University of Glasgow | School of Computing Science

Probabilistic Topic Modelling for Reducing the Impact of 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

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 aim to tackle the issue of noise, which arises from the hardware limitation of the data collection devices.

The branch of biology we are expecting to carry out the research is metabolomics. In brief, metabolomics is the scientific study of chemical processes which involve small molecules – metamobilites. These chemical processes can be predicted by the features of metabolites; and metabolites are these low level biological entities. By using mass spectrometers, we can obtain the masses and intensities of metabolites. Note that the mass suggests different types of metabolites, whereas the intensity displays the level of concentration. Thereby, a mass spectrum can be used to distinguish and visualise the observed metabolites; and combinations of specific metabolites types may suggest a specific pattern for the chemical process leading to a particular formation.

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. For this reason, we can choose a specific metabolite type and produce pixels reflecting such metabolite activity. Further, we can choose some specific combination (pattern) of metabolites and produce an image of such pattern concentrations. The problem arises upon finding such valid patterns of metabolites, therefore, we are applying machine learning to find the perspective patterns for us.

The objective to discover new patterns makes it a problem of unsupervised machine learning. That is, we have no prior settings in how these patterns would look like thus we are utilising machine learning in order to structure the data. The problems of this flavour can be tackled using topic modelling. This branch of unsupervised machine learning is a statistical process. In other words, we are looking into probabilistic models – the hidden structure of the data is inferred by the frequencies of its instances. To be more specific, we will be focusing on enhancing a particular probabilistic machine learning model – Latent Dirichlet Allocation (LDA). This model dictates the state-of-the-art of topic modeling and is widely applied in the current research on metabolomics [insert citation]. Since the process of the data collection in metabolomics is prone to noise, we will carry a research on whether Latent Dirichlet Allocation could be applied as a method of error correction.

The structure of this proposal is set-up 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

The current mass spectrometry (MS) equipment display a significant accuracy error upon reflecting the exact metabolite construction in a given sample. Rather than relying upon the advances of such equipment, we can investigate the prospect of enhancing the data processing techniques. As described in Section 1, topic modelling is capable to distinguish the hidden structures of metabolites. However, at this point of time, there is no well set generative approach on how to configure Latent Dirichlet Allocation to reduce the impact of noise in a given dataset.

The primary goal of this research is to expand the consensus on whether topic modelling is a valid of approach to smooth datasets. To be more specific, the objectives we will be looking into are listed as follows:

1. Adapt and utilise the generative Latent Dirichlet Allocation models in metabolomics;
2. Reduce the impact of noise in the datasets of metabolomics;
3. Give a basis to the generative Latent Dirichlet Allocation model in error correction.

The first objective is expected to investigate the perspective applications of topic modelling which are successfully utilised in other domains. Effectively, we want to assess the general models (or models displaying significant results in other fields). The second objective would adapt, and thus enhance, these general models for the particular use of error correction in metabolomics. Finally, the third objective can be treated as a general contribution to the field of machine learning.

We are setting a hypothesis that topic modelling, or more specifically Latent Dirichlet Allocation, can be utilised to correct the errors in the datasets of metabolomics. Both cases of proving and disproving this hypothesis would be a contribution in metabolomics. The successful outcome would lead to the potential applications provided in the following paragraph, whereas the unsuccessful outcome would set-up additional requirements to tackle this issue.

The potential applications of this research can be induced directly induced from the set-up objectives. First of all, we would show that models from other domains can be applied in metabolomics and, thus, generalised. Further, the developed error correction methodology for the datasets of Metabolomics would improve the accuracy of the chemical processes prediction. Finally, the generalisation of the developed error correction methodologies would provide an alternative tool-kit for other domains as well as expand the potential applications of the results derived from metabolomics.

3 Background Survey

The provided literature review on latent Dirichlet allocation assesses its underlying methodology and proposes several different implementation variations. The review is structured in progressing order: to start with, we look into the original latent Dirichlet allocation paper and define the terminology which will be used throughout this proposal; then, we familiarise with two different approaches used for the inference of latent Dirichlet allocation; finally, we review the perspective advances which could be used for error correction.

3.1 Terminology

Recall that in topic modelling we are trying to induce some hidden structure within the studied data. For instance, if we were analysing a book, it would be likely to address several different topics. Further, the chapters of the book itself would have unique concentrations of such topics. In order to cluster these topics, a probabilistic topic modelling method would look into the structure of the words used in the book. Also, the performance of the method would depend on its ability to optimise this clustering, for example, we could dismiss the commonly used words such as: and, or, the, etcetera. Note that in we will assess only the topic modelling where we do not know the naming of topics. That is, we will be looking into a *unsupervised learning* problem.

The terminology used in this proposal is equivalent to the one used in the survey on probabilistic topic models carried by Blei et al. [citation]. Figure 1 below represents the latent Dirichlet allocation model in plate notation. We will use it as a guide to familiarise with the terminology.

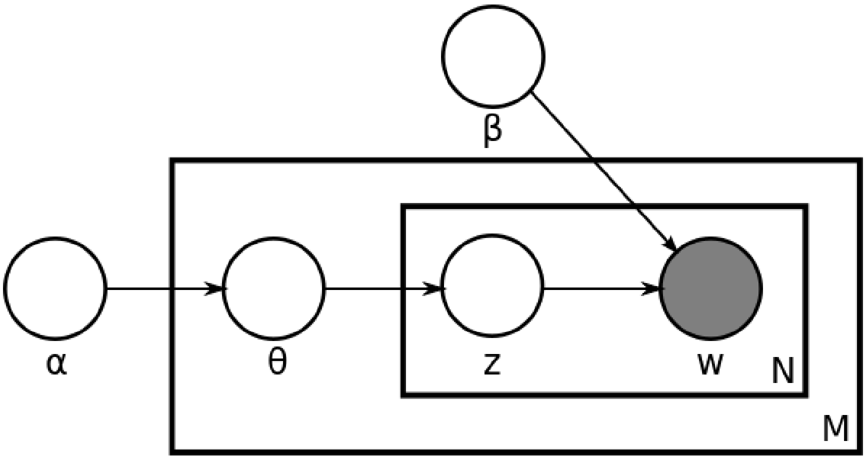

Figure 1: Latent Dirichlet allocation. COPYRIGHT. CHANGE M TO D

To start with, we will familiarise with the domain of the model in a high level. The grey circle represents a single word, whereas a collection of words (document) is represented by the inner plate, and the outer plate corresponds to a collection of documents (corpus). In the model, words are the only *observable variables* – this is indicated by the grey colour of the circle, whereas the white coloured circles indicate the *hidden variables.* As suggested by the figure, a *word* is denoted by w; a *document* (sequence of words) is denoted by $\mathbf{w} = \{w_1, w_2, \ldots, w_N\}$, where N indicates the number of words in the document; and a *corpus* is denoted by $\mathbf{D} = \{\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_M}\}$, where D is the number of documents in the corpus.

Now, we will introduce the lower level entities – the hidden variables. Recall that latent Dirichlet allocation attempts to induce an arbitrary mixture of topics. The corpus and each document is expected to vary in the proportions of the induced topics; and each document is likely to have a different number of word. Let's denote the number of such arbitrary topics by K, and the number of unique words in the corpus by V. This notation will simplify the explanation on the remaining entities in Figure 1. To start with, β is the hyper-parameter (a matrix of dimension $V \times K$) referring to the prior distribution of the topics over the vocabulary. The hyper-parameter α (a vector of length K) refers to the prior distribution of the topics over the documents. Further, the latent variable θ (a matrix of dimension $D \times K$) is the topic distribution over documents. Finally, z is the latent variable on the topic assignments of each word in every document, effectively, it is a matrix of dimension $M \times V$.

3.2 Preliminaries

The relationship 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 intuition in the inference process and provide the notation of the variable instances. By looking into Figure 1, we can see that α_k (the prior weight of the kth topic in a document) influences $\theta_{:,k}$ (the latent kth topic distribution over M documents). It follows that $\theta_{d,k}$ (the probability of the kth topic occurrence in the document d) influences $z_{d,:}$ (the topic assignment over V vocabulary terms in document d). Note that $z_{d,v}$ would refer to the assignment of some specific vocabulary term v. Now, by taking $z_{d,v}$ and $\beta_{v,k}$ (the prior weight of the kth topic on the vocabulary v) we describe the origins of word $w_{d,v}$. The relationship of the terms, effectively, induces the generation of the corpus, this can be expressed as the following joint probability distribution

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

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

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

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 $p(w)$. By reviewing the literature on latent Dirichlet allocation [CITATION], we will also familiarise with the numerical methods used in approximating $p(w)$.

3.3 Latent Dirichlet allocation

Latent Dirichlet allocation is a probabilistic topic modelling technique proposed by Blei et al [citation]. Since this subsection is dedicated for the origins of the technique, it will be broader in detail. This will cover aspects addressed in the following subsections covering the advances on latent Dirichlet Allocation.

The initial latent Dirichlet allocation model sets some assumptions for the data which will be processed by the model. First of all, it is assumed that both documents and words in the documents are *exchangeable*. That is, the order in which these will be processed does not matter. More formally, this assumption means that the model follows the *bag-of-words* principle. The next assumption is that the model works on discrete data. Effectively, the original paper does not cover the treatment of continuous features. Another assumption is on the dimensionality of the topic variables: the number of topics is set to be fixed during the complete run of inference.

As discussed in the preliminaries subsection, the process of latent Dirichlet allocation assumes that the corpus can be generated by a probabilistic process. The generation of the corpus is made of 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. The generative algorithm given in the original paper 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:     $z_n \sim \text{Multinomial}(\theta)$
5:     $w_n \sim \text{p}(w_n|z_n, \beta)$
6: end for
```

Note that the algorithm describes the generation of a single document. In order to generate the corpus, we would run the algorithm several times. To start with, in the line 1 we draw the number of words in the document N, where ξ is an ancillary variable suggesting the mean for the Poisson Distribution. In the line 2 we draw the topic distribution θ from the Dirichlet distribution on α. Now in the lines 3–6 we generate the words: in the line 4 we draw topic z_n from the multinomial distribution on θ; and, finally, in the line 5 we obtain the word w_n from the conditional probability on z_n and the distribution over the words in the vocabulary β.

The inference is, effectively, the computation of the evidence for the posterior distribution.

The original latent Dirichlet allocation paper uses the method of variational inference. Basically, the authors introduce the variational parameters γ and ϕ (computation of these set-ups a problem of optimisation). Note that the original paper provides pseudo code for solving the optimisation problem.

Additionally, the authors emphasise the estimations of the parameters α and β. They suggest a two-step EM procedure utilising the variational parameters γ and ϕ: E-step is the optimisation problem addressed in the previous paragraph; and M-step updates the parameters α and β. Note that the authors provide an analytical update for β and a Newton–Raphson method for updating α. Apart from that, the authors also introduce the issue of a sparse corpus. Such corpus would be expected upon having a large vocabulary or a significant amount of documents. The authors suggest addressing this problem by introducing smoothing. They introduce an additional parameter η which impacts the estimation of β.

The conducted experiments display the improvement performance over the predecessor models and suggest the application domain of latent Dirichlet allocation. The first experiment addresses document modelling. The latent Dirichlet allocation model is reported to display balanced hidden topic proportions in the set of test documents. That is, the test set has replicated the proportions of the training set. The second experiment addresses document classification. In order to display better performance compared to a support vector machine (SVM) model, the authors address the dimensionality reduction of the tested documents. This means that the documents would possess a lower amount of features. It follows that the latent Dirichlet allocation model has successfully reduced the number of features and displayed an improvement in accuracy compared to the support vector machine model.

The review on the initial latent Dirichlet allocation model has set-up a basis for introducing the successor variations. These will be provided in the following sections. We have introduced the initial methodology for the following reasons: (1) the novel models will display the relaxation of the initial assumptions; (2) the provided document generation methodology will be compared to the document generation in batches; (3) we will cover different inference and parameter estimation approaches. Also, note that the paper is relatively old in time, and the experiments do not display the state-of-the-art performance. Rather than that, they were introduced to address the application domains of the model.

3.4 Finding scientific topics

The paper on "Finding scientific topics" [CITATION] suggests an alternative approach to the inference in latent Dirichlet allocation. To be more specific, the paper provides a readable introduction on how to apply a Markov chain Monte Carlo (MCMC) algorithm as a sampling-based method for the inference in topic modelling. Further, the authors discuss the method's application settings in vision and conduct an experiment on deriving topics from the abstracts of papers published in "Proceedings of the National Academy of

Sciences" (PNAS). Finally, the authors compare the performance of the suggested sampling-based inference method to variational inference methods.

The inference method using a Markov chain Monte Carlo algorithm can be referred by *Gibbs sampling.* The authors claim that the method provides a first-order approximation in deriving the topics on the given corpus. That is, we can establish quantitative reasoning on the document correlation in terms of content and content's change over time. Recall that by θ we denote the topic distribution over documents and by ϕ we denote the topic distribution over words. In Gibbs sampling, these values can be obtained by examining the posterior distribution. Essentially, the variables are sequentially sampled from their respective distributions. The process draws topic assignment z_i and keeps a track of this assignment with respect to each word and document. Note that z_i is drawn from the following distribution:

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

It follows that this distribution is utilise the previously mentioned topic assignments. For a brief familiarisation with the notation used in the previous expression, note the following: n^d_j and n^w_j are the counts of topic j assignments in document d and word w; W is the number of words; and T is the number of topics. Ultimately, the left side of the expression is ϕ and the right side is θ. That is,

$$\phi_j = \frac{n^{(w_i)}_{-i,j} + \beta}{n^{(\cdot)}_{-i,j}},$$

$$\theta_j = \frac{n^{(d_i)}_{-i,j} + \alpha}{n^{(d_i)}_{-i,\cdot} + T\alpha}.$$

This process of topic assignment would be executed for every word in a document and for all documents in the corpus. The procedure would be repeated until the distributions displayed fluctuations below some set threshold. Notice that the procedure does not require direct assignment of θ and ϕ. If needed, the values of these can be sampled upon converge by using the previously provided representations.

The authors have executed Gibbs sampling in the domain of vision. They have represented an image as a document and each pixel as a word. This means that they were able to recover the underlying basis from which the images could be constructed. The performance was compared to the latent Dirichlet allocations models which use variational inference. Note that we define the performance in terms of perplexity, which indicates the uncertainty of the word assignment to a topic. It appears that, with relatively small data sets, Gibbs sampling displays superior performance. Further, the authors have investigated the performance upon varying the number of topics. The results indicate that there exists a trade-off between the model being able to cover the significant underlying topics (low number of topics) and start deriving irrelevant branches (high number of topics).

The main experiment was carried in the PNAS data set, note that the processed papers were published from 1991 to 2001. By having a varying notion of time, the authors were able to investigate the prevailing topics. Thus, establish a sense in dynamic topic modelling. Even though the model was not adjusted to smooth the time-series, from the derived results we can deduce the varying relevance of underlying science domains.

We have reviewed an alternative method for the inference in latent Dirichlet allocation. The idea behind the inference process was explained by considering the primary aspects: the notion of counts and the implicit use of hidden variables. Also, we have introduced the experiment settings its results. This allowed to familiarise with an example of topic modelling application and vision. Further, the experiment on PNAS has emphasised the relevance of time series

3.5 Dynamic Topic Models

Now we review the paper on Dynamic Topic Models [CITATION]. Effectively, the dynamic topic model can be treated as an extension of static topic models (such as latent Dirichlet allocation). The main idea behind a dynamic topic model is that it induces the topic evolution the processed corpus. That is, the documents in the corpus evolve over time. In this subsection, we review the key assumptions of dynamic topic models, go over the generative process of documents, review the suggested inference methods, and familiarise with the results of the carried experiment.

We will emphasise the assumptions given on the dynamic topic models. First of all, even though we are discussing a model of time series suggesting the use of continuous variables, the data remains categorical. The distinction between continuous and categorical data is that continuous data is strictly numeric and infinite, whereas categorical data can take only particular values. The second assumption relaxes the ex-changeability on documents. That is, the documents are expected to form a sequence. Or in other words, the corpus is divided into time-slices. Note that in such time-slice documents are exchangeable.

The generative process of the models updates the parameters α and β. This means that the topic distributions over documents and words change over time. The visualisation of the process is given in Figure 2 below.

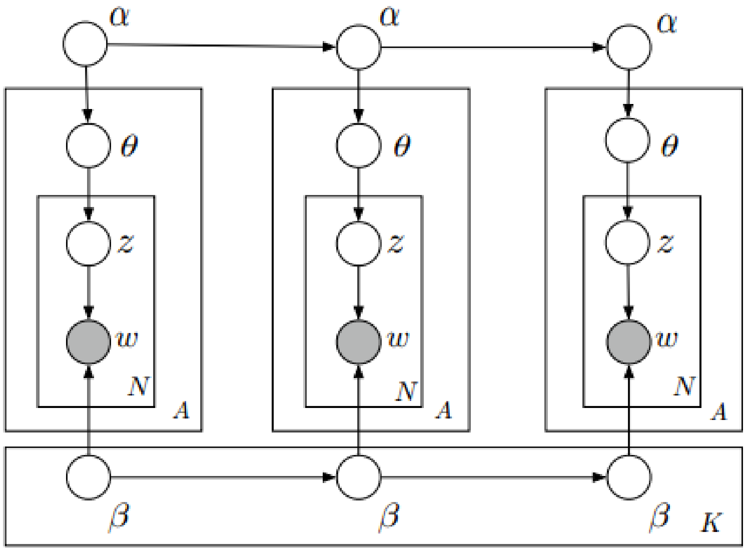

Figure 2: Dynamic topic model. COPYRIGHT.

Notice that each time-slice is equivalent to the model shown in Figure 1. Also, the parameters α and β are induced by their predecessors in the previous time-slice. More formal representation of the generative process is given in Algorithm 2.

Algorithm 2 Dynamic document generation.

```
$\beta_t|\beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \sigma^2)$
$\alpha_t|\alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2)$
for $d \leftarrow 1, D$ do
    $\eta_d \sim \mathcal{N}(\alpha_t, a^2)$
    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
```

The given algorithm describes the generative process for the time-scale t. It starts by drawing the parameters α_t and β_t from the Gaussian distribution with the means given by the previous parameter values. For now, we will not discuss the meaning of the variance values. Further, for each document in the time-scale, we carry the following procedure: we draw the topic distribution in a document η_d from the Gaussian distribution with the mean given by α_t; then, for each word in the document, we draw the topic $z_{d,n}$ from the Multinomial distribution on the parametrised η_d, and, finally, we draw the word $w_{t,d,n}$ from the Multinomial distribution on the parametrised β_t which is conditioned on $z_{d,n}$. By the use of *parametrisation*, we are expressing the result of the Gaussian distributions into a format suitable for Multinomial distribution. This mapping is expressed as

$$\pi(x)_n = \frac{\exp{(x_n)}}{\sum_{i=1}^{n} \exp{(x_i)}}. \tag{4}$$

The authors suggest using variational methods to approximate the inference posterior. It is claimed that stochastic simulation would be unable to scale with large data sets. Two techniques of variational methods are provided: Kalman Filtering and Wavelet Regression. The implementation details of these are provided in a brief manner suggesting the computation of the parameter β_t.

The dynamic topic model is evaluated by conducting an experiment on the journals of *Science*. The experiment is expected to establish the time series on the topic development with the articles. Indeed, the experiment is able to produce results displaying the rises and falls of the topics indicating scientific fields. Note that the experiment is performed using both variational inference methods.

The review on the dynamic topic models has familiarised us with more realistic generative process: the documents are induced by time series. Further, we have provided the visualisation of the model and introduced the algorithm for the generative process. Also, we have captured the techniques used for inference. Finally, we have familiarised with the experiment settings.

4 Proposed Approach

4.1 Overview

4.2 Approach

4.3 Risk Management

5 Work Plan

5.1 Schedule

5.2 Deliverables

References

[1] C. Baier and J.-P. Katoen. *Principles of Model Checking.* MIT Press, 2008.