ABSTRACT

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

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 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 contains information about ionised metabolites. Note that metabolites are molecules produced by the chemical process of metabolism; whereas by ionisation, we refer to the method used to sample metabolites. 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-over-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 a topic model is a statistical approach to perceive real metabolomics data, the utilised topic models are tuned to reflect the metabolomics environment as good 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 types 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 results; 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. [1]. 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 the 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 can be expressed in the form of posterior probability distributions. However, note that we can not analytically compute such posterior expressions. Instead, we can estimate the 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 [2]. The authors have utilised a collapsed Gibbs sampling – a Markov chain Monte Carlo (MCMC) algorithm – in order integrate the uncertainty out and sample only the parameters of interest. 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; and 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 the 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: *the corpus* is the whole image of a sample; *the document* is an MSI image's pixel; and *the word* is an interval of mass-over-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 corresponds 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 T number of documents, and each document has N number of words.

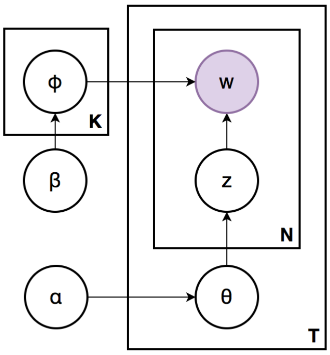

Figure 1: The initial LDA model's architecture.

Before providing a listing with the definitions of the LDA variables, we will briefly 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 to tune 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 document t;
- ϕ_k is the vocabulary term distribution over 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 will introduce the qualities of the raw MSI data. Furthermore, we will 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, we establish the basis of our MSI data by working on the data in the mzML format. Conveniently, we apply the `pymzml` Python library to parse the MSI data.

The raw MSI data contains a collection mass spectrum with the mass and the intensity values of each captured ion. Effectively, each mass spectrum corresponds to a particular pixel of an MSI image. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

3. STATEMENT OF PROBLEM

To start with, we set the hypothesis of this research project to *'The noisiness of MSI pixels can be reduced by applying a topic model tuned for spatial smoothing'*. By spatial smoothing, it is meant that the topic model would have an auto-regressive treatment among the pixels. 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;
- Reducing the noisiness of MSI data;
- Motivating the further research in applying spatial smoothing to MSI data.

Speaking of the overlapping topic detection and the reduced noisiness, both contributions would improve the performance of MSI pattern recognition. Additionally, we measure the changes in the performances upon varying the complexity of the data. Ultimately, if a naive spatial smoothing application displayed performance improvements, we would set a basis to apply state-of-the-art auto-regression approaches on 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 will serve as an exploratory assessment on the spatial smoothing application: we will introduce the rationale behind the applied methodology; also, we will clearly define the range of the experiment settings. To give a brief intuition about the methodology, the study will assess the domain-specific parameter tuning and its impact on a diverse range of synthetic datasets.

4. RELEVANT RESEARCH

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

5. METHODOLOGY

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

6. EXPERIMENTS

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

7. CONCLUSION

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

8. REFERENCES

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

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