Probabilistic Topic Modelling for Spatial Smoothing in Mass Spectrometry Imaging

Arijus Pleska

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

Masters project proposal

December 18, 2016

Contents

1 Introduction

This proposal is focused on applying machine learning techniques in order to improve the analysis of biology data sets. We intend to enhance the quality of *mass spectrometry imaging* (MSI) data. Briefly, by the use of MSI, the researchers are able to interpret the formation of chemical processes. However, as suggested by a recent overview by Smith et al. [7], the pre-processing of molecular-level entity data sets is lagging in terms of the current noise reduction algorithm effectiveness. We will attempt to tackle this issue of noise by utilising spatial smoothing.

In this project, the application domain of MSI is set to be *metabolomics* – it is the scientific study of chemical processes which involve metabolites (small molecules) and metabolomes (sets of small molecules). In the research settings of metabolomics, we utilise mass spectrometry (MS) to obtain the distribution of chemical entities within a sample. By utilising MSI, we can visualise the distributions of the ions (or ion sets) of different masses. For instance, Figure 1 below represents a captured sample as an image, where each picture's pixel is a small region within the sample.

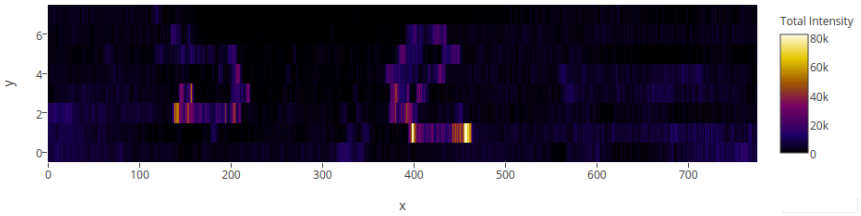

Figure 1: MSI application on a sample with the letters b and e.

Note that each pixel contains a distribution of chemical entities as well. In Figure 1 we have chosen the range of mass to reflect the ink used to inscribe the letters. That is, for each pixel, we have summed the intensities of the chemical entities meeting the mass criteria. Then, each pixel has been assigned an intensity value. Therefore, the sample could be expressed as the latter image. An open problem in metabolomics is to find the sets of chemical entities inducing valid visual patterns. Ultimately, the discovered biological patterns could be used to suggest and measure the likelihood of a particular chemical process occurrence. For a rigorous discussion on machine learning applications in metabolomics, the reader can consult the recent survey conducted by Alonso et al. [1].

The challenges on working with MSI data arise from the data capturing process. In metabolomics, the MS equipment requires the ionisation of metabolites. Effectively, a metabolite is compounded with an appropriate chemical entity to produce an ion. It follows that the mass spectrometers are capable to capture some properties of the ionised metabolites. Since the mass of an ion belongs to these properties, the collection of captured ions can be represented in terms of mass spectra. However, the following issues have to be considered upon reasoning from the MSI data: (1) the ionisation of a metabolite can be performed using different chemical entities, therefore the ions of same metabolite

type would have different masses; (2) a metabolite might be fragmented during the ionisation; (3) the metabolites of same type might have different atom structure, which suggests a different mass value. The latter issues, effectively, corresponds to a problem of error correction.

Since this proposal is focused on utilising spatial smoothing to reduce the impact of noise in MSI data sets, we will familiarise with the machine learning techniques and problems tackled in the MSI domain. The objective to discover novel patterns can be treated as a problem of *unsupervised* machine learning. That is, we have no initial settings in how the patterns are expected to look like. However, upon appropriate optimisation, the unsupervised machine learning techniques are capable to learn the patterns on their own. To be more specific, we will utilise the techniques of *topic modelling* (in this case, a pattern is referred by a topic). Also, we are focusing on the *probabilistic* (statistical) models, more specifically, *latent Dirichlet allocation* (LDA). First of all, the successor LDA-like models are treated as a state-of-the-art method in tackling semantic analysis type problems. For more detailed review on LDA applications, the reader can consult the survey conducted by Blei [2]. Also, Hooft et al. [8] has recently addressed the applicability of LDA in MSI data sets. This proposal is directly influenced by the latter study: we will utilise the LDA-like models to reduce the impact of noise in the MSI data sets.

The structure of this proposal is given as follows: in Section 2 we present the issue of noise in MSI data sets; in section 3 we cover the literature in topic modelling which focuses on the variations of LDA-like models; in Section 4 we discuss the approach to tackle the problem (we discuss the reviewed literature, address design and evaluation details); finally, in Section 5, we set and schedule the tasks required to complete the project.

2 Statement of Problem

Currently, MS equipment is incapable to fully represent the metabolite construction. As discussed by Palmer et al. [6], the current hardware limit in differentiating between the ions in millidalton precision does not mitigate the issue of the false positive data. Thereby, rather than relying upon the advances in 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 patterns in MSI data sets. For this reason, we raise a hypothesis that the degeneracies of MSI data sets can be detected using LDA-like models.

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

1. Optimise the general topic modelling methods for MSI data sets;
2. Successfully apply spatial smoothing in MSI data sets;

3. Give a basis for the development of a general topic modelling method for error correction.

Briefly, for the first objective, we will conduct a survey on LDA-like topic modelling methodology. This survey will distinguish the potential models for the applications in MSI. The second objective focuses on optimising these potential models for the particular task of spatial smoothing in MSI data sets. Finally, the third objective can be treated as a general contribution to the field of machine learning – the discovered spatial smoothing techniques might be applicable to other domains.

The main project hypothesis – the LDA-like models can detect and correct the degeneracies of MSI data sets – can be interpreted as a spatial smoothing application in MSI data sets. Both cases of proving and disproving this hypothesis would contribute to the research in MSI. 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 MSI data sets.

The potential applications of the successful research outcome 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 MSI data analysis. Further, the developed spatial smoothing methodology would improve the accuracy of the chemical processes prediction. Finally, the generalisation of the developed spatial smoothing model might suggest different approaches in tackling error correction in general; finally, the successful research outcome would suggest the applicability prospect of topic modelling in MSI data sets; reversely, it would also impact the prospect to derive machine learning generalisations from the research in MSI data sets.

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 show the potential to be utilised as a technique of error correction.

3.1 Terminology

Recall that topic modelling infers hidden patterns from the processed data sets. For example, if we were analysing a book, its chapters might address different topics or address the topics at different extent. Further, the chapters itself would have unique distributions of

the topics. Note that we are assuming that there is no knowledge about the actual topics covered in the book. In order to derive those 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. The method's performance would depend on its ability to optimise the inference of the patterns. For example, commonly used words (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 those words to improve the time performance of the method.

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

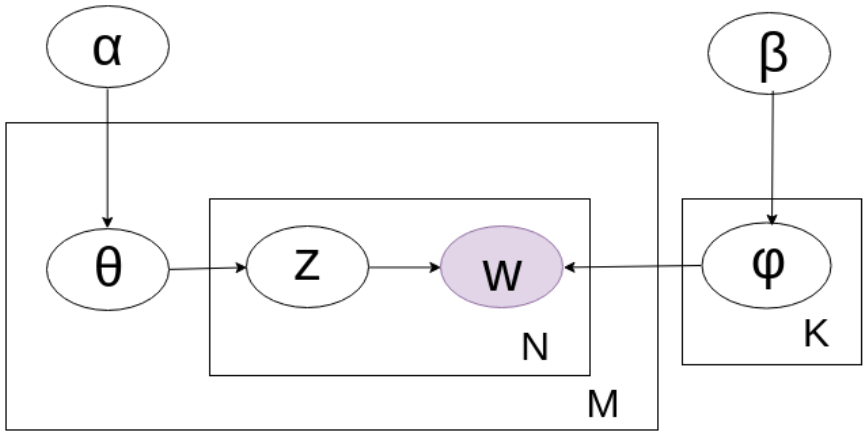

Figure 2: The design of the original LDA model.

To start with, we will familiarise with the higher level entities of the model. The circles represent particular entities, whereas the plates indicate the number of entities. For example, there would be $N \times M$ entities of z and M entities of θ. Further, LDA is a three-level hierarchical structure: the smallest entity (grey circle) is referred by a *word*; then a collection of words (inner plate) is referred by a *document*; and a collection of documents (outer plate) is referred by a *corpus*. Note that Figure 2 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 2, let's denote a word by w; then, the sequence of words in document d would be denoted by

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

and the sequence of documents 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 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. 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 distribution over topics
- w is the observable variable referring to the words in documents;
- 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 $M \times N$.

3.2 Preliminaries

The relationship of 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; also, we will familiarise with the lower level notation of the previous listed variables. Figure 2 illustrates 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 assignments of of the words in document d). More specifically, $z_{d,v}$ refers to the topic assignment of vocabulary term v. By taking $z_{d,v}$ and $\beta_{k,v}$ (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

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

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 $p(w)$. The approximation of $p(w)$ will be considered during the literature review on LDA-like models. Note that 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 in order 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, the model follows the *bag-of-words* principle. The next assumption states that we should use discrete data. Therefore, note that the basic LDA model 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 also set 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 in a document N(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 by the last 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, the authors suggest 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

2 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 3 below.

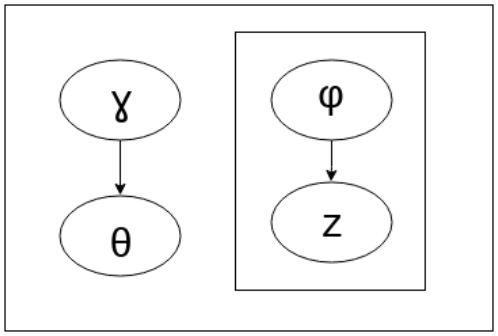

Figure 3: Variational parameters.

Note that the authors provide pseudo code to develop 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 an issue of *sparsity* in the 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. To be more specific, LDA has displayed balanced hidden topic proportions in the set of tested documents. That is, the tested documents have replicated the topic proportions of the training documents. The second experiment is directed towards the problem of document classification. The authors have addressed the *dimensionality reduction* in order to surpass the performance of the support vector machine (SVM) models. To be more specific, the documents with reduced dimensionality would possess 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 SVM-like models.

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 the word-topic correspondence over time; (3) the discussion over alternative inference methods. Apart from that, the experiments introduced

in this section have reflected the LDA performance during the time of the initial model release. Current state-of-the-art techniques overcome the shown results. Nevertheless, the carried experiments suggest the application fields of LDA-like models.

3.4 Finding Scientific Topics

The research paper 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 mixture. Further, the authors discuss the method's application settings for the machine learning problems in vision. Also, the authors conduct an experiment on the topic inference in the abstracts of the research papers published in 'Proceedings of the National Academy of Sciences' (PNAS). Finally, the paper is concluded by assessing the sampling-based method's 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 for the inference of 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 terminology: n_j^d and n_j^w are the counts of topic j assignments in document d and word w, respectively; 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 have shown an application of Gibbs sampling in a problem of vision. In their experiment settings, it is assumed that a collection of images represents a corpus; an image represents 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 to a topic. It has also been shown that Gibbs sampling displays superior performance in small data sets, whereas the performance in larger data sets 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 patterns. That is, the low number of topics might result in mashing significant patterns, whereas the high number of topics is likely to infer irrelevant patterns.

The main experiment of the research paper 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 in the future.

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 research paper on utilising the notion of time-series in LDA-like models. That is, we review the initial paper on utilising time-series in LDA-like models – '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 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. By reviewing this research paper, 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 sets 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 are strictly numeric and infinite, whereas categorical data take fixed values 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 4 below.

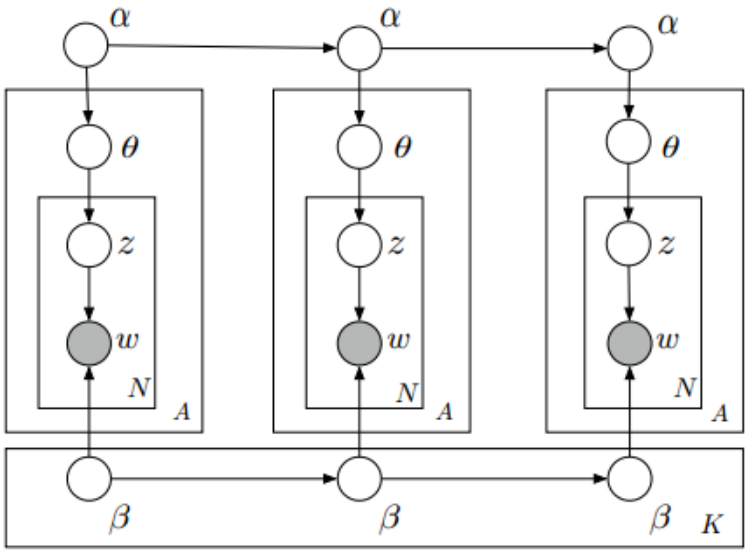

Figure 4: The design of a dynamic topic model. ©Dynamic Topic Models [4].

Effectively, Figure 4 is a sequential combination of the basic LDA model illustrated in Figure 2. In Figure 4, 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.

$\beta_t|\beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \sigma^2 I)$
2: $\alpha_t|\alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$
for $d \leftarrow 1, M$ **do**
4: $\eta_d \sim \mathcal{N}(\alpha_t, a^2 I)$
for $n \leftarrow 1, N$ **do**
6: $z_{d,n} \sim \text{Multinomial}(\pi(\eta_d))$
$w_{t,d,n} \sim \text{Multinomial}(\pi(\beta_t|z_{d,n}))$
8: **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 special case. That is, these parameters can be initialised randomly. 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 upcoming 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 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*. To be more specific, the 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 need of supplementary literature.

The implemented dynamic topic model has been evaluated by conducting an experiment on the journals of 'Science'. The experiment was 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 those latent 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 can be proved by comparing the inferred rises and falls of the scientific fields to the

actual history of science. Note that the experiment has been performed using 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 suggested by the authors. Finally, we have familiarised with the experiment settings which were used to infer the time series in the articles of 'Science'.

4 Proposed Approach

In this section, we will set up the approach of the project's execution. To start with, we will go over the literature reviewed in Section 3 in order to derive the promising techniques in terms of applicability to MSI data sets. Then, we will go over the high-level design and implementation requirements. Finally, we will set the basis for the model's evaluation and assess the potential risks.

4.1 Discussion

Throughout the literature review, we have looked into the papers establishing the mathematical underlying of the LDA-like models. However, the approach in utilising those models in MSI-like data sets have not been reviewed. Nevertheless, two sources of relevant research have been provided (the paper by Hooft et al. [8] and the survey by Alonso et al. [1]). Ultimately, this section is focused on emphasising the superior approaches of the previously reviewed papers; also, we will establish the appropriate techniques for the generative and inference processes.

The reviewed generative processes describe how the initial data is expected to be formed. Recall that we have reviewed two generative processes: static and dynamic. The static generative process is addressed in the initial LDA paper – 'Latent Dirichlet Allocation', and the dynamic generative process is addressed in the 'Dynamic Topic Models' paper. The applications of these generative processes are expressed respectively in Figure 5 and Figure 6 below.

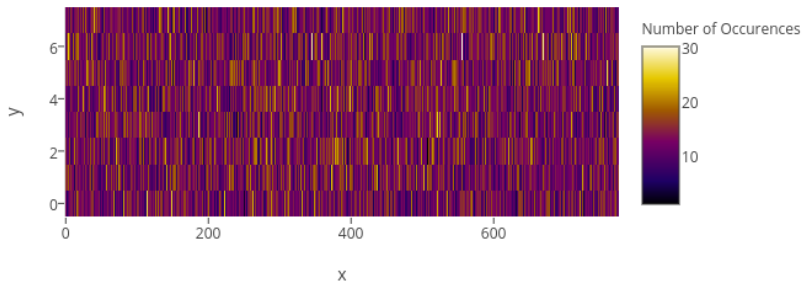

Figure 5: Static generative process.

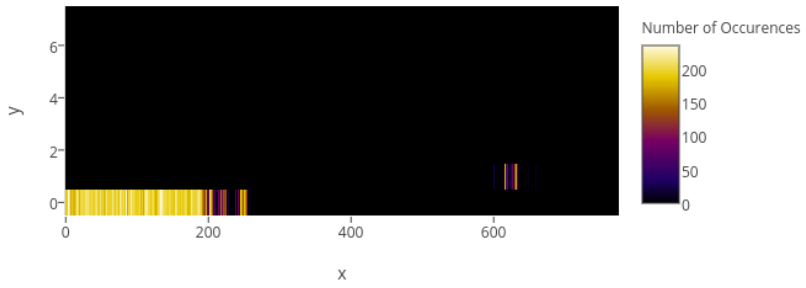

Figure 6: Dynamic generative process.

In the previous plots, we have visualised the occurrences of a particular entity within a corpus. The static generative process suggests a random distribution of the entity, whereas the dynamic generative process induces a particular pattern. Since we expect the ions of the same type to be gathered in clusters (rather than to be randomly distributed across the sample), the dynamic generative process is more appropriate to reflect the nature of the chemical entity patterns.

The initial dynamic topic modelling paper suggests to induce the time-series into the generative process of data sets. However, the way we capture the ions is done instantaneously for the whole corpus. Therefore, we are relaxing the assumption of time-series. Instead of that, we will use the dynamic topic modelling techniques to establish the relationship between the time-slices (even though we are not taking time into account, we will stick to the same terminology to preserve the consistency). For example, if we assume that the data is generated in a 1-dimensional manner, then time-slice t would depends on its adjacent time-slices $t-1$ and $t+1$. However, since the previous illustration of the sample is 2-dimensional, there are more adjacent time-slices. This relationship is shown in Figure 7 below.

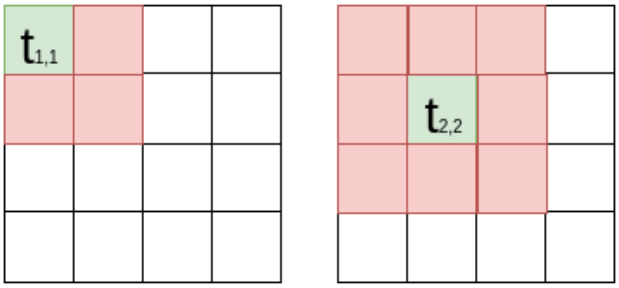

Figure 7: 2-dimensional generative process.

In the boundary case shown on the left, time-slice $t_{1,1}$ is related to three surrounding time-slices, whereas, in the general case indicated on the right side, time-slice $t_{2,2}$ relates to eight surrounding time-slices. Ultimately, the notion of dynamic topic modelling can be preserved even upon relaxing the assumption of time-series.

The reviewed inference techniques have covered the mathematical underlying of two approaches: stochastic (sampling-based) and optimisation (variational). The variational methods are discussed in the 'Latent Dirichlet Allocation' and 'Dynamic Topic Models' papers, and the sampling-based method is discussed in the 'Finding Scientific Topics' paper. Since both methods have been utilised in the state-of-the-art topic modelling applications, we will implement both of them in order to set up a discussion for the developed model's evaluation. However, since the authors of 'Dynamic Topic Models' have utilised only the variational methods, we have to address the compatibility of the sampling-based methods in terms of the applicability to the dynamic topic modelling.

4.2 Design

By covering the design of the model, we will set up the requirements for the project's execution and propose methods which would supplement the techniques derived from the reviewed literature. To start with, we will address the dimensionality reduction of the MSI data sets. Then, we will introduce the additional data capturing requirements shifting the set-up of the generative process. Finally, we will suggest an approach for utilising sampling-based methodology in the dynamic topic modelling.

In order to address the scalability of the model in terms of the large MSI data sets, we will present two techniques for the dimensionality reduction: utilisation of a dictionary and pre-processing of the data. Both of these techniques reduce the sparsity of the data sets. To start with, the use of dictionary will eliminate the irrelevant use of memory by not storing the words which do not occur in the documents. For example, let's assume that the masses of the ions in our data can be represented as the set $\{1, 2, 3, 4, 5\}$, also assume that the ions can be represented in the form of a (mass, intensity) tuple. Then, if our sample contained two following documents: $\mathbf{w_0} = \{(1, 0.1), (2, 200), (3, 500)\}$ and $\mathbf{w_1} = \{(5, 300)\}$, the sample representation in matrix and dictionary forms would be interchangeable with the expressions given in Table 1 below.

	1	2	3	4	5
$\mathbf{w_0}$	0.1	200	500	0	0
$\mathbf{w_1}$	0	0	0	0	300

	1	2	3	4	5
$\mathbf{w_0}$	0.1	200	500	Ø	Ø
$\mathbf{w_1}$	Ø	Ø	Ø	Ø	300

Table 1: Matrix and dictionary representation of the data set.

Note that, in these settings, the empty set symbol Ø represents a value which is not assigned to the memory; and the matrix and dictionary forms are respectively given on the left and right hand sides. It follows that the use of dictionary requires about as half as much memory. Another dimensionality reduction technique can be introduced by setting the minimum threshold value. That is, we can limit the data set by dismissing values which do not reach the sufficient significance criteria. If we set the threshold of intensity to be 1, then the dictionary could be represented in Table 2 below.

	1	2	3	4	5
$\mathbf{w_0}$	Ø	200	500	Ø	Ø
$\mathbf{w_1}$	Ø	Ø	Ø	Ø	300

Table 2: The dictionary of the data set with the minimum intensity threshold of 1.

As discussed in Section 1, the captured ions of the same type are likely to have fluctuations in their masses. For this reason, we are applying another pre-processing step: by setting another threshold for the maximum fluctuation between the masses of the ions, we would group the ions which are likely to represent the same type. As a result, we would reduce the dimensionality of the data set. By taking the example of the previous paragraph and setting the fluctuation tolerance to 1, we obtain the data structure given in Table 3 below.

	[2,3]	[5,5]
$\mathbf{w_0}$	700	Ø
$\mathbf{w_1}$	Ø	300

Table 3: The dictionary of the data set with the fluctuation threshold of 1.

Notice that the masses are now represented by intervals, and the intensities below the threshold are summed together. Ultimately, by utilising the latter dimensionality reduction and ion grouping techniques, we establish a data structure which would scale better compared to the naive data capturing settings.

The generative process of MSI data sets, which was addressed in the discussion over the reviewed literature, has utilised 2-dimensional data sets. However, we are expecting to work on 3-dimensional data sets. Therefore, the established relationship between time-slices is illustrated by Figure 8 below.

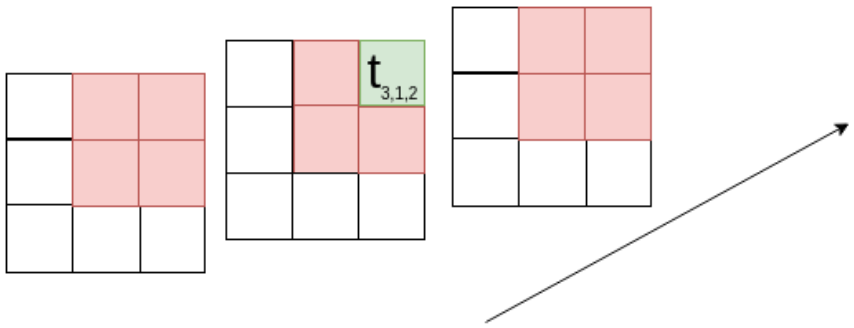

Figure 8: 3-dimensional generative process.

The time-slice indicated by the green colour, which is one of the boundary cases, is related to other elven time-slices indicated by the red colour. Note that in the general case, the time-slice would have twenty-six relating documents. Effectively, we have introduced the application settings of the dynamic topic modelling in the 3-dimensional space.

In Section 3, we have addressed a sampling-based inference method – Gibbs sampling. However, for the dynamic topic models, the latter method has to be reconsidered due to the change from Dirichlet to Gaussian distributions. Therefore, we will look into another MCMC based algorithm – Metropolis–Hastings.

4.3 Evaluation

The developed technique addressing the spatial smoothing of the MSI data sets will evaluated by measuring the improvement in performance. That is, the performance will be assessed by investigating the quality of the inferred patterns. Also, we will utilise two different inference methods: sampling-based and variational. By having two model variations, we will be able to compare the time performance. Hence, we will be able to deduce which variation scales better with the size of the data sets. Finally, the section will be finished by investigating the possible risks; also, we will provide potential strategies mitigating the risks.

The implemented model will be evaluated using synthetic and real data sets. However, at the start, we will use the synthetic data sets. Note that the performance will be measured in the following way: (1) take two equivalent data sets; (2) apply the spatial smoothing on one of the data sets; (3) infer the underlying structure of the data; (4) check whether the applied smoothing has inferred the underlying structures better. By working with synthetic data sets, we had the control over inducing the underlying structure of the data. Hence, we will able to give a basis in measuring the performance of the applied spatial smoothing.

Even though we have induced a strategy for the project's evaluation, the described process assumes that we can mitigate the potential risks. Note that, at the initial stage of the project, we have identified the following potential risks:

- The implementation complexity of the inference algorithms;
- The access to the real data.
- The model's scalability in large data sets;
- The inability to measure the performance of the applied spatial smoothing.

The first three risks would slow down the project execution thus raise a concern of finishing the project on time, whereas the fourth one would prevent the confirmation of denial of the initial project's hypothesis – 'the degeneracies of MSI data sets can be detected using LDA-like models'.

The risk of complex (in terms of development) algorithms has been already mitigated by familiarising with basic sampling-based and optimisation-based inference models. That is, we have implemented the models utilising Gibbs sampling and variational inference. Also, note that the main model will firstly be developed using Gibbs sampling, as this method is easier to implement compared to the variational one.

Further, the risk of model's scalability can be mitigated by working on the small data sets. In case the final model displays an improvement in the inference after the application of the spatial smoothing, this project would bring motivation for another research project on optimising the method for larger data sets.

The risk of the real data access might be partially resolved, as the University has a laboratory for the research in Metabolomics. However, in case this options is not possible, we can obtain the data sets from the Web.

Finally, the risk of the inability to measure the performance of the applied smoothing can impact the outcome of the project the most. That is, it would be difficult to reason about the improvement in the inference of the real data sets, because we do not possess the full knowledge of the data set underlying structure. We will mitigate this issue by establishing a performance metric from the synthetic data sets.

By having the defined boundaries of the expected improvement in performance, we can attempt to tackle the real data sets. That is, the expected performance improvement would serve as an indicator whether the inferred real data deviates in the expected limit. To be more specific, the data sets with applied spatial and the data sets without applied smoothing will display some differences in the inferred topics. Hence, we will check whether those differences meet the performance boundaries suggested by the inference of the synthetic data sets.

5 Work Plan

In this section, we will outline the project milestones and deliverables. That is, we will outline the most important steps and provide the estimations in a form of Gantt chart.

Further, the deliverables will be directly induced from the project's objectives and hypothesis (these are addressed in Section 2). The deliverables will contain both successful and successful outcomes of the set hypothesis.

5.1 Schedule

The high-level plan of the project's execution include the following milestones: design, implementation, evaluation, and documentation/presentation. To establish a basis for the design, we will continuously review the literature on spatial smoothing, MSI, and dynamic topic models. Then, the implementation will require the utilisation of spatial smoothing in the generative process of the dynamic topic model; also, for the inference, we will implement both: sampling-based and variational methods. Further, the evaluation will be conducted by assessing the performance of the developed models. Finally, the results will be documented in a form of weekly reports, dissertation, and presentation.

The outlined tasks will not be necessarily executed in sequential manner. For some of the tasks, we will use the parallel approach as given in Figure 9 below.

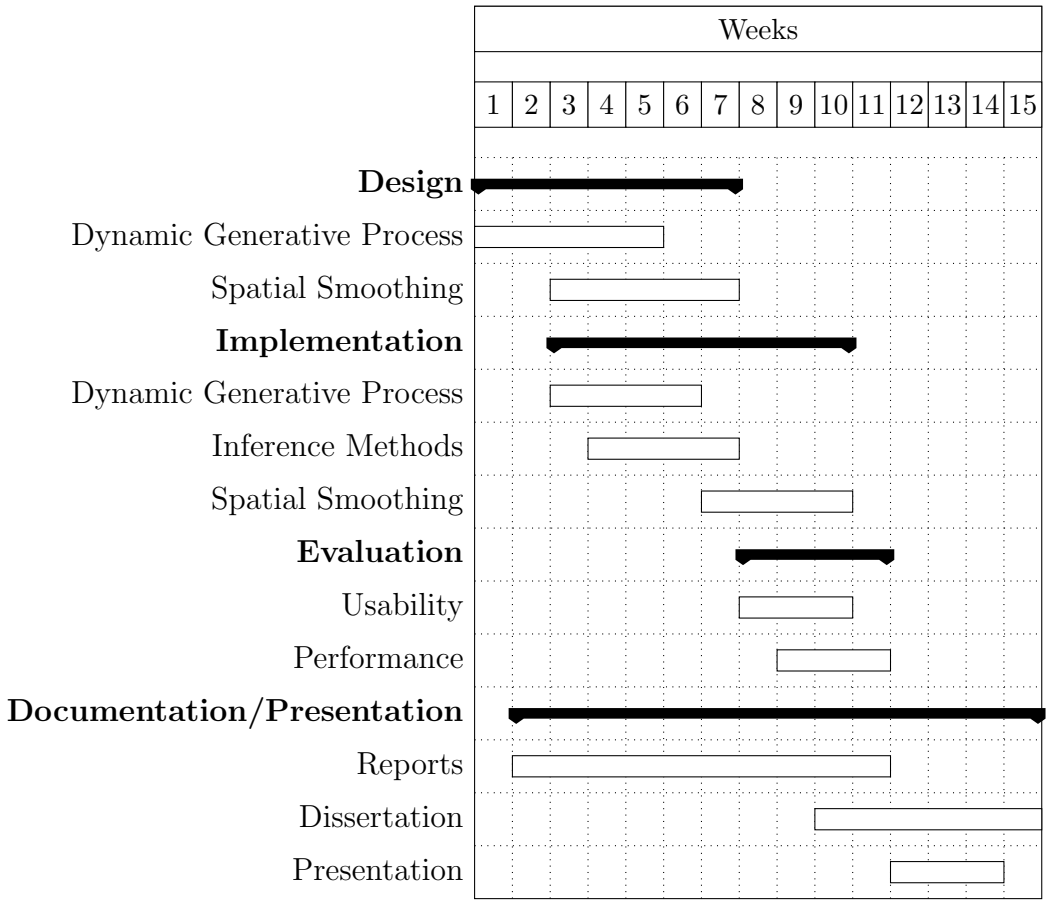

Figure 9: The Gantt chart estimating the project's execution.

5.2 Deliverables

The deliverables of the project will depend on the outcome of the initial hypothesis – 'the degeneracies of MSI data sets can be detected using LDA-like models'. Nevertheless, in any of the project's outcomes, we will provide the following deliverables: the presentation; the dissertation; and the spatial smoothing model, which will be written in Python. However, the main project's contribution to the research in MSI will depend on the results: if the model displays an improvement in performance, then we could utilise the model in pre-processing MSI data sets; in the case of unsuccessful outcome, we would assess the techniques used and provide details on why the model does not work. Both outcomes would bring the motivation for further research in improving MSI data sets: the successful outcome would bring attention to the applicability of LDA-like models; the unsuccessful outcome would suggest the need of a different perspective required to tackle the issue of noise.

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] Rob Smith, Andrew D Mathis, Dan Ventura, and John T Prince. Proteomics, lipidomics, metabolomics: a mass spectrometry tutorial from a computer scientist's point of view. *BMC bioinformatics*, 15(7):1, 2014.

[8] 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.