\phi}))] = E_q\left[\log\left(\prod_{i=1}^{M}\prod_{j=1}^{N_i}\prod_{k=1}^{K}\prod_{v=1}^{V}\boldsymbol{\phi}_{k,v}^{\mathbf{1}_{\{z_{i,j}=k\cap w_{i,j}=v\}}}\right)\right] \tag{A.60}$$

$$= E_q\left[\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{K}\sum_{v=1}^{V}\mathbf{1}_{\{z_{i,j}=k\cap w_{i,j}=v\}}\log(\boldsymbol{\phi}_{k,v})\right] \tag{A.61}$$

$$= \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{K}\sum_{v=1}^{V}E_q\left[\mathbf{1}_{\{z_{i,j}=k\}}\mathbf{1}_{\{w_{i,j}=v\}}\log(\boldsymbol{\phi}_{k,v})\right] \tag{A.62}$$

$$= \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{K}\sum_{v=1}^{V}\boldsymbol{\pi}_{i,j,k}\mathbf{1}_{\{w_{i,j}=v\}}\left(\Psi(\boldsymbol{\lambda}_{k,v}) - \Psi\left(\sum_{u=1}^{V}\boldsymbol{\lambda}_{k,u}\right)\right). \tag{A.63}$$

(5)
The fifth expectation is identical to **(3)** except that we vary the word priors from $\boldsymbol{\beta}$ to $\boldsymbol{\lambda}$.

$$E_q\left[\log\left(q(\boldsymbol{\phi};\boldsymbol{\lambda})\right)\right] = \sum_{k=1}^{K}\log\left(\frac{\Gamma\left(\sum_{v=1}^{V}\boldsymbol{\lambda}_{k,v}\right)}{\prod_{v=1}^{V}\Gamma\left(\boldsymbol{\lambda}_{k,v}\right)}\right) + \sum_{v=1}^{V}(\boldsymbol{\lambda}_{k,v}-1)\left(\Psi(\boldsymbol{\lambda}_{k,v}) - \Psi\left(\sum_{u=1}^{V}\boldsymbol{\lambda}_{k,u}\right)\right). \tag{A.64}$$

(6)
The sixth expectation is identical to **(1)** except that we vary the topic priors from

$\boldsymbol{\alpha}$ to $\boldsymbol{\gamma}$, giving us

$$E_q[\log(q(\boldsymbol{\theta};\boldsymbol{\gamma}))] = \sum_{i=1}^{M} \log\left(\frac{\Gamma\left(\sum_{k=1}^{K}\boldsymbol{\gamma}_{i,k}\right)}{\prod_{k=1}^{K}\Gamma\left(\boldsymbol{\gamma}_{i,k}\right)}\right) + \sum_{k=1}^{K}(\boldsymbol{\gamma}_{i,k}-1)\left(\Psi\left(\boldsymbol{\gamma}_{i,k}\right) - \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_{i,l}\right)\right). \tag{A.65}$$

(7)

$$\begin{aligned}
E_q\left[\log\left(q(\mathbf{z};\boldsymbol{\pi})\right)\right] &= E_q\left[\log\left(\prod_{i=1}^{M}\prod_{j=1}^{N_i}\prod_{k=1}^{K}\boldsymbol{\pi}_{i,j,k}^{\mathbf{1}_{\{z_{i,j}=k\}}}\right)\right] && \text{(A.66)}\\
&= E_q\left[\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{K}\mathbf{1}_{\{z_{i,j}=k\}}\log\left(\boldsymbol{\pi}_{i,j,k}\right)\right] && \text{(A.67)}\\
&= \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{K}E_q\left[\mathbf{1}_{\{z_{i,j}=k\}}\log\left(\boldsymbol{\pi}_{i,j,k}\right)\right] && \text{(A.68)}\\
&= \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{K}\boldsymbol{\pi}_{i,j,k}\log\left(\boldsymbol{\pi}_{i,j,k}\right). && \text{(A.69)}
\end{aligned}$$

Combining **(1)**-**(7)** we get that L as follows

$$\begin{aligned}
L(\boldsymbol{\gamma},\boldsymbol{\pi},\boldsymbol{\lambda};\boldsymbol{\alpha},\boldsymbol{\beta}) = &\sum_{i=1}^{M}\log\left(\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)}\right) + \sum_{k=1}^{K}(\alpha_k-1)\left(\Psi\left(\boldsymbol{\gamma}_{i,k}\right) - \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_{i,l}\right)\right)\\
&+ \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{K}\boldsymbol{\pi}_{i,j,k}\left(\Psi\left(\boldsymbol{\gamma}_{i,k}\right) - \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_{i,l}\right)\right)\\
&+ \sum_{k=1}^{K}\log\left(\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v\right)}\right) + \sum_{v=1}^{V}(\beta_v-1)\left(\Psi(\boldsymbol{\lambda}_{k,v}) - \Psi\left(\sum_{u=1}^{V}\boldsymbol{\lambda}_{k,u}\right)\right)\\
&+ \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{K}\sum_{v=1}^{V}\boldsymbol{\pi}_{i,j,k}\mathbf{1}_{\{w_{i,j}=v\}}\left(\Psi(\boldsymbol{\lambda}_{k,v}) - \Psi\left(\sum_{u=1}^{V}\boldsymbol{\lambda}_{k,u}\right)\right)\\
&- \sum_{k=1}^{K}\log\left(\frac{\Gamma\left(\sum_{v=1}^{V}\boldsymbol{\lambda}_{k,v}\right)}{\prod_{v=1}^{V}\Gamma\left(\boldsymbol{\lambda}_{k,v}\right)}\right) + \sum_{v=1}^{V}(\boldsymbol{\lambda}_{k,v}-1)\left(\Psi(\boldsymbol{\lambda}_{k,v}) - \Psi\left(\sum_{u=1}^{V}\boldsymbol{\lambda}_{k,u}\right)\right)\\
&- \sum_{i=1}^{M}\log\left(\frac{\Gamma\left(\sum_{k=1}^{K}\boldsymbol{\gamma}_{i,k}\right)}{\prod_{k=1}^{K}\Gamma\left(\boldsymbol{\gamma}_{i,k}\right)}\right) + \sum_{k=1}^{K}(\boldsymbol{\gamma}_{i,k}-1)\left(\Psi\left(\boldsymbol{\gamma}_{i,k}\right) - \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_{i,l}\right)\right)\\
&- \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{K}\boldsymbol{\pi}_{i,j,k}\log\left(\boldsymbol{\pi}_{i,j,k}\right).
\end{aligned} \tag{A.70}$$

The update equations for $\boldsymbol{\gamma}_{i,k}$ and $\boldsymbol{\pi}_{i,j,k}$ and $\boldsymbol{\lambda}_{k,v}$ can now be obtained by maximizing L with respect to these parameters.

We begin by examining $\boldsymbol{\gamma}_i$, letting $L_{\boldsymbol{\gamma}_i}$ denote the proportionality of L w.r.t. $\boldsymbol{\gamma}_i$.

$$\begin{aligned}
L_{\boldsymbol{\gamma}_i} = &\sum_{k=1}^{K}\left(\alpha_k + \sum_{j=1}^{N_i}\boldsymbol{\pi}_{i,j,k} - \boldsymbol{\gamma}_{i,k}\right)\left(\Psi(\boldsymbol{\gamma}_{i,k}) - \Psi\left(\sum_{k=1}^{K}\boldsymbol{\gamma}_{i,k}\right)\right)\\
&- \log\left(\Gamma\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_{i,l}\right)\right) + \sum_{l=1}^{K}\log\left(\Gamma\left(\boldsymbol{\gamma}_{i,l}\right)\right).
\end{aligned} \tag{A.71}$$

Differentiating w.r.t. $\boldsymbol{\gamma}_{i,k}$ we obtain

$$\begin{aligned}\frac{\partial L_{\boldsymbol{\gamma}_i}}{\partial \boldsymbol{\gamma}_{i,k}} &= \left(\alpha_k + \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,j,k} - \boldsymbol{\gamma}_{i,k}\right)\left(\Psi'(\boldsymbol{\gamma}_{i,k}) - \Psi'\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_{i,l}\right)\right) \\ &\quad - \sum_{c=1}^{K}\left(\alpha_c + \sum_{j=1}^{N_i}\boldsymbol{\pi}_{i,j,c} - \boldsymbol{\gamma}_{i,c}\right)\Psi'\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_{i,l}\right) \qquad (A.72) \\ &\quad - \left(\Psi(\boldsymbol{\gamma}_{i,k}) - \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_{i,l}\right)\right) - \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_{i,l}\right) + \Psi\left(\boldsymbol{\gamma}_{i,k}\right) \\ &= \left(\alpha_k + \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,j,k} - \boldsymbol{\gamma}_{i,k}\right)\left(\Psi'(\boldsymbol{\gamma}_{i,k}) - \Psi'\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_{i,l}\right)\right) \\ &\quad - \sum_{c=1}^{K}\left(\alpha_c + \sum_{j=1}^{N_i}\boldsymbol{\pi}_{i,j,c} - \boldsymbol{\gamma}_{i,c}\right)\Psi'\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_{i,l}\right). \qquad (A.73)\end{aligned}$$

Setting $\nabla L_{\boldsymbol{\gamma}_i} = \mathbf{0}$ we obtain the following

$$\boldsymbol{\gamma}_{i,k} = \alpha_k + \sum_{j=1}^{N_i}\boldsymbol{\pi}_{i,j,k}. \qquad (A.74)$$

We now examine $\boldsymbol{\pi}_{i,j,k}$, letting $L_{\boldsymbol{\pi}_{i,j,k}}$ denote the proportionality of L w.r.t. $\boldsymbol{\pi}_{i,j,k}$. We also note that since $\boldsymbol{\pi}_{i,j,k}$ is the probability that word j in document i has topic k we have the constraint $\sum_{k=1}^{K}\boldsymbol{\pi}_{i,j,k} = 1$.

$$\begin{aligned}L_{\boldsymbol{\pi}_{i,j,k}} &= \boldsymbol{\pi}_{i,j,k}\left(\Psi\left(\boldsymbol{\gamma}_{i,k}\right) - \Psi\left(\sum_{k=1}^{K}\boldsymbol{\gamma}_{i,k}\right)\right) \\ &\quad + \sum_{v=1}^{V}\boldsymbol{\pi}_{i,j,k}\mathbf{1}_{\{w_{i,j}=v\}}\left(\Psi(\boldsymbol{\lambda}_{k,v}) - \Psi\left(\sum_{u=1}^{V}\boldsymbol{\lambda}_{k,u}\right)\right) \qquad (A.75) \\ &\quad - \boldsymbol{\pi}_{i,j,k}\log\left(\boldsymbol{\pi}_{i,j,k}\right) + \lambda\left(\sum_{k=1}^{K}\boldsymbol{\pi}_{i,j,k} - 1\right).\end{aligned}$$

Differentiating the above expression w.r.t. $\boldsymbol{\pi}_{i,j,k}$ we obtain

$$\begin{aligned}\frac{\partial L_{\boldsymbol{\pi}_{i,j,k}}}{\partial \boldsymbol{\pi}_{i,j,k}} &= \left(\Psi\left(\boldsymbol{\gamma}_{i,k}\right) - \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_{i,l}\right)\right) \\ &\quad + \sum_{v=1}^{V}\mathbf{1}_{\{w_{i,j}=v\}}\left(\Psi(\boldsymbol{\lambda}_{k,v}) - \Psi\left(\sum_{u=1}^{V}\boldsymbol{\lambda}_{k,u}\right)\right) \qquad (A.76) \\ &\quad - \left(\log\left(\boldsymbol{\pi}_{i,j,k}\right) + 1\right) + \lambda \\ &= \left(\Psi\left(\boldsymbol{\gamma}_{i,k}\right) - \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_{i,l}\right)\right) + \left(\Psi(\boldsymbol{\lambda}_{k,w_{i,j}}) - \Psi\left(\sum_{u=1}^{V}\boldsymbol{\lambda}_{k,u}\right)\right) \qquad (A.77) \\ &\quad - \left(\log\left(\boldsymbol{\pi}_{i,j,k}\right) + 1\right) + \lambda.\end{aligned}$$

Setting the above to zero we get that

$$\boldsymbol{\pi}_{i,j,k} \propto \exp\left\{\Psi\left(\boldsymbol{\gamma}_{i,k}\right) + \Psi(\boldsymbol{\lambda}_{k,w_{i,j}}) - \Psi\left(\sum_{v=1}^{V}\boldsymbol{\lambda}_{k,v}\right)\right\}. \qquad (A.78)$$

Note that $\Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_{i,l}\right)$ disappears when we obtain the proportionality. This is since we only need $\boldsymbol{\pi}_{i,j,k}$ in proportion to all other $\boldsymbol{\pi}_{i,j,\cdot}$, i.e. the probabilities for the

same document-word pairs (and for the same i, j we have that $\Psi\left(\sum_{l=1}^{K} \boldsymbol{\gamma}_{i,l}\right)$ is the same).

Lastly we examine $\boldsymbol{\lambda}_{k,v}$ and let $L_{\boldsymbol{\lambda}_k}$ denote the proportionality of L w.r.t. $\boldsymbol{\lambda}_{k,v}$.

$$L_{\boldsymbol{\lambda}_k} = \sum_{v=1}^{V} \left(\beta_v + \sum_{i=1}^{M} \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,j,k} \mathbf{1}_{\{w_{i,j}=v\}} - \boldsymbol{\lambda}_{k,v} \right) \left(\Psi(\boldsymbol{\lambda}_{k,v}) - \Psi\left(\sum_{u=1}^{V} \boldsymbol{\lambda}_{k,u} \right) \right) - \log\left(\Gamma\left(\sum_{u=1}^{V} \boldsymbol{\lambda}_{k,u} \right) \right) + \sum_{u=1}^{V} \log\left(\Gamma\left(\boldsymbol{\lambda}_{k,u} \right) \right). \tag{A.79}$$

Differentiating the above expression w.r.t. $\boldsymbol{\lambda}_{k,v}$ we obtain

$$\begin{aligned} \frac{\partial L_{\boldsymbol{\lambda}_k}}{\partial \boldsymbol{\lambda}_{k,v}} &= \left(\beta_v + \sum_{i=1}^{M} \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,j,k} \mathbf{1}_{\{w_{i,j}=v\}} - \boldsymbol{\lambda}_{k,v} \right) \left(\Psi'(\boldsymbol{\lambda}_{k,v}) - \Psi'\left(\sum_{u=1}^{V} \boldsymbol{\lambda}_{k,u} \right) \right) \\ &\quad - \sum_{u=1}^{V} \left(\beta_u + \sum_{i=1}^{M} \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,j,k} \mathbf{1}_{\{w_{i,j}=u\}} - \boldsymbol{\lambda}_{k,u} \right) \Psi'\left(\sum_{t=1}^{V} \boldsymbol{\lambda}_{k,t} \right) \\ &\quad - \left(\Psi'(\boldsymbol{\lambda}_{k,v}) - \Psi'\left(\sum_{u=1}^{V} \boldsymbol{\lambda}_{k,u} \right) \right) - \Psi'\left(\sum_{u=1}^{V} \boldsymbol{\lambda}_{k,u} \right) + \Psi'(\boldsymbol{\lambda}_{k,v}) \end{aligned} \tag{A.80}$$

$$\begin{aligned} &= \left(\beta_v + \sum_{i=1}^{M} \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,j,k} \mathbf{1}_{\{w_{i,j}=v\}} - \boldsymbol{\lambda}_{k,v} \right) \left(\Psi'(\boldsymbol{\lambda}_{k,v}) - \Psi'\left(\sum_{u=1}^{V} \boldsymbol{\lambda}_{k,u} \right) \right) \\ &\quad - \sum_{u=1}^{V} \left(\beta_u + \sum_{i=1}^{M} \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,j,k} \mathbf{1}_{\{w_{i,j}=u\}} - \boldsymbol{\lambda}_{k,u} \right) \Psi'\left(\sum_{t=1}^{V} \boldsymbol{\lambda}_{k,t} \right). \end{aligned} \tag{A.81}$$

Per the same argument as for $\boldsymbol{\gamma}_{i,k}$ we get that the update equation below (since $\boldsymbol{\lambda}_{k,v} \neq 0$ for any v).

$$\boldsymbol{\lambda}_{k,v} = \beta_v + \sum_{i=1}^{M} \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,j,k} \mathbf{1}_{\{w_{i,j}=v\}}. \tag{A.82}$$

In all the update equations are

$$\boldsymbol{\gamma}_{i,k} = \alpha_k + \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,j,k}, \tag{A.83}$$

$$\boldsymbol{\pi}_{i,j,k} \propto \exp\left\{ \Psi\left(\boldsymbol{\gamma}_{i,k} \right) + \Psi(\boldsymbol{\lambda}_{k,w_{i,j}}) - \Psi\left(\sum_{k=1}^{K} \boldsymbol{\lambda}_{k,w_{i,j}} \right) \right\}, \tag{A.84}$$

$$\boldsymbol{\lambda}_{k,v} = \beta_v + \sum_{i=1}^{M} \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,j,k} \mathbf{1}_{\{w_{i,j}=v\}}. \tag{A.85}$$

Note that both *Geigle (2017)* and *Blei et al. (2003)* claim that the update equation, under the expanded model where we endow $\boldsymbol{\phi}$ with a Dirichlet prior, is as follows

$$\boldsymbol{\pi}_{i,j,k} \propto \boldsymbol{\phi}_{k,w_{i,j}} \exp\left\{ \Psi\left(\boldsymbol{\gamma}_{i,k} \right) - \Psi\left(\sum_{l=1}^{K} \boldsymbol{\gamma}_{i,l} \right) \right\}. \tag{A.86}$$

However in the reduced variational distribution q, $\boldsymbol{\phi}$ is random (specifically Dirichlet with parameter $\boldsymbol{\lambda}$), and $E_q\left[\log\left(\boldsymbol{\phi}_{k,v} \right) \right]$ is not equal to $\log\left(\boldsymbol{\phi}_{k,v} \right)$ as in the original definition in *Blei et al. (2003)* but instead equal to $\left(\Psi(\boldsymbol{\lambda}_{k,v}) - \Psi\left(\sum_{u=1}^{V} \boldsymbol{\lambda}_{k,u} \right) \right)$ (see **(3)** and **(4)** above). Hence the update equations for $\boldsymbol{\pi}_{i,j,k}$ are not identical between the two definitions as claimed by both. Instead they take the form presented above.

A.3 Derivation of a Collapsed Gibbs Sampler for DMM

In a very similar fashion to LDA we begin by collapsing the Dirichlet priors.

$$
\begin{aligned}
P(\mathbf{w}, \mathbf{z}; \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \int\int P(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\Phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) d\boldsymbol{\theta} d\boldsymbol{\Phi} && \text{(A.87)} \\
&= \int\int P(\boldsymbol{\theta}; \boldsymbol{\alpha}) \prod_{k=1}^{K} P(\boldsymbol{\phi}_k; \boldsymbol{\beta}) \prod_{i=1}^{M} P(z_i|\boldsymbol{\theta}) \prod_{j=1}^{N_i} P(w_{i,j}|\boldsymbol{\phi}_{z_i}) d\boldsymbol{\theta} d\boldsymbol{\Phi} && \text{(A.88)} \\
&= \underbrace{\int P(\boldsymbol{\theta}; \boldsymbol{\alpha}) \prod_{i=1}^{M} P(z_i|\boldsymbol{\theta}) d\boldsymbol{\theta}}_{(1)} \\
&\quad \times \underbrace{\prod_{k=1}^{K} \int P(\boldsymbol{\phi}_k; \boldsymbol{\beta}) \prod_{i=1}^{M} \prod_{j=1}^{N_i} P(w_{i,j}|\boldsymbol{\phi}_{z_i}) d\boldsymbol{\phi}_k}_{(2)} . && \text{(A.89)}
\end{aligned}
$$

As for LDA we examine both integrals separately

(1)

$$
\begin{aligned}
(1) &= \int P(\boldsymbol{\theta}; \boldsymbol{\alpha}) \prod_{i=1}^{M} P(z_i|\boldsymbol{\theta}) d\boldsymbol{\theta} && \text{(A.90)} \\
&= \int \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \boldsymbol{\theta}_k^{\alpha_k - 1} \prod_{i=1}^{M} \prod_{k=1}^{K} \boldsymbol{\theta}_k^{\mathbf{1}_{\{z_i=k\}}} d\boldsymbol{\theta} && \text{(A.91)} \\
&= \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \int \prod_{k=1}^{K} \boldsymbol{\theta}_k^{\alpha_k - 1} \prod_{k=1}^{K} \boldsymbol{\theta}_k^{\sum_{i=1}^{M} \mathbf{1}_{\{z_i=k\}}} d\boldsymbol{\theta} && \text{(A.92)} \\
&= \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \int \prod_{k=1}^{K} \boldsymbol{\theta}_k^{\sum_{i=1}^{M} \mathbf{1}_{\{z_i=k\}} + \alpha_k - 1} d\boldsymbol{\theta}. && \text{(A.93)}
\end{aligned}
$$

We now let $m^{(k)}$ be the number of documents assigned topic k. Formally we can define it as

$$
m^{(k)} = \sum_{i=1}^{M} \mathbf{1}_{\{z_i=k\}}. \tag{A.94}
$$

Using this formulation we obtain

$$
\begin{aligned}
(1) &= \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \int \prod_{k=1}^{K} \boldsymbol{\theta}_k^{\sum_{i=1}^{M} \mathbf{1}_{\{\mathbf{z}_i=k\}} + \alpha_k - 1} d\boldsymbol{\theta} && \text{(A.95)} \\
&= \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \int \prod_{k=1}^{K} \boldsymbol{\theta}_k^{m^{(k)} + \alpha_k - 1} d\boldsymbol{\theta} && \text{(A.96)} \\
&= \frac{\frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)}}{\frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k + m^{(k)}\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k + m^{(k)}\right)}} \underbrace{\int \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k + m^{(k)}\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k + m^{(k)}\right)} \prod_{k=1}^{K} \boldsymbol{\theta}_k^{m^{(k)} + \alpha_k - 1} d\boldsymbol{\theta}}_{=1} && \text{(A.97)} \\
&= \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\prod_{k=1}^{K} \Gamma\left(\alpha_k + m^{(k)}\right)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k + m^{(k)}\right)}. && \text{(A.98)}
\end{aligned}
$$

We obtain the 1 on the right hand side in (A.97) as we integrate the density function of a $Dir\left(\alpha_1 + m^{(1)}, ..., \alpha_K + m^{(K)}\right)$ over its entire support.

(2)

$$(2) = \prod_{k=1}^{K} \int P(\boldsymbol{\phi}_k; \boldsymbol{\beta}) \prod_{i=1}^{M} \prod_{j=1}^{N_i} P(w_{i,j}|\boldsymbol{\phi}_{z_i}) d\boldsymbol{\phi}_k \tag{A.99}$$

$$= \prod_{k=1}^{K} \int \frac{\prod_{v=1}^{V} \Gamma(\beta_v)}{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)} \prod_{v=1}^{V} \boldsymbol{\phi}_{k,v}^{\beta_v - 1} \prod_{i=1}^{M} \prod_{j=1}^{N_i} P(w_{i,j}|\boldsymbol{\phi}_{z_i}) d\boldsymbol{\phi}_k. \tag{A.100}$$

We now define $m_{i,v}^{(k)}$ as the number of times the vth word in a vocabulary appears in document i when document i has topic k. Formally we define it as

$$m_{i,v}^{(k)} = \sum_{j=1}^{N_i} \mathbf{1}_{\{w_{i,j}=v \cap z_i = k\}}. \tag{A.101}$$

We also adopt the $\bullet$ notation which indicates a sum over the corresponding index, e.g.

$$m_{\bullet,v}^{(k)} = \sum_{i=1}^{M} n_{i,v}^{(k)} \tag{A.102}$$

$$= \sum_{i=1}^{M} \sum_{j=1}^{N_i} \mathbf{1}_{\{w_{i,j}=v \cap z_i = k\}}. \tag{A.103}$$

Using the formulation we can rewrite $P(w_{i,j}|\boldsymbol{\phi}_{z_i})$ as follows

$$\prod_{i=1}^{M} \prod_{j=1}^{N_i} P(w_{i,j}|\boldsymbol{\phi}_{z_i}) = \prod_{i=1}^{M} \prod_{j=1}^{N_i} \prod_{v=1}^{V} \boldsymbol{\phi}_{z_i,v}^{\mathbf{1}_{\{w_{i,j}=v\}}} \tag{A.104}$$

$$= \prod_{k=1}^{K} \prod_{v=1}^{V} \boldsymbol{\phi}_{k,v}^{\sum_{i=1}^{M} \sum_{j=1}^{N_i} \mathbf{1}_{\{w_{i,j}=v \cap z_i = k\}}} \tag{A.105}$$

$$= \prod_{k=1}^{K} \prod_{v=1}^{V} \boldsymbol{\phi}_{k,v}^{m_{\bullet,v}^{(k)}}. \tag{A.106}$$

Note here that $m_{\bullet,v}^{(k)}$ is the number of times word v in a vocabulary appears in a document with topic k. We now continue where we left off.

$$(2) = \prod_{k=1}^{K} \int P(\boldsymbol{\phi}_k; \boldsymbol{\beta}) \prod_{i=1}^{M} \prod_{j=1}^{N_i} P(w_{i,j}|\boldsymbol{\phi}_{\mathbf{z}_i}) d\boldsymbol{\phi}_k \tag{A.107}$$

$$= \prod_{k=1}^{K} \int \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \prod_{v=1}^{V} \boldsymbol{\phi}_{k,v}^{\beta_v - 1} \prod_{k=1}^{K} \prod_{v=1}^{V} \boldsymbol{\phi}_{k,v}^{m_{\bullet,v}^{(k)}} d\boldsymbol{\phi}_k \tag{A.108}$$

$$= \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \int \prod_{v=1}^{V} \boldsymbol{\phi}_{k,v}^{m_{\bullet,v}^{(k)} + \beta_v - 1} d\boldsymbol{\phi}_k \tag{A.109}$$

$$= \prod_{k=1}^{K} \frac{\frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma(\beta_v)}}{\frac{\Gamma\left(\sum_{v=1}^{V} \beta_v + m_{\bullet,v}^{(k)}\right)}{\prod_{v=1}^{V} \Gamma\left(\beta_v + m_{\bullet,v}^{(k)}\right)}} \underbrace{\int \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v + m_{\bullet,v}^{(k)}\right)}{\prod_{v=1}^{V} \Gamma\left(\beta_v + m_{\bullet,v}^{(k)}\right)} \prod_{v=1}^{V} \boldsymbol{\phi}_{k,v}^{m_{\bullet,v}^{(k)} + \beta_v - 1} d\boldsymbol{\phi}_k}_{=1} \tag{A.110}$$

$$= \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \frac{\prod_{v=1}^{V} \Gamma\left(\beta_v + m_{\bullet,v}^{(k)}\right)}{\Gamma\left(\sum_{v=1}^{V} \beta_v + m_{\bullet,v}^{(k)}\right)}. \tag{A.111}$$

We once again obtain the 1 on the right hand side in (A.110) as we integrate the density function of a $Dir\left(\beta_1 + m^{(k)}_{\bullet,1}, ..., \beta_V + n^{(k)}_{\bullet,V}\right)$ over its entire support.

We now return to the initial integral and combine integrals (1) and (2).

$$\begin{aligned} P(\mathbf{w}, \mathbf{z}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = &\frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k\right)} \frac{\prod_{k=1}^{K} \Gamma\left(\alpha_k + m^{(k)}\right)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k + m^{(k)}\right)} \\ &\times \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma\left(\beta_v\right)} \frac{\prod_{v=1}^{V} \Gamma\left(\beta_v + m^{(k)}_{\bullet,v}\right)}{\Gamma\left(\sum_{v=1}^{V} \beta_v + m^{(k)}_{\bullet,v}\right)}. \end{aligned} \tag{A.112}$$

What remains is to derive $P(z_a|\mathbf{z}_{\neg a}, \mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})$ where z_a indicates the topic of document a and $\mathbf{z}_{\neg a}$ the topics of all documents except a [11].

$$P(z_a|\mathbf{z}_{-a}, \mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{P(z_a, \mathbf{z}_{\neg a}, \mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})}{P(\mathbf{z}_{\neg a}, \mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})} \tag{A.113}$$

$$\propto P(\mathbf{z}, \mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta}). \tag{A.114}$$

Using the above proportionality we get

$$P(z_a|\mathbf{z}_{\neg a}, \mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \frac{\prod_{k=1}^{K} \Gamma\left(\alpha_k + m^{(k)}\right)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k + m^{(k)}\right)} \times \prod_{k=1}^{K} \frac{\prod_{v \in \mathbf{w}_a} \Gamma\left(\beta_v + m^{(k)}_{\bullet,v}\right)}{\Gamma\left(\sum_{v=1}^{V} \beta_v + m^{(k)}_{\bullet,v}\right)}. \tag{A.115}$$

We now let $m^{(k)\neg a}_{\bullet,v}$ denote the words counts $m^{(k)}_{\bullet,v}$ in documents, excluding document a, of topic k. We also let $m^{(k)\neg a}$ be the number of documents of topic k, excluding document a. Formally we define these counts as

$$m^{(k)\neg a}_{\bullet,v} = \sum_{i \neq a} m^{(k)}_{i,v} \tag{A.116}$$

$$= \sum_{i \neq a} \sum_{j=1}^{N_i} \mathbf{1}_{\{w_{i,j}=v \cap z_i=k\}}, \tag{A.117}$$

$$m^{(k)\neg a} = \sum_{i \neq a} \mathbf{1}_{\{z_i=k\}}. \tag{A.118}$$

Using this formulation we obtain

$$\begin{aligned} P(z_a = \kappa|\mathbf{z}_{\neg a}, \mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto &\frac{\prod_{k \neq \kappa} \Gamma\left(\alpha_k + m^{(k)\neg a}\right) \times \Gamma\left(1 + \alpha_\kappa + m^{(\kappa)\neg a}\right)}{\Gamma\left(1 + \sum_{k=1}^{K} \alpha_k + m^{(k)}\right)} \\ &\times \prod_{k \neq \kappa} \frac{\prod_{v \in \mathbf{w}_a} \Gamma\left(\beta_v + m^{(k)\neg a}_{\bullet,v}\right)}{\Gamma\left(\sum_{v=1}^{V} \beta_v + m^{(k)\neg a}_{\bullet,v}\right)} \\ &\times \frac{\prod_{v \in \mathbf{w}_a} \Gamma\left(\beta_v + m^{(\kappa)\neg a}_{\bullet,v} + m^{(\kappa)}_{a,v}\right)}{\Gamma\left(m^{(\kappa)}_{a,\bullet} + \sum_{v=1}^{V} \beta_v + m^{(\kappa)\neg a}_{\bullet,v}\right)}. \end{aligned} \tag{A.119}$$

We now use that $m^{(\kappa)\neg a} = m^{(\kappa)} - 1$ which is true under the assumption that document a is of topic κ. The same subtraction cannot be made for $m^{(k)\neg a}_{\bullet,v}$ since a document can have multiple occurrences of the same word. We do however note that $m^{(\kappa)\neg a}_{\bullet,v} = m^{(\kappa)\neg a}_{\bullet,v} + m^{(\kappa)}_{a,v}$. Now we use that the Gamma function satisfies $\Gamma(x) = (x-1)\Gamma(x-1)$

and keep in mind that this property can be used multiple times such that $\Gamma(x) = \Gamma(x-n)\prod_{i=1}^{n}(x-n+i-1)$.

$$
\begin{aligned}
P(z_a = \kappa|\mathbf{z}_{\neg a}, \mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto & \frac{\prod_{k\neq\kappa}\Gamma\left(\alpha_k + m^{(k)\neg a}\right) \times \Gamma\left(\alpha_\kappa + m^{(\kappa)\neg a}\right)}{\Gamma\left(\sum_{k=1}^{K}\alpha_k + m^{(k)}\right)} \\
& \times \frac{\left(\alpha_\kappa + m^{(\kappa)\neg a}\right)}{\sum_{k=1}^{K}\alpha_k + m^{(k)}} \\
& \times \prod_{k\neq\kappa}\frac{\prod_{v\in\mathbf{w}_a}\Gamma\left(\beta_v + m_{\bullet,v}^{(k)\neg a}\right)}{\Gamma\left(\sum_{v=1}^{V}\beta_v + m_{\bullet,v}^{(k)\neg a}\right)} \\
& \times \frac{\prod_{v\in\mathbf{w}_a}\Gamma\left(\beta_v + m_{\bullet,v}^{(\kappa)\neg a}\right)}{\Gamma\left(\sum_{v=1}^{V}\beta_v + m_{\bullet,v}^{(\kappa)\neg a}\right)} \\
& \times \frac{\prod_{v\in\mathbf{w}_a}\prod_{n=1}^{m_{a,v}^{(\bullet)}}\left(\beta_v + m_{\bullet,v}^{(\kappa)\neg a} - m_{a,v}^{(\bullet)} + n - 1\right)}{\prod_{n=1}^{N_a}\left(\sum_{v=1}^{V}\left(\beta_v + m_{\bullet,v}^{(\kappa)\neg a}\right) - N_a + n - 1\right)} \\
= & \underbrace{\frac{\prod_{k=1}^{K}\Gamma\left(\alpha_k + m^{(k)\neg a}\right)}{\Gamma\left(\sum_{k=1}^{K}\alpha_k + m^{(k)}\right)}}_{(1)} \times \frac{\left(\alpha_\kappa + m^{(\kappa)\neg a}\right)}{\sum_{k=1}^{K}\alpha_k + m^{(k)}} \\
& \times \underbrace{\prod_{k=1}^{K}\frac{\prod_{v\in\mathbf{w}_a}\Gamma\left(\beta_v + m_{\bullet,v}^{(k)\neg a}\right)}{\Gamma\left(\sum_{v=1}^{V}\beta_v + m_{\bullet,v}^{(k)\neg a}\right)}}_{(2)} \\
& \times \frac{\prod_{v\in\mathbf{w}_a}\prod_{n=1}^{m_{a,v}^{(\bullet)}}\left(\beta_v + m_{\bullet,v}^{(\kappa)\neg a} - m_{a,v}^{(\bullet)} + n - 1\right)}{\prod_{n=1}^{N_a}\left(\sum_{v=1}^{V}\left(\beta_v + m_{\bullet,v}^{(\kappa)\neg a}\right) - N_a + n - 1\right)}.
\end{aligned}
\tag{A.120}
\tag{A.121}
$$

Now we can drop (1) and (2) as they do not depend of κ to obtain the following. Note that $m_{i,v}^{(\bullet)}$ corresponds to the number of occurrences of the vth word in document i (regardless of topic).

$$
\begin{aligned}
P(z_a = \kappa|\mathbf{z}_{\neg a}, \mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto & \frac{\left(\alpha_\kappa + m^{(\kappa)\neg a}\right)}{\sum_{k=1}^{K}\alpha_k + m^{(k)}} \\
& \times \frac{\prod_{v\in\mathbf{w}_a}\prod_{n=1}^{m_{a,v}^{(\bullet)}}\left(\beta_v + m_{\bullet,v}^{(\kappa)\neg a} - m_{a,v}^{(\bullet)} + n - 1\right)}{\prod_{n=1}^{N_a}\left(\sum_{v=1}^{V}\left(\beta_v + m_{\bullet,v}^{(\kappa)\neg a}\right) - N_a + n - 1\right)}
\end{aligned}
\tag{A.122}
$$

As for LDA we can see the upper fraction in equation (A.122) as the document-topic probability and the lower fraction as the topic-word probabilities. These results coincide with those obtained in [5].

A.4 Variational Bayesian Inference for DMM

Recall that

$$
p(\mathbf{z}, \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = P(\boldsymbol{\theta}; \boldsymbol{\alpha})\prod_{k=1}^{K}P(\boldsymbol{\phi}_k; \boldsymbol{\beta})\prod_{i=1}^{M}P(z_i|\boldsymbol{\theta})\prod_{j=1}^{N_i}P(w_{i,j}|\boldsymbol{\phi}_{z_i}). \tag{A.123}
$$

These derivations are inspired by those for LDA in *Geigle (2017)* and in *Blei et al. (2003)*.

Like for LDA in in *Blei et al. (2003)* we reduced the model to the variational distribution $q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi})$ as in Figure 2.5.

$$q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}) = q(\boldsymbol{\theta}; \boldsymbol{\gamma}) \prod_{k=1}^{K} q(\boldsymbol{\phi}_k; \boldsymbol{\lambda}_k) \prod_{i=1}^{M} q(z_i; \boldsymbol{\pi}_i). \tag{A.124}$$

We then find the update equations for $\boldsymbol{\gamma}$, $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$ by minimizing the Kullback-Leibler Divergence between the variational distribution q and the true posteriors p.

$$D(q||p) = \int\int \sum_{\mathbf{z}} q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}) \log\left(\frac{q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda})}{p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})}\right) d\boldsymbol{\theta} d\boldsymbol{\phi} \tag{A.125}$$

$$= \int\int \sum_{\mathbf{z}} q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}) \log\left(\frac{q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda})}{p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})} p(\mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})\right) d\boldsymbol{\theta} d\boldsymbol{\phi} \tag{A.126}$$

$$\begin{aligned} &= \int\int \sum_{\mathbf{z}} q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}) \log\left(\frac{q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda})}{p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})}\right) d\boldsymbol{\theta} d\boldsymbol{\phi} \\ &+ \int\int \sum_{\mathbf{z}} q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}) \log\left(p(\mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})\right) d\boldsymbol{\theta} d\boldsymbol{\phi}. \end{aligned} \tag{A.127}$$

In (A.127) we note that $\log(p(\mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta}))$ does not depend on neither $\mathbf{z}$, $\boldsymbol{\theta}$ nor $\boldsymbol{\phi}$. As such we obtain

$$\begin{aligned} D(q||p) = &\int\int \sum_{\mathbf{z}} q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}) \log\left(\frac{q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda})}{p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})}\right) d\boldsymbol{\theta} d\boldsymbol{\phi} \\ &+ \log\left(p(\mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})\right) \underbrace{\int\int \sum_{\mathbf{z}} q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}) d\boldsymbol{\theta} d\boldsymbol{\phi}}_{=1} \end{aligned} \tag{A.128}$$

$$\begin{aligned} &= \int\int \sum_{\mathbf{z}} q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}) \log\left(\frac{q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda})}{p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})}\right) d\boldsymbol{\theta} d\boldsymbol{\phi} \\ &+ \log\left(p(\mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})\right). \end{aligned} \tag{A.129}$$

As $\log\left(p(\mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\phi})\right)$ does not depend on q we simply have to minimize the first part of equation (A.129).

$$\int\int \sum_{\mathbf{z}} q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}) \log\left(\frac{q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda})}{p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})}\right) d\boldsymbol{\theta} d\boldsymbol{\phi} \tag{A.130}$$

$$\begin{aligned} &= \int\int \sum_{\mathbf{z}} q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}) \log\left(q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda})\right) d\boldsymbol{\theta} d\boldsymbol{\phi} \\ &- \int\int \sum_{\mathbf{z}} q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}) \log\left(p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})\right) d\boldsymbol{\theta} d\boldsymbol{\phi} \end{aligned} \tag{A.131}$$

$$= E_q\left[\log\left(q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda})\right)\right] - E_q\left[\log\left(p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})\right)\right]. \tag{A.132}$$

We can then define L as follows.

$$\begin{aligned} L(\boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}; \boldsymbol{\alpha}, \boldsymbol{\beta}) &= E_q \left[\log \left(p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{w}; \boldsymbol{\alpha}, \boldsymbol{\beta})\right)\right] \\ &\quad - E_q \left[\log \left(q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda})\right)\right] \end{aligned} \tag{A.133}$$

$$\begin{aligned} &= E_q \left[\log(p(\boldsymbol{\theta}; \boldsymbol{\alpha}) p(\mathbf{z}; \boldsymbol{\theta}) p(\boldsymbol{\phi}; \boldsymbol{\beta}) p(\mathbf{w}; \mathbf{z}, \boldsymbol{\phi}))\right] \\ &\quad - E_q \left[\log(q(\boldsymbol{\phi}; \boldsymbol{\lambda}) q(\boldsymbol{\theta}; \boldsymbol{\gamma}) q(\mathbf{z}; \boldsymbol{\pi}))\right] \end{aligned} \tag{A.134}$$

$$\begin{aligned} &= \underbrace{E_q[\log(p(\boldsymbol{\theta}; \boldsymbol{\alpha}))]}_{(1)} + \underbrace{E_q[\log(p(\mathbf{z}; \boldsymbol{\theta}))]}_{(2)} \\ &\quad + \underbrace{E_q[\log(p(\boldsymbol{\phi}; \boldsymbol{\beta}))]}_{(3)} + \underbrace{E_q \left[p(\mathbf{w}; \mathbf{z}, \boldsymbol{\phi})\right]}_{(4)} \\ &\quad - \underbrace{E_q[\log(q(\boldsymbol{\phi}; \boldsymbol{\lambda}))]}_{(5)} - \underbrace{E_q[\log(q(\boldsymbol{\theta}; \boldsymbol{\gamma}))]}_{(6)} - \underbrace{E_q[\log(q(\mathbf{z}; \boldsymbol{\pi}))]}_{(7)}. \end{aligned} \tag{A.135}$$

We can now evaluate the seven expectations above individually.
(1)

$$E_q[\log(p(\boldsymbol{\theta}; \boldsymbol{\alpha}))] = E_q \left[\log \left(\frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma\left(\alpha_k\right)} \prod_{k=1}^K \boldsymbol{\theta}_k^{\alpha_k - 1}\right)\right] \tag{A.136}$$

$$= \log \left(\frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma\left(\alpha_k\right)}\right) + \sum_{k=1}^K (\alpha_k - 1) E_q[\log(\boldsymbol{\theta}_k)]. \tag{A.137}$$

In the reduced model $\boldsymbol{\theta} \sim Dir(\boldsymbol{\gamma})$ which is in the exponential family of distributions as is shown in Appendix A.6. Using the fact that $\boldsymbol{\theta}$ is of an exponential family we, as derived in Appendix A.7, have that

$$E_q[\log(\boldsymbol{\theta}_k)] = \Psi\left(\boldsymbol{\gamma}_k\right) - \Psi\left(\sum_{l=1}^K \boldsymbol{\gamma}_l\right) \tag{A.138}$$

where Ψ is the digamma function, i.e. $\frac{\Gamma'}{\Gamma}$. Now returning to where we left off we get

$$E_q[\log(p(\boldsymbol{\theta}|\boldsymbol{\alpha}))] = \log \left(\frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma\left(\alpha_k\right)}\right) + \sum_{k=1}^K (\alpha_k - 1) \left(\Psi\left(\boldsymbol{\gamma}_k\right) - \Psi\left(\sum_{l=1}^K \boldsymbol{\gamma}_l\right)\right). \tag{A.139}$$

(2)

$$E_q[\log(p(\mathbf{z}; \boldsymbol{\theta}))] = E_q \left[\log \left(\prod_{i=1}^M \prod_{k=1}^K \boldsymbol{\theta}_k^{\mathbf{1}_{\{z_i = k\}}}\right)\right] \tag{A.140}$$

$$= E_q \left[\sum_{i=1}^M \sum_{k=1}^K \mathbf{1}_{\{z_i = k\}} \log\left(\boldsymbol{\theta}_k\right)\right] \tag{A.141}$$

$$= \sum_{i=1}^M \sum_{k=1}^K E_q[\mathbf{1}_{\{z_i = k\}}] E_q\left[\log\left(\boldsymbol{\theta}_k\right)\right] \tag{A.142}$$

$$= \sum_{i=1}^M \sum_{k=1}^K \boldsymbol{\pi}_{i,k} \left(\Psi\left(\boldsymbol{\gamma}_k\right) - \Psi\left(\sum_{l=1}^K \boldsymbol{\gamma}_l\right)\right). \tag{A.143}$$

(3)

$$E_q\left[\log\left(p(\boldsymbol{\phi};\boldsymbol{\beta})\right)\right] = E_q\left[\log\left(\prod_{k=1}^{K}\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v\right)}\prod_{v=1}^{V}\boldsymbol{\phi}_{k,v}^{\beta_v-1}\right)\right] \tag{A.144}$$

$$= \sum_{k=1}^{K}\log\left(\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v\right)}\right) + \sum_{v=1}^{V}(\beta_k-1)E_q\left[\log\left(\boldsymbol{\phi}_{k,v}\right)\right]. \tag{A.145}$$

Like in **(1)** we use that $\boldsymbol{\phi}_k \sim Dir(\boldsymbol{\lambda})$ to obtain

$$E_q\left[\log\left(p(\boldsymbol{\phi};\boldsymbol{\beta})\right)\right] = \sum_{k=1}^{K}\log\left(\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v\right)}\right) + \sum_{v=1}^{V}(\beta_v-1)\left(\Psi(\boldsymbol{\lambda}_{k,v}) - \Psi\left(\sum_{u=1}^{V}\boldsymbol{\lambda}_{k,u}\right)\right). \tag{A.146}$$

(4)

$$E_q[\log(p(\mathbf{w}|\mathbf{z},\boldsymbol{\phi}))] = E_q\left[\log\left(\prod_{i=1}^{M}\prod_{j=1}^{N_i}\prod_{k=1}^{K}\prod_{v=1}^{V}\boldsymbol{\phi}_{k,v}^{\mathbf{1}_{\{z_i=k\cap w_{i,j}=v\}}}\right)\right] \tag{A.147}$$

$$= E_q\left[\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{K}\sum_{v=1}^{V}\mathbf{1}_{\{z_i=k\cap w_{i,j}=v\}}\log(\boldsymbol{\phi}_{k,v})\right] \tag{A.148}$$

$$= \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{K}\sum_{v=1}^{V}E_q\left[\mathbf{1}_{\{z_i=k\}}\mathbf{1}_{\{w_{i,j}=v\}}\log(\boldsymbol{\phi}_{k,v})\right] \tag{A.149}$$

$$= \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{K}\sum_{v=1}^{V}\boldsymbol{\pi}_{i,k}\mathbf{1}_{\{w_{i,j}=v\}}\left(\Psi(\boldsymbol{\lambda}_{k,v}) - \Psi\left(\sum_{u=1}^{V}\boldsymbol{\lambda}_{k,u}\right)\right). \tag{A.150}$$

(5)
The fifth expectation is identical to **(3)** except that we vary the parameters of the word priors from $\boldsymbol{\beta}$ to $\boldsymbol{\lambda}$.

$$E_q\left[\log\left(\boldsymbol{\phi};\boldsymbol{\lambda}\right)\right] = \sum_{k=1}^{K}\log\left(\frac{\Gamma\left(\sum_{v=1}^{V}\boldsymbol{\lambda}_{k,v}\right)}{\prod_{v=1}^{V}\Gamma\left(\boldsymbol{\lambda}_{k,v}\right)}\right) + \sum_{v=1}^{V}(\boldsymbol{\lambda}_{k,v}-1)\left(\Psi(\boldsymbol{\lambda}_{k,v}) - \Psi\left(\sum_{u=1}^{V}\boldsymbol{\lambda}_{k,u}\right)\right). \tag{A.151}$$

(6)
The sixth expectation is identical to **(1)** except that we vary the parameters topic priors from $\boldsymbol{\alpha}$ to $\boldsymbol{\gamma}$, giving us

$$E_q[\log(q(\boldsymbol{\theta};\boldsymbol{\gamma}))] = \log\left(\frac{\Gamma\left(\sum_{k=1}^{K}\boldsymbol{\gamma}_k\right)}{\prod_{k=1}^{K}\Gamma\left(\boldsymbol{\gamma}_k\right)}\right) + \sum_{k=1}^{K}(\boldsymbol{\gamma}_k-1)\left(\Psi\left(\boldsymbol{\gamma}_k\right) - \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_l\right)\right). \tag{A.152}$$

(7)

$$E_q\left[\log\left(q(\mathbf{z}|\boldsymbol{\pi})\right)\right] = E_q\left[\log\left(\prod_{i=1}^{M}\prod_{k=1}^{K}\boldsymbol{\pi}_{i,j,k}^{\mathbf{1}_{\{z_i=k\}}}\right)\right] \tag{A.153}$$

$$= E_q\left[\sum_{i=1}^{M}\sum_{k=1}^{K}\mathbf{1}_{\{z_i=k\}}\log\left(\boldsymbol{\pi}_{i,k}\right)\right] \tag{A.154}$$

$$= \sum_{i=1}^{M}\sum_{k=1}^{K}E_q\left[\mathbf{1}_{\{z_i=k\}}\log\left(\boldsymbol{\pi}_{i,k}\right)\right] \tag{A.155}$$

$$= \sum_{i=1}^{M}\sum_{k=1}^{K}\boldsymbol{\pi}_{i,k}\log\left(\boldsymbol{\pi}_{i,k}\right). \tag{A.156}$$

Combining **(1)**-**(7)** we get that L as follows

$$\begin{aligned}
L(\boldsymbol{\gamma},\boldsymbol{\pi},\boldsymbol{\lambda};\boldsymbol{\alpha},\boldsymbol{\beta}) = {} & \log\left(\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)}\right) + \sum_{k=1}^{K}(\alpha_k-1)\left(\Psi\left(\boldsymbol{\gamma}_k\right) - \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_l\right)\right) \\
& + \sum_{i=1}^{M}\sum_{k=1}^{K}\boldsymbol{\pi}_{i,k}\left(\Psi\left(\boldsymbol{\gamma}_k\right) - \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_l\right)\right) \\
& + \sum_{k=1}^{K}\log\left(\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v\right)}\right) + \sum_{v=1}^{V}(\beta_v-1)\left(\Psi(\boldsymbol{\lambda}_{k,v}) - \Psi\left(\sum_{u=1}^{V}\boldsymbol{\lambda}_{k,u}\right)\right) \\
& + \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{K}\sum_{v=1}^{V}\boldsymbol{\pi}_{i,k}\mathbf{1}_{\{w_{i,j}=v\}}\left(\Psi(\boldsymbol{\lambda}_{k,v}) - \Psi\left(\sum_{u=1}^{V}\boldsymbol{\lambda}_{k,u}\right)\right) \\
& - \sum_{k=1}^{K}\log\left(\frac{\Gamma\left(\sum_{v=1}^{V}\boldsymbol{\lambda}_{k,v}\right)}{\prod_{v=1}^{V}\Gamma\left(\boldsymbol{\lambda}_{k,v}\right)}\right) + \sum_{v=1}^{V}(\boldsymbol{\lambda}_{k,v}-1)\left(\Psi(\boldsymbol{\lambda}_{k,v}) - \Psi\left(\sum_{u=1}^{V}\boldsymbol{\lambda}_{k,u}\right)\right) \\
& - \log\left(\frac{\Gamma\left(\sum_{k=1}^{K}\boldsymbol{\gamma}_k\right)}{\prod_{k=1}^{K}\Gamma\left(\boldsymbol{\gamma}_k\right)}\right) + \sum_{k=1}^{K}(\boldsymbol{\gamma}_k-1)\left(\Psi\left(\boldsymbol{\gamma}_k\right) - \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_l\right)\right) \\
& - \sum_{i=1}^{M}\sum_{k=1}^{K}\boldsymbol{\pi}_{i,k}\log\left(\boldsymbol{\pi}_{i,k}\right).
\end{aligned} \tag{A.157}$$

The estimates for $\boldsymbol{\gamma}_k$ and $\boldsymbol{\pi}_{i,k}$ and $\boldsymbol{\lambda}_{k,v}$ can now be obtained by maximizing L with respect to these parameters.

We begin by examining $\boldsymbol{\gamma}_k$, letting $L_{\boldsymbol{\gamma}}$ denote the proportionality of L w.r.t. $\boldsymbol{\gamma}$.

$$\begin{aligned}
L_{\boldsymbol{\gamma}_k} = {} & \sum_{k=1}^{K}\left(\alpha_k + \sum_{i=1}^{M}\boldsymbol{\pi}_{i,k} - \boldsymbol{\gamma}_k\right)\left(\Psi(\boldsymbol{\gamma}_k) - \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_l\right)\right) \\
& - \log\left(\Gamma\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_l\right)\right) + \sum_{k=1}^{K}\log\left(\Gamma\left(\boldsymbol{\gamma}_k\right)\right).
\end{aligned} \tag{A.158}$$

Differentiating w.r.t. $\boldsymbol{\gamma}_k$ we obtain

$$\begin{aligned}\frac{\partial L_{\boldsymbol{\gamma}}}{\partial \boldsymbol{\gamma}_k} &= \left(\alpha_k + \sum_{i=1}^{M} \boldsymbol{\pi}_{i,k} - \boldsymbol{\gamma}\right)\left(\Psi\left(\boldsymbol{\gamma}_k\right)\Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_l\right)\right) \\ &\quad - \sum_{c=1}^{K}\left(\alpha_c + \sum_{i=1}^{M}\boldsymbol{\pi}_{i,c} - \boldsymbol{\gamma}_c\right)\left(\Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_l\right)\right) \\ &\quad - \left(\Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_l\right) - \Psi\left(\boldsymbol{\gamma}_k\right)\right) + \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_l\right) - \Psi\left(\boldsymbol{\gamma}_k\right) \end{aligned} \tag{A.159}$$

$$\begin{aligned}&= \left(\alpha_k + \sum_{i=1}^{M} \boldsymbol{\pi}_{i,k} - \boldsymbol{\gamma}_k\right)\left(\Psi\left(\boldsymbol{\gamma}_k\right)\Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_l\right)\right) \\ &\quad - \sum_{c=1}^{K}\left(\alpha_c + \sum_{i=1}^{M}\boldsymbol{\pi}_{i,c} - \boldsymbol{\gamma}_c\right)\left(\Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_l\right)\right). \end{aligned} \tag{A.160}$$

We then have that $\nabla\boldsymbol{\gamma} = \mathbf{0}$ (since $\boldsymbol{\gamma}_k \neq 0$ for all k)

$$\boldsymbol{\gamma}_k = \alpha_k + \sum_{i=1}^{M}\boldsymbol{\pi}_{i,k} \tag{A.161}$$

for all k.

We now examine $\boldsymbol{\pi}_{i,k}$, letting $L_{\boldsymbol{\pi}_{i,k}}$ denote the proportionality of L w.r.t. $\boldsymbol{\pi}_{i,k}$. We also note that since $\boldsymbol{\pi}_{i,k}$ is the probability that document i has topic k we have the constraint $\sum_{k=1}^{K}\boldsymbol{\pi}_{i,k} = 1$.

$$\begin{aligned}L_{\boldsymbol{\pi}_{i,k}} &= \left(\alpha_k + \sum_{i=1}^{M}\boldsymbol{\pi}_{i,k} - \boldsymbol{\gamma}_k\right)\left(\Psi\left(\boldsymbol{\gamma}_k\right)\Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_l\right)\right) \\ &\quad - \sum_{c=1}^{K}\left(\alpha_c + \sum_{i=1}^{M}\boldsymbol{\pi}_{i,c} - \boldsymbol{\gamma}_c\right)\left(\Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_l\right)\right) \\ &= \boldsymbol{\pi}_{i,k}\left(\Psi\left(\boldsymbol{\gamma}_k\right) - \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_l\right)\right) \end{aligned} \tag{A.162}$$

$$\begin{aligned}&+ \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{K}\sum_{v=1}^{V}\boldsymbol{\pi}_{i,k}\mathbf{1}_{\{w_{i,j}=v\}}\left(\Psi(\boldsymbol{\lambda}_{k,v}) - \Psi\left(\sum_{u=1}^{V}\boldsymbol{\lambda}_{k,u}\right)\right) \\ &- \boldsymbol{\pi}_{i,k}\log\left(\boldsymbol{\pi}_{i,k}\right) + \lambda\left(\sum_{k=1}^{K}\boldsymbol{\pi}_{i,k} - 1\right). \end{aligned} \tag{A.163}$$

Differentiating the above expression w.r.t. $\boldsymbol{\pi}_{i,k}$ we obtain

$$\begin{aligned}\frac{\partial L_{\boldsymbol{\pi}_{i,k}}}{\partial \boldsymbol{\pi}_{i,k}} &= \left(\Psi\left(\boldsymbol{\gamma}_k\right) - \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_l\right)\right) \\ &\quad + \sum_{j=1}^{N_i}\sum_{v=1}^{V}\mathbf{1}_{\{w_{i,j}=v\}}\left(\Psi(\boldsymbol{\lambda}_{k,v}) - \Psi\left(\sum_{u=1}^{V}\boldsymbol{\lambda}_{k,u}\right)\right) \\ &\quad - \left(\log\left(\boldsymbol{\pi}_{i,j,k}\right) + 1\right) + \lambda \end{aligned} \tag{A.164}$$

$$\begin{aligned}&= \left(\Psi\left(\boldsymbol{\gamma}_k\right) - \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_l\right)\right) \\ &\quad + \sum_{j=1}^{N_i}\left(\Psi(\boldsymbol{\lambda}_{k,w_{i,j}}) - \Psi\left(\sum_{u=1}^{V}\boldsymbol{\lambda}_{k,u}\right)\right) \\ &\quad - \left(\log\left(\boldsymbol{\pi}_{i,j,k}\right) + 1\right) + \lambda. \end{aligned} \tag{A.165}$$

Setting the above to zero we get that

$$\boldsymbol{\pi}_{i,k} \propto \exp\left\{ \Psi\left(\boldsymbol{\gamma}_k\right) + \sum_{j=1}^{N_i} \Psi(\boldsymbol{\lambda}_{k,w_{i,j}}) - \Psi\left(\sum_{u=1}^{V} \boldsymbol{\lambda}_{k,u}\right) \right\}. \tag{A.166}$$

Lastly we examine $\boldsymbol{\lambda}_{k,v}$ and let $L_{\boldsymbol{\lambda}_{k,v}}$ denote the proportionality of L w.r.t. $\boldsymbol{\lambda}_{k,v}$.

$$\begin{aligned} L_{\boldsymbol{\lambda}_k} = &\sum_{v=1}^{V} \left(\beta_v + \sum_{i=1}^{M} \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,k} \mathbf{1}_{\{w_{i,j}=v\}} - \boldsymbol{\lambda}_{k,v} \right) \left(\Psi(\boldsymbol{\lambda}_{k,v}) - \Psi\left(\sum_{u=1}^{V} \boldsymbol{\lambda}_{k,u} \right) \right) \\ &- \log\left(\Gamma\left(\sum_{v=1}^{V} \boldsymbol{\lambda}_{k,v} \right) \right) + \sum_{v=1}^{V} \log\left(\Gamma\left(\boldsymbol{\lambda}_{k,v} \right) \right). \end{aligned} \tag{A.167}$$

Differentiating the above expression w.r.t. $\boldsymbol{\lambda}_{k,v}$ we obtain

$$\begin{aligned} \frac{\partial L_{\boldsymbol{\lambda}_k}}{\partial \boldsymbol{\lambda}_{k,v}} = &\left(\beta_v + \sum_{i=1}^{M} \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,k} \mathbf{1}_{\{w_{i,j}=v\}} - \boldsymbol{\lambda}_{k,v} \right) \left(\Psi'(\boldsymbol{\lambda}_{k,v}) - \Psi'\left(\sum_{u=1}^{V} \boldsymbol{\lambda}_{k,u} \right) \right) \\ &- \sum_{u=1}^{V} \left(\beta_u + \sum_{i=1}^{M} \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,k} \mathbf{1}_{\{w_{i,j}=u\}} - \boldsymbol{\lambda}_{k,u} \right) \Psi'\left(\sum_{u=1}^{V} \boldsymbol{\lambda}_{k,u} \right) \\ &- \left(\Psi'(\boldsymbol{\lambda}_{k,v}) - \Psi'\left(\sum_{u=1}^{V} \boldsymbol{\lambda}_{k,u} \right) \right) - \Psi'\left(\sum_{u=1}^{V} \boldsymbol{\lambda}_{k,u} \right) + \Psi'(\boldsymbol{\lambda}_{k,v}) \end{aligned} \tag{A.168}$$

$$\begin{aligned} = &\left(\beta_v + \sum_{i=1}^{M} \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,k} \mathbf{1}_{\{w_{i,j}=v\}} - \boldsymbol{\lambda}_{k,v} \right) \left(\Psi'(\boldsymbol{\lambda}_{k,v}) - \Psi'\left(\sum_{k=1}^{K} \boldsymbol{\lambda}_{k,v} \right) \right) \\ &- \sum_{u=1}^{V} \left(\beta_u + \sum_{i=1}^{M} \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,k} \mathbf{1}_{\{w_{i,j}=u\}} - \boldsymbol{\lambda}_{k,u} \right) \Psi'\left(\sum_{u=1}^{V} \boldsymbol{\lambda}_{k,u} \right). \end{aligned} \tag{A.169}$$

Per the same argument as for $\boldsymbol{\gamma}_k$ we obtain (since $\boldsymbol{\lambda}_{k,v} \neq 0$ for any v)

$$\boldsymbol{\lambda}_{k,v} = \beta_v + \sum_{i=1}^{M} \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,k} \mathbf{1}_{\{w_{i,j}=v\}} \tag{A.170}$$

for all k and v.

In all the update equations are

$$\boldsymbol{\gamma}_k = \alpha_k + \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,k}, \tag{A.171}$$

$$\boldsymbol{\pi}_{i,k} \propto \exp\left\{ \Psi\left(\boldsymbol{\gamma}_k\right) + \sum_{j=1}^{N_i} \Psi(\boldsymbol{\lambda}_{k,w_{i,j}}) - \Psi\left(\sum_{u=1}^{V} \boldsymbol{\lambda}_{k,u}\right) \right\}, \tag{A.172}$$

$$\boldsymbol{\lambda}_{k,v} = \beta_v + \sum_{i=1}^{M} \sum_{j=1}^{N_i} \boldsymbol{\pi}_{i,k} \mathbf{1}_{\{w_{i,j}=v\}}. \tag{A.173}$$

A.5 Derivation for Empirical Bayes for LDA

We remind ourselves that the minimization of the KL divergence is equivalent to the maximization of $L(\boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\lambda}; \boldsymbol{\alpha}, \boldsymbol{\beta})$ where

$$L_{\boldsymbol{\alpha},\boldsymbol{\beta}} = \sum_{i=1}^{M} \log\left(\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)}\right) + \sum_{k=1}^{K}(\alpha_k - 1)\left(\Psi\left(\boldsymbol{\gamma}_{i,k}\right) - \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_{i,l}\right)\right) + \sum_{k=1}^{K} \log\left(\frac{\Gamma\left(\sum_{v=1}^{V}\beta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\beta_v\right)}\right) + \sum_{v=1}^{V}(\beta_v - 1)\left(\Psi(\boldsymbol{\lambda}_{k,v}) - \Psi\left(\sum_{u=1}^{V}\boldsymbol{\lambda}_{k,u}\right)\right). \tag{A.174}$$

corresponds to the components of L that involve $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$.

We start of by examining $\boldsymbol{\alpha}$. As when finding the optimal values for the variational parameters we let

$$L_{\boldsymbol{\alpha}} = \sum_{i=1}^{M}\left(\log\left(\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)}\right) + \sum_{k=1}^{K}(\alpha_k - 1)\left(\Psi\left(\boldsymbol{\gamma}_{i,k}\right) - \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_{i,l}\right)\right)\right). \tag{A.175}$$

Derivating the above w.r.t. α_k yields

$$\frac{\partial L_{\boldsymbol{\alpha}}}{\partial \alpha_k} = M\left(\Psi\left(\sum_{l=1}^{K}\alpha_l\right) - \Psi\left(\alpha_k\right)\right) + \sum_{i=1}^{M}\left(\Psi\left(\boldsymbol{\gamma}_{i,k}\right) - \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_{i,l}\right)\right). \tag{A.176}$$

We do however note that the above depends on α_l where $l \neq k$, meaning that have to use an iterative approach to finding the optimal values for $\boldsymbol{\alpha}$ [3]. In *Blei et al (2003)* the proposed method is to use Newton-Raphson with the following elements in the hessian.

$$\frac{\partial L_{\boldsymbol{\alpha}}}{\partial \alpha_k \alpha_h} = M\left(\delta_{k,h}\Psi'\left(\alpha_k\right) - \Psi'\left(\sum_{l=1}^{K}\alpha_l\right)\right). \tag{A.177}$$

Where $\delta_{k,h}$ is the Kronecker delta. In an almost identical manner we can examine $\boldsymbol{\beta}$, where $L_{\boldsymbol{\beta}}$ is once again identical in the two models, to obtain

$$\frac{\partial L_{\boldsymbol{\beta}}}{\partial \beta_v} = K\left(\Psi\left(\sum_{u=1}^{V}\beta_u\right) - \Psi\left(\beta_v\right)\right) + \sum_{k=1}^{K}\left(\Psi\left(\boldsymbol{\lambda}_{k,v}\right) - \Psi\left(\sum_{u=1}^{V}\boldsymbol{\lambda}_{k,u}\right)\right) \tag{A.178}$$

as well as

$$\frac{\partial L_{\boldsymbol{\beta}}}{\partial \beta_v \beta_s} = K\left(\delta_{v,s}\Psi'\left(\beta_v\right) - \Psi'\left(\sum_{u=1}^{V}\beta_u\right)\right). \tag{A.179}$$

Under the assumption of a symmetric priors, i.e. $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K) = (\alpha, ..., \alpha)$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_V) = (\beta, ..., \beta)$ we obtain

$$\frac{\partial L_{\alpha}}{\partial \alpha} = MK\left(\Psi(K\alpha) - \Psi(\alpha)\right) + \sum_{i=1}^{M}\sum_{k=1}^{K}\left(\Psi\left(\boldsymbol{\gamma}_{i,k}\right) - \Psi\left(\sum_{l=1}^{K}\boldsymbol{\gamma}_{i,l}\right)\right), \tag{A.180}$$

$$\frac{\partial^2 L_{\alpha}}{\partial \alpha^2} = MK\left(K\Psi'\left(K\alpha\right) - \Psi'\left(\alpha\right)\right), \tag{A.181}$$

$$\frac{\partial L_{\beta}}{\partial \beta} = KV\left(\Psi\left(V\beta\right) - \Psi\left(\beta\right)\right) + \sum_{k=1}^{K}\sum_{v=1}^{V}\left(\Psi\left(\boldsymbol{\lambda}_{k,v}\right) - \Psi\left(\sum_{u=1}^{V}\boldsymbol{\lambda}_{k,u}\right)\right), \tag{A.182}$$

$$\frac{\partial L_{\alpha}}{\partial \alpha} = KV\left(V\Psi'\left(V\beta\right) - \Psi'\left(\beta\right)\right). \tag{A.183}$$

A.6 The Dirichlet Distribution Being in the Exponential Family of Distributions

We can express the density function of a Dirichlet distributed random variable $\mathbf{X} \sim Dir(\underbrace{\alpha_1, ..., \alpha_k}_{=\boldsymbol{\alpha}})$ as below [30].

$$f(\mathbf{x}|\boldsymbol{\alpha}) = \underbrace{\frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k\right)}}_{a(\boldsymbol{\alpha})} \prod_{k=1}^{K} \mathbf{x}_k^{\alpha_k - 1} \tag{A.184}$$

$$= a(\boldsymbol{\alpha}) \underbrace{\prod_{k=1}^{K} \mathbf{x}_k^{-1}}_{t(\boldsymbol{\alpha})} \underbrace{\exp\left\{\sum_{k=1}^{K} \alpha_k \log(\mathbf{x}_k)\right\}}_{\exp\left\{\boldsymbol{\gamma}_i^T t(\mathbf{x})\right\}} \tag{A.185}$$

$$= a(\boldsymbol{\alpha}) h(\mathbf{x}) \exp\left\{\boldsymbol{\alpha}^T t(\mathbf{x})\right\}. \tag{A.186}$$

A.7 Expectation of the Logarithm of a Dirichlet Random Variable

From Appendix A.5 we have that the Dirichlet distribution is of the exponential family of random variables. As such we can use the following property where $\mathbf{X} \sim Dir(\underbrace{\alpha_1, ..., \alpha_k}_{=\boldsymbol{\alpha}})$ [30].

$$E\left[\log\left(\mathbf{X}_k\right)\right] = \frac{\partial}{\partial \alpha_k} \log\left(\frac{1}{a(\boldsymbol{\alpha})}\right) \tag{A.187}$$

$$= \frac{\partial}{\partial \alpha_k} \log\left(\frac{\prod_{k=1}^{K} \Gamma\left(\alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}\right) \tag{A.188}$$

$$= \frac{\partial}{\partial \alpha_k} \sum_{k=1}^{K} \log\left(\Gamma(\alpha_k)\right) - \log\left(\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)\right) \tag{A.189}$$

$$= \Psi\left(\alpha_k\right) - \Psi\left(\sum_{k=1}^{K} \alpha_k\right). \tag{A.190}$$

Above we have that $\Psi(x)$ is the digamma function, i.e. $\frac{\Gamma'(x)}{\Gamma(x)}$ which is the first derivative of $\log(\Gamma(x))$.

A.8 Alternate Definition of LDA

Below is the unsmoothed definition of LDA as introduced in the original paper by Blei et al. [3]. By unsmoothed we mean when $\boldsymbol{\phi}$ is seen as a parameter as is not endowed with a Dirichlet prior. The generative process for an entire corpus $\mathbf{w}$ is described as follows

1. For each document $i = 1, ..., M$
   (a) Choose $N_i \sim \text{Po}(\xi)$
   (b) Choose $\boldsymbol{\theta}_i \sim \text{Dir}(\boldsymbol{\alpha})$
   (c) For each word $j = 1, ..., N_i$
      i. Choose a topic $z_{i,j} \sim \text{Cat}(\boldsymbol{\theta}_i)$
      ii. Choose a word $w_{i,j} \sim \text{Cat}(\boldsymbol{\phi}_{z_{i,j}})$

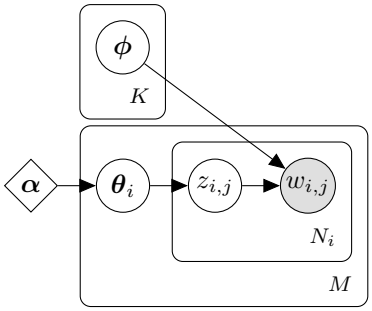


Figure A.1: Plate representation of unsmoothed LDA