University of Glasgow | School of Computing Science

Spatial Smoothing in Mass Spectrometry Imaging

Arijus Pleska (2019828P)

April 19, 2017

ABSTRACT

In this paper, we target a data modelling approach used in computational metabolomics; to be specific, we assess whether spatial smoothing improves the topic term and noise identification. By assessing mass spectrometry imaging data, we design an enhancement for latent Dirichlet allocation-based topic models. For both data pre-processing and model design, we survey relevant research. Further, we focus on presenting our methodology in detail providing the preliminaries to the basic models and guiding through the performed adjustments in the proposed model variations. To assess the application's performance, we run the model comparison on a number of diverse synthetic datasets.

1. INTRODUCTION

In this research paper, we assess an application of spatial smoothing in visual data; that is, we induce continuity among data elements. The spatial smoothing application is particularly targeted to be applied for unsupervised pattern recognition. To be more specific, our focus is to model a biomedical application in the field of metabolomics. Furthermore, we limit the scope of the applied unsupervised machine learning techniques to the branch of topic modelling.

The characteristics of our utilised metabolomics data are expressed in the form of mass spectrometry imaging (MSI). Effectively, we use MSI to visualise the metabolomics data in the form of spatial distribution. Speaking of the metabolomics data, it contain information about ionised metabolites. Note that metabolites are molecules produced by the chemical process of metabolism, whereas by ionisation we refer to the information capture method. In other words, MSI data is a visualisation of ion (sampled metabolite) distributions: the complete dataset is the whole image; the image's pixel is a particular sampling region; and each region contains intensities of ions with unique mass-to-charge m/z values.

Speaking of our machine learning application, topic modelling is a technique used to infer unlabelled topic distributions based on data's underlying semantic structure. With respect to MSI data, we can model the topic distributions over an image and its every pixel; furthermore, topic modelling can express the types of ions corresponding to particular topics. Since we use a topic model to model metabolomics data in a statistic manner, the utilised topic models are tuned to reflect the metabolomics environment as realistically as possible. Relating to our research targets, spatial smoothing is one of such environment settings.

The basis of the project's research problems comes from the limitations of current metabolite sampling techniques. The main limitation is the loss of information caused by the metabolite ionisation. As a consequence, the MSI data is noisy. To expand on the noisiness, it is caused by metabolite fragmentation and the limitation to tune different metabolite retention times. By metabolite fragmentation, we refer to a metabolite split; the split would cause the captured ions to possess unexpected values. Speaking of the retention time, the intensity value of each ion type varies with respect to time; therefore, since each ion is captured at a different state of its retention, each ion type would possess some variance with respect to its intensity. Finally, note that the loss of information might be also caused by overlapping ion topics. Since different ion topics can contain same ion types, the topic possessing a lower ion intensity value would be overwhelmed and, thus, not reflected in the MSI data.

In this paper, we contribute to the research in MSI by carrying an extensive assessment of the spatial smoothing application. The assessment is carried in both quantitative and qualitative manners: we assess the performance on a number of diverse datasets; also, our experiments are designed to reflect the nature of the MSI data reflecting computational metabolomics. Furthermore, we provide a Python implementation of a tuned topic model; also, we establish maintainable experiment settings. By the tuned topic model, we mean that the model's implementation is particularly designed to meet the characteristics of the MSI data. Speaking of the experiment settings, note that we are carrying the experiments using Jupyter notebooks. Effectively, the use of the notebooks creates a portable and well-documented environment to initialise the experiment settings and execute the topic modelling. As a result, external parties could run the released notebooks and reproduce the experiment results in a swift manner.

The paper is organised in the following order: in Section 2, we discuss the background of the research project; in Section 3, we provide a formal definition of the assessed research problems; in Section 4, we review the results of the relevant research; in Section 5, we establish the rationale of the applied methodology; in Section 6, we introduce the experiments; finally, in Section 7, we conclude the findings.

2. BACKGROUND

The background section covers the basis of the concepts used throughout the paper. At the start, we provide a high-level overview of the general topic modelling concepts; then, we define the terminology used throughout the paper; finally, we introduce the characteristic qualities of the MSI data.

2.1 Topic Modelling Preliminaries

The research project targets a specific branch of topic models. The branch consists of Latent Dirichlet Allocation (LDA) derivatives. Note that the initial LDA model was introduced by Blei et al. [4]. One of the model's key characteristics is the three-level hierarchical treatment of the data. In the context of the utilised MSI data, the hierarchical structure can be perceived as follows: in the highest level, we have an MSI image; in the middle level, we have a pixel of an MSI image; and in the lowest level, we have the intensities of particular ions in a pixel.

Another model's key characteristic is the generative treatment of the data. By a generative model, we mean that the latent data instances are treated as a result of a mixture of underlying parameters drawn from probability distributions. In other words, the generative data treatment induces randomness in the end products of the data; however, note that the source of the data – the lowest level parameters of the probability distributions – remain the same. The key aspect of the generative model is the degree of freedom in the connections of random variables; this notion allows modelling more realistic, thus, more complex data relations. Ultimately, the rationale of the generative model is based on recovering the underlying semantic structure. Effectively, in order to recover the generative process of the data, we delve into the applications of Bayesian methods.

By applying Bayes' rule, we can express the underlying semantic structure in the form of a posterior probability distribution. However, note that we can not analytically compute such posterior expression. Instead, we can estimate the posterior distributions using optimisation or direct sampling. In this project, we particularly focus on the underlying semantic structure's inference using direct sampling. Speaking of the sampling-based inference method applications to LDA-like models, the ground-work has been established by Griffiths and Steyvers [9]. The authors have utilised the collapsed Gibbs sampling – a Markov chain Monte Carlo (MCMC) algorithm. Effectively, the method integrates the uncertainty out; so that we sample the entities of interest directly. More specific details about the rationale of the collapsed Gibbs sampling application will be provided in Section 4.

Relating to the initial LDA model, the authors have set the following assumptions for the utilised data: exchangeability among the inner components of the lower and middle data hierarchy levels; also, a discrete treatment of the lower-level data. By exchangeability, it is meant that the components follow the bag-of-words principle; that is, the order of the data has no correlation with the underlying semantic structure. Speaking of the discrete data treatment, it is assumed that the lower-level components have no spatial connection. With respect to our project, both assumption types – exchangeability and discreteness – do not correspond to the characteristics of the MSI data. Therefore, we will establish a methodology to relax the latter assumptions.

2.2 Topic Modelling Terminology

In this subsection, we introduce our topic modelling terminology. In order to put the MSI data in the topic modelling context, the components of LDA's three-level hierarchical structure are defined as follows: *a corpus* is the whole image of a sample; *a document* is an MSI image's pixel; and *a word* is an interval of mass-to-charge values corresponding to a particular ion. Starting from now on, we will use the latter topic modelling concepts to introduce the LDA model's architecture and notation.

Since LDA-like models are also graphical models, the dependence of random variables can be illustrated using graphs. In order to familiarise with the architecture and the variables of LDA-like models, we provide Figure 1 illustrating the initial LDA model in the plate notation. The circles indicate the model's variables: the coloured circle corresponds to the observable variable, whereas the uncoloured circles correspond to the hidden variables. Effectively, the hidden variables define the model's underlying structure. Speaking of the plates, the plate denoted by K corresponds to the number of topics; the plate denoted by T corresponds to the number of documents, and the plate denoted by N corresponds to the number of words. Effectively, the letters in the bottom right corners indicate the total number of variables. Therefore, a corpus has a T number of documents, and each document has an N number of words.

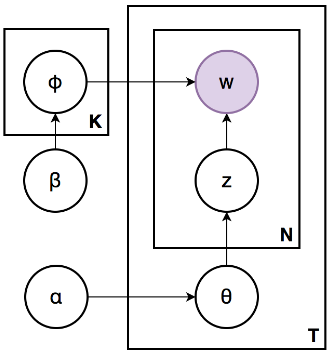


Figure 1: The initial LDA model's architecture.

Before providing a listing with the definitions of the LDA variables, we introduce the purpose of the variables. The variables denoted by θ and ϕ correspond to the underlying probability distributions (e.g., in the case of the initial LDA model, we use Dirichlet distributions). It follows that the variables denoted by α and β act as the parameters of the latter probability distributions; note that in the context of machine learning, such auxiliary parameters are called hyper-parameters. Effectively, *a hyper-parameter* allows tuning a machine learning model for a particular dataset application. As a side reference, note that by *a vocabulary* we refer to a collection of terms reflecting a fixed range of the words. At this point, we provide the following list containing the definitions of the initial LDA model's variables:

- K is the number of topics;
- T is the number of documents;
- N is the number of words per document;
- V is the size of a vocabulary;
- w is a word;
- z is a word's topic assignment;
- θ_t is the topic distribution over the document t;
- ϕ_k is the vocabulary term distribution over the topic k;
- α is the hyper-parameter for the topic distributions;
- β is the hyper-parameter for the vocabulary term distributions.

2.3 MSI Data Characteristics

In this subsection, we introduce the qualities of the raw MSI data. Furthermore, we set the requirements for the raw data's pre-processing; note that the pre-processing serves as an auxiliary method making the raw MSI data compatible for a scalable topic modelling application. As a side note, the basis of our applied MSI data characteristics is established from the mzML data format. Conveniently, we parse mzML data using the `pymzml` Python library.

Essentially, raw MSI data contain mass spectra of a sample – the sample's every pixel is expressed in the form of a mass spectrum. By a mass spectrum, we refer to a map from a mass-to-charge value to an intensity value. Ultimately, we can establish a continuous notion of the MSI data by ordering the mass-to-charge values. Note that since the sampling equipment can detect the mass-to-charge values in the millidalton (mDa) precision, the MSI data is sparse (i.e., a large portion of mass-to-charge values are mapped to the zero intensity).

In the context of topic modelling, the raw MSI data possess a large vocabulary (above 5000 terms). Note that we consider every mass-to-charge value as a word; whereas every intensity value is perceived as a word's occurrence count. Further, since the intensity values could spike up above 1000, the time complexity of the latent topic inference would require pro-longed runs. To overcome the introduced scalability issues, we carry data pre-processing – the applied techniques are discussed in the upcoming methodology section.

3. STATEMENT OF PROBLEM

To start with, we set the hypothesis of this research project to *'The spatial smoothing application induces more realistic representation of the visual computational metabolomics data – MSI'*. By spatial smoothing, it is meant that the topic model would have an auto-regressive treatment among the nearby MSI instances. As an example, we assume that adjacent pixels would have similar latent topic distributions. This assumption corresponds to the nature of our datasets – a metabolite construction (i.e., a topic) is continuous throughout nearby regions (i.e., sets of adjacent pixels).

The impact of proving the hypothesis would bring the following contributions:

- Improving the detection of overlapping topics and vocabulary terms;
- Reducing the noisiness of the MSI data;
- Motivating a further research in the spatial smoothing application in MSI data.

To expand on the overlapping topic detection, the issue arises when two underlying topics are made of similar vocabulary terms: instead of a separate representation, the topics are merged. We intend to identify the flow of distinct topics by the application of spatial smoothing; as a consequence, the spatial smoothing would also impact the data noisiness reduction. Ultimately, if a naive spatial smoothing application displayed a performance improvement, the contribution would set a basis to utilise state-of-the-art auto-regression techniques on the MSI data.

To my knowledge, the impact of the spatial smoothing application to the MSI domain has not yet been thoroughly studied. For the latter reason, this research project serves as an exploratory assessment of the spatial smoothing application: we introduce the rationale behind the applied methodology; also, we clearly define the range of the experiment settings. To give a brief intuition about the methodology, the study assesses the domain-specific parameter tuning and its impact on a diverse range of synthetic datasets.

4. RELEVANT RESEARCH

In this section, we review the ground-work carried on the following aspects: the pre-processing of MSI data; the rationale of the utilised topic models; and novel approaches exploiting the characteristics of MSI data. Note that the covered groundwork is selected to reflect the rationale of the utilised techniques. In order to establish the basis of the state-of-the-art computational metabolomics methodology, we consult the survey carried by Alonso et al. [1]; speaking of the key probabilistic topic modelling branches, we consult the survey carried by Blei [2].

4.1 MSI Data Pre-processing

In order to establish a scalable topic inference, we review the following MSI data pre-processing techniques: data normalisation; feature binning; and noise reduction. To start with, data normalisation serves as a method establishing an adaptable data structure. Bolstat et al. [5] have proposed a data normalisation method called linear baseline scaling. Note that the method is particularly targeted at sparse datasets. Effectively, the method is applicable to numerical features; note that the method's work-flow is carried as follows: we find the largest numerical feature of all data instances; then, we calculate the scaling factors by aligning the largest values to a pre-set upper threshold; finally, we align the remaining numerical features based on the established scaling factors. Relating linear baseline scaling to applications on MSI data, Kohl et al. [10] have shown that the method's application does not produce a significant loss of information to establish a well-performing MSI data inference.

Speaking of feature binning, the technique is used to merge the instances of a feature space; as a result, feature binning reduces the data dimensionality. Note that by the MSI data feature space, we refer to distinct m/z values. Since raw MSI data is sparse, a successful application of feature binning is based on identifying appropriate m/z boundaries reflecting unique metabolite types. To introduce some examples, the research carried by Craig et al. [6] have utilised an equally spaced feature binning. Even though the authors have succeeded to reduce the data dimensionality, they have encountered a loss of information by splitting the metabolite topics into arbitrary regions. As an alternative approach, De et al. [7] have performed a feature binning based on the spectral peak identification. Effectively, the technique induces a dynamic notion of bin boundaries which are based on the identified intensity peak regions. As a result, the merged m/z values serve as a better representation of metabolite terms.

The last reviewed MSI data pre-processing techniques addresses the MSI data noisiness. Even though the MSI data noise reduction is an open research problem going well beyond data pre-processing, Smith et al. [12] have shown that the application of a general data pre-processing routine displays performance improvements. For example, the authors have induced a lower intensity bound. Effectively, the intensity values below the intensity threshold would be treated as insignificant and/or as a product of data capturing device imperfections. By applying this procedure, the authors have successfully reduced the data dimensionality and, thus, increased the data scalability.

4.2 Prospective Topic Models

In this subsection, we review the key characteristics of the prospective topic modelling inference techniques and model variations. Speaking of the inference techniques, we employ the rationale of a sampling-based LDA model. Griffiths and Steyvers [9] have displayed a successful application of the collapsed Gibbs sampler for the topic inference of both textual and visual data. To expand on the collapsed sampling, the technique allows skipping the estimation of θ and ϕ values; instead, the inference is based on the notion of the assignment counts. In order to estimate a topic assignment's probability, the authors suggest using the following expression:

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

note that i refers to the current iteration ($-i$ refers to the previous iteration); k refers to a particular topic; and n refers to the count of the instances indicated by the term's superscript. If required, θ and ϕ can be sampled from the following distributions:

$$\phi_{k,w} \sim \text{Dirichlet}\left(\frac{n_k^{(w)} + \beta}{n_k^{(\cdot)} + W\beta}\right),$$
$$\theta_{t,k} \sim \text{Dirichlet}\left(\frac{n_k^{(t)} + \alpha}{n^{(t)} + K\alpha}\right).$$

Essentially, more accurate representations of θ and ϕ are obtained by running the Gibbs sampler until a sufficient convergence: we preserve the θ and ϕ values of each iteration; when the sampling is finished, the average of the preserved values is an accurate approximation of the underlying θ and ϕ values.

In order to relax the topic modelling assumptions of the initial LDA model, we review the model's derivatives. One of the assumptions – word exchangeability – is addressed by Blei and Lafferty [3]. The authors have proposed dynamic topic model (DTM) inducing the notion of change in the topic and vocabulary distributions; the model's architecture is illustrated in Figure 2. Ultimately, the documents of a corpus are assigned to segments with unique θ and ϕ values. Note that the changes in θ and ϕ values are impacted by their hyper-parameter updates. Since in the DTM paper the authors use variational inference, we provide a reference for an alternative dynamic topic modelling variation – sequential LDA – proposed by Du et al. [8]. Note that in the sequential LDA paper, the authors display a rigorous application of the collapsed Gibbs sampler.

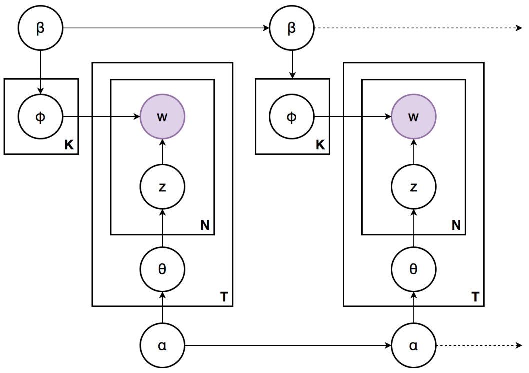


Figure 2: The DTM architecture.

4.3 Prospective Characteristics of MSI Data

At this point, we look into the prospective topic modelling applications exploiting the characteristics of MSI data and employing spatial smoothing. To start with, Hooft et al. [13] have utilised an LDA-like model to infer metabolite substructures from MSI data. The authors have established a novel approach utilising a favourable LDA's property – the option to assign a unique vocabulary term to multiple topics. Effectively, this approach allows identifying metabolite substructures which are made of overlapping elements.

Speaking of the spatial smoothing application, to my knowledge, the idea has not yet been widely spread among the computational metabolomics community. Nevertheless, a recent study by Palmer et al. [11] have attempted to quantify spatial chaos among the partitions of MSI data. The authors have reported that the established notion of spatial chaos has improved the speed and accuracy of continuous metabolite pattern identification.

5. METHODOLOGY

In this section, we cover the rationale of a topic model tuned for the spatial smoothing application. To start with, we provide details on how to establish the auto-regressive notion among MSI data; then, we show how to apply the auto-regression to a topic inference based on the collapsed Gibbs sampling. Further, we provide a list of the applied data preprocessing techniques for MSI data. Finally, by considering the data format induced by the data pre-preprocessing, we introduce the generative MSI data process. Effectively, we apply the generative process to generate synthetic data for our experiments.

5.1 Spatial Smoothing

We establish the spatial smoothing among MSI data by inducing the auto-regressiveness among the pixels (documents). Before going into details, note that we cover the established methods using the previously introduced topic modelling notation. To start with, recall that by auto-regressiveness we refer to the smooth topic development among the nearby instances of an MSI corpus. In our settings, the auto-regressiveness is established by assuming that the joint probability distribution of the α priors is given by the following expression:

$$p(\alpha_1, \dots, \alpha_T) = p(\alpha_1) \prod_{t=2}^{T} p(\alpha_t|\alpha_{t-1}), \quad \text{where}$$
$$p(\alpha_0) = \mathcal{N}(\alpha_0; 0, \sigma_0^2 I), \quad \text{and}$$
$$p(\alpha_t) = \mathcal{N}(\alpha_t; \alpha_{t-1}, \sigma^2 I), \quad t > 0.$$

To introduce the previous expression, note that the index t refers to a particular pixel (a document). This means that every pixel of an MSI corpus has a unique underlying topic distribution induced by a unique α. Further, the variances

σ_0^2 and σ^2 correspond to the initialisation variance and the smoothness variance, respectively. Effectively, σ_0^2 is used to create larger gaps among different topics, whereas σ^2 preserves the smoothness. Therefore, we set σ_0^2 to possess a higher value compared to σ^2.

The previously introduced α priors serve as the initial point in estimating the true α values. To estimate the true α values, we utilise the Metropolis–Hastings (MH) algorithm. The work-flow of the MH algorithm is started by drawing the proposed state:

$$x' \sim q(x, \delta^2 I).$$

Note that x denotes the current state, q denotes the proposal distribution, and δ^2 denotes the proposal variance; also, note that if δ^2 is large, the proposal state converges to the true posterior in larger yet random increments; alternatively, if δ^2 is small, the convergence is performed in small yet uniform increments. Note that the settings for an optimal convergence are unique for diverse datasets; as an example, we can find the optimal values using cross validation. Going back to the work-flow, for the second step, we consider the acceptance distributions (these are denoted by A) and derive the formula for the acceptance rate:

$$\frac{A(x'|x)}{A(x|x')} = \frac{p(x'|x)}{p(x|x')} \cdot \frac{q(x'|x)}{q(x|x')} = \frac{p(x',x)}{p(x)} \cdot \frac{p(x')}{p(x,x')} = \frac{p(x')}{p(x)}.$$

Note that the proposal distributions cancel out as

$$q = \mathcal{N} \implies q(x'|x) = q(x|x').$$

For the MH algorithm's final step, we make sure that the acceptance rate does not overflow the probability boundaries; that is, we obtain the acceptance rate denoted by r using the following procedure:

$$r = \min\left(1, \frac{p(x')}{p(x)}\right).$$

At this point, we apply the rationale of the introduced MH algorithm to the topic modelling context. Since we utilise the MH algorithm to update a single value at a time (i.e., we update $\alpha_{t,k}$), we make use of the following notation:

$$\alpha^{-tk} = \alpha \setminus \alpha_{t,k}.$$

Having the previous notation in mind, the MH algorithm's application to update α is given as follows:

$$\begin{aligned}
&\frac{p(z, \alpha^{-tk}, \alpha'_{t,k}|X)}{p(z,\alpha|X)} = \dots \\
&\dots = \frac{p(X|z, \alpha^{-tk}, \alpha'_{t,k})}{p(X|z,\alpha)} \cdot \frac{p(z|\alpha^{-tk}, \alpha'_{t,k})}{p(z|\alpha)} \cdot \frac{p(\alpha^{-tk}, \alpha'_{t,k})}{p(\alpha)} \\
&\dots = \frac{\prod_{k=1}^{K} \pi(\alpha'_{t,k})^{z_{t,k}} \cdot p(\alpha'_t|\alpha_{t-1}) \cdot p(\alpha_{t+1}|\alpha'_t)}{\prod_{k=1}^{K} \pi(\alpha_{t,k})^{z_{t,k}} \cdot p(\alpha_t|\alpha_{t-1}) \cdot p(\alpha_{t+1}|\alpha_t)}; \quad t \notin \{1, T\}.
\end{aligned}$$

For completeness, the expressions at the boundaries take the following form:

$$\begin{aligned}
t = 1 &\implies \frac{p(\alpha^{-1k}, \alpha'_{1,k})}{p(\alpha)} = \frac{p(\alpha'_1) \cdot p(\alpha_2|\alpha'_1)}{p(\alpha_1) \cdot p(\alpha_2|\alpha_1)}, \\
t = T &\implies \frac{p(\alpha^{-Tk}, \alpha'_{T,k})}{p(\alpha)} = \frac{p(\alpha'_T|\alpha_{T-1})}{p(\alpha_T|\alpha_{T-1})}.
\end{aligned}$$

Also, note that by π we denote the softmax function which is expressed as follows:

$$\pi(\alpha_{t,k}) = \frac{\exp(\alpha_{t,k})}{\sum_{k'=1}^{K} \exp(\alpha_{t,k'})}$$

5.2 Auto-regressive Dynamic Topic Model

Our auto-regressive topic model is based on the rationale of the reviewed dynamic topic model. However, based on the MSI data characteristics and the application of spatial smoothing, the auto-regressive model possesses the following aspects:

- The static treatment of the β hyper-parameter;
- The Gibbs sampler utilising spatial smoothing;
- The application of logarithmic space to perform calculations.

In the following paragraphs, we introduce each of the previous listings.

Even though we utilise the rationale of DTM in order to establish the dynamic notion of the topic development, we preserve a static β hyper-parameter. This assumption comes from the characteristics of the metabolomics-based MSI data: we expect the metabolite patterns (i.e., a topic's vocabulary) remain constant. Therefore, in our model, we utilise the architecture illustrated in Figure 3 below.

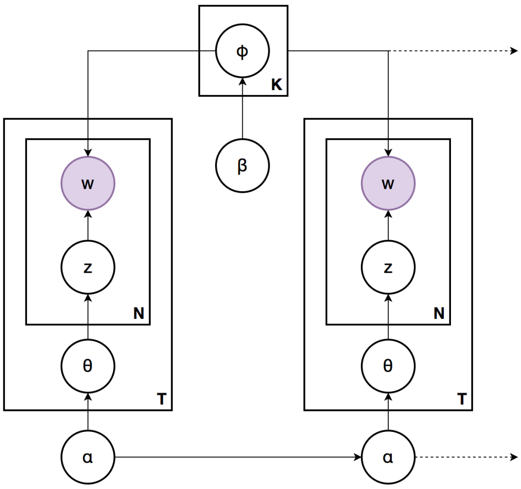


Figure 3: The auto-regressive topic model architecture.

Speaking of the Gibbs sampler's enhancement, the updated topic assignment formula takes the following form:

$$P(z_i = k | z_{-i}, w, t) \propto \frac{n_{-i,k}^{(w_i)} + \beta}{n_{-i,k}^{(\cdot)} + V\beta} \cdot \pi(\alpha_{t,k}).$$

As a consequence, the sampling of θ also changes; now, we obtain the topic distribution as follows:

$$\theta_{t,k} = \pi(\alpha_{t,k}).$$

Finally, we address the computational stability by performing calculations in logarithmic space. Effectively, the application of logarithmic space mitigates the susceptibility to numerical underflow. Note that numerical underflow is especially relevant in the context of probabilistic models:

calculations involve large products of probabilities. In logarithmic space, however, the products are transformed into sums. Speaking of our model, we apply logarithmic space for both the auto-regressive α update and the sampling-based inference. The updated expression for the auto-regressive α update is given as follows:

$$\begin{aligned}
&\log\left[\frac{p(z, \alpha^{-tk}, \alpha'_{t,k}|X)}{p(z, \alpha|X)}\right] = \dots \\
&\dots = \log\left[p(z, \alpha^{-tk}, \alpha'_{t,k}|X)\right] - \log\left[p(z, \alpha|X)\right] \\
&\dots = z_t \sum_{k=1}^{K} \log\left[\pi(\alpha'_{t,k})\right] + \log\left[p(\alpha'_t|\alpha_{t-1})\right] + \log\left[p(\alpha_{t+1}|\alpha'_t)\right] \\
&\qquad - z_t \sum_{k=1}^{K} \log\left[\pi(\alpha_{t,k})\right] + \log\left[p(\alpha_t|\alpha_{t-1})\right] + \log\left[p(\alpha_{t+1}|\alpha_t)\right].
\end{aligned}$$

As a result, the acceptance rate takes the following expression:

$$r_{t,k} = \exp\left[\min\left(0, \log\left[p(z, \alpha^{-tk}, \alpha'_{t,k}|X)\right] - \log\left[p(z, \alpha|X)\right]\right)\right].$$

Speaking of the updated expression for the inference, it is updated in the following way:

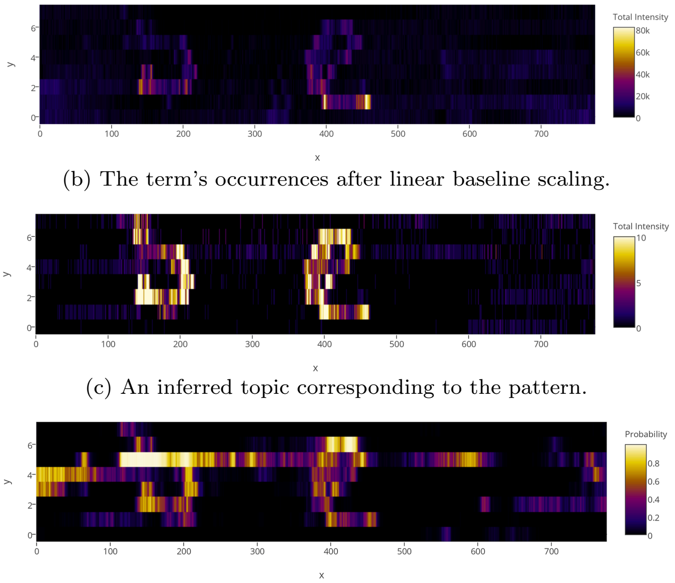
$$\begin{aligned}
&P(z_i = k|z_{-i}, w, t) \propto \dots \\
&\dots \propto \log\left[n^{(w_i)}_{-i,k} + \beta\right] - \log\left[n^{(\cdot)}_{-i,k} + V\beta\right] + \log\left[\pi(\alpha_{t,k})\right].
\end{aligned}$$

5.3 Data Pre-processing and Generative Process

In this subsection, we introduce the MSI data format used in the experiments. At the start, we introduce an example of real data. Effectively, the example displays an application of the pre-processing techniques presented in the literature review section. Further, we transfer the qualities of real MSI data into our synthetic data generation module. To be more specific, we provide an algorithm for the generative data process.

Before carrying the experiments, we familiarise with the raw MSI data characteristics and assess their scalability. To be more specific, we define the characteristics of our synthetic data by pre-processing a real MSI data sample. To introduce the pre-processing details, we dismiss the words below the intensity threshold of 10; then, we apply the following bucketisation strategy: merge adjacent vocabulary terms which differ less than 7 mDA. Effectively, the bucketisation strategy is based on the spectral peak identification. Finally, we apply linear baseline scaling to align the highest intensities to 25. Most importantly, note that these settings are unique with every dataset; however, the provided values allow carrying experiments in a scalable manner (i.e., a single experiment run on one dataset would take approximately 60 minutes).

At this point, we take an exemplary sample. Note that, in the sample, there are two letters inscribed with the ink corresponding to a particular mass-to-charge value. In Figure 4, we compare the visualisation of the sample with and without the applied pre-processing.



(a) The term's occurrences before linear baseline scaling.

(b) The term's occurrences after linear baseline scaling.

(c) An inferred topic corresponding to the pattern.

Figure 4: The term's visual comparison.

Having a basis for a scalable inference, we transfer the identified data properties into the synthetic corpus generation. Before introducing the generative process, recall that our dynamic topic treatment is unique with respect to every document. Therefore, contrary to the reviewed dynamic topic models, our dynamic segment consists of only one document. Applying the latter aspects, we establish our utilised generative process given in Algorithm 1 below:

Algorithm 1 The generative process for a synthetic corpus.

```
for $t \leftarrow 1, T$ do
    $N \sim \text{Poisson}(\xi)$
    for $n \leftarrow 1, N$ do
        $z_{t,n} \sim \text{Multinomial}(\pi(\alpha_t))$
        $k = \{i : z_{t,n,i} = 1\}$
        $w_{t,n} \sim \text{Multinomial}(\phi_k)$
    end for
end for
```

In practical settings, the rationale of the generative process is defined as follows: the ξ term represents an approximate number of words per document; α_t is the pre-defined auto-regressive hyper-parameter for document t; and ϕ_k is the pre-defined vocabulary term distribution for topic k.

6. EXPERIMENTS

In this section, we assess the research problems introduced in Section 3:

- The spatial smoothing application for recovering the underlying vocabulary term distributions;
- The auto-regressive model's performance in terms of identifying the noise topic.

At the start of the section, we define the settings for tuning the topic models; then, we look into the settings for generating the synthetic datasets; afterwards, we introduce the scope of our experiments. Having defined the settings, we provide several illustrative examples of the experiment execution; and finally, we show the results of all experiments.

6.1 Pre-experiment Settings

The performance of the experiments is assessed by running the auto-regressive and non-auto-regressive topic models in parallel. That is, we run the topic models with and without the pre-set assumption of spatial smoothing. For both models, we tune the variances corresponding to the α update introduced in Section 5. Recall that the variance δ^2 is used to propose a new α hyper-parameter's state; the variance σ_0^2 is used to initialise α_0; and the variance σ^2 controls spatial smoothing. Effectively, in the non-auto-regressive model, we do not have the σ^2 term as all α terms are initialised using σ_0^2 (this notion relaxes the assumption of spatial smoothing).

In order to run the experiments in an efficient manner, we need to identify optimal values of the previously noted variances. One reason behind the variance tuning corresponds to the rate of convergence upon the application of the MH algorithm. Based on the algorithm's rationale – a low acceptance rate indicates a slow and stable convergence, whereas a high acceptance rate indicates a random and unstable convergence – we would find the variance δ^2 inducing the acceptance rate of around 30%. Another reason behind the variance tuning is related to the spatial smoothing application. Most importantly, we want to keep the variance σ^2 in tact with the rate of change of the topic smoothing throughout our data. Furthermore, since our topic development is captured in discrete space, we want to make sure that the discretisation step induced by the generative data process is smaller than the σ^2 variance; otherwise, we would fail to capture the high rate of change induced by steep topic changes.

Speaking of the datasets generated for the experiments, these are designed to reflect three following aspects: the effect of overlapping topics; the effect of overlapping vocabulary terms; and the effect of noise. Note that our assessment is based on an intuitive 3 topic scenario: 2 topics model distinct metabolite entities, and the remaining topic models the noise topic. Speaking of the dataset size, we set $T = 50$ for the number of documents per corpus and $\xi = 100$ for the number of words per documents: the choice of T surpasses the discretisation concern; also, as suggested by the generative algorithm given in Subsection 5.3, the use of the ξ parameter establishes a slightly varying number of words in each document. However, in order to speed up the inference, we normalise the number of words per document to possess the maximum value of 50.

In the preceding figures, we illustrate the variations of the data generation settings: in Figure 5, we display the setting controlling the topic overlap; in Figure 6, we display the setting controlling the topic term overlap; and in Figure 7, we display the setting controlling the error overlap. Note that the red and green colourings indicate the synthetic metabolite topics; the green colouring indicates the noise topic; and, for the term names set in the horizontal axes of Figures 6 and 7, we use arbitrary, unique numbers.

Relating the latter settings to our experiments, we assessed their all eight possible permutations. To give an example of a permutation, one of the experiments would assess the ability to recover the underlying topic term distributions with the enabled topic overlap, the disabled topic term over-

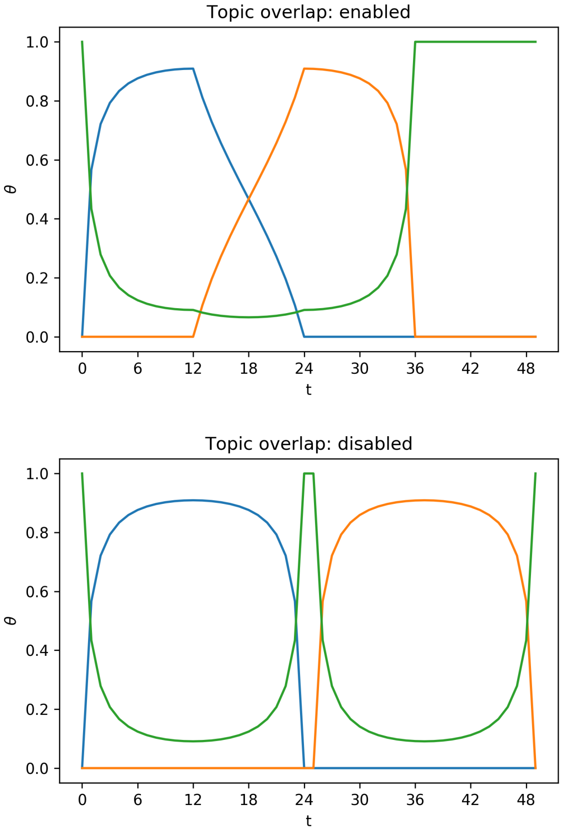


Figure 5: Controlling the topic overlap.

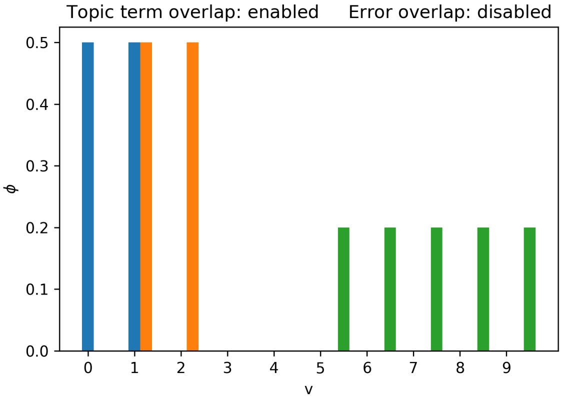


Figure 6: Controlling the topic term overlap.

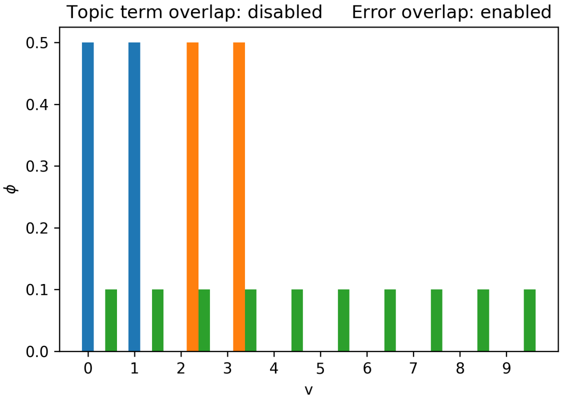


Figure 7: Controlling the error overlap.

lap, and the enabled error overlap. Speaking of the data generation, note that the latter settings directly reflect the θ and ϕ values of the synthetic datasets (we do not use the α and β hyper-parameters). By following the latter principle, we establish a clearer representation of the synthetic data; thus, simplify the performance assessment.

6.2 Experiment execution

Before going into the experiment execution, note that we assess the performance based on the models' ability to recover the underlying synthetic corpus generation settings. To wrap this assessment into a more concise terminology, the true solution corresponds to the underlying synthetic corpus generation settings; and the approximate solution corresponds to the inference results. As a result, the performance is measured by taking the difference between the true and approximate solutions.

In order to introduce the rationale behind the performance assessment, we look into one of the eight experiments in more detail. Just like for all our experiments, we run both auto-regressive and non-auto-regressive models for 5000 Gibbs sampling iterations, 1000 of which are dedicated to the burn-in process. For each of the remaining 4000 iterations, we sample the corresponding θ and ϕ values; afterwards, in every 100th iteration, we average the stored θ and ϕ values, respectively; then, this average is compared to the true solution. As an example, in the 1100th iteration, we would take the average of 100 samples; in the 1200th iteration, we would take the average of 200 samples. In a single experiment, we would have 40 of such batches indicating the performance – this is illustrated in Figure 8 and Figure 9.

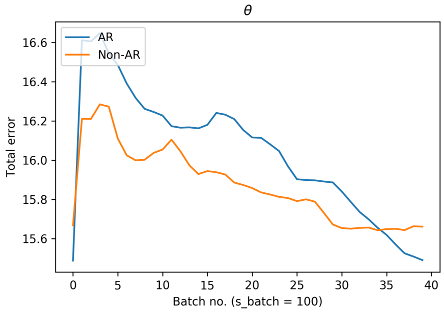


Figure 8: The θ recovery performance.

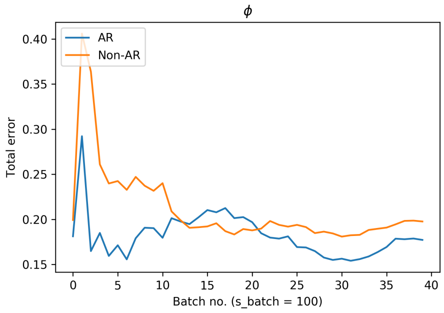


Figure 9: The ϕ recovery performance.

Relating to the previous example, the θ and ϕ values corresponding to the last iteration are illustrated in Figure 10 and Figure 11, respectively. Also, note that we relax the colour coding of our figures as the topics are inferred in unsupervised manner.

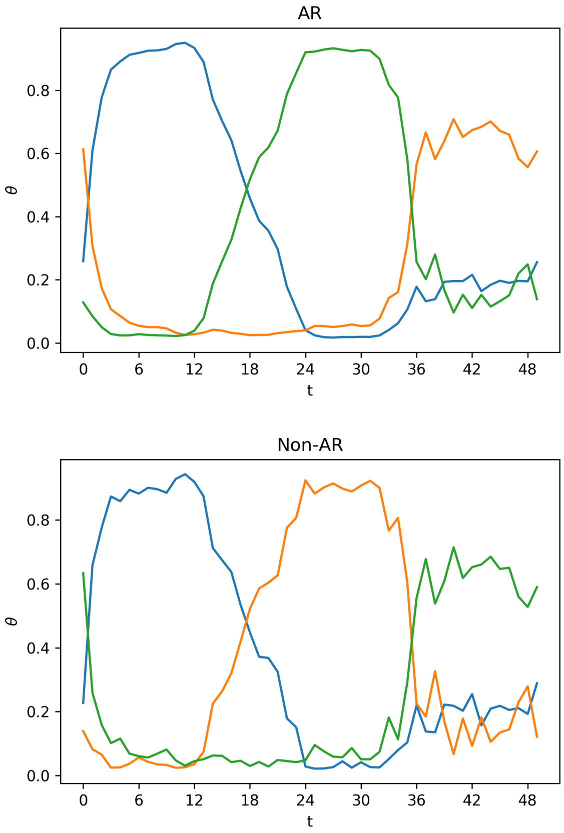


Figure 10: The comparison of the θ values.

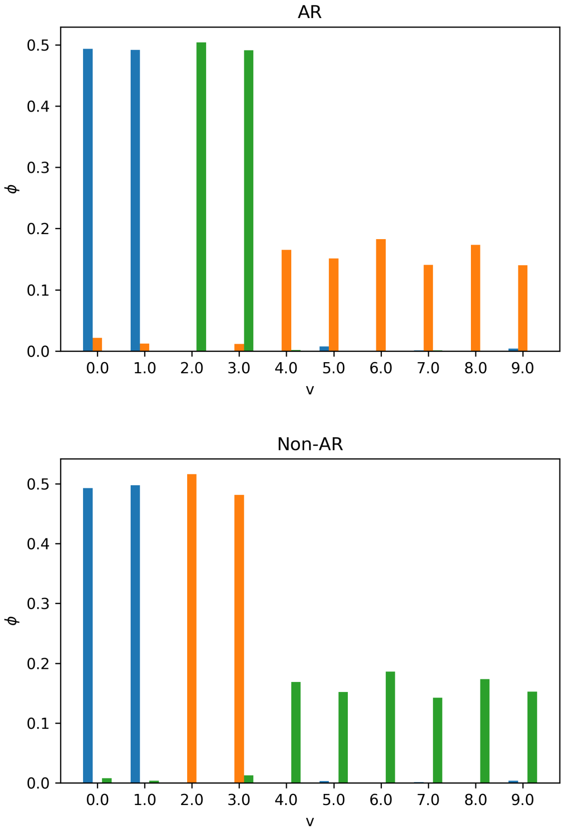


Figure 11: The comparison of the ϕ values.

To assess the auto-regressive model's performance, we have generated 10 distinct datasets for each previously introduced corpus generation setting. Recall that we assess the following conditions: the topic overlap; the vocabulary term overlap; and the error term overlap. For every setting permutation, we perform the two-tailed t-test: the t-statistic suggests the difference in performance; and the p-value suggests whether the result is significant. To be more specific, a negative t-statistic indicates the auto-regressive model's superior performance; whereas the results are determined to be significant if the p-value is below 0.05. The t-statistics and p-values of all eight performed experiments are provided in Table 1 below.

Overlap			t-statistic		p-value	
Topic	Term	Error	θ	ϕ	θ	ϕ
False	True	True	-5.41	0.47	0.00	0.65
False	True	False	-9.46	-0.38	0.00	0.71
False	False	True	4.74	3.93	0.00	0.00
False	False	False	-2.91	-3.24	0.02	0.01
True	True	True	-5.99	-1.71	0.00	0.12
True	True	False	-1.78	-1.53	0.11	0.16
True	False	True	1.12	1.17	0.29	0.27
True	False	False	0.52	1.10	0.61	0.30

Table 1: The results of the t-test.

Based on the obtained results, the most significant changes (the p-values vary from 0.00 to 0.01) occur upon only switching the error overlap setting (the remaining settings are disabled). If the error overlap is disabled, the spatial smoothing application displays an improved performance (i.e., 3.24 lower error in recovering ϕ); however, if the setting is enabled, the auto-regressive model performs poorly (i.e., 3.93 higher error in recovering ϕ). Further, by relaxing the significance threshold, we can also consider the experiment instances where the p-values vary from 0.12 to 0.16. Conveniently, in this experiment pair, we again consider only the switch of the error overlap setting; however, in this case, all other settings are enabled. To comment on the respective performance, the spatial smoothing application is superior in both cases: 1.71 and 1.53 lower error rates in recovering ϕ.

Interestingly, we can group the experiment listings in four pairs: the pairs are centred around the ϕ p-values of 0.01, 0.16, 0.30, or 0.71. The first are pairs presented in the previous paragraph; however, for the last two pairs, the p-values are well beyond the significance threshold. By noting that, in every pair, only the error setting varies, the insignificant results occur when one of the topic and term settings is disabled and another is enabled. By looking into the ϕ recovery plots related to the insignificant results, we noticed that both auto-regressive and non-auto-regressive models infer similar latent ϕ distributions. Effectively, in the case of the enabled error setting, both models simplify the dataset complexity; that is, the models assign the overlapping error terms to the main topics. Alternatively, in the case when the error setting is disabled, the problem is too simple – both models recover the ϕ values equally well. To give an example of a similar performance, the reader can consider the previously introduced Figure 9 and Figure 11. However, by considering Figure 9, note that the ϕ value of the auto-regressive model converges faster.

7. CONCLUSION

In this research paper, we reviewed an attempt to induce spatial smoothing in MSI data: the research problematic was supported and inspired by covering the relevant literature; the model's design was introduced by providing the preliminary knowledge covering LDA, spatial smoothing, and MSI data pre-processing; finally, the experiment settings were designed to identify both superior and inferior spatial smoothing application prospects.

Our main objectives were to identify the spatial smoothing application's prospect in recovering the ϕ values used upon the generative data process and the ability to separate the noise topic. Speaking of the ϕ values, only a half of the carried experiments displayed significant performance compared to the topic model without the spatial smoothing application. We report an improved ϕ recovery performance on the synthetic datasets with the enabled topic and terms overlap settings; alternatively, when both topic and term overlap settings are disabled, the performance is superior if the error overlap is disabled and inferior if the error overlap is enabled.

Speaking of the overlapping noise topic's identification, the auto-regressive model – just like the non-auto-regressive model – assigns the overlapping error terms to the main topics. For this reason, we conclude that the spatial smoothing application's impact in improving the overlapping noise topic terms is negligible. However, looking into the statistical test on the θ values, 6 out 8 experiments display a significant performance in recovering the synthetic corpus generation settings. In 5 out 6 cases, the θ values are recovered with a lower error; this result is mostly impacted by the model's ability to reflect the shape of the noise topic with a better accuracy.

Since some of the experiment settings display a performance improvement, the spatial smoothing application can be considered for a further research. We would recommend looking into the application of the undirected graphical model – Markov random field. Effectively, the application would establish more complex spatial smoothing settings: the spatial treatment of neighbouring entities could be improved from 1-dimensional to 2- or 3-dimensional. Effectively, the spatial dimensionality escalation would reflect the visual aspect of MSI data better.

Speaking of the alternative research directions in assessing the spatial smoothing application, we propose a research problem on investigating bucketisation enhancements. In other words, the spatial smoothing application might have an impact upon the feature extraction from MSI data. To be more specific, spatial smoothing might establish more appropriate bucket size ranges in concatenating raw mass-to-charge values. If successful, this application would improve the quality of MSI data features and, consequently, improve the performance of the research problems addressed in this research project.

To give a final verdict on the spatial smoothing application's performance, we consider the synthetic data settings which reflect the metabolomics-like environment the best. Effectively, if more overlap settings are enabled, then the synthetic corpus reflects the environment better. As shown by Table 1, the auto-regressive model tends to perform better than the non-auto-regressive model when most of the overlap settings are enabled. Ultimately, spatial smoothing is effective upon considering more complex data.

8. REFERENCES

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

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

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

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

[5] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[6] A. Craig, O. Cloarec, E. Holmes, J. K. Nicholson, and J. C. Lindon. Scaling and normalization effects in nmr spectroscopic metabonomic data sets. *Analytical chemistry*, 78(7):2262–2267, 2006.

[7] T. De Meyer, D. Sinnaeve, B. Van Gasse, E. Tsiporkova, E. R. Rietzschel, M. L. De Buyzere, T. C. Gillebert, S. Bekaert, J. C. Martins, and W. Van Criekinge. Nmr-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm. *Analytical Chemistry*, 80(10):3783–3790, 2008.

[8] L. Du, W. Buntine, H. Jin, and C. Chen. Sequential latent dirichlet allocation. *Knowledge and information systems*, 31(3):475–503, 2012.

[9] T. L. Griffiths and M. Steyvers. Finding scientif ic topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[10] S. M. Kohl, M. S. Klein, J. Hochrein, P. J. Oefner, R. Spang, and W. Gronwald. State-of-the art data normalization methods improve nmr-based metabolomic analysis. *Metabolomics*, 8(1):146–160, 2012.

[11] A. Palmer, P. Phapale, I. Chernyavsky, R. Lavigne, D. Fay, A. Tarasov, V. Kovalev, J. Fuchser, S. Nikolenko, C. Pineau, et al. Fdr-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nature Methods*, 2016.

[12] R. Smith, A. D. Mathis, D. Ventura, and J. T. Prince. Proteomics, lipidomics, metabolomics: a mass spectrometry tutorial from a computer scientist's point of view. *BMC bioinformatics*, 15(7):S9, 2014.

[13] J. J. J. van der Hooft, J. Wandy, M. P. Barrett, K. E. Burgess, and S. Rogers. Topic modeling for untargeted substructure exploration in metabolomics. *Proceedings of the National Academy of Sciences*, page 201608041, 2016.