University of Glasgow | School of Computing Science

Probabilistic Topic Modelling to Reduce the Noise in Metabolomics

Arijus Pleska

School of Computing Science
Sir Alwyn Williams Building
University of Glasgow
G12 8QQ

Masters project proposal

Date of submission placed here

Contents

1 Introduction **2**

2 Statement of Problem **3**

3 Background Survey **4**

3.1 Terminology 4

3.2 Preliminaries 6

3.3 Latent Dirichlet Allocation 6

3.4 Finding Scientific Topics 9

3.5 Dynamic Topic Models 10

4 Proposed Approach **13**

4.1 Overview 13

4.2 Approach 13

4.3 Risk Management 13

5 Work Plan **13**

5.1 Schedule 13

5.2 Deliverables 13

1 Introduction

This research proposal suggests an application of machine learning in biology. We intend to study the scans of low level biological entities in order to derive their patterns. To be more specific, we will be tackling a problem of noise. This issue of noisiness arise from the hardware limitation of the high precision data collection devices.

The application domain of this project is metabolomics – a branch of computational biology. In brief, metabolomics is the scientific study of chemical processes which involve small molecules – metamobilites. Ultimately, these chemical processes can be predicted by the features of metabolites. We intend to set the basis of our study on data sets acquired by *mass spectrometry* (MS). By using this technique, we obtain the masses and intensities of metabolites within a sample of interest. This data representation is known as mass spectrum. Note that a different mass indicates different types of metabolites, whereas intensity suggests the level of concentration. Thereby, a mass spectrum is used to distinguish and visualise the observed metabolites. By deriving the formulae of specific metabolites types, we can induce patterns leading to the prediction of chemical processes.

In the terms of machine learning applicability, we are looking into a problem of vision. The metabolites of some biological tissue are sequentially captured within regions of fixed size. Effectively, this process is generating an image: each scan is a pixel, therefore, the collections of scans produce an image. However, each pixel contains a large amount of metabolites, so the visualisation of these in a single image would be rather complex. Thereby, we can choose a specific metabolite type and produce pixels reflecting such metabolite activity. For a rigorous discussion on machine learning applications in metabolomics, the reader can consult the recent survey conducted by Alonso et al. [1]. Going back to the visualisation of metabolites, we can choose a specific combination (structure) and produce an image where activated pixels would suggest the concentration of the chosen metabolite structure. The problem arises upon discovering valid matabolite structures. Therefore, the techniques of machine learning, or more specifically topic modelling, are used to induce the prospective structures.

The objective to discover novel structures is treated as a problem of *unsupervised* machine learning. That is, we have no initial settings in how these structures are expected to look like: the structures will be inferred by machine learning methodology. The inference of unknown structures is studied within *topic modelling*. Further, note that we are concentrating on statistical topic modelling. In other words, we are putting the emphasis on the *probabilistic* models: the structures are inferred from the data instance distributions. To be more specific, we are focusing to enhance a particular probabilistic machine learning model – *latent Dirichlet allocation* (LDA). First of all, the successor LDA models are treated as a state-of-the-art method in tackling semantic-analysis type problems. For more detailed review on the LDA applications, the reader can consult the survey [2] conducted by Blei. Further, Hooft et al. [7] has recently addressed the applicability of LDA in metabolomics. Note that this proposal is directly influenced by the latter prospect: we will attempt to utilise the LDA-like models to reduce the impact of noise in the data sets of the metabolomics domain.

The structure of this proposal is given as follows: in Section 2 we present the issue of noise in Metabolomics; in section 3 we cover the literature in topic modeling which focuses on the variations of Latent Dirichlet Allocation; in Section 4 we provide the metric of success of the proposed research; finally, in Section 5, we go into details on how the project will be executed.

2 Statement of Problem

Currently, MS equipment is incapable to fully represent the metabolite construction. As discussed by Palmer et al. [6], even though the current machinery differentiates metabolites in millidalton precision, we still obtain a significant amount of false positives. Thereby, rather than relying upon the advances of MS equipment, we can investigate the prospect of enhancing the data processing techniques. As described in Section 1, LDA derivatives are promising methods in discovering the hidden structures in metabolomics data sets. For this reason, we raise a hypothesis that the degeneracies of metabolomics data sets can be detected using LDA-like models. Note that currently there is no set approach on how to configure LDA-like models to reduce the impact of noise in metabolomics data sets.

The primary goal of this research is to expand the consensus on whether topic modelling is a valid approach to smooth metabolomics data sets. To be more specific, we set our objectives to the following:

1. Utilise general topic modelling methods in metabolomics;
2. Reduce the impact of noise in metabolomics data sets;
3. Give a basis to a general topic modelling method for error correction.

Briefly, the first objective is expected to utilise the general state-of-the-art topic modelling methodology to distinguish the techniques displaying better performance in metobolomics applications. In other words, we will conduct a survey on LDA-like topic modelling methodology. The second objective is focusing on enhancing the general models for the particular task of error correction in metabolomics. Finally, the third objective can be treated as a general contribution to the field of machine learning, since the discovered techniques might be applicable to other domains.

We are setting a hypothesis that LDA-like topic modelling can be successfully utilised to correct errors in metabolomics data sets. Both cases of proving and disproving this hypothesis would contribute to metabolomics. The successful outcome would lead to the potential applications discussed in the following paragraph, whereas the unsuccessful outcome would set-up additional requirements to tackle the issue of noisiness in metabolomics data sets.

The potential applications of this research can be directly induced from the previously listed objectives. First of all, we would show that the general models or the models applied in other domains can be utilised in metabolomics. Further, the developed error correction methodology would improve the accuracy of the chemical processes prediction. Finally, the generalisation of the developed error correction model might suggest different approaches in tackling error correction in general; also, it would impact the awareness in metabolomics and the field's potential in developing machine learning generalisations.

3 Background Survey

The provided literature review on topic modelling is directed towards familiarising with LDA derivatives. Also, the review assesses the methodology of LDA-like models and proposes several implementation variations. The review is structured in progressing order: at the start, we look into the original LDA paper and define the terminology (it will be used throughout the proposal); then, we familiarise with two different techniques used for the inference of LDA-like models; finally, we review LDA derivatives which shows the potential to be utilised in designing the error correction technique.

3.1 Terminology

Recall that in topic modelling induces a hidden structure within the studied data sets. We will provide an example providing an intuition in this process. Say, if we were analysing a book, its chapters might address different topics or address the topics at a different extent. Further, the chapters itself would have unique distributions of the topics. Note that we are assuming that we do not know what the topics addressed in the book. In order to derive these hidden topics, a probabilistic topic modelling method would analyse the structure of the words used in the book. In other words, the model would induce patterns which would refer to a particular topic. Regarding the method's performance, it would depend on the method's ability to optimise the inference of the patterns. For example, commonly used words (such as and, or, the, etc.) would be relevant in all topics. Therefore, their semantic impact in defining the topics would be negligible. For this reason, we could dismiss these to improve the time performance of the method.

The terminology used in this proposal is equivalent to the terminology used by Blei [2]. Figure 1 below represents the original LDA model in plate notation. We will use it as a guide to familiarise with the terminology.

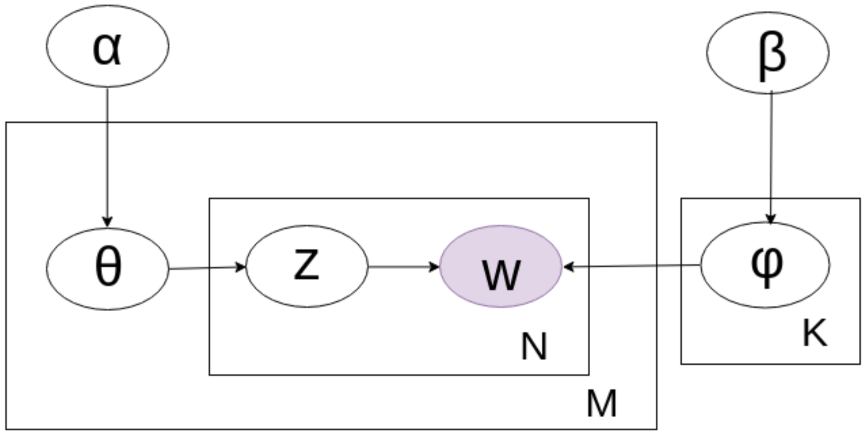

Figure 1: The design of the original LDA model

To start with, we will familiarise with the higher level entities of the model. The circles represent an entity, whereas the plates indicate the number of entities. For example, there would $N \times M$ entities of z and M entities of θ. Further, LDA is a three-level hierarchical structure: the smallest entity (grey circle) is defined as a *word*; then a collection of words (inner plate) is defined as a *document*; and a collection of documents (outer plate) is a *corpus*. Note that Figure 1 represents M documents and N words per document, it follows that in the corpus there would be $M \times N$ words. Also, as suggested by Figure 1, let's denote a word by w. Then, let's denote the sequence of words in document d by

$$\mathbf{w}_d = \{w_1, w_2, \ldots, w_N\},$$

then the sequence of words in the corpus would be denoted by

$$\mathbf{D} = \{\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_M}\}.$$

Now, we will introduce the lower level entities. The grey circle indicates that the *observable* variables, whereas the white coloured circles indicate the *hidden* variables. Note that in Figure 1 the words are the only observable variables; all other entities are hidden, i.e., they describe the underlying structure of the model. Further, recall that topic modelling is applied in order to discover a mixture of topics; or in other words, topic modelling could describe a corpus and its documents in terms of topic distributions. This intuition will simplify the definitions of the remaining entities in Figure 1, these are given as follows:

- K is the number of topics;
- V is the size of the vocabulary, i.e. the number of unique words in the corpus;
- α is the parameter referring to the prior distribution of the topics over documents;
- β is the parameter referring to the prior distribution of the vocabulary over topics;
- θ is the latent variable referring to the topic distribution over documents;

- ϕ is the latent variable referring to the vocabulary over topics
- w is the observable variable referring to the words in documents;
- and z is the latent variable of topic assignments over the words.

Further, note that the listed entities can be perceived as matrices of the following dimensions: α is $1 \times K$; β is $K \times V$; θ is $M \times K$; ϕ is $K \times V$; w is $M \times N$; and z is $M \times V$.

3.2 Preliminaries

The relationship the between observable and hidden variables, effectively, describes the inference of the underlying topic structure. We will go through an arbitrary example in order to give an intuition in the inference process and familiarise with the lower level notation of the previous listed variables. Figure 1 shows see that α_k (the prior for the kth topic in a document) influences $\theta_{:,k}$ (the latent variable describing the distribution of topic k over all documents). It follows that $\theta_{d,k}$ (the probability of the kth topic occurrence in document d) influences $z_{d,:}$ (the topic assignment over all vocabulary terms for document d). Further, $z_{d,v}$ refers to the topic assignment of vocabulary term v. By taking $z_{d,v}$ and $\beta_{v,k}$ (the prior of topic k on vocabulary term v), we obtain word $w_{d,v}$. In the complete perspective, the generative process of the corpus is given by the following joint probability distribution

$$p(\beta, \theta, z, w) = \prod_{k=1}^{K} p(\beta_{:,k}) \prod_{d=1}^{D} p(\theta_{d,:}) \Bigg(\prod_{v=1}^{V} p(z_{d,v}|\theta_{d,:}) p(w_{d,v}|\beta, z_{d,v}) \Bigg). \quad (1)$$

By using Bayes Theorem, we derive the computation of the conditional distribution

$$p(\beta, \theta, z|w) = \frac{p(\beta, \theta, z, w)}{p(w)}. \quad (2)$$

This equation expresses the computation of the posterior – the derivation of the hidden corpus structure given observable variable w. The problem arises from inefficient analytical computation of evidence $p(w)$. The approximation of it will be considered during the literature review on LDA-like models. The naming of the following subsections corresponds to the reviewed papers.

3.3 Latent Dirichlet Allocation

Latent Dirichlet allocation is a probabilistic topic modelling technique proposed by Blei et al. [4]. Since this paper is the origin of LDA-like models, it will be discussed in more detail to introduce the basis of LDA-like methodology. As a result, this coverage will simplify the review on the LDA derivatives discussed in the upcoming subsections.

The original LDA model sets some assumptions for the data processed by the model. First of all, it is assumed that documents and words are *exchangeable*. That is, the order in which these entities are processed does not matter. In other words, we can say the model follows the *bag-of-words* principle. The next assumption states that we should use discrete data. Effectively, the basic LDA does not cover continuous features. Another assumption is on the number of topics: it is set to be fixed throughout the complete run of inference.

As discussed in the preliminaries subsection, we are also taking an assumption that the corpus is induced by a probabilistic process. The generation of the corpus has three levels: the parameters α and β are sampled once; the variable θ is sampled once per document; and the variables z and w are sampled once per word. This hierarchical process relates to the generative algorithm given in the original LDA paper, it is equivalent to Algorithm 1 below.

Algorithm 1 Document Generation

1: $N \sim \text{Poisson}(\xi)$
2: $\theta \sim \text{Dir}(\alpha)$
3: **for** $n \leftarrow 1, N$ **do**
4: $\quad z_n \sim \text{Multinomial}(\theta)$
5: $\quad w_n \sim \text{p}(w_n | z_n, \beta)$
6: **end for**

Note that the algorithm describes the generation of a single document. In order to generate the whole corpus, we would run the algorithm for each document. Going back to Algorithm 1, a brief intuition behind it is given as follows: in line 1 we draw the number of words N in a document (ξ is an ancillary variable used as the mean for the Poisson Distribution); in line 2 we draw topic distribution θ from the Dirichlet distribution on parameter α; further, in lines 3–6 we generate the words: in line 4 we draw topic z_n from the multinomial distribution on θ and, finally, in line 5 we obtain word w_n from the conditional probability on z_n and β, recall that β is the prior on the vocabulary terms correspondence to topics.

As suggested in the ending paragraph of the preliminaries subsection, the inference relates to the computation of evidence $p(w)$ (it is the denominator term given in Equation 2). Even though the original paper mentions the inference methods such as Markov chain Monte Carlo, it suggests using the method of variational inference. In order to develop an intuition over the method, the authors introduce the variational parameters γ and ϕ. It is also mentioned that the computation of γ and ϕ is an optimisation problem. That is, we are assessing the minimum bound over the latter parameters. Further, recall that in Figure 1 we explicitly express the parameter ϕ in describing the basic LDA model. However, the authors have chosen to show it separately – upon describing the variational method. Their expression is given in Figure 2 below.

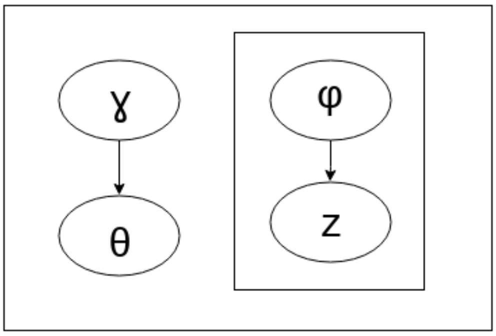

Figure 2: Variational parameters

Note that the authors provide pseudo code to build an intuition in tackling this optimisation problem.

Additionally, the authors emphasise the estimation of the parameters α and β. They use a two-step EM procedure which involves the previously introduced variational parameters γ and ϕ. E-step is the solution of the variational parameters optimisation problem; and M-step is an update of the parameters α and β. Note that the authors provide an analytical approach to estimate β which is proportional to the changes of ϕ; and for α, they provide an approach utilising the Newton–Raphson method. Apart from that, the authors introduce the an issue of *sparsity* in a corpus structure. A sparse corpus is expected upon using a large vocabulary or a large amount of documents. This issue is addressed by introducing smoothing: an additional parameter η is utilised to smooth the estimation of β.

The conducted experiments display an improvement in performance over the predecessor models and suggest the application domains of LDA-like models. The first experiment addresses document modelling. LDA has displayed balanced hidden topic proportions in the set of tested documents. That is, the tested documents has replicated the topic proportions of the training documents. The second experiment is directed towards the problem of document classification. They have addressed the *dimensionality reduction* in order to surpass the performance of the support-vector-machine-like models. To be more specific, the documents with reduced dimensionality would possess a lower amount of features. The results of the experiment indicate that the basic LDA model has successfully reduced the number of features. Also, it has displayed an improvement in accuracy compared to the support-vector machine-like model.

The review on the original LDA model has set up a basis required to understand the model's derivatives. That is, the introduced methodology will be useful in considering the following improvements: (1) the relaxation of the exchangeability assumption; (2) the adaptation of time-series suggesting word correspondence to topic over time; (3) the discussion over alternative inference methods. Apart from that, the experiments introduced in this section has reflected the LDA performance upon the model's introduction to the public. The current state-of-the-hard topic modelling techniques overcome the shown results. The experiments were addressed in order to suggest the application fields of LDA-like models.

3.4 Finding Scientific Topics

The article titled as 'Finding Scientific Topics' [5] suggests an alternative approach to the inference which is suitable for LDA-like models. To be more specific, the article provides a readable introduction on the application of the Markov chain Monte Carlo (MCMC) algorithm. That is, the MCMC algorithm is applied as a sampling-based method to infer an underlying topic structure. Further, the authors discuss the method's application settings for the machine learning problems in vision. Also, the authors conduct an experiment on inferring the structure of topics in the abstracts of papers published in 'Proceedings of the National Academy of Sciences' (PNAS). Finally, the paper is concluded by a discussion on assessing the sampling-based method usability. To be more specific, the performance of the sampled-based method is compared to the performance of variational inference methods.

The MCMC-based inference method is often referred by *Gibbs sampling*. The authors claim that the method provides a first-order approximation upon inferring the hidden structure of the data. That is, the performance is sufficient to establish quantitative reasoning in assessing the correlating sections and semantic structure of the data. Recall that by θ we denote the topic distribution over the documents and by ϕ we denote the vocabulary terms distribution over the topics. In Gibbs sampling, the latter variables are obtained by examining the posterior distribution, whereas topic assignments z are sampled sequentially. That is, we sample topic assignment z and update the distribution depending on the obtained value (this will impact the next sample of z); we repeat this process until the topic distribution converges. To be more specific, each drawn z_i (i denotes the iteration) is recorded as a count. That is, we keep a track of topic assignments for particular words and documents. The value of drawn topic z_i is proportional to the distribution given below:

$$P(z_i = j | z_{-i}, w) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + V\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + K\alpha}. \tag{3}$$

For a brief familiarisation with the notation used in the expression, note the following: n_j^d and n_j^w are the counts of topic j assignments in document d and word w; recall that V is the number of words in the vocabulary and K is the number of topics. Also, note that the counts do not include the last assignment of the topic, this is indicated by $-i$. Ultimately, the left side of the expression is ϕ and the right side is θ. That is,

$$\phi_j = \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta}, \tag{4}$$

$$\theta_j = \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + K\alpha}. \tag{5}$$

This process of the topic assignment is executed for every word in the document and for all documents in the corpus. The procedure is repeated until convergence or, more pragmatically, until the distribution displays the assigned topic fluctuations below the set threshold. In the Gibbs sampling procedure, there is no need to explicitly assign θ and ϕ values. If needed, these values are obtained directly from the topic distribution, as shown by the two previous equations.

The authors has shown a use of Gibbs sampling on a problem of vision. In their experiment settings, it is assumed that a collection of images represents a corpus; an image represent a document; and a pixel of an image is a word. The experiment has shown that the model is able to infer the underlying basis (topics) from which the images could be constructed. The performance of the model was compared to the variational inference based LDA models. Note that the performance was assessed in terms of *perplexity* – the uncertainty in assigning a word assignment to a topic. It has been shown that Gibbs sampling displays superior performance in the small data sets, whereas the performance in larger data tends to fall off. Furthermore, the authors have assessed the performance upon varying the number of topics. The results of this experiment indicate a trade-off in deriving high-level and low-level structures. That is, the low number of topics might result in mashing several structures, whereas the high number of topics is likely to infer irrelevant structures.

The main experiment of the article is conducted on the PNAS data set containing abstracts of the articles published from 1991 to 2001. The authors have noticed that the topic distribution over the vocabulary terms change over time. However, the model had not yet been optimised to take this notion of time-series into account upon the inference. Nevertheless, the authors have emphasised the necessity to assess the dynamic topic modelling.

The review has introduced an alternative method used in the process of inference of LDA-like models. We have familiarised with the intuition in the implementation requirements of the sampling-based inference method. Also, we have reviewed experiments suggesting an approach in tackling problems of vision. Further, we have introduced the relevance of dynamic topic modelling (topic modelling with the notion of time-series).

3.5 Dynamic Topic Models

At this point, we have set up the preliminaries to cover the article on utilising the notion of time-series in LDA-like models. We review the initial article on utilising time-series – 'Dynamic Topic Models' [3]. Note that by a *static* model we mean that the assumed generative process does not take the development/change of the vocabulary terms into account, whereas in *dynamic* topic modelling we assume the notion of that the documents were generated in a sequential process. That is, in dynamic topic modelling, the topics evolve in terms of their distribution over the vocabulary terms. Upon reviewing the dynamic topic modelling article, we emphasise the following aspects: the fundamental dynamic topic modelling assumptions and their differences compared to the basic LDA model; the generative process; the suggested inference methods; and the results displayed by the experiments on the initial dynamic topic model.

The assumptions set for the dynamic topic model emphasises its strengths over the basic LDA model. Most importantly, the dynamic topic model relaxes the assumption of the document exchangeability. Since the documents are generated over time, we assign them to specific time-slices. The combination of these time-slices would represent the corpus. Note that the documents in each time-slice are exchangeable. This treatment of time set-ups a basis to tackle the problems which involve the sequential development of the data.

On the other hand, note that another assumption of discrete data remains the same as in the basic LDA model. Even though we are inducing time series (which would suggest the use of *continuous* variables), the authors use only *categorical* (discrete) data. The distinction between continuous and categorical data is the following: continuous data is strictly numeric and infinite, whereas categorical data takes a fixed value from a finite set of values.

In the dynamic topic modelling, the generative process updates the parameters α and β for each new time-slice. This means that the topic distribution over documents and the vocabulary terms distribution over topics change over time. The visualisation of this process is given in Figure 3 below.

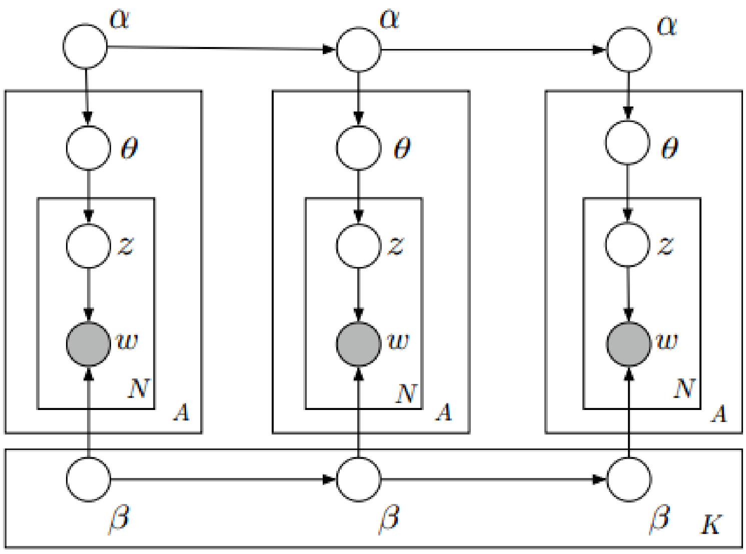

Figure 3: Dynamic topic model. COPYRIGHT.

Effectively, Figure 3 is a sequential combination of the basic LDA model illustrated in Figure 1. In Figure 3, one basic LDA model represents one time-slice. Also, the parameters α and β are induced by their predecessors. More formal representation of the generative process is given in Algorithm 2 below.

Algorithm 2 Dynamic document generation.

$\beta_t|\beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \sigma^2 I)$
$\alpha_t|\alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$
for $d \leftarrow 1, M$ **do**
    $\eta_d \sim \mathcal{N}(\alpha_t, a^2 I)$
    for $n \leftarrow 1, N$ **do**
        $z_{d,n} \sim \text{Multinomial}(\pi(\eta_d))$
        $w_{t,d,n} \sim \text{Multinomial}(\pi(\beta_t|z_{d,n}))$
    end for
end for

Algorithm 2 describe the generative process of the documents in time-slice t. The algorithm is initialised by drawing the parameters α_t and β_t from the Gaussian distribution with the means given by the previous parameter values. Note that the parameters α_0 and β_0 are the the special case. They can be initialised randomly, i.e. the mean of the Gaussian distribution is random. However, for the parameters α_0 and β_0 we take larger variance values compared to α_t and β_t. The larger variance values induce the initial time-slice with larger fluctuations in the topic and vocabulary term distributions; and the smaller variance values in the following time-slices induce the smoothness. Further, in lines 3–9, for every document in time-slice t we execute the following procedure: for document d, we draw the topic distribution η_d (the meaning of the parameter is equivalent to the previously introduced θ_d); then, in lines 5–7, for every word in document d we draw topic assignment $z_{d,n}$ and, finally, we draw word $w_{t,d,n}$. Note that the means for the Multinomial distributions are normalised by the use of *parametrisation*. Parametrisation mapping π is expressed as

$$\pi(x)_n = \frac{\exp{(x_n)}}{\sum_{i=1}^{N} \exp{(x_i)}}. \tag{6}$$

The authors have used variational methods to approximate the inference of the posterior. It is claimed that stochastic simulation (sampling-based variance methods) would be unable to scale with larger data sets. Further, the authors discuss two methods used for the variational inference: Kalman Filtering and Wavelet Regression. The implementation details of these methods can be found in the 'Dynamic Topic Modelling' article [3]. Note that the implementation details are provided in a brief manner suggesting the use of additional literature, these are also provided in the article.

The dynamic topic model is evaluated by conducting an experiment on the journals of 'Science'. The experiment is expected to infer the prevalent science fields and the prevalent vocabulary used in these fields. Note that the relevance of the science fields can be respectively obtained from the latent variables θ and ϕ. That is, the authors would compare these variables throughout all time-slices in order to induce the reasoning over the prevalent science fields (by using θ) and the prevalent vocabulary terms (by using ϕ). The experiment has successfully established this notion of time-series. This is proved by comparing the inferred rises and falls of the scientific fields to the history of science. Note that the experiment has been performed both Kalman Filtering and Wavelet Regression methods.

By reviewing the original dynamic topic model, we have familiarised with the generative process which takes time-series into account. We have provided the visualisation of the model and introduced the algorithm for the generative process. Further, we have outlined the inference methods suggests be the authors. Finally, we have familiarised with the experiment settings which were used to infer time series in the articles of 'Science'.

4 Proposed Approach

4.1 Overview

4.2 Approach

4.3 Risk Management

5 Work Plan

5.1 Schedule

5.2 Deliverables

References

[1] Arnald Alonso, Sara Marsal, and Antonio Julià. Analytical methods in untargeted metabolomics: state of the art in 2015. *Frontiers in bioengineering and biotechnology*, 3:23, 2015.

[2] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[3] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[5] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[6] Andrew Palmer, Prasad Phapale, Ilya Chernyavsky, Regis Lavigne, Dominik Fay, Artem Tarasov, Vitaly Kovalev, Jens Fuchser, Sergey Nikolenko, Charles Pineau, et al. Fdr-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nature Methods*, 2016.

[7] Justin Johan Jozias van der Hooft, Joe Wandy, Michael P Barrett, Karl EV Burgess, and Simon Rogers. Topic modeling for untargeted substructure exploration in metabolomics. *Proceedings of the National Academy of Sciences*, page 201608041, 2016.