Imperial College London
Department of Surgery and Cancer

Applications of generative probabilistic models for information recovery in ^{1}H NMR metabolomics

Edward Leon Zukowski

Submitted in part fulfilment of the requirements
for the degree of Doctor of Philosophy at
Imperial College London, August 2018

I declare that this thesis and the research in it are my own work. Any work of others is acknowledged and appropriately referenced.

Abstract

Metabolomics is a well-established approach for investigation of the metabolic state of an organism usually conducted via high-throughput methods and focusing on quantification and identification of small molecules. A popular analytical technique used in metabolomics is ^{1}H NMR spectroscopy. The data obtained in NMR experiments contains a wealth of information on metabolites in a sample and their chemical structure. To help uncover this information and find patterns in the data, statistical and machine learning methods must be applied.

The work presented in this thesis demonstrates applications of probabilistic generative modelling, with particular focus in Latent Dirichlet Allocation (LDA), as a tool for information recovery in ^{1}H NMR data sets obtained in metabolomics research. LDA is an example of a topic model. The model is based on a generative process which can be thought of as a source of the data. Topics are latent variables which select co-occurring metabolites in a sample. In turn, NMR spectra can be represented in the latent variable space.

We present applications of LDA in three scenarios. (1) How LDA can be used to simulate NMR spectra; such spectra demonstrate that LDA is a valid model for NMR data and also provide synthetic data for evaluation of statistical models. (2) Unsupervised learning with LDA to uncover patterns in the NMR data; we use synthetics and real NMR data with knowledge of key biomarkers from a prior study and conclude that LDA was successful in the recovery of useful topics. (3) Supervised learning with SLDA and combined latent variable models with ElasticNet regression where we investigate NMR data from The Multi-Ethnic Study of Atherosclerosis (MESA) study which is paired with clinical variables such as BMI. The goal was to examine if topics can be informative about clinical outcomes.

Acknowledgements

I would like to thank my supervisors Tim Ebbels and Naomi Chayen. Tim has my gratitude for many hours of our discussions and his enthusiasm to share his knowledge, all of which had a genuinely positive effect on my research. Naomi was an invaluable source of continuous support, encouragement and feedback, especially with the final version of this thesis. Also, I would like to thank members of the Computational and Systems Medicine Section, especially Jia, Caroline, Kirill and Gonçalo. My friends Kitty, Andy, Albert, Joe, Bogdan and Steve gave me great support and advice. Last but not least, I thank my family for encouragement, love and faith in my academic venture, especially Mum, Króliczek and Aunt Grażyna.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this chapter, we provide background material referred to or otherwise used in the subsequent chapters. We begin by introducing metabolomics and NMR spectroscopy, followed by probabilistic and latent variable models. We close with the aims and objectives of this thesis as we all a short overview of the following chapters.

1.1 Metabolomics

Metabolic profiling, also called metabolomics and metabonomics refer to a subdomain of systems biology which investigates the chemistry of metabolism (Nicholson et al., 1999; Fiehn, 2002; Nicholson and Lindon, 2008). Definitions of metabolomics and metabonomics originally varied slightly but in recent years the terms are used interchangeably, and this is how we treat them in this thesis. The investigation is usually conducted via high-throughput methods, focusing on quantification and identification of small molecules. The small molecule is defined as having a molecular mass less than 1500Da. A defining strength of metabolomics is an abil-

ity to capture a snapshot of all metabolic activities of an organism. A metabolome is a complete set of metabolites in a specific type of sample (usually bio-fluid) of a particular organism, for example, The Human Urine Metabolome (Bouatra et al., 2013). Some of the metabolic activities are governed by genes, some by interactions with the environment (for instance, ethanol in human bio-fluids is present if alcohol was consumed) and some by biological conditions such as a disease. Additionally, metabolomics can be used to capture dynamic changes in metabolism over multiple time-points, although due to cost the number of time-points is usually small (<20). In summary, metabolomics adds to genomics and other omics another view at the inner workings of a biological system, and it is a valuable tool in multiple domains, including but not limited to toxicology (Coen et al., 2008), molecular epidemiology (Holmes et al., 2007), personalised medicine (Nicholson, 2006), plant genomics (Roessner et al., 2001; Fiehn et al., 2000), and clinical research for diagnostic applications (Gowda et al., 2008).

The most common analytical methods in metabolomics are mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy. In this thesis, we focus exclusively on ^{1}H NMR spectroscopy. ^{1}H NMR is very well suited for biological samples, usually bio-fluids because it detects molecules containing protons (^{1}H) which are to be found everywhere in biochemistry. NMR is very good at providing structural information about the molecules. Usually NMR spectroscopy is used as an untargeted technique i.e. there is no filtering to some set of interesting metabolites and NMR resonances are coming from all metabolites present in a sample. One of the major disadvantages of NMR is low sensitivity, a sample must be provided of a certain mass or sometimes concentrated. There are many challenges in

NMR metabolomics data. Some resonances come from known metabolites that we expect to be present in a sample, and others come from unidentified molecules. This makes it challenging to identify peaks in NMR spectra. The problem of identification is magnified by the fact that peaks from different molecules can overlap. A high number of peaks[1], among those many, are correlated, and noise present in the signal is inherent to the nature of NMR spectroscopy. Another challenge can be a relatively small number of subjects. This is usually caused by the cost and complexity of setting up experiments with live subjects. High variability in the responses from individuals can be observed even if the replicates are highly homogenous (Nicholson et al., 2002).

Metabolomics and NMR data mining methods

To gain biological insight from complex patterns found in NMR data, both unsupervised and supervised machine learning methods are frequently employed. Principal component analysis (PCA) is an unsupervised method which is routinely applied to metabolomics data. PCA is one of the oldest unsupervised methods; its origin can be found in Pearson (1901). A modern description can be found in Wold et al. (1987) and Jackson (2003). We give an overview of PCA later in this chapter. Among supervised methods, partial least squares PLS (Wold et al., 2001) remains a popular choice. In this thesis, in Chapter 4, we use another supervised method called ElasticNet (Zou and Hastie, 2005) which is a doubly regularised regression method that combines penalties of Lasso and Ridge regression. This thesis proposes to expand unsupervised and supervised tool-kits used in NMR metabolomics

[1]Around 200 metabolites can be identified in human urine using NMR, Bouatra et al. (2013).

by applications of topic models in particular Latent Dirichlet Allocation.

Challenges in metabolomic NMR data processing

The primary source of variation in NMR data comes from the proton resonance when it relaxes from the excited state to the equilibrium. We describe the principles of NMR in the next section. However, sources of variation can also be technical, coming from NMR equipment, or biological factors, e.g. pH of the sample. The former are usually corrected by the equipment software, the latter need careful consideration and preprocessing. For example, Nicholson et al. (2002); Dieterle et al. (2006) note that in urine samples, concentrations can vary significantly between individuals even if they are not subjected to any biological challenge. To reduce the concentration variability between samples, we apply normalisation. The standard procedure is to divide all spectral bins in a sample by a sum of all the intensities. Another variant of normalisation could be achieved by dividing each sample by the area under the spectrum. The area can be approximated by use of the trapezium rule. There are more sophisticated methods of normalisation in NMR such as probabilistic quotient normalisation described in Dieterle et al. (2006). Another preprocessing technique common in metabolomics is centring and scaling. Those operations are performed on spectral bins across all samples. The most common centring technique is by subtracting the mean from each bin. We apply this approach when we fit PCA models. Scaling is usually applied to account for different concentrations in biological samples by normalising the NMR spectra to a unified, virtual concentration. The usual method of scaling is to divide all the values in given spectral bins across all samples by the standard

deviation. Lastly, metabolomics samples, urine, in particular, will expose shifts in peaks between samples due to, mainly, differences in pH but also due to other factors like temperature. This problem can be addressed but defining broad spectral bins and calculating the area under the spectrum or applying a special procedure on full resolution data to move peaks, so they are aligned between samples. Two examples of such algorithms are the Recursive Segment-Wise Peak Alignment by Veselkov et al. (2009) and metabolite deconvolution and quantification from complex NMR spectra by using the Bayesian Automated Metabolite Analyser for NMR (BATMAN) by Hao et al. (2014).

1.2 ^{1}H NMR spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy is a broadly used chemical analysis technique with applications in medicine and biochemistry (Hausser and Kalbitzer, 1991) including metabolomics (Nicholson et al., 1999). NMR is possible with any nucleus which has a non-zero nuclear spin quantum number, for example, ^{1}H or ^{13}C. ^{1}H NMR spectroscopy is one of the major[2] experimental techniques in metabolomics as hydrogen is universally present in metabolic molecules. The methods described later in this thesis have only been applied to ^{1}H NMR data, therefore, this section focuses only on ^{1}H NMR spectroscopy.

NMR is a mature technique. The method for measuring magnetic nuclear moment, which is a cornerstone of NMR, was first described by Rabi et al. (1938). The significance of the discovery was quickly recognised as NMR produces a lot of information on chemical structure. Rabi was

[2]Mass spectrometry (MS) is another major one.

made the Nobel Prize laureate in physics in 1944. Later, Bloch and Purcell worked on the extension of NMR to liquids and solids which earned them the Nobel Prize in Physics in 1952. Since then, NMR has been continuously perfected. Contemporary apparatus is capable of performing 1GHz NMR. Such a strong static magnetic field allows obtaining high-resolution spectra which is particularly useful for investigating complex molecules as it reveals distinct signals which otherwise would overlap.

NMR requires minimal sample preparation, usually only pH buffering as signal shifts can occur due to sample acidity. NMR is a non-destructive technique allowing samples to be reused in subsequent experiments. Moreover, live samples are allowed, for example, most hospitals will routinely perform magnetic resonance imaging (MRI) scans on humans. This section focuses on the fundamental principles of NMR, which is mostly based on textbooks by Hore (2015) and Keeler (2011).

Nuclei have a mass, a charge and a spin[3]. Atomic spins come in different forms, but for our purposes, the interesting quantum spin number is ½. Examples of nuclei with $I = ½$ are ^{1}H and ^{13}C. According to $2I + 1$ rule[4], a nucleus with the quantum spin number $I = ½$ has only two energy levels labelled with the magnetic quantum number $m = -½$ and $m = ½$. Some nuclei do have $I = 0$ and thus will not have a magnetic spin. These can not be used in NMR (there are no distinct energy levels), a notable example is ^{12}C. The protons at each energy level have a unique spin orientation (see left panel in fig. 1.3).

Let us now address the behaviour of an atomic nucleus in a magnetic

[3]We use spin, angular momentum and magnetic dipole moment as interchangeable terms.

[4]$2I + 1$ rule determines the number of energy levels for nuclei with a quantum spin number I.

field. We start with a simple analogy. A compass needle without the earth magnetic field would point in a random direction. The earth's field applies energy to the needle, so it points along the direction of the earth's magnetic field. We can apply some external energy (e.g. with a finger) to change the needle's direction. If we let go of the needle, it will return to its original state ("ground state"). A proton could be thought of as a tiny magnet. Without a static external magnetic field, its spin would point in a random direction. After application of some external static magnetic field usually called B_0 in NMR literature, the proton's spin will align with the direction of B_0. There are only two possible spin orientations for a proton: parallel or anti-parallel to B_0.

The intensities of the NMR signal depend on the differences in the populations of the energy levels. Those differences are determined by the Boltzmann distribution which depends on the sample temperature, the applied field B_0 and the gyromagnetic ratio γ specific for a proton. If we fix temperature to, for example, room temperature, an increase in B_0 will add to spin excess (more nuclei in ground state). In NMR, the parallel and anti-parallel spins cancel out, on average, and only an excess of nuclei in the ground state is detectable.

A basic NMR experiment consists of three main phases: initial static magnetisation, radio frequency (RF) pulse and spin relaxation. Firstly, a sample at room temperature is placed in an external homogeneous magnetic field B_0. This will make the protons arrange themselves in two split energy states. The spins will either be oriented parallel (ground state) to B_0 or will be oriented anti-parallel (excited state) to B_0. The nuclei precession is about the axis of B_0, and its frequency is called Larmor frequency.

In the next step, the sample is hit with an electromagnetic pulse of a broad range of radio frequencies (RF). This event will equalise the number of protons in each energy state. The aggregate magnetisation vector **m** is a sum of all the parallel and anti-parallel spins. To get **m** in a 90° plane we need to apply RF pulse for the right amount of time, usually very short, e.g. 1ms. Because the RT pulse aims to force the spins to precess in a 90° plane to the B_0 it is called 90° pulse.

After the RF pulse, there is a relaxation period in which the protons go back to the equilibrium arrangement. There will be slightly more protons in the ground state, in order of one in 10^5 at room temperature. If populations were equal, we would describe the system as saturated or demagnetised, and no NMR signal would be observed[5]. Those "returning" to equilibrium protons will emit energy in the form of photons of a characteristic frequency (known as resonances) which can be detected by the spectrometer with a coil which detects induced voltage from the procession of spins in the ground state. This voltage gives a free induction decay (FID) signal, see Figure 1.1.

The human eye is not a good instrument to see the frequencies in FID plot. Fourier transform is used to convert the signal from the time domain to the frequency domain. The result can be plotted as a spectrum, i.e. frequencies are on the x-axis, and intensities are on the y-axis. In NMR x-axis is a chemical shift with zero chosen arbitrarily, usually a frequency of a special internal standard molecule. The intensities are proportional to the number of protons in particular chemical environments, but the values on the y-axis are not of any natural meaning as the area under the curve is proportional to a number of nuclei at each frequency. Also, NMR

[5]We could think of all parallel and all anti-parallel spins cancelling out each other.

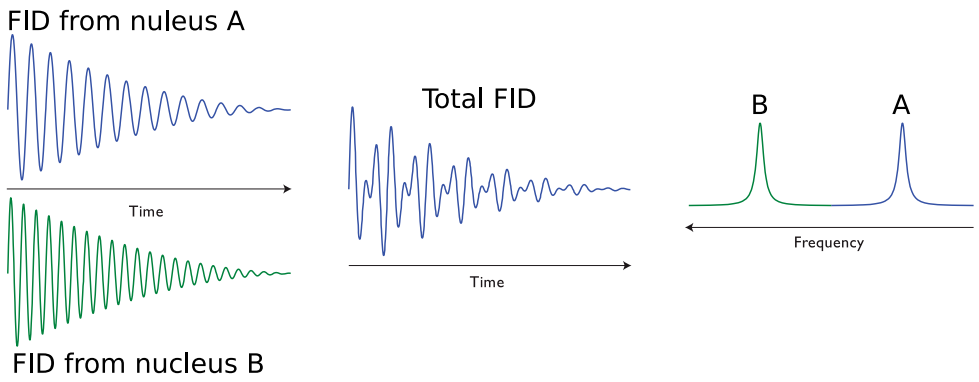

Figure 1.1: Illustration of a free induction decay transform from the time domain to frequency domain. On the left, there are FID from two hypothetical nuclei A and B. Nucleus A has slightly lower frequency then nucleus B. In this case, the frequencies are exaggerated for illustration purposes, real FIDs are not readable by the human eye. The middle plot illustrates combined, total FID from all nuclei in a sample. The right panel illustrates the signal converted from the time domain to the frequency domain. The transformation is obtained via Fourier transform. FID is a complex number with real and imaginary parts; here we plot only a real part. We can observe a higher frequency from nucleus B on the left and lower frequency A on the right. Figure adapted from Pearce (2010)

spectra are often normalised therefore y-values will bear different meaning depending on the preprocessing.

With time the procession induced by 90° pulse dies away, and the nuclei come to their equilibrium. This process is called relaxation. There are two types of relaxation: T1 and T2 relaxation. The former is known as spin-lattice relaxation and it is related to nuclei returning to the equilibrium state; T1 affects the height of a peak. The latter is known as spin-spin relaxation; it is a random process in which the exchange of energy between spins causes loss of magnetisation; T2 affects the width of a signal.

Chemical shift is essential to NMR applications in metabolomics and chemistry in general. A proton bonded to another nucleus will have electrons around it from its chemical bond. Those electrons will induce electronic currents leading to a tiny local magnetic field. This local field shields or

de-shields the proton from external field B_0. This effect will change the resonance frequency of such proton to be slightly different from just a bare proton, and it may be higher or lower depending on the electron density resulting from a bond configuration. The phenomenon of shifting of the resonance frequency is called chemical shifts[6].

Chemical shift gives information about the chemical structure, and it is a key to the identification of molecules. All protons from identical environments are chemically equivalent. Their specific bonding configuration makes them indistinguishable from each other; for example, it would be impossible to number them in any meaningful way. All protons from a particular environment will give one NMR signal. In practice, chemical shifts are frequently considered as empirical parameters. NMR experts can recognise a "fingerprint" of a molecule just by looking at a spectrum and identify it without analysing why a chemical shift occurred precisely in this unique way.

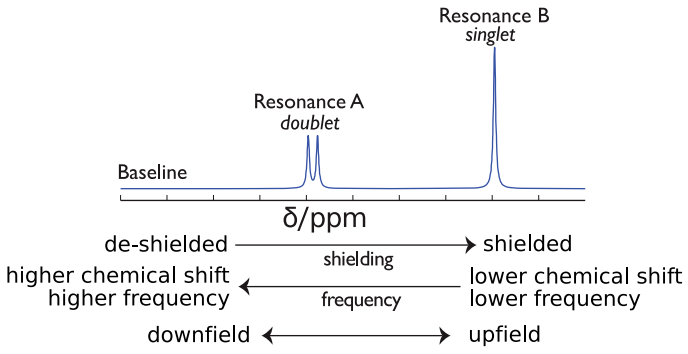

Figure 1.2: NMR spectrum and associated terms. It is important to note that the NMR spectrum is labelled from right to left (i.e. zero is on the right). Terms downfield and upfield come from the physics interpretation of NMR. Figure adapted from Pearce (2010)

NMR spectrum x-axis represents resonance frequencies of protons present

[6]Also called nuclear shielding because the electrons shield B_0.

in the sample. Those depend on the chemical environment of each proton. NMR apparatus has a characteristic frequency at which it operates and with it an associated strength of B_0; both parameters affect the horizontal axis scale. The x-axis could be labelled with the frequencies expressed in Hz, but this would be impractical. The chemical shifts are very small compared to the characteristic frequency of the apparatus so comparison of the spectra from machines operating at different frequencies would be inconvenient. A remedy for these issues is to express x-axis as δ, a dimensionless scalar which traditionally, is labelled in ppm (parts per million). δ quantifies the extent of nuclear shielding. To emphasise that ^{1}H proton is used as the NMR isotope, we sometimes write the x-axis label as $\delta(^1\text{H})$.

Chemical shift scale is always expressed in relation to a reference molecule and adjusted to zero at a resonance frequency of the reference molecule. Two molecules, tetramethylsilane (TMS) and 3-(Trimethylsilyl)propanoic acid (TSP) are popular choices for ^{1}H NMR. Other nuclei will have different standards. What makes TMS and TSP good choices for a zero point is that the protons on both molecules are more shielded than in many organic compounds. The more shielding, the lower the frequency (see fig. 1.2) so TMS and TSP show far on the right of the δ scale. When reading NMR spectra, it is worth noting what reference molecule was used as this will affect the values of δ, but fortunately only slightly as TMS and TSP are related therefore they are close on the chemical shift scale.

The formula for chemical shift is $\delta = 10^6(\frac{\nu_0 - \nu_{ref}}{\nu_{ref}})$ where ν_0 is a frequency of molecule of interest and ν_{ref} is a frequency of reference molecule. For example, let us consider two compounds TMS and benzene. If we use 300MHz spectrometer, a proton on benzene absorbs a frequency of 2181Hz

more than the protons on TMS. If we use 60MHz spectrometer then the protons on benzene absorb 436Hz more than the protons on TMS. Let us now convert both benzene frequencies to ppm scale: $\delta = 10^6 \frac{2181Hz}{300\times10^6 Hz} \approx 10^6 \frac{436Hz}{60\times10^6 Hz} \approx 7.27$

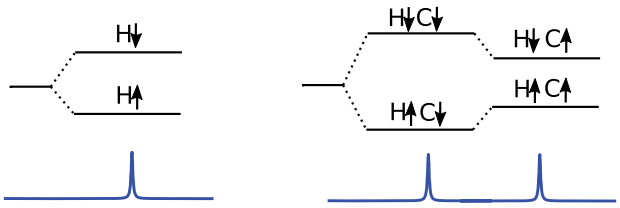

Figure 1.3: Spin-spin coupling principle. On the left, a bare proton signal. The proton is not bound, so it will give a single line. On the right, a proton which is bound to ^{13}C. The signal from the bound proton will be split into two lines, a doublet. The line split because the resonance frequency of the proton depends on the spin direction of the bound carbon nuclei. If the carbon spin is anti-parallel, ΔE will increase, and the signal will move to the left. For the proton bound to the carbon with parallel spin, the ΔE will decrease, and the signal will move to the right. Figure inspired by Hore (2015)

Lines in a spectrum always come from the energy level transitions, see Figure 1.3. Let us consider the formate ion $H^{13}COO^-$, where the proton from formic acid $H^{13}COOH$ is dissociated in aqueous solution. Energy level splitting for $HC\downarrow$ ($C\downarrow$ means ^{13}C carbon in spin configuration anti-parallel to B_0) is larger than splitting for bare proton H. When ^{13}C is in parallel spin configuration $HC\uparrow$ energy level splitting is smaller than for ^{1}H alone. Higher energy ΔE moves the NMR line to the left (higher frequency), lower ΔE moves NMR line to the right (lower frequency). In the formate ion $H^{13}COO^-$, this results in the ^{1}H NMR line from its single proton to split into a doublet.

There is a rule of thumb for determining the type of multiplet arising from a particular chemical environment of a given proton. This rule is called n+1, and it refers to a number of protons involved in a neighbouring

environment. For example, let us consider ethanol. CH_2 is split into quadruplet because it has a neighbour CH_3; CH_3 is split into a triplet because its neighbour is CH_2.

The area under the curve (AUC) is proportional to the number of protons in the signal. The amplitudes of the signal or more accurately integrals are proportional to the number of nuclei going back to the equilibrium (after an RF pulse is applied) which in turn is proportional to the total number of molecules of that type. AUC remain constant as the number of molecules in the sample did not change. If AUC is fixed, a broader line forces a smaller amplitude. It is worth remembering that the signal gets closer to the background noise when the amplitude becomes smaller. This makes the overall spectrum look much noisier.

The chemical exchange is another important phenomenon which affects the pattern of NMR peaks. There are two regimes: slow exchange and fast exchange. Free induction decay (FID) happens on a certain timescale which is quite slow, e.g. one second. A fast exchange takes place when the exchange rate (molecules per second) is orders of magnitude faster than the FID time. A spectrometer will register only an average signal in the fast exchange regime. In contrast, we observe slow exchange when reaction time is below an FID time scale, e.g. 10^{-1} (1/10 per second, i.e. 1 molecule per 10 seconds). In this regime, two distinct populations of molecules in two different states are present. Over the FID time scale, these two populations will give two distinct signals. In the fast exchange, in the timescale of FID, the spectrometer cannot distinguish between the two populations as the reaction is too fast. This results in observing a single, average signal.

NMR is often praised for its advantages such as minimal sample prepara-

tion, the non-destructive nature of the NMR experiment, and the ability to analyse live samples. However, there are some drawbacks to the technique. Firstly, there is the low sensitivity of NMR which is caused by the way that NMR works; the apparatus can only detect the excess of nuclei in the energy ground state. This very small difference in populations of the two energy states (in the order of one per hundred thousand at room temperature) makes NMR a low sensitive method. A solution to overcome this limitation is the use of concentrated samples. Another approach is to repeat the experiment many times in order to achieve the required signal and sufficient information.

Another limitation of 1-dimensional ^{1}H NMR is that the spectra can have heavily overlapping signals thereby making the identification of molecules very challenging or even impossible. This problem is particularly pronounced in some biological samples where large molecules are present, giving vastly complicated spectra with significant overlap between peaks. Those areas of overlap make it hard to identify particular peaks, whereas smaller peaks may be overshadowed entirely and lost to the researcher.

1.3 Probabilistic and Latent variable models

Latent variable models are statistical models in which we assume that the observed data arises from the interplay of unobserved and latent variables. In the context of ^{1}H NMR metabolomics, we observe spectra which are a manifestation of metabolites present in a sample. We know that those metabolites and their concentration are not random but a result of biological processes of an organism.

Probabilistic modelling is a framework to express uncertainty about

models in a language of probability theory. Bayesian statistics methods are used to infer latent variables. The cornerstone of Bayesian statistics is the Bayes rule, which in the context of modelling, can be written as:

$$p(model|data) = \frac{p(data|model) \times p(model)}{p(data)}$$

The interpretation of the above equation is that the probability of the model given observed data (posterior probability) is equal to the likelihood of the data given the model, times a prior probability of the model normalised by the probability of data (also called evidence). The prior probability of the model is our current belief that the model is correct. Upon observing new data, we calculate new posterior probability using previous posterior as a prior. In practice, the model usually refers to the parameters of the model which are frequently expressed as probabilities. It is, of course, possible to come up with an entirely new model but this would leave the existing model unexplored.

As a concrete example let us consider modelling of a collection of text documents where each document exhibits a number of themes or topics. A part of a model is a collection of parameters which describe how much each document exhibits each topic. For each document, we have a vector of probabilities across topics which sums up to one. Another part of the model is a collection of topics; each topic can be represented as a vector of probabilities across the whole vocabulary (again, probabilities sum up to one). In this example, the data point is a text document consisting of a list of words. Such a list can be represented as a vector of the counts of all the document's words across the whole vocabulary existing in all documents. As a very first step, we assign some, perhaps using our intuition, probabilities

for our model – we define the prior $p(model)$. Next, we calculate how likely is our document given our model $p(data|model)$ and how likely is our document under all the possible[7] model's parameters $p(data)$. Now we have all the components to calculate the posterior probability of the model. We repeat this process for all documents using the posterior from the previous step as the prior in the current step. Now we have the posterior after processing all the documents. To obtain a different (and possibly better) model, we change the parameters of the model and repeat the whole procedure. This process is repeated multiple times, each time giving us the posterior $p(model|data)$ for the specific values of the model's parameters. Finally, we choose the best model using a criterion of our choice, e.g. maximum a posteriori (MAP).

The above example is a very high-level description of Bayesian inference with its guiding principle that the models which are not consistent with data lose credibility whilst the models consistent with data gain credibility.

1.3.1 Latent Dirichlet Allocation

Topic modelling is a case of applying hierarchical Bayesian models (Jordan et al., 1999; Wainwright and Jordan, 2008) to grouped data. In this context, grouped data means a collection, for example, a collection of text documents, images, or NMR spectra. Latent Dirichlet Allocation (LDA) is the simplest topic model.

LDA is also a probabilistic model. Probabilistic modelling assumes the existence of a generative process which shapes our data. The process

[7]There can be exponentially many possible model's parameters. In practice, the marginal probability $p(data)$ might not be possible to calculate exactly, but many methods of approximation exist.

is probabilistic and random; it includes hidden (latent) variables. The latent variables drive the underlying pattern of the data. For example, in text documents, the hidden variables reflect the thematic structure of the collection. Our goal is to find this hidden structure. The way of inferring the latent variables is by using posterior probability. Probabilistic modelling is a standard tool in scientific data analysis. A recent review of the field is given by Ghahramani (2015).

It is natural to see that the topics exist outside of the document collection (also known as the corpus). Per-corpus topic distribution over the vocabulary remains constant, but per-document topic distribution varies from one document to another. A document exhibits multiple topics, and a topic can be used in many documents thus LDA is a mixed membership model (Erosheva et al., 2004). This is different from mixture models where one document can only be associated with one topic, in this case, called a component or a cluster. Mixed membership models are particularly suitable for grouped data such as a collection of documents or a collection of NMR spectra.

The LDA model was introduced by Blei et al. (2003), but a more accessible presentation can be found in the author's later review paper (Blei, 2012). The original context for LDA was natural language processing (NLP) for modelling text document collections. Topics in text documents can be intuitively understood as they answer the question of what the documents are about. For example, a document can be about metabolomics and data analysis, while another document is about genetics and evolution. Despite NLP origins, LDA is a general purpose probabilistic model, i.e. there are no underlying assumptions which would bind LDA exclusively to NLP. LDA

and its variants can be and have been, applied to all kinds of data, for example, to toxicogenomics data (Chung et al., 2015; Lee et al., 2016), bioinformatics (Liu et al., 2016), healthcare data (Lu et al., 2016), image classification and annotation (Chong et al., 2009) to name just a few. A thorough and recent review by Boyd-Graber et al. (2017) provides even more examples.

Previous to LDA, Pritchard et al. (2000) constructed a similar model for populations genetics, but their paper did not present the model as a universal method. In the paper, individuals are assigned to populations depending on their genotypes, but they could be members of multiple populations depending on their genetic heredity. Credit must be given to the LDA authors (Blei, Ng and Jordan) for their ability to recognise the broad appeal of topic modelling and, eventually, for establishing a sizeable niche around LDA in a broader field of probabilistic models and machine learning.

Figure 1.4 shows an LDA model in three perspectives, on top we have an intuitive view of a generative process for LDA, the middle represents a view from a data representation point of view (as matrices), and lastly at the bottom there is a plate notation for LDA which is a standard notation describing probabilistic graphical models (PGM).

Generative process for LDA

The generative process for LDA is a random process which serves as an assumption for the LDA model. For simplicity, let us illustrate the mechanics of generative LDA using context from the original paper by Blei et al. (2003) describing a collection of text documents which are treated as a sequence

of words. In topic modelling, the order of words in the documents is not deemed necessary. What is important is only what words are present and how often they occur in a given document. This approach is called a bag-of-words model. This is a simplification, but we do not attempt to generate human-readable text documents. Bag-of-words is an assumption which allows creating a simple model with the potential for successful Bayesian inference. Topics encapsulate co-occurring words in documents. The order of words is not essential to detect the co-occurrence.

The LDA generative process is a random process which is visually represented at the top part of Figure 1.4. The process can be defined in the form of an algorithm which is represented as a flowchart in Figure 1.5. We assume that an observed text document was created via the mechanics of this process. We further assume that each document will contain a fixed number of words N.

K topics (K=3 in Figure 1.4) are defined outside of the document collection. Figure 1.4 illustrates the process with only one document but to generate a collection of D documents, it would be repeated D times with the same K topics. Each topic is a list of all vocabulary words with probabilities differentiating between topics. Formally, a topic is a probability distribution $\boldsymbol{\beta_k}$ over the fixed vocabulary of V words. The vocabulary contains all possible words to be used in all generated documents. In topic modelling literature, topics are usually depicted as incomplete lists of words showing only most probable words. When we think about simulation, defining topics would be an initial step. Row vectors $\boldsymbol{\beta_k}$ are stacked vertically to form the matrix $\boldsymbol{\beta}$ of size $K \times V$ (rightmost in the middle of Figure 1.4).

The proportions of topics per document $\boldsymbol{\theta_d}$ (rows of the leftmost matrix

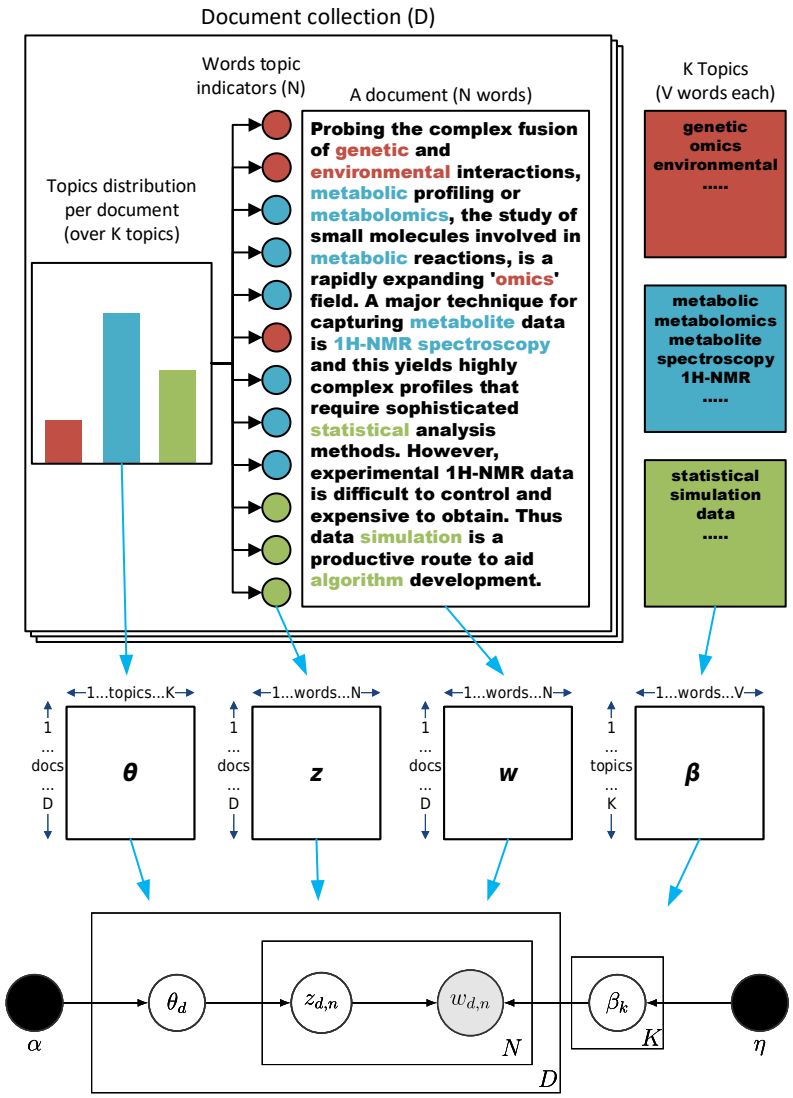

Figure 1.4: LDA model. The text from Muncey et al. (2010)

in the middle of Figure 1.4) drive the selection of topics for words in the said document. We use a Dirichlet distribution with a parameter α to obtain$\boldsymbol{\theta_d}$ for each document[8]. The histogram on the top left in Figure 1.4 represents $\boldsymbol{\theta_d}$, which is a K-vector. There are D such vectors, and they are stored as rows in matrix $\boldsymbol{\theta}$. The different proportions of topic per each generated

[8]Sometimes Dirichlet distribution is referred to as distribution over distributions.

document is a distinguishing feature of the LDA model.

Once per-document topic distribution $\boldsymbol{\theta_d}$ is determined we use it to generate topic indicators for words. In Figure 1.4, for simplicity, we show only a dozen of topic indicators which are colour coded with a topic colour. In reality, there are N topic indicators, each per every word of a document. To obtain z_n we sample from a categorical distribution with probabilities $\boldsymbol{\theta_d}$. Sampling from the categorical distribution with probabilities $\boldsymbol{\theta_d}$ can be compared to rolling K-sided die which is not fair, distribution $\boldsymbol{\theta_d}$ defines the probability of each side. Each z_n is an integer between one and K indicating a topic for each word in a document.

Topic indicator z_n gives a row index in matrix $\boldsymbol{\beta}$. Selecting k-th row gives us $\boldsymbol{\beta_k}$ which is a probability distribution itself. Sampling from this distribution yields w_n which is an integer between one and V indicating the n-th word in a document. The generative process for LDA can be expressed more formally as an algorithm. Figure 1.5 shows a flowchart for this algorithm.

Graphical models

The goal of the LDA model is to discover topics in a collection of data. The data is observed but topic structure, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are hidden. We want to infer topic structure from the observed data using a computational procedure. If we look at the generative process for LDA as a way of "constructing" the data, the inference is akin to reversing this process. We want to know the topic structure, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ which plausibly could generate the observed data.

Probabilistic modelling framework allows us to express the generative process for LDA as a joint probability of all variables, both observed and

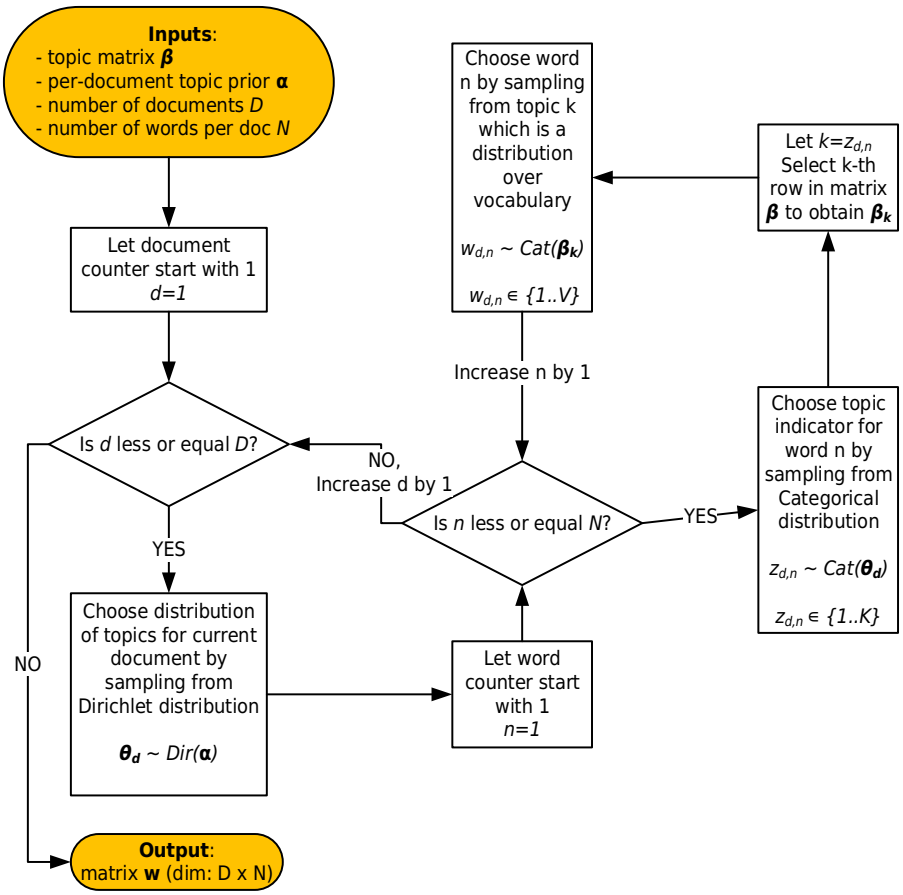

Figure 1.5: Generative process for LDA flowchart

hidden. The joint probability allows us to compute the posterior distribution which is a conditional probability of hidden variables given the observed data. The remainder of this section will give details about how to obtain the joint probability in probabilistic graphical models (PGM) and LDA in particular.

The LDA model can be described using plate notation which is a standard method of describing PGM. Joint distribution of random variables is represented as a directed graph where nodes represent random variables and edges represent conditions. Simple cases for two and three variables are

shown in Figure 1.6. A random variable at the end of an arrow is conditioned on a random variable at the start of an arrow. Plate notation helps to describe how joint probability factors out to a product of conditionals probabilities using the probability product rule also called the general product rule or chain rule (not to be confused with chain rule from calculus).

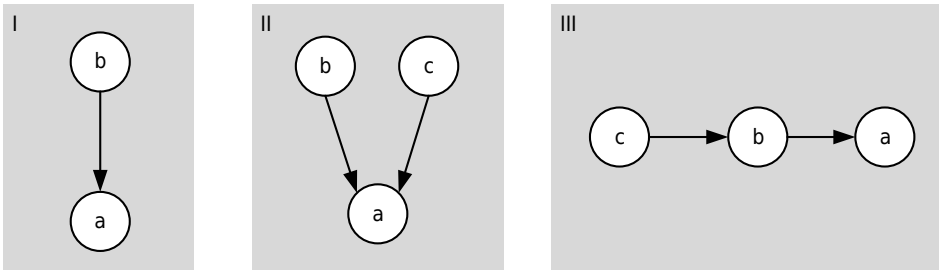

Figure 1.6: Simple graphical models. Nodes represent random variables, edges represent conditioning. A random variable at the end of an arrow is conditioned on a random variable at the start of an arrow. Joint probability of a model is given by the product rule: (I) $p(a,b) = p(a|b)p(b)$ (II) $p(a,b,c) = p(a|b,c)p(b)p(c)$ (III) $p(a,b,c) = p(a|b)p(b|c)p(c)$

Figure 1.7 illustrates the key concept of plate notation. A plate around a random variable is a shorthand for repetition. For example, in a document with N words, instead of repeating a_n for each word, we can put a plate around variable a. Plate notation makes it possible to express models with a large number of random variables concisely. The joint probability of the model in Figure 1.7 is given by Equation 1.1. Note how plate notation naturally translates into a product of conditional probabilities.

$$\begin{aligned} p(a_1, a_2, ..., a_{N-1}, a_N, b) &= p(a_1|z)p(a_2|z)...p(a_{N-1}|z)p(a_N|z)p(b) \\ &= p(b)\prod_{n=1}^{N} p(a_n|b) \end{aligned} \tag{1.1}$$

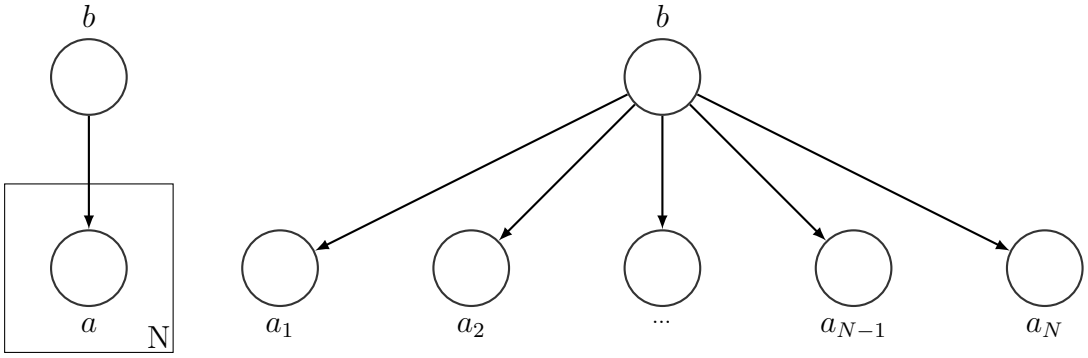

Figure 1.7: Plate notation on the left is a shorthand for a tree on the right. Joint probability can be expressed as a product of $p(b)$ and conditional probabilities of a_n. The joint probability for this model is given by Equation 1.1

Figure 1.8 illustrates a more complex model, where a set of N random variables is conditionally dependent on a random variable which itself is repeated D times. An example of such an arrangement could be D documents, each with N words. The two plate arrangement translates into a double product of conditional probabilities:

$$p(\{a\}, \{b\}) = \prod_{n=1}^{N} \prod_{d=1}^{D} p(a_{n,d}|b_d)p(b_d) \tag{1.2}$$

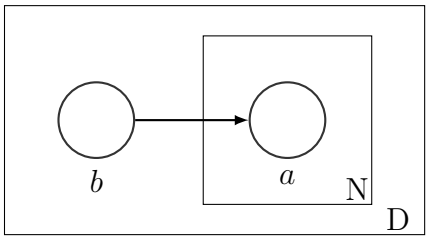

Figure 1.8: Plate notation with two plates, one embedded in another. Embedded plates translate to a double product of conditional probabilities. The joint probability for this model is given by equation 1.2

The LDA model in plate notation is presented in the bottom part of Figure 1.4. The black circles are fixed values (not random variables), and in this case they denote Dirichlet priors. They are hyper-parameters of the model. α is the prior of per-document topics, η is the prior of per-corpus

topic distribution over vocabulary. The grey circle means that words $w_{d,n}$ are observable variables, i.e. the input data. The white circles are the latent variables which are to be inferred. $\boldsymbol{\theta_d}$ is the topic distribution for a document d, $\boldsymbol{\beta_k}$ is the word distribution for topic k and $z_{d,n}$ is the topic indicator for the n-th word in document d. The joint probability for the LDA model is:

$$p(\{\beta\},\{\theta\},\{z\},\{w\}) = \prod_{k=1}^{K} p(\beta_k|\eta) \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N} p(z_{n,d}|\theta_d) p(w_{d,n}|z_{d,n},\beta_{z_{d,n}}) \tag{1.3}$$

Approximate posterior inference

The first step in inferring the topic structure from a document collection is computing posterior distribution for the LDA model:

$$p(\{\beta\},\{\theta\},\{z\}|\{w\}) = \frac{p(\{\beta\},\{\theta\},\{z\},\{w\})}{p(\{w\})} \tag{1.4}$$

The numerator in Equation 1.4 is a joint probability as given by Equation 1.3. The denominator, $p(\{w\})$, is the marginal probability of the data (also known as evidence) which is the probability of observing all the documents in our collection under all potential topic models. All potential topic models give rise to a large space which makes the computation of the marginal probability intractable, i.e. in practice we can not calculate the evidence $p(\{w\})$. In consequence, it is not possible to calculate an exact posterior probability. This problem is not unique to LDA; it affects most of the Bayesian models. However, we are rarely interested in calculating the posterior of just one model. Usually we want to compare models in order

to select the one which most likely represents the data. When comparing two models the evidence cancels out because $p(\{w\})$ does not depend on a model so its value is the same in both cases. We still need to calculate the joint probability for each model. Here the idea of Bayesian approximation proves useful.

The approximation of the posterior is based on an idea of finding a probability distribution which is similar to the true posterior. The two most common types of approximation methods applied to the inference in LDA are (a) sampling-based algorithms such as Gibbs sampling (Casella and George, 1992; Resnik and Hardisty, 2010) and (b) variational methods (Jordan et al., 1999; Wainwright and Jordan, 2008). In this thesis, we used only Gibbs sampler which is a Markov chain Monte Carlo algorithm (MCMC, Gilks et al. (1996)).

Gibbs sampling is based on an idea of construction of a Markov chain (Neal, 1993) whose stationary distribution is an approximation of the posterior of interest. The Markov chain is defined as a sequence of random variables; each is depending only on its predecessor. These random variables represent latent topic structure for our data. Steyvers and Griffiths (2007) provide a readable description of a collapsed Gibbs sampler (CGS) for LDA. Further details can be found in Griffiths and Steyvers (2004) and Griffiths (2002). Derivation of CGS equations for LDA is a challenging task for non-statisticians; full derivations are provided in Carpenter (2010) and Heinrich (2008). A tutorial on the practical implementation of CGS for LDA is given by Darling (2011).

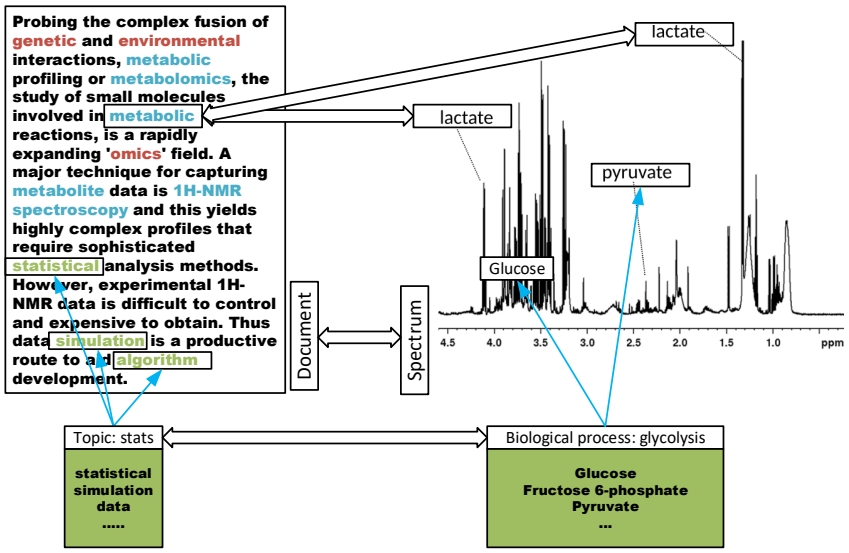

Figure 1.9: Topic models and NMR

LDA and NMR

Mapping of terminology between topic modelling and NMR is given in Table 1.1 and Figure 1.9. The basic idea is that a collection of NMR spectra is an analogue to a collection of documents. Each spectrum consists of peaks whose amplitude and location come from metabolites present in a sample. A metabolite here means a set of NMR peaks; it is a counterpart of a word in a text document. If a word is present multiple times in a text document, we would state its count. In the NMR domain this is analogous to a level of a metabolite.

We must avoid the word *concentration* because in chemistry its meaning is specific; it is an amount of a compound in a total volume of a mixture, and it is always expressed in some units. In this thesis, we use the term *relative concentration* which is a unit-less number expressing a virtual amount of a compound in a simulated NMR spectrum. Relative concentrations only make sense compared to each other. For example, let us assume that we

simulate using the generative process for LDA, an NMR spectrum with only two compounds: X and Y. In a text document world, we would say that we generated a document of the length of 1000. There are 200 Xs and 800 Ys. In the NMR world, we would state that the relative concentrations between compounds X and Y were 200 and 800, meaning that there is four times more Y than X.

There is a potential ambiguity when we talk about words in documents and metabolites in NMR. A phrase *number of words* in a document means document length in words, understanding that some words will be repeated many times. In the NMR spectrum, *a number of metabolites* means how many different molecules (regardless of their levels) are there. Therefore, *document length* does not have a good translation to the NMR domain. When this ambiguity occurs in this thesis, we use the term *word* even if we talk about NMR spectra. For example, "we simulate 1000 words NMR spectrum" means the total relative concentration of all metabolites will sum up to 1000.

The topics are a thematic pattern running within a set of documents. In NMR metabolomics, the underlying biological process in the organism determines a composition of metabolites in a sample which is investigated. A biological process is an analogue of a topic. As with topic in text documents, a biological topic is a list of all compounds of interest with assigned probabilities. In this thesis, topics in NMR are sets of metabolites (often identified with KEGG IDs) and assigned probabilities. The metabolites associated with a biological process represented by a topic have high probabilities while remaining compounds have low probabilities (zero or close to zero).

The generative model for LDA and the bag-of-words model assumption would not produce human-readable text documents. However, in the NMR domain, the words are entries for metabolites from HMDB (Wishart et al., 2018), including peaks' amplitude, location and number of NMR observable protons. This allows us to place simulated peaks in the correct places and with plausible amplitudes thus resulting in realistic looking spectra. This makes using LDA an excellent choice for simulating NMR spectra.

Table 1.1: Mapping of terminology used in text documents and in NMR metabolomics

Text documents	**NMR metabolomics**
a word	a metabolite e.g. Glucose
a word count	a metabolite's relative concentration or level
a document	an NMR spectrum
a length of a document	total concentration of all metabolites in a sample
a topic	a biological condition, process or set of metabolites
a vocabulary	a list of all metabolites of interest

1.4 Aims and objectives

^{1}H NMR data contains a wealth of information about the metabolic state of an organism. However, unveiling this information is always a challenge. Development and applications of novel statistical methods to find patterns in ^{1}H NMR data is an ongoing effort which aims at providing better tools for the metabolomics research community. In this thesis, we propose and evaluate applications of generative probabilistic models for information recovery in ^{1}H NMR metabolomics. We focus on a very successful, in other domains, but novel for ^{1}H NMR metabolomics, the Latent Dirichlet Allocation model. Our overall aim is to assess the applicability of LDA to analyse ^{1}H NMR metabolomics data. In particular, we set out to investigate

the following:

1. Can the generative process for LDA be used as a basis to simulate ^{1}H NMR metabolomics data?
2. Can LDA as an unsupervised method recover interesting topics to drive biological discovery?
3. Can LDA as a supervised method predict response variables associated with clinical outcome and highlight topics relevant to the response variables?

The remainder of this thesis is structured as follows.

Chapter 2 focuses on examining the generative process for LDA as a method for simulating ^{1}H NMR data. The goal is to show that the generative process for LDA can produce realistic looking spectra. This gives a foundation for using LDA to infer topics from existing NMR data. Simulating NMR data is also useful on its own. We often need simulated spectra with known parameters, something which is not always feasible using data from NMR experiments, to test and evaluate statistical models. Simulating with LDA adds to the existing toolkit (for example Muncey et al. (2010)) available to metabolomics researchers.

In Chapter 3 we turn to use the LDA model for inference of topics in synthetic and real NMR data sets. The inference is in unsupervised learning mode where no metadata about samples (spectra) is fed into the model; the only input is a matrix of spectra. We are able to evaluate the results as we know the ground truth for the simulated data and have prior publications with key results for the real data. In both cases, LDA was successful in recovering interesting topics which were consistent with the ground truth and with previous results.

Chapter 4 seeks to investigate applications of topic models in supervised learning. We use a variant of LDA which caters for modelling a response variable with each spectrum, called SLDA. We also use topics (components) from LDA, SLDA and PCA as a latent representation of NMR data and combine those with ElasticNet which is a linear regression model. We apply all those methods to a data set which comes from The Multi-Ethnic Study of Atherosclerosis (MESA) study. In this data set each spectrum is paired with multiple clinical variables such as BMI, glucose levels and cholesterol levels. The goal of the modelling was to examine if inferred topics can be informative about clinical outcomes. We successfully showed that the topics were indicative of glucose levels and HDL cholesterol levels.

Chapter 2

Simulation of ^{1}H NMR metabolomic spectra using Latent Dirichlet Allocation

2.1 Introduction

NMR spectra of bio-fluids are complex because of a large number of compounds in a sample, and the resonances overlap, noise and variation between samples (e.g. peak shifts). To unravel biologically meaningful information, advanced statistical methods and machine learning algorithms must be employed. Such an approach has a chance to make sense of large amounts of data coming from high-throughput experiments. Development of new methods requires training and testing data of specific, pre-programmed characteristics. Using data from NMR experiments for the development of new methods is problematic because of the untargeted (NMR detects all metabolites), nature of NMR. Artificial mixtures of pure compounds could be prepared, e.g. Takis et al. (2017), and used as samples to produce

"clean" data, however, such an approach is time-consuming and expensive. A solution to this problem is to simulate NMR spectra. Such simulation can be implemented as a software tool which allows specifying a list of metabolites and their concentrations. The fundamental physics of NMR (see Section 1.2) makes it easy to simulate multi-compound spectra as a linear combination of spectra of pure compounds. This will ensure that peaks are present in correct positions, and where overlapping they will add up forming the right shape, and their amplitudes will be guided by the parameters of the simulation.

Another point we would like to emphasise is that our simulations are not meant to compete with other simulators like MetAssimulo by Muncey et al. (2010). We do not want to generate the most realistic looking spectrum. We aim to simulate spectra which can be used for verification of machine learning algorithms and assessment of inference in probabilistic models. We do not add features which would add more realism to the simulated spectra such as peak shifts or noise. Such phenomena are deemed unwanted, and data preprocessing steps are taken to remove them. In the next chapter, we propose LDA as machinery to infer topics from NMR spectra. We need means for testing those tasks by providing data of known characteristics. We hope that the models we propose can infer information embedded in the data, and thereby enables us to move on to other types of data, e.g. experimental data from metabolomics studies.

Earlier we defined that the topics are distributions over metabolites, i.e. each metabolite will have assigned some probability. Each topic has the same list of metabolites, but the probabilities will be different. In practice, we use only top metabolites from a topic; this allows us to treat topics as

sets of metabolites, which simplifies describing, comparing and reasoning about topics. For example, in a simulation with 40 metabolites we defined four topics, and we assign the first ten metabolites to the first topic. What it means in technical terms is that the first topic is a vector of length 40 (all metabolites present in the simulation), but only the first ten get non zero probabilities, say 0.1, the rest are assigned zero probabilities.

Topics can be thought of as themes running across NMR samples. We could imagine topics representing metabolites characteristic for usual experimental groups, i.e. treatment topic and control topic or disease topic and healthy topic. It is important to realise that those topics contain all the metabolites of interest. The topics differ in probabilities assigned to the metabolites. The topics could also broadly be related to biological processes. For example, a biological process could be associated with a metabolic pathway or some of their combination. The interpretation of topics is very flexible. We can talk about very broad topics, e.g. a normal urine topic which contains metabolites present in normal human urine or a very narrow topic, the one consisting of a single metabolite. If we could discover topics in existing data sets, it would help to explore patterns in those data sets and perhaps be used as a driver for further analysis. This will be addressed in the two other chapters which build on the tools and simulated data developed here.

In this chapter, we develop a simulation method for NMR spectra based on the generative process for LDA. We describe the generative process for LDA and how it maps into NMR spectroscopy. We give details of how spectra are assembled from pure compounds spectra and how those are simulated using Lorentzian with parameters retrieved from HMDB.

Finally, we investigate simulated spectra with visualisations and Principal Component Analysis (PCA). We also simulate complex spectra based on 40 normal human urine metabolites and compare this to a real spectrum from the INTERMAP project. We conclude that our method gives good results. We can simulate groups of spectra exhibiting different topics and detect those groups with PCA. Comparison of spectra simulated with normal human urine metabolites to the real NMR spectra shows that our method can produce realistically looking NMR spectra.

2.2 Data sets

2.2.1 HMDB data

Human Metabolome Database (HMDB, Wishart et al. (2018)) was our primary source of information about metabolites and their properties. We also used human urine standard concentrations data included in the MetAssimulo package (Muncey et al., 2010) which were in turn derived from HMDB. We build a local database with the following information about each metabolite: KEGG ID, InChIKey, standard name, list of peaks (each peak data is a pair: ppm position and amplitude), number of protons, standard concentration. Although HMDB is already limited to human metabolism, we add a filter in our internal metabolites database to include only molecules found in any *H. sapiens* metabolic pathways in KEGG database (Kanehisa and Goto, 2000).

KEGG IDs are well-established identifiers for small molecules present in metabolic pathways. KEGG IDs are frequently present in HMDB, and we use them when they are there. If they are absent, we use Chemical Translation Service (Wohlgemuth et al., 2010) developed by The Fiehn

laboratory at UC Davis to translate between InChIKey and KEGG ID. We do not import molecules for which we could not establish KEGG ID.

Peak data are read from XML files provided by HMDB. For each metabolite, there will usually be multiple spectra present. We always choose ^{1}H NMR spectra from samples with the following conditions: the temperature of a sample must be between 20C and 30C; the solvent must be H_2O and pH between 6.9 and 7.9. If there are no spectra within such parameters, we do not include this molecule in our database. We chose those because they are the conditions that relate to normal human urine which is the primary example of bio-fluid which we simulate in this chapter.

Standard concentrations are sourced from HMDB and from MetAssimulo which was also derived from HMDB and reviewed by an expert. Usually, there are multiple entries per metabolite for its concentrations, and we always manually select an entry for a normal healthy adult.

Another required parameter for ^{1}H NMR simulations is the number of NMR observable protons[1] in a given metabolite. This information is not readily accessible in HMDB. We manually inspect the multiplets tables for a given molecule and sum up the values in Hs columns. This process is not easy to automate because there are many syntactic or otherwise obvious errors in the multiplets tables. We attempt to correct those errors to the best of our ability. If reading from multiplet tables fails, we count the protons manually from the chemical structure diagrams. We collected standard concentrations and number of protons for 40 metabolites which are characteristic for normal human urine and match with those available in MetAssimulo.

Finally, for demonstration and visualisation purposes we use three virtual

[1]Some protons will be missing due to the aqueous solution of the sample.

molecules. They are toy molecules designed to be easily spotted by eye in the spectra plots. The first, Metabolite Singlet (A), consisting of a singlet at 1.0 ppm. The second, Metabolite Doublet (B) consists only of a doublet at 2.00 and 2.025 ppm. The third one, Metabolite Triplet (C), consists only of a triplet at 3.00, 3.025 and 3.05 ppm. Metabolites B and C are purely for visualisation purposes as they could not arise from a real NMR experiment as triplets and doublets never exist on their own.

2.2.2 INTERMAP data

The INTERnational study of MAcro/micronutrients and Blood Pressure (INTERMAP) study (Stamler et al., 2003; Holmes et al., 2008) is a source of real NMR spectra that we use to compare against the simulated spectra (see Section 2.4.3). The study was an epidemiologic study to investigate how nutrition and which nutrients in particular influence blood pressure. There were 4680 participants from 17 diverse population samples in China, Japan, UK, and the USA. The study produced NMR data from urine samples from each participant, and we obtained ∼4000 spectra.

2.3 Methods

2.3.1 Simulating NMR spectra

The basic idea is to combine the spectra of pure compounds. Figure 2.1 illustrates the process. We used a subset of Human Metabolome Database (HMDB, Wishart et al. (2018)) to obtain information on pure compounds, in particular, peak positions (chemical shifts) and amplitudes, and the number of NMR detectable protons. Those are the critical parameters

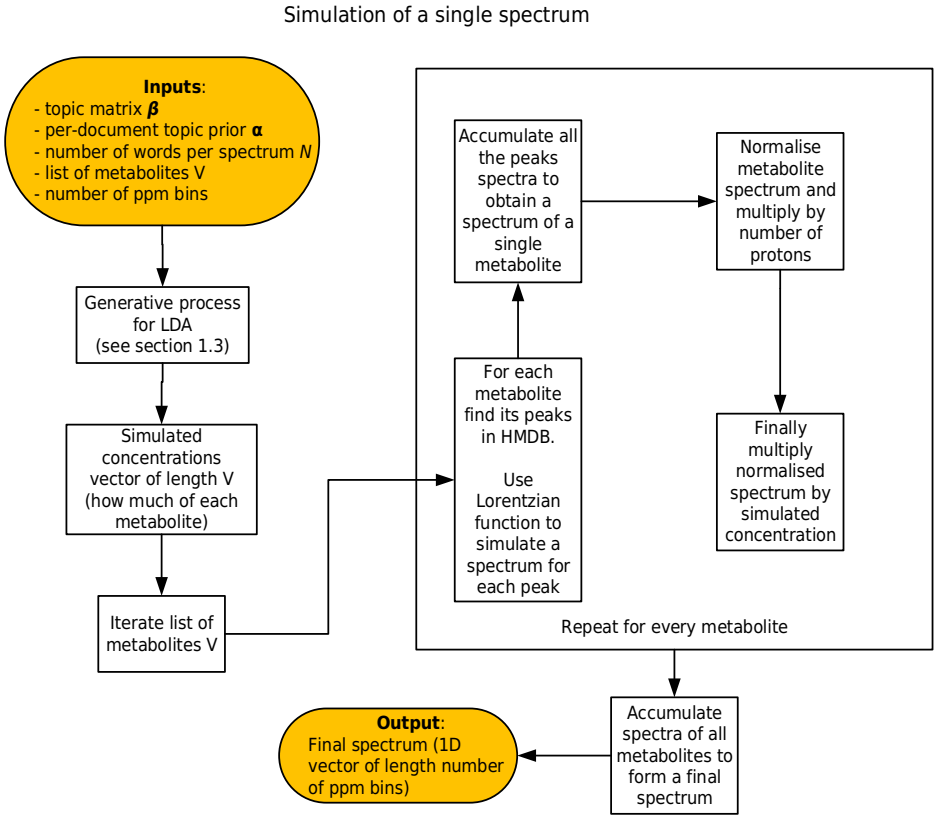

Figure 2.1: A spectrum simulation flowchart

for simulating an NMR spectrum of a pure compound. Using such an approach, the multiplet information, e.g. J-couplings, is not required as spectra simulated with Lorentzian function produce equivalent output at all peak locations. An NMR peak is modelled using a Lorentzian function (see details below). For each molecule, spectra representing its peaks are modelled and concatenated. The result is a spectrum with one or more peaks representing a metabolite. Each spectrum is normalised by dividing it by the area under the curve. Finally, each spectrum is multiplied by the number of protons in the molecule. This ensures that pure compounds' spectra are on the same relative scale and can be combined to create spectra representing a mixture of metabolites.

In a real NMR experiment, a sample will contain a particular mixture of molecules. Their concentrations (usually measured in mM) are directly reflected in the spectrum. In our simulations, we do not replicate real concentrations, only relative concentrations. The terms "levels" and "simulated concentrations" are also used to distinguish the outcomes of simulations from real chemical concentrations present in real samples. For example, a simple two molecule simulation could produce metabolites levels of 100 and 200. Their relative concentration is 1:2. As we do not use any units, such as mM, we only know that there was twice as much of the second molecule compared to the first one. The relative concentrations are obtained from a generative process which is based on a random process underlying the Latent Dirichlet Allocation(LDA) model. We construct a linear combination of the pure compounds spectra using, as coefficients, relative concentrations obtained from the LDA-based generative process described later in this chapter.

Our model for a single peak is a Lorentzian function is given by the equation:

$$L(x; x_0, H, \delta) = \frac{H}{1 + (\frac{x - x_0}{\delta})^2}$$

where x_0 is a position of the centre of a peak, H is an amplitude (a maximum at x_0), and δ is half of the peak's width at half of its intensity. The use of Lorentzian function comes from the application of Fourier transform to the signal from free induction decay (FID) which is in the time domain. Fourier transform provides a spectrum in frequency domain. FID is exponentially decaying, and Fourier transform of an exponential function is a Lorentzian function, so it makes a good choice for a peak model.

In our simulations, we always use a fixed value for the half-width parameter δ which is 0.00166(6) ppm; it translates to 1Hz assuming 600MHz spectrometer frequency i.e. $0.00166(6) \times 600MHz \div 10^6 = 1Hz$. The intensity at $x_0 \pm 1Hz$ will have half of the H.

To calculate a spectrum of a pure compound, we obtain all the parameters of its peaks (location x_0 and amplitude H) from HMDB and use them to evaluate the Lorentzian function on the whole ppm range, usually from 0.0 to 10.0 ppm. All the spectra representing peaks of a metabolite are added up to form its spectrum. The whole spectrum is normalised, by using approximation via the trapezium rule, to have the area under the curve equal to one. The last step in forming a single metabolite spectrum is scaling it by the number of protons in the molecule to ensure that the spectra can be later combined with some other metabolites' spectra. For example, if we had only two compounds in the sample, in equal proportions, but one has twice as many protons as the other, the area under the peaks from the former should be twice as large as the area under the peaks under the latter. Let us represent a spectrum with just one peak as a vector $\mathbf{L}$ (we use boldface for vectors). The length of such vector is the number of spectral bins B. For example, let $B = 20,000$ bins representing for values between $b_0 = 0.0$ ppm and $b_{19,999} = 10.0$ ppm form the bins vector $\mathbf{B}$. So now, each element of the vector $\mathbf{L}$ which we denote as L_i is given by Lorentzian function with parameters x_0, H and δ:

$$L_i(b_i; x_0, H, \delta) = \frac{H}{1 + (\frac{b_i - x_0}{\delta})^2}$$

where b_i are elements of $\mathbf{B}$ for $i = 1 \dots B$. We defined single peak spectrum vector $\mathbf{L_i}$, we can formulate a spectrum of a metabolite (there

will be M metabolites in final full spectrum) containing P peaks, $\mathbf{S_n}$:

$$\mathbf{S_m} = \frac{\sum_{i=1}^{P} \mathbf{L_i}}{\int_{x_{min}}^{x_{max}} \sum_{i=1}^{P} \mathbf{L_i}} \times \#protons_m$$

The integral in the denominator is approximated using the trapezium rule.

Finally, we want to assemble a spectrum consisting of all the M metabolites in a simulation. Each metabolite has a specific relative concentration. To calculate the final full spectrum $\mathbf{S_{all}}$ consisting of M metabolites, a linear combination of individual metabolites spectra $\mathbf{S_m}$ using relative concentrations C_m as coefficients is applied:

$$\mathbf{S_{all}} = \sum_{m=1}^{M} C_m \mathbf{S_m}$$

Section 1.3.1 explained how the generative process for LDA works. We apply this process to simulate the relative concentrations C_m which produce plausible looking spectra. As an example, let us focus on metabolites related to normal human urine, see Table 2.1. We arrange the 40 metabolites in four topics, ten metabolites per topic. This means that all 40 metabolites are present in each topic, but only ten will have non-zero probabilities i.e. each topic has ten different metabolites with non-zero probabilities. We use those topics, along with θ and β to generate documents where words are metabolites, in particular, our words are KEGG IDs of the normal urine metabolite set. These documents, or rather their word counts are used to simulate NMR spectra by first constructing spectra for each metabolite in the document, and then combining them, using word counts as relative concentrations. We repeat the process if we want to construct a data set of spectra sharing similar patterns. As mentioned before, each document-

spectrum will be different but still exposing the same underlying topics. Multiple data sets can be produced from the same topics setup by changing θ for each batch of spectra. We will see this approach later in this chapter.

In summary, we simulate NMR spectra by a linear combination of pure compounds. We use a subset of HMDB to obtain critical characteristics of pure compound spectra, i.e. a list of peaks with their positions and amplitudes and the number of protons in the metabolites. We use this information to model pure compounds with the Lorentzian function. The next step is to simulate the relative concentrations of each metabolite using the LDA generative process with predefined topics, θ and β. This allows us to obtain a random composition of metabolites but with a level of control via topic proportions per document and topics as distributions over all metabolites of interest. In the following section, we will present detailed simulations, explore their properties and compare simulated spectra with real NMR data.

2.4 Results

2.4.1 Single spectrum simulations

Ethanol is frequently used as a molecule to demonstrate the basics of NMR. We start with the most straightforward simulation possible, just with one metabolite, ethanol, no topics, no relative concentrations. The result is shown in Figure 2.2. Ethanol in aqueous solution shows only two multiplets in the NMR spectrum: a triplet and a quadruplet. There is no resonance from the OH group at physiological pH as the proton from the OH group is exchanged between the molecule and the solvent. At a more acidic pH, the

Table 2.1: Normal, human urine metabolites. The list was derived from the MetAssimulo package (Muncey et al., 2010) and used with permission from the authors.

	KEGG-ID	Name	Concentration (μM)
1	C00791	Creatinine	13200.00
2	C00047	L-Lysine	4220.40
3	C01586	Hippuric acid	2640.00
4	C00158	Citric acid	2022.00
5	C00062	L-Arginine	1130.50
6	C00037	Glycine	1029.00
7	C00135	L-Histidine	948.00
8	C00245	Taurine	834.24
9	C00160	Glycolic acid	752.00
10	C00058	Formic acid	583.00
11	C00064	L-Glutamine	485.80
12	C00186	L-Lactic acid	441.00
13	C00065	L-Serine	396.00
14	C00417	cis-Aconitic acid	393.00
15	C00082	L-Tyrosine	361.02
16	C00300	Creatine	343.00
17	C00581	Guanidoacetic acid	300.00
18	C00041	L-Alanine	290.00
19	C00031	D-Glucose	264.00
20	C01904	D-Arabitol	250.80
21	C00311	Isocitric acid	250.00
22	C00719	Betaine	245.52
23	C01004	Trigonelline	223.00
24	C00033	Acetic acid	200.00
25	C00042	Succinic acid	166.32
26	C01620	Threonic acid	132.00
27	C00642	p-Hydroxyphenylacetic acid	92.40
28	C02336	D-Fructose	85.00
29	C01026	Dimethylglycine	81.84
30	C02918	1-Methylnicotinamide	80.50
31	C00026	Oxoglutaric acid	77.00
32	C00984	D-Galactose	58.08
33	C00386	Carnosine	46.20
34	C00123	L-Leucine	39.60
35	C00149	L-Malic acid	34.32
36	C05984	2-Hydroxybutyric acid	32.27
37	C00073	L-Methionine	30.00
38	C00025	L-Glutamic acid	22.59
39	C00327	Citrulline	14.26
40	C00881	Deoxycytidine	8.58

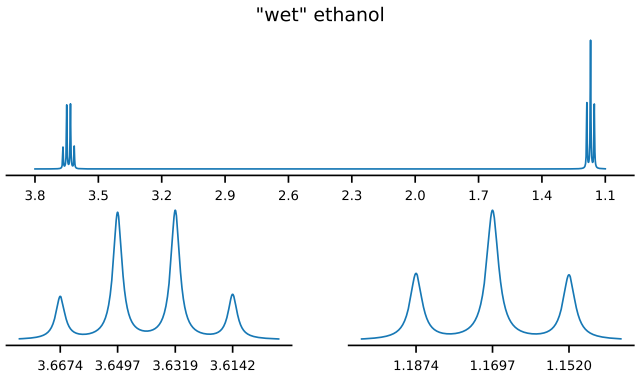

Figure 2.2: Ethanol in aqueous solution shows two multiplets in the NMR spectrum: a triplet and a quadruplet

proton might be visible.

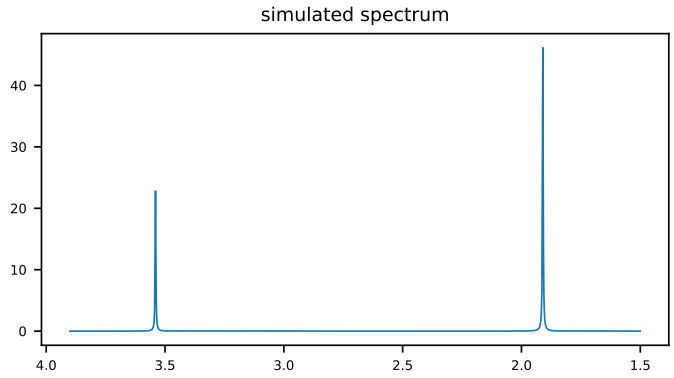

Figure 2.3: Glycine and Acetic acid

Next, we show how a simulated spectrum is constructed in a slightly more complicated scenario, so all moving parts are present, but the setup is still simple enough to see what is going on by eye. Our setup consists of only two metabolites: Glycine (KEGG ID C00037) and Acetic acid (KEGG ID C00033). There are two topics: topic A, where Glycine has a probability 1.0, and Acetic acid has a probability of 0.0 and topic B, where the probabilities are reversed. For simplicity, in this experiment both topics were set to be equally probable. Relative concentrations of 503 for Glycine and 497 for Acetic acid (1000 in total) were obtained.

We purposely chose these two metabolites because they have only one resonance each. Glycine has a singlet at 3.54 ppm and Acetic acid a singlet at 1.91 ppm. In NMR spectroscopy the concentration of a molecule is proportional to the area under the curve (AUC) shown in its spectrum. However, if a molecule has only one singlet resonance, then AUC can be approximated by this singlet's amplitude. In our simulated spectrum, the Acetic acid's amplitude is 46 and Glycine's amplitude is 22.8. When we take the number of protons in each molecule (Glycine 2 protons, Acetic acid 4 protons) into account, the ratio $\frac{503*4}{497*2}$ is approximately equal to the ratio of amplitudes $\frac{46}{22.8}$ which is ~ 2, as expected.

2.4.2 Multiple spectra simulations

So far we have reported only single spectra simulations. The concept of topics, which is a key feature of LDA and its generative process, was not playing any significant part. In this section, we set out to investigate how topics can be used in multi-spectra simulations. We use Principal component analysis (PCA), which is a standard tool in NMR metabolomics, to test if it is possible to distinguish between sets of spectra generated with distinct topic proportions.

Let us start with a simulation of 100 spectra using two topic system (Figure 2.4). We use two virtual molecules which are useful for visualisations as they are extremely easy to recognise. Molecule A contains only a single at 1.0 ppm. Molecule B contains only a doublet at 2.0 ppm. There are two topics: topic A where metabolite A has a probability of 1.0 and metabolite B has a probability of 0.0; and topic B where the probabilities are reversed. The distribution of topics for each spectrum is determined by sampling

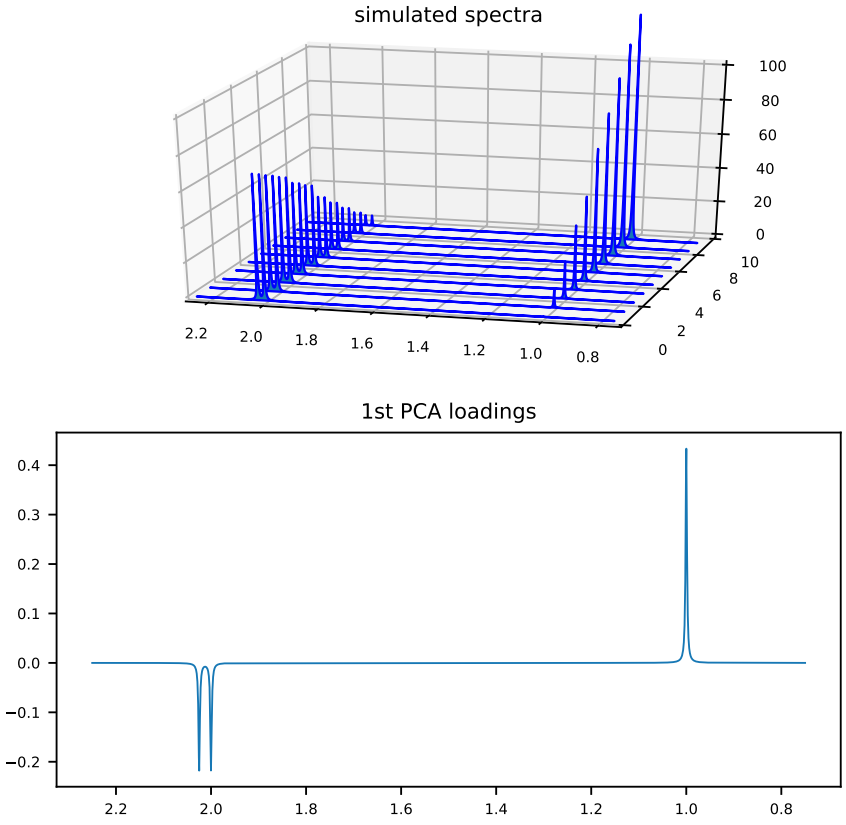

Figure 2.4: Two topic system simulation of 100 spectra. Topic A always selects a singlet, while topic B always selects a doublet. Top panel: Every 10th spectrum out of 100 simulated (spectra were sorted). Bottom panel: first and only PCA loadings visualisation confirm that the topics are anti-correlated.

from the Dirichlet distribution with parameter $\alpha = [1, 1]$. Setting a uniform α will result in samples which are, on average, equally distributed, i.e. all pairs are equally probable and they sum up to one thus no topic is dominant. For each spectrum, we simulate 1000 words so the relative concentrations will sum up to 1000. The probability of each metabolite is in this case determined by the document level of topic distribution as the probabilities of words within the topics are set to 1.0 for the topic metabolite, and zero

for the other metabolite. To give a concrete example, let $\theta = [0.73, 0.27]$. For the first (out of 1000) word, we sample a topic indicator z_n, and it comes as topic A with a probability of 0.73 and as topic B with a probability of 0.27. We then choose topics according to those probabilities. If A is chosen, then we select metabolite A with probability 1.0 because metabolite B has zero probability in topic A. This leads to a significant variability between simulated spectra, but the key observation is that if the relative concentration of metabolite A goes up then the concentration of metabolite B must go down (and vice versa). We could say that the occurrence of metabolites A and B among 100 spectra is anti-correlated. An average simulated concentration will be about the same as no topic was dominant. Once all 100 spectra are simulated, we run a standard PCA on them. In the case of the two topic system, the first principal component explains all the variance as the system had only one degree of freedom. In the principal component space, all 100 spectra lie on a straight line.

Figure 2.5 illustrates the results of an experiment with three simple topics. Each topic consists of only one virtual metabolite with non zero probability. The virtual metabolites are used only for demonstration and visualisation purposes. The metabolites are called Singlet, Doublet and Triplet. They consist of resonances suggested by their names, and they are located at 1.0, 2.0 and 3.0 ppm. Three hundred spectra are simulated in total, three batches of 100 spectra where one of the virtual metabolites is dominant. As before, each spectrum has the topic proportions sampled from the Dirichlet distribution with a specific parameter α. We use non-uniform α, always heavily skewed towards the dominant topic for 100 spectra batch. The first batch is simulated using $\alpha_{Singlet} = [10, 1, 1]$, those spectra will

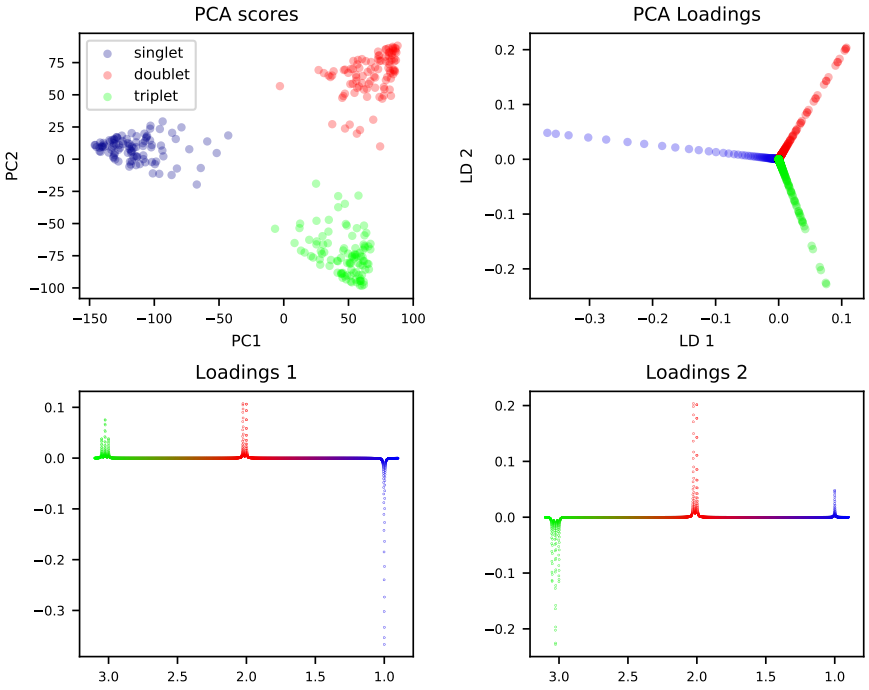

Figure 2.5: Three topic PCA

mostly consist of the topic Singlet, which also assigned a probability of 1.0 to a virtual metabolite Singlet. The two other batches of spectra are simulated with the analogous parameters, namely $\alpha_{Doublet} = [1, 10, 1]$ and $\alpha_{Triplet} = [1, 1, 10]$. With α constructed that way and 1000 words per spectrum, the average relative concentrations are ∼800 for the principal metabolite and ∼ 100 and ∼ 100 for the other two metabolites.

We use PCA to check if the batches of spectra simulated with varying parameters can be differentiated. PCA transforms the peaks' coordinate system to a new coordinate system called principal components (PC). The first PC axis points in the direction of the highest variance in the data (in this case 300 spectra simulated with three topics). The second PC axis points in the direction of the second highest variance, etc. For the three

topics setup, the results from PCA are as follows. 66.5% of the variance in all 300 spectra is explained by the first principal component (PC1). 33.5% is explained by PC2. As expected, PC1 and PC2 explain all the variance; thus we can project all 300 spectra to a 2D space and retain all of the information. The top left panel of Figure 2.5 is entitled "PCA scores". It shows all spectra in the PC space where each point represents a spectrum. Because we know the ground truth about the underlying topics, we can colour each data point, a spectrum, in the PC space by its dominant topic. The scores plot here does not look similar to the real NMR data because the parameters for the topic proportions per spectrum were chosen to be quite extreme for demonstration purposes. We observe that the scatter plot of the PCA scores resembles a simplex triangle which is characteristic of the Dirichlet distribution. We conclude that PCA can differentiate spectra generated for distinct topics.

The PCA loadings plots, the top right panel (Figure 2.5) illustrate how many spectral variables (bins) are in the original coordinate system. Peaks in ppm, contribute to the PC space. We observe that the majority of variables are around point (0, 0), i.e. they do not contribute to the PC space, those are regions between peaks. There are some points away from the loadings origin; those bins contribute to the PC space.

Next, instead of looking at all the spectral variables, we focus on multiplets. By recording a multiplet's top value, we approximate how much a given virtual metabolite contributed to the PC space (see Figure 2.5 bottom panels Loadings 1 and Loadings 2). In this case, all variables which constitute a Singlet at 1.0 ppm contribute about -0.36 to PC1 and 0.04 to PC2. Analogously, the variables that form a Doublet at 2.0 ppm contributed

about 0.1 to PC1 and 0.2 to PC2. The variables of the Triplet at 3.0 ppm contributed about 0.04 to PC1 and -0.13 to PC2. This is consistent with the scores plot: mostly positive values of PC1 for topics of the Doublet and Triplet and negative PC1 values for the topic Singlet. Similarly, positive values of PC2 for the Singlet and Doublet topics and negative values for the PC2 for the Triplet topic.

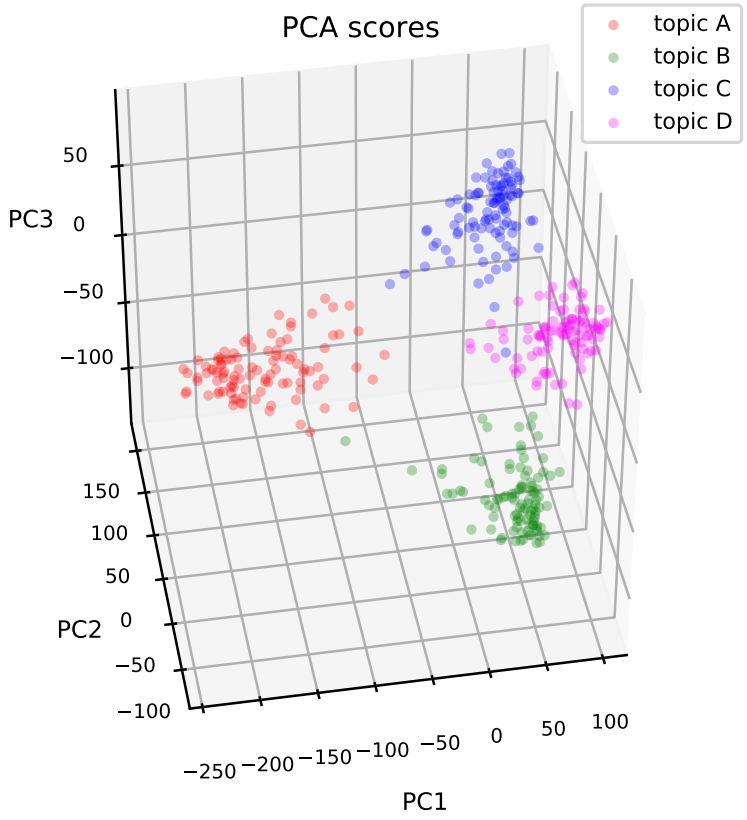

Figure 2.6: Four topic PCA

We move on to a next experiment which increases the complexity of the simulated spectra. Forty normal human urine metabolites were arranged in four topics, ten metabolites per topic. Within each topic, the ten

principal metabolites have equal probabilities of 0.1 (the remaining 30 have probabilities zero). Similarly, as before, we simulate batches of 100 spectra where a single topic is dominant. For topic proportions per spectrum, we use the Dirichlet distribution with α where the position of the dominant topic is set to 10 and the rest is set to 1, e.g. for first 100 spectra batch, the first topic is dominant, so $\alpha = [10, 1, 1, 1]$. Repeating this procedure for each topic will result in 400 spectra. We run PCA on all the spectra as before. The PC components explain the variance in this data as follows: PC1 0.51, PC2: 0.29, PC3: 0.18 and PC4: 0.01. The four PCs explain 98% of the variance in the data. We conclude that PCA explains the data well, as expected. We visualise the PCA results with 3D scatter of the PCA scores, see Figure 2.6. The scatter plot resembles a tetrahedron with points, spectra in PC space, gravitating towards its corners, depending on which topic was dominant for a particular spectrum. This result shows that topic driven NMR spectra simulations can be separated into groups based on their topic composition using techniques widely used on real metabolomics data.

2.4.3 Comparing simulated spectra to the real NMR data

In our final simulation in this chapter, we compare the simulated spectrum with a real NMR spectrum from the INTERMAP study (Stamler et al., 2003). The simulated spectrum is based on just one topic which contains 40 metabolites found in normal human urine (see Table 2.1). The topic distribution, in this case, is not applicable as we have only one topic. The metabolites probabilities within this topic are based on standard concentra-

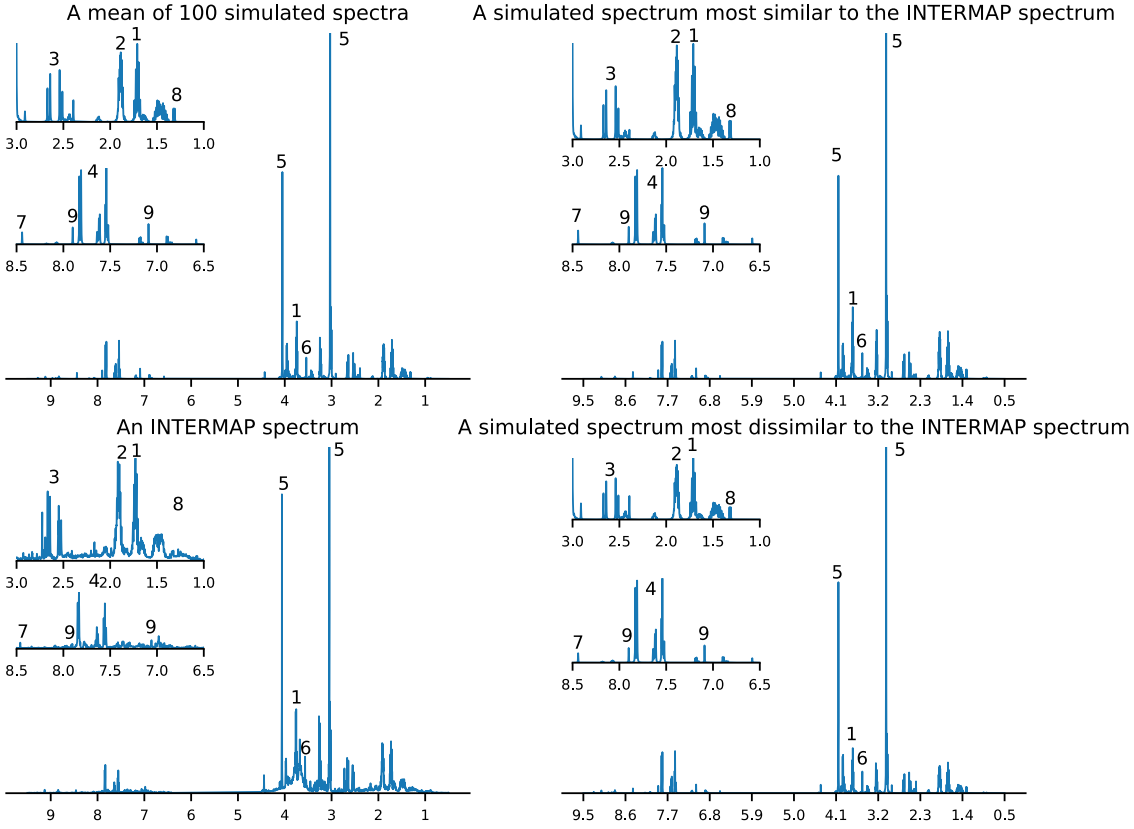

Figure 2.7: 100 hundred spectra were simulated using normal human urine metabolites (see Table 2.1). We compare simulated spectra with the real NMR data from the INTERMAP study (Stamler et al., 2003). Top left: a mean spectrum out of 100 simulated spectra. Bottom left: an INTERMAP spectrum which is the most similar to the mean simulated spectrum. Top right: a single simulated spectrum which is the most similar to the INTERMAP spectrum. Bottom right: a single simulated spectrum which is the most dissimilar to the INTERMAP spectrum. The similarity was determined by computing a distance between spectra using the correlation coefficient as a metric. We identified nine peaks which help to compare the spectra by eye: (1) Lysine, (2) Arginine, (3) Citric acid, (4) Hippuric acid, (5) Creatinine, (6) Glycine, (7) Formic acid, (8) Lactic acid, (9) Histidine.

tions as reported in HMDB (Wishart et al., 2018) and the Human Urine Metabolome study by Bouatra et al. (2013), also see column "concentrations" in Table 2.1.

One hundred spectra were simulated, and their mean was compared with a real spectrum from the INTERMAP study, see figure 2.7. We note many similarities, most notably Lysine (1) and Arginine (2) which are at a high level in the real spectrum. We simulated those high levels of Lysine and Arginine by increasing the values of standard concentrations by the tenfold as we wanted to match the real spectrum which came from an individual who had those high levels. The other metabolites worth highlighting are Citric Acid (3), Hippuric Acid (4), Creatinine (5), Glycine (6), Formic acid (7), Lactic acid (8) and Histidine (9). It is worth noticing that we only used 40 metabolites while the real INTERMAP spectrum comes from an untargeted study and contains at least ~ 100 metabolites. Despite this, both simulated and INTERMAP spectra display similarities in many places. Adding more metabolites to our simulation would improve this resemblance.

2.5 Discussion

In this chapter, we set out to test if NMR spectra, as encountered in metabolomics, could be simulated using a generative process based on Latent Dirichlet Allocation (LDA). LDA can infer topics for a variety of types of data (Blei, 2012). It was not clear that the generative process for LDA can produce reasonable looking NMR spectra which contain predefined internal patterns. After all, the bag-of-words approach in text document rules out generating documents in any human language. In this chapter, we demonstrate that the generative process for LDA was successful in producing such NMR spectra. The benefit is twofold: (a) we know that LDA can model NMR spectra, (b) we have tools to test LDA inference of

topics in the next chapter. The latter is essential as LDA by its nature is unsupervised, so the evaluation of its performance on real NMR data is difficult, if not impossible, because the true underlying topics are unknown.

We claim that the benefit of our method is its simplicity and also that it provides a reasonable degree of control over simulated data. Its simplicity is that the inputs for the simulations are just basic information about the metabolites used, lists of their peaks locations and amplitudes; Θ and β to define topics and their parameters. Having multiple Θ allows to easily generate data sets simulating experimental groups, i.e. not only treatments and controls but more complex scenarios. Real NMR data comes from untargeted experiments, and the ground truth is unknown. Our approach addresses this problem by providing simulated NMR spectra with a known composition.

2.5.1 Potential limitations

Fixed variance

There is no control over the variance of relative concentrations in our generative model. Let us illustrate this with an example. Our last simulation was with only one topic of a normal urine simulation. In this case, our generative model reduces to simple sampling from a multinomial distribution. Each spectrum consists of 1000 samples from the same multinomial distribution (because we have a single topic) which uses probabilities $\beta_{\mathbf{1}} = [p_1, p_2, \ldots p_m]$, $m = 40$ in this case, for each metabolite which are derived from standard concentrations. This simplification highlights a potential limitation of our simulation method, namely, the variance in a multinomial distribution, similar to a binomial distribution, depends only on $\beta_{\mathbf{1}}$, so it is fixed because $\beta_{\mathbf{1}}$

is constant. Therefore there is no option to specify variance in our simulated spectra. Each metabolite's variance is defined by $np_i(1-p_i)$ where p_i is a probability of i-th metabolite. For example, the probability of Citric acid is $p_{citric} = 0.06$. When a spectrum is generated with $n = 1000$ words, Citric acid's theoretical standard deviation is $\sqrt{1000 * 0.06 * (1 - 0.06)} \approx 7.51$. In the actual simulation, 100 spectra, the mean level of Citric acid was 61.32 (minimum was 44.0 and maximum was 79.0), the standard deviation was 7.66 which is very close to the theoretical value of 7.51. We could change n and p_i to influence the variance of the i-th metabolite, but we can not directly specify it as we would be able in the case of sampling from, for example, a Gaussian distribution.

Semi-realistic spectra

Our aim with the simulation is to produce semi-realistic spectra with known ground truth to test the probabilistic models. Our simulation does not produce some artefacts found in real NMR data, such as noise or peak shifts. Those are usually unwanted in statistical modelling and as such are removed by data preprocessing steps like de-noising and peak alignment (Veselkov et al., 2009; Hao et al., 2014). Producing highly realistic spectra is not critical for our purposes. Also, there are simulators with the goal of generating realistic data, for example, MetAssimulo (Muncey et al., 2010).

Future work

The method that we have developed here should be considered as a proof of principle. Future work would include turning this project into a tool which has some user interface, to allow researchers to simulate data rapidly and

iterate the results. At the moment the simulation process is cumbersome to set up and require knowledge of the implementation internals. Running-time of the simulation is another issue which could be improved. As mentioned earlier, the simulation could be extended with some features of real spectra like the addition of noise and peak shifts. The simulation software could also use more parameters which are fixed at the moment, such as the width of a peak. Finally, the database of metabolites could potentially be extended beyond what is presented here.

Chapter 3

Unsupervised Latent Dirichlet Allocation for information recovery in ^{1}H NMR metabolomics

3.1 Introduction

In the previous chapter, we showed how the generative process could be used for the simulation of NMR spectra. We concluded that the LDA-based generative process produces realistic-looking spectra. In this chapter we investigate the opposite approach, i.e. can LDA successfully infer topics in NMR spectra?

Researchers in NMR metabolomics often face the following challenge. There is a matrix of data, where rows are spectra and columns are spectral bins. Samples can originate from individuals who might be divided into groups by some key, e.g. disease, clinical outcome and similar. The research

question is if there is interesting biology to be revealed from this data? Metabolites frequently participate in several metabolic pathways, and they play different roles ranging from being a key component in the system to having only some auxiliary function. This leads us to propose mixed membership models such as Latent Dirichlet Allocation as a good model for such biological constructs. LDA has not been applied in modelling of NMR metabolomics; therefore, we hope that this new application will be a useful addition to a toolbox for metabolomics practitioners. We propose to use LDA as an exploratory tool to identify interesting patterns in data without any prior knowledge. This is an unsupervised learning task, and it could be thought of as a step in early stage analysis. In a fitted LDA model, some topics might be important for the underlying biological process or for differentiating between groups. Such topics can indicate metabolites for further analysis. For example, a pairwise investigation of selected metabolites to demonstrate what exactly drives the differentiation between the groups (or not). We will follow this approach to analyse *S. mansoni* data in this chapter.

LDA is useful in this scenario as it provides interpretable results. Other latent variable methods such as PCA can also perform this task, but their strength is not in interpretability of the model. Once the latent variable model is fitted to the data, it is possible to use it to drive further analysis. We demonstrate how we approached such analysis in the case of *S. mansoni* data set (see Sections 3.2.2 and 3.4.2).

To assess the usefulness of LDA in NMR metabolomics, we tested the method on sets of simulated data with increasing complex structure but of known ground truth and real NMR data from *S. mansoni* study by Li

et al. (2011). The authors of the study found discriminating biomarkers using linear modelling which we use as a baseline.

Evaluation of the results was by relating to prior knowledge about the data and also by using the Jaccard index when applicable. Jaccard index is used to compare topics pairwise. It measures the similarity of two sets, and it is calculated by dividing the size of the intersection over the size of the union. For two set A and B the equation is:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

A Jaccard of 1.0 means two sets are identical, a Jaccard of 0.0 means two sets do not share elements. In the case of simulated data, we know the ground truth in the form of topic compositions and probabilities of each topic; we can compare parameters of simulations with inferred parameters of the model. We compared results from LDA to PCA, comparing topics and components. PCA is a popular unsupervised method in metabolomics (Ebbels et al., 2011). PCA can also be seen as a latent variable model but its goal and assumptions are different from the LDA model.

A note about terminology. In LDA, a topic is a distribution over vocabulary or in our case a set of all bins in the spectra of interest. Let V be the size of this set. Each topic can be thought of as a list of length V where each metabolite has been assigned a probability. The probability is what makes topics different between them. The list can be sorted in descending order by probability and shortened to top N elements (usually $N = 10$). In this text, we frequently use the term "topic" in reference to the top N list. Sometimes we may mean the full list, but it always should be clear from the context which one is it. We use this simplification because

it helps to think about, compare and visualise topics.

3.2 Data sets

3.2.1 Simulated data

We used two toy data sets described in the previous chapter: two and three topics systems (see Section 2.4.2). Although those are very simple, we want to be consistent and follow the pattern of testing by increasingly complex simulated data sets.

Four topic simulations data set was constructed using normal urine metabolites arranged in four sets, forty metabolites in total (see Table 3.1). Each topic can be thought of as a set of metabolites characteristic to some biological condition. The four topic simulation represents an experiment with four biologically different groups of individuals each associated with one topic. We simulate 100 samples[1] per group. The metabolites were selected, so each topic had some metabolites with high standard concentration and some with low standard concentration. This way each topic could produce spectra with some high and some smaller peaks (see Table 3.2).

There are two variants of four topic simulation: (a) the topics do not share any metabolites resulting in non-overlapping arrangement (Table 3.2) and (b) there is some overlap between the sets, i.e. some metabolites are shared between topics forming an overlapping arrangement (Table 3.3). Those two arrangements represent an idea of having increasingly complex data sets therefore increasing the challenge for modelling tools.

Simulated topics are distributions over all metabolites[2], we assign equal

[1] In the context of this chapter, samples and NMR spectra are used interchangeably.
[2] Technically a topic is a vector of length V, where V is a total number of metabolites,

Table 3.1: Normal, human urine metabolites used for four topic simulations. The list was derived from the MetAssimulo package (Muncey et al., 2010) and used with permission from the authors.

	KEGG-ID	Name	Concentration (μM)
1	C00791	Creatinine	13200.00
2	C00047	L-Lysine	4220.40
3	C01586	Hippuric acid	2640.00
4	C00158	Citric acid	2022.00
5	C00062	L-Arginine	1130.50
6	C00037	Glycine	1029.00
7	C00135	L-Histidine	948.00
8	C00245	Taurine	834.24
9	C00160	Glycolic acid	752.00
10	C00058	Formic acid	583.00
11	C00064	L-Glutamine	485.80
12	C00186	L-Lactic acid	441.00
13	C00065	L-Serine	396.00
14	C00417	cis-Aconitic acid	393.00
15	C00082	L-Tyrosine	361.02
16	C00300	Creatine	343.00
17	C00581	Guanidoacetic acid	300.00
18	C00041	L-Alanine	290.00
19	C00031	D-Glucose	264.00
20	C01904	D-Arabitol	250.80
21	C00311	Isocitric acid	250.00
22	C00719	Betaine	245.52
23	C01004	Trigonelline	223.00
24	C00033	Acetic acid	200.00
25	C00042	Succinic acid	166.32
26	C01620	Threonic acid	132.00
27	C00642	p-Hydroxyphenylacetic acid	92.40
28	C02336	D-Fructose	85.00
29	C01026	Dimethylglycine	81.84
30	C02918	1-Methylnicotinamide	80.50
31	C00026	Oxoglutaric acid	77.00
32	C00984	D-Galactose	58.08
33	C00386	Carnosine	46.20
34	C00123	L-Leucine	39.60
35	C00149	L-Malic acid	34.32
36	C05984	2-Hydroxybutyric acid	32.27
37	C00073	L-Methionine	30.00
38	C00025	L-Glutamic acid	22.59
39	C00327	Citrulline	14.26
40	C00881	Deoxycytidine	8.58

V=40 in this case.

Table 3.2: Topics which are the basis for the four non-overlapping topics simulation data set. Each topic is a set of 10 metabolites. The metabolites are represented by their KEGG IDs. Prior probability for each metabolite in a topic is 0.1.

	Topic 0	Topic 1	Topic 2	Topic 3
1	C00791	C00047	C01586	C00158
2	C00062	C00037	C00135	C00245
3	C00160	C00058	C00064	C00186
4	C00065	C00417	C00082	C00300
5	C00581	C00041	C00031	C01904
6	C00311	C00719	C01004	C00033
7	C00042	C01620	C00642	C02336
8	C01026	C02918	C00026	C00984
9	C00386	C00123	C00149	C05984
10	C00073	C00025	C00327	C00881

probabilities to metabolites specific to a topic and zero probability to the rest. For example, in Topic 0 (see Table 3.2), the ten metabolites have a probability of 0.1 and all others have a probability of 0.0.

Each sample most likely contains all forty metabolites specified in the simulation, but only those metabolites which constitute a topic will result in higher concentrations. This is because the way we define the topics' probabilities for simulation. The proportions of topics for each sample are not fixed, but they are sampled from the Dirichlet distribution with a non-uniform parameter vector α. For example, if topic zero is the chosen topic for a group, we sample from Dirichlet with $\alpha = [10;\ 1;\ 1;\ 1]$. A possible sample looks like this: [0.901; 0.073; 0.014; 0.014]. Indeed, the first probability is large, 0.9, and the rest are small but non zero. Intuitively we would expect this sample from such α. However, sampling from Dirichlet with this particular α can also yield probabilities like this[3]: [0.455; 0.214; 0.087; 0.243]. The first component is much lower, less than 0.5 but more

[3]Those are real samples from Dirichlet distribution with $\alpha = [10;\ 1;\ 1;\ 1]$ as implemented in NumPy.

interestingly the second and the last probabilities are more than one in five. This sample highlights that it cannot be expected that the non-driver topics will have minuscule probabilities. Therefore metabolites from all topics will have a non-zero contribution in each sample and each sample is most likely to contain all forty metabolites.

Table 3.3: Topics which are the basis for the four overlapping topics simulation data set. Each topic is a set of 12 metabolites. The metabolites are represented by their KEGG IDs. Prior probability for each metabolite in a topic is $\frac{1}{12}$. Metabolites which are shared among topics are marked in red.

	Topic 0	Topic 1	Topic 2	Topic 3
1	C00791	C00047	C00123	C00149
2	C00062	C00037	C00025	C00327
3	C00160	C00058	C01586	C00158
4	C00065	C00417	C00135	C00245
5	C00581	C00041	C00064	C00186
6	C00311	C00719	C00082	C00300
7	C00042	C01620	C00031	C01904
8	C01026	C02918	C01004	C00033
9	C00386	C00123	C00642	C02336
10	C00073	C00025	C00026	C00984
11	C00047	C01586	C00149	C05984
12	C00037	C00135	C00327	C00881

3.2.2 *S. mansoni* data set

The study was designed to investigate biomarkers in a mouse model infected with *Schistosoma mansoni*. Schistosomiasis is a tropical disease caused by contact with larvae of *S. mansoni* which is present in water reservoirs in Africa, Asia and South America. When a human consumes the contaminated water, they become a host for a parasite and an adult worm of *S. mansoni* develops.

The experimental design of the study was as follows. Ten mice were infected with *S. mansoni*. Another ten mice remained uninfected and acted

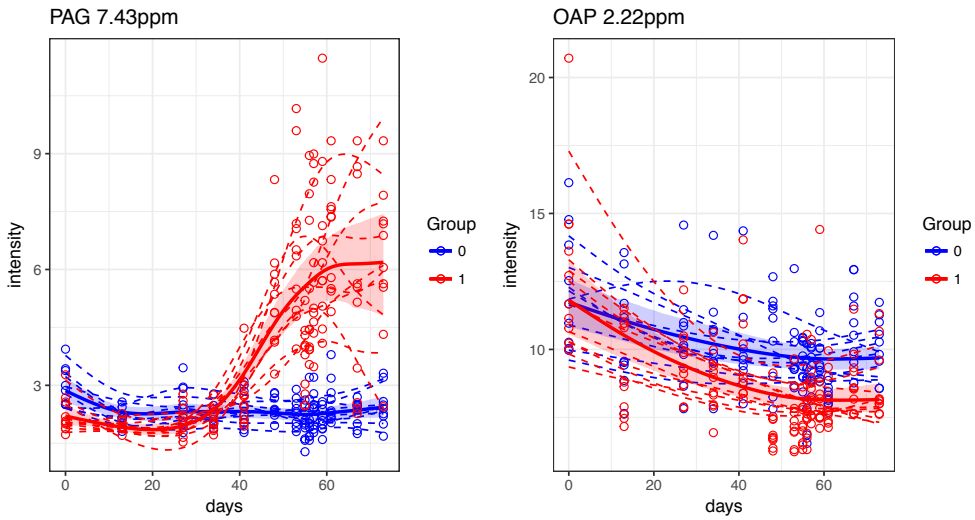

Figure 3.1: Two metabolites from the *S. mansoni* study are presented as a time course data. Left panel: phenylacetylglycine (PAG) at 7.43ppm, right panel: 2-oxoadipate (OAP) at 2.22ppm. Thick solid lines represent fitted smooth splines for each group mean. Dashed lines represent time courses of individuals fitted as smooth splines, circles are raw data points. Shaded regions are confidence intervals calculated by resampling of individual curves with replacement.

as a control group. The study collected urine, plasma and faecal samples. We used only the urine data. The samples were collected on day 0 at which point the treatment group was infected with the parasite. Samples of urine were subsequently collected on days: 13, 27, 34, 41, 48, 53, 55, 56, 57, 59, 61, 67 with the last sample collected on day 73. Time courses for two selected metabolites phenylacetylglycine (PAG) at 7.43ppm and 2-oxoadipate (OAP) at 2.22ppm are visualised in Figure 3.1. The sample collection was done at the same time points for both experimental groups. One mouse in the control group died mid-experiment, and it was not included in the data.

^{1}H NMR spectroscopy was used as the analytical technique to determine the metabolic content of the samples. The data were analysed using multivariate techniques such as PCA and a series of linear models were constructed at each time point to find significant differences between treat-

Table 3.4: Identified metabolites in the S. mansoni study (urine specific). Signal multiplicities - s: singlet; d: doublet; t: triplet; m: multiplet; q: quadruplet. Adapted from Li et al. (2011)

	Full name	Short name	Chemical shift and signal multiplicity
1	hippurate	Hip	3.97(d);7.84(d);7.55(t);7.64(t)
2	3-methyl-2-oxovalerate	MOV	2.93(m);1.1(d);1.7(m);1.46(m);0.9(t)
3	2-oxoadipate	OAP	2.77(t);1.82(m);2.22(t)
4	2-oxoisocaproate	OIC	2.61(d);2.1(m); 0.94(d)
5	2-oxoisovalerate	OIV	3.02(m);1.13(d)
6	p-cresol glucuronide	p-CG	7.06(d);7.23(d);2.3(s)
7	phenylacetylglycine	PAG	7.43(m);7.37(m);3.75(d);3.68(s)
8	taurine	Tau	3.43(t);3.27(t)
9	trimethylamine N-oxide	TMA-N	3.28(s)
10	3-ureidopropionic acid	UPA	2.38(t);3.31(t)
11	acetate	Ace	1.93(s)
12	arginine	Arg	3.78(t);1.92(m);1.65(m);3.20(t)
13	citrate	Cit	2.66(d);2.54(d)
14	3-carboxy-2-methyl-3-oxopropanamine	CMOPA	2.49(m);1.08(d);3.19(m);3.56(m);3.72(m)
15	creatine	CRE	3.03(s);3.92(s)
16	creatinine	CRT	3.03(s);4.05(s)
17	dimethylamine	DMA	2.72(s)
18	lactate	Lac	4.11(q);1.32(d)
19	lysine	Lys	3.78(t);1.92(m);1.47(m);3.03(t);1.72(m)
20	N-acetyl glycoprotein fraction	N-AG	2.06(s)
21	2-oxoglutarate	OGT	3.01(t);2.45(t)
22	pyruvate	Pyr	2.36(s)
23	scyllo-inositol	S-In	3.33(s)
24	succinate	Suc	2.41(s)

ments and control groups. The study found differences between experimental groups starting on day 41. Urinary biomarkers of the infection were found with hippurate, phenylacetylglycine (PAG) and 2-oxoadipate topping the list. The list of all identified metabolites in urine (only some of them show

differences between the groups) is given in table 3.4 (Li et al., 2011).

3.3 Methods

This chapter uses models and methods described earlier in this thesis. LDA was introduced in Chapter 1. NMR spectra simulations were described in Chapter 2.

From spectra to high-level variables

To improve interpretability of outputs from LDA and PCA models we use coarse grain variables such as metabolites (as in the simulated data sets) or bins associated with particular NMR resonances as in the case of the *S. mansoni* data set (multiple bins per metabolite are possible). We did not attempt to run LDA and PCA on high resolution spectra[4]. Topics formed with such a high number of strongly correlated variables would not be informative.

The *S. mansoni* data set was manually processed to obtain high-level variables (see Section 3.3.1 for details). Whilst the simulated data sets were processed algorithmically. The simulation produced high-resolution data but we processed each spectrum in the data set to produce a vector of concentration per metabolite. The details of this algorithm are as follows.

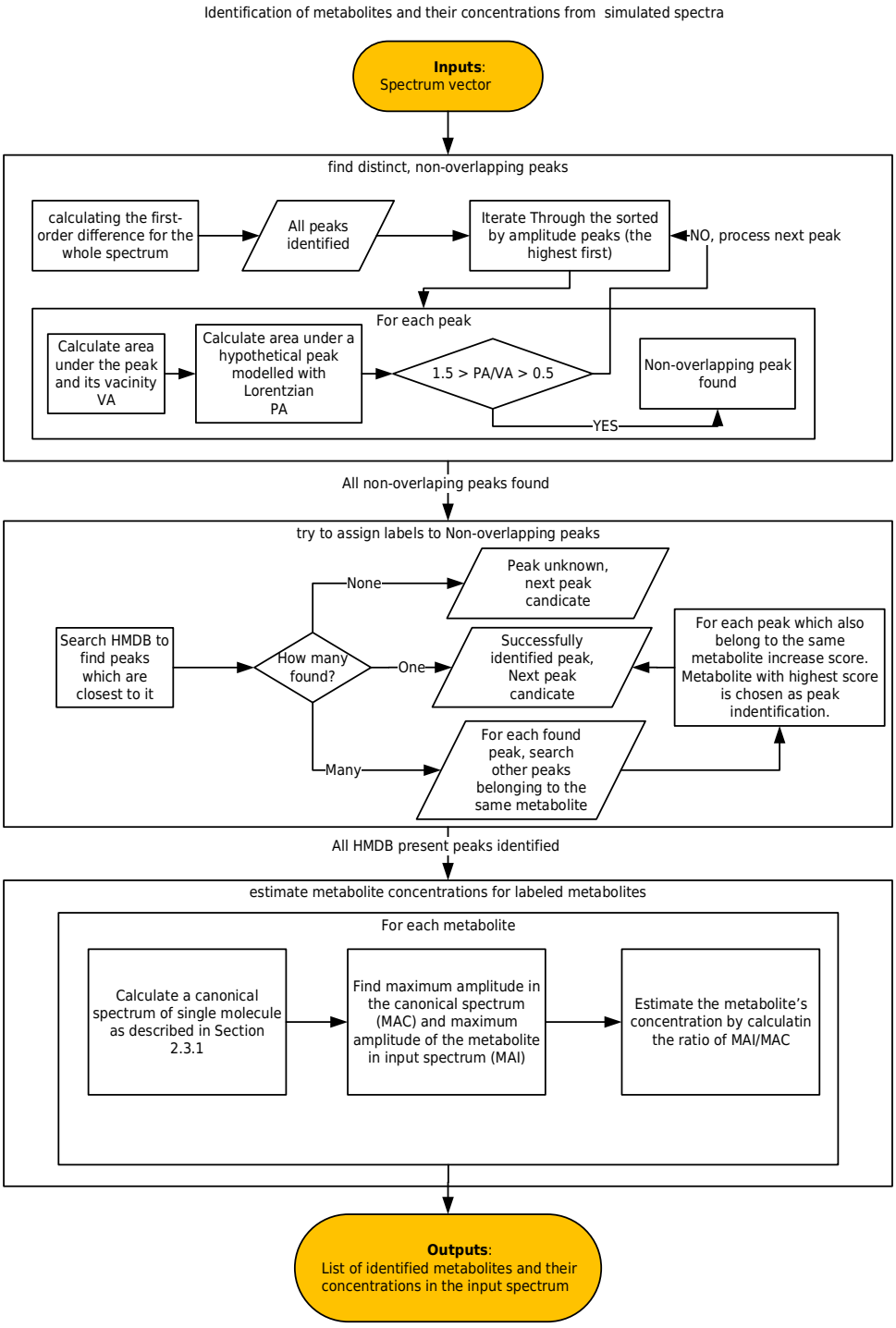

Figure 3.2: Identification of metabolites and their concentrations from simulated spectra flowchart

Identification of metabolites and their concentrations from simulated spectra

Our primary goal is not to solve the problem of identifying metabolites and their concentrations in a spectrum. Our goal is to find sets of co-occurring metabolites in samples. Peak identification is only an auxiliary task. Our method consists of three main tasks which are performed for each spectrum (see Figure 3.2):

1. Construction of a list of candidate peaks. Those peaks should be distinct and non-overlapping to increase the chance of identification of a metabolite.
2. Attempt to identify the metabolites for all candidate peaks. This step produces a list of identified metabolites in the spectrum. The list is usually shorter than the list of candidate peaks due to a possible failure of identification of some peaks.
3. Estimate levels of each identified metabolite.

The first task is to find distinct, non-overlapping peaks. We are interested in non-overlapping peaks because eventually we want to estimate levels. Peaks where there is overlap from multiple metabolites would make the task of level estimation hard if not impossible. Firstly, we identify all the peaks. We achieve this by calculating the first-order difference for the whole spectrum, looking at where it is zero and checking if it changed from positive before the plateau point to negative after[5]. This procedure is a discrete analogue of finding maxima of continuous functions with the first derivative. Once all the peaks are found, they were sorted by amplitude (the highest

[4]High resolution means $\geq$ 10,000 spectral bins.
[5]We used Python package PeakUtils v1.3

first). Next, we iterate through this sorted list to find non-overlapping peaks. A decision if a peak is non-overlapping is made by calculating a ratio of two areas: (a) the vicinity area (VA) around the peak using the trapezium rule and (b) the peak area (PA) under the Lorentzian function modelling a hypothetical peak with the same amplitude as the candidate. If the ratio $\frac{PA}{VA}$ is between 0.5 and 1.5, then the peak is classified as non-overlapping.

Once the candidate peaks are identified, an attempt is made to identify the metabolites from those peaks. This is achieved by using peak information from the Human Metabolome Database (Wishart et al., 2018). HMDB is a database of known metabolites and their properties including the standard chemical shifts. Finding a peak in HMDB means we know the metabolite. For each candidate peak, we search HMDB to find peaks which are closest to it. We limit our search to find the peaks only in a particular vicinity around the candidate peak. This search can result in one of three cases: (a) there are no peaks found in HMDB, (b) there is precisely one peak found and (c) there are multiple peaks found. In (a) we mark the candidate peak as unknown and move on to the next candidate. If (b), we assume that we successfully identified the metabolite. Lastly, in the case of (c), we apply the following procedure to decide which one to choose. For each peak found, we look at the metabolite and all its peaks. We try to locate those metabolite's peaks in the spectrum under investigation. We keep a count score which is increased by one for each metabolite's peak also found in the spectrum. The metabolite with the highest count score is assigned to the candidate peak.

The last task is to determine the identified metabolites' levels. For each metabolite we simulate a new spectrum with just this one metabolite. This

simulation is based on peak information from HMDB and uses the same methods, based on the Lorentzian function, as described before. Let us call this single molecule spectrum a canonical spectrum of the metabolite. A single molecule spectrum gives us values of peaks' heights for one unit of the molecule. Next, we look for a peak, belonging to the metabolite, with maximum amplitude in the spectrum under investigation. A ratio of the maximum amplitude of a metabolite's peak from the spectrum of interest and its counterpart peak in the canonical spectrum is the relative concentration we are looking for. We repeat this procedure for all identified metabolites in the spectrum.

3.3.1 *S. mansoni* data set preprocessing

We obtained raw data generated in *S. mansoni* study from Li et al. (2011). It consisted of 260 NMR spectra in the spectral range of 0-10 ppm with the resolution of 0.0005 ppm, 20,000 variables in total. This data was provided as a MATLAB file with a matrix of which the rows were samples and columns were spectral bins. Metadata was included as three vectors for sample identification: group ID, animal ID and day of measurement. The spectra were already preprocessed, so TSP, urea and water regions were removed. Further, we processed the data in the following way:

1. Rows of the matrix were sorted by group, day and animal ID (this helped with the manual part of the alignment procedure as it is easier to look at sorted samples).
2. Automatic alignment was performed using the method by Veselkov et al. (2009)
3. Manual alignment was carried out to optimise the output from the

previous step (in-house Matlab tool based on the method by Veselkov et al. (2009))

4. The data was downsampled by a factor of 10 to reduce computational run-time. The trapezium rule (MATLAB's `trapz` function) was applied to every ten spectral bins of the original data resulting in one new spectral bin accumulating the original values corresponding to a bin size of 0.005 ppm.
5. The probabilistic quotient method was applied to normalise the data (in-house MATLAB script based on the method of Dieterle et al. (2006))
6. A single spectrum (sample 91) was removed as it had bad water suppression resulting in high negative values.
7. Finally, the data were manually inspected, and high-level variables were created based on NMR resonances which we could manually identify. This resulted in a matrix of 260 rows and 38 variables corresponding to metabolites in Table 3.4. This matrix constitutes an input for the software implementation of LDA and PCA.

3.4 Results

3.4.1 Simulated data results

This section will cover the results for simulated data sets for which we know the ground truth. The data sets will be progressively more complicated as we want to raise the difficulty of inference.

Two and three topics systems

The two and three topics simulated data is the same as in the previous chapter (see Section 2.4.2) where we applied PCA to it. Here we use LDA, and have obtained the following results. For the two topic system, an average (over 100 spectra) Jaccard index between the identified metabolites in the spectra and all metabolites is 0.97. The simulation topics are also well recovered.

For the three topic system, the identification of metabolites is solid with a Jaccard index of 0.93. The simulation topics of the three topic system are also recovered well. In summary, we established that LDA could successfully infer topics from those toy data sets. Now let us move on to more complicated data.

Four topics simulations

We run four experiments using four topics simulation data (see section 3.2.1). Recall that the simulated topics were arranged in two ways: (a) topics do not share metabolites and (b) there are some shared metabolites between the topics. We refer to the former as non-overlapping topics and to the latter as overlapping topics, see panels A in figures 3.3 and 3.4. For visualisation in this section we use binary heat maps where the top 10 (or 12 in case of overlapping topics) metabolites are assigned a value of one and shown in light colour while the remaining metabolites in topics (rows) are assigned zero and are shown in dark colour. This technique helps us to focus on comparing inferred topics (panels B, C and D) with simulated topics (panels A).

Using those two data sets, with overlapping and non-overlapping topics,

we run two types of simulations: a full simulation including generating NMR spectra and a simplified version where we omit the NMR spectra simulation step and run LDA inference on just simulated concentrations matrix as obtained from the LDA-based generative process described in the previous chapter. The motivation to do this simplification is to probe how the loss of information arising from peak overlap in NMR spectra will affect topic inference. Peak overlap makes it harder to determine concentrations and also to identify smaller peaks which can be buried in the overlap area.

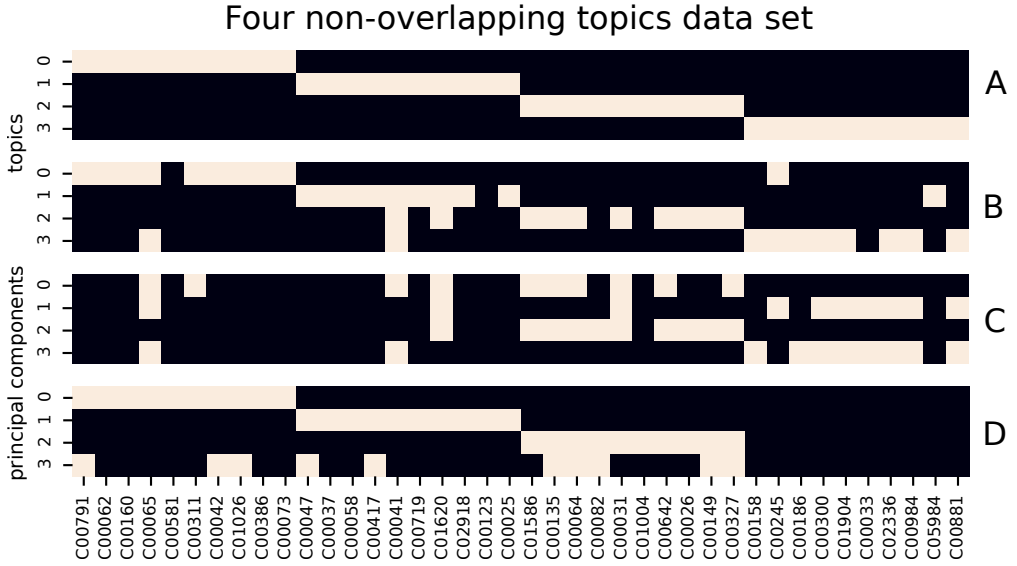

Figure 3.3: Results for the Four non-overlapping topics. Binary heat maps with the top 10 metabolites are shown in light colour. The metabolites are labelled by their KEGG IDs. (A) simulated topics, (B) LDA inferred topics for a data set with NMR spectra, (C) PCA components for a data set with NMR spectra, (D) PCA components for a data set with no NMR spectra (only simulated concentrations).

First we look at the results of simplified experiments with no spectra simulation, both non-overlapping and overlapping, see Figure 3.3, panel D and Figure 3.4, also panel D. LDA can recover all topics perfectly. The heat maps, in this case, are not informative as the inferred topics are identical as simulated topics. An interesting point to note is that PCA (visualised in

panels D) can recover only three topics and fails to recover the last one. The variance explained by principal components is [0.33; 0.32; 0.32; 0.0] for the non-overlapping case and [0.39; 0.37; 0.21; 0.0] for the overlapping case. All the data can be explained by the first three principal components. In particular, the last topic is missed altogether by PCA in the non-overlapping case, i.e. no principal component contains metabolites from simulated topic 3, see Figure 3.3, panel D. This is an example of the data where LDA can recover more information than PCA.

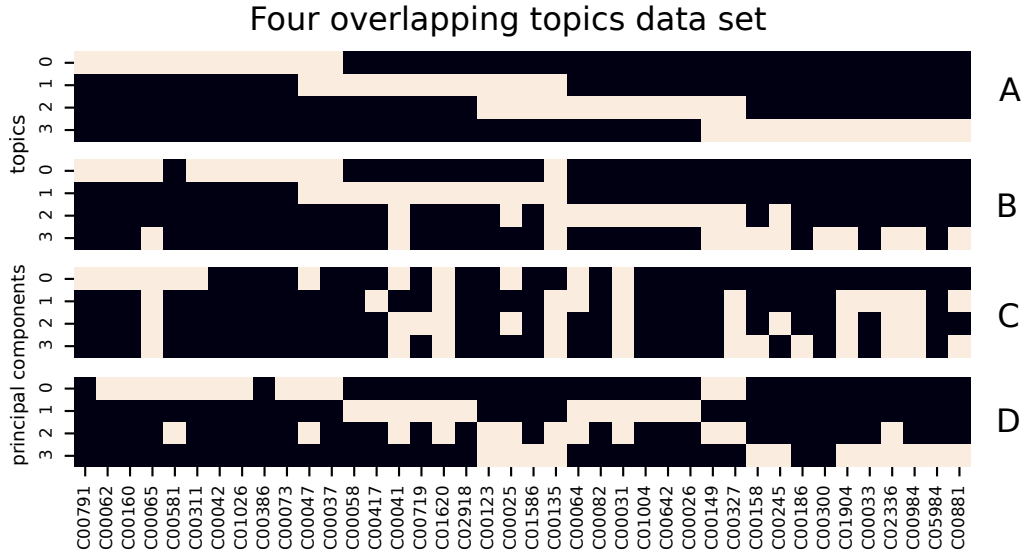

Figure 3.4: Results for Four overlapping topics, results. Binary heat maps with the top 12 metabolites shown in light colour. The metabolites are labelled by their KEGG IDs.(A) simulated topics, (B) LDA inferred topics for a data set with NMR spectra, (C) PCA components for a data set with NMR spectra, (D) PCA components for a data set with no NMR spectra (only simulated concentrations).

Now we move on to the non-overlapping data set with NMR simulations and metabolites identification from the spectra. Topic recovery for the non-overlapping topics experiment is illustrated in figure 3.3, panel B for LDA and panel C for PCA. In the case of LDA, note that topic 0 and topic 1 are recovered well, one out of ten metabolites are missing, resulting in a Jaccard index of 0.82. Topic 2 and topic 3 are recovered slightly

worse. Two out of ten metabolites are missing, producing a Jaccard index of 0.67. Moving on to PCA results, the variance explained by principal components is [0.65; 0.1; 0.06; 0.05]. Components 0 and 1 do not map well to the first two simulated topics, but component 2 misses only one metabolite of simulated topic 2 resulting in a Jaccard index of 0.82. Lastly, component 3 misses two metabolites of simulated topic 3 (Jaccard 0.67). We conclude that LDA performed well in topic recovery especially given that metabolites identification in the simulated spectra is an imperfect process. PCA performed visibly worse, with only two components giving a Jaccard index greater than 0.5.

Finally, we turn to our most complex experiment in this section: overlapping topics with NMR spectra simulation. Figure 3.4 depicts simulated topics in the top panel A (each topic consists of 12 metabolites), inferred topics for LDA in panel B and PCA in panel C. Topic 0 is recovered very well, only one metabolite is missed (Jaccard 0.85), topic 1 is recovered perfectly (Jaccard 1.0), topic 2 is missing two metabolites (Jaccard 0.71) and the last topic is missing 3 metabolites (Jaccard 0.6). This is a good performance from LDA in this complex scenario. PCA variance explained by the principal components is [0.52; 0.19; 0.13; 0.04]. The highest Jaccard index, 0.41, is for the first and the last component only, the two other topics are not recovered (Jaccard index less than 0.26). Overall PCA performs worse in this experiment than in the non-overlapping case. LDA seems to perform marginally better in the overlapping scenario than in the non-overlapping scenario.

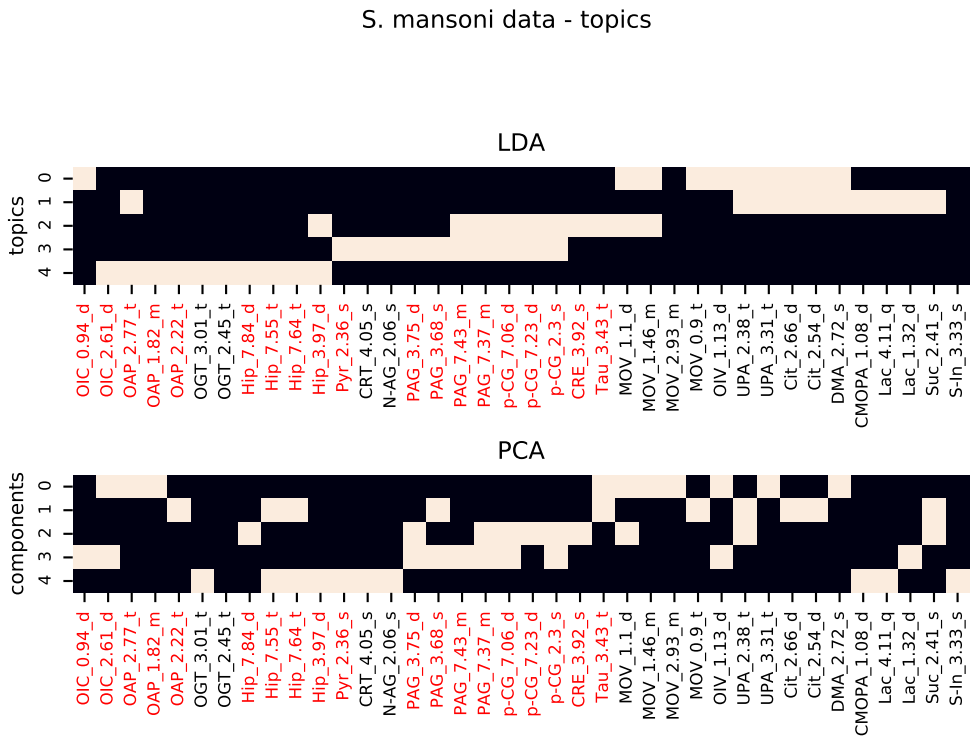

Figure 3.5: S. mansoni: Binary heat maps of inferred topics with the top 10 metabolites shown in light colour. (top) LDA inferred topics, (bottom) PCA components. In red, metabolites previously found (Li et al., 2011) to differentiate between experimental groups (see Table 3.4).

3.4.2 *S. mansoni* data set results

In this section, we will cover the results of a real NMR data set from the *S. mansoni* study described in section 3.2.2. We will use the findings from Li et al. (2011) which describes the study in detail, to act as the reference point as there is no ground truth in this case. Hippurate, phenylacetylglycine (PAG) and 2-oxoadipate (OAP) were showing the strongest connection with disease. p-cresol glucuronide (p-CG), creatine (CRE), 3-ureidopropionate (UPA), 2-oxoisovalerate (OIV), 2-oxoisocaproate (OIC), taurine (Tau), trimethylamine (TMA-N) and pyruvate (Pyr) were found to be involved in *S. mansoni* infection. In figure 3.5 and table 3.5, the metabolites which were found useful for differentiation between the experimental

groups are marked red.

In figure 3.5, we use the idea of binary heat map again to visualise the topics inferred by LDA and PCA. In the LDA case, we observe that topics three and four include hippurate, 2-oxoadipate(OAP), phenylacetylglycine(PAG), the three most reliable differentiators between the groups found previously in Li et al. (2011). For PCA, the results are less consistent: components 0 and 1 cover OAP bins, components 2 and 4 cover hippurate bins and component 3 cover PAG bins. We notice that the same metabolite bins are split between components while in LDA they are mostly members of the same topic. Variance explained by the principal components is [0.56; 0.21; 0.12; 0.05; 0.02].

A useful view of the LDA model is to plot NMR spectra in topic space, see Figure 3.6. Each spectrum is reduced to five latent dimensions. A spectrum is obtained from a sample from an individual at a time point, therefore a given animal will have multiple points in the plot, corresponding to 14 time points. The spectra are visualised in scatter plots, each plot for a pair of topics. Additionally, the samples are marked by shape, depending on the group, and colour depending on the time. Diamond shape represents controls and circles represent treatments; lighter shades indicate early samples and darker for late samples. The samples were split into early and late with day 41 acting as a division line, so each individual has five early data points and nine late points. The idea behind early and late data points is that the *S. mansoni* infection is not immediately symptomatic, the effects of infection are present only in the later stage. Figure 3.5 reveals that topics 3 and 4 contain many metabolites previously shown as involved in the separation between the groups. Topic 2 overlaps with topic 3, so it

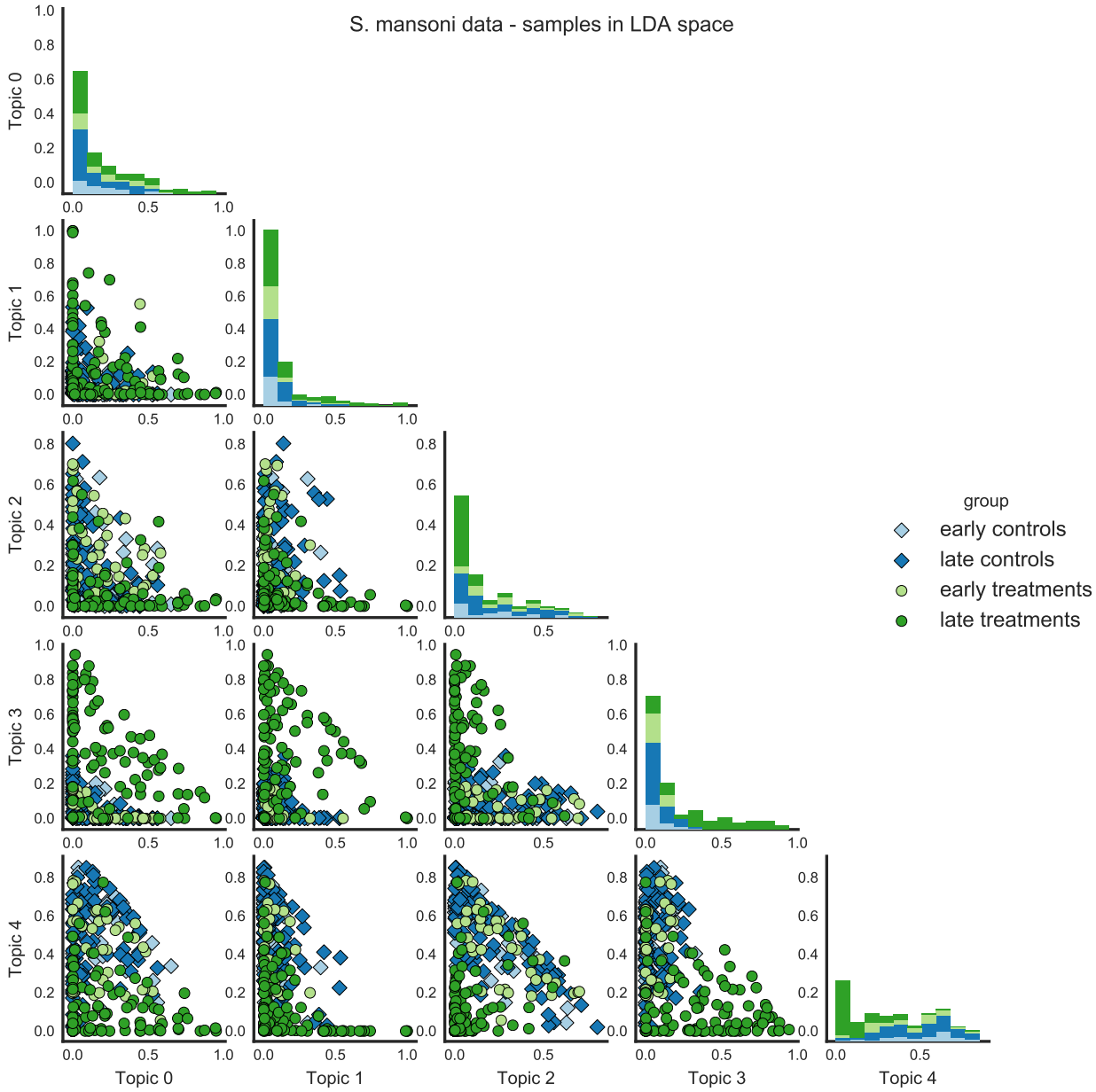

Figure 3.6: *S. mansoni*: NMR spectra in LDA topic space. Each animal has multiple points in a pair plot, corresponding to 14 time points. The samples were split into early and late with day 41 acting as a division line, so each individual has five early data points and nine late points. The late treatment samples seem to be well separated from the other points for Topic 3 and Topic 4. Topic 2 shares five metabolites with Topic 3 so the separation of late treatment samples can be observed as well. T-test and Kolmogorov-Smirnov test for late treatments were run on the topics, see Tables A.1 and A.2 in Appendix A

is also expected to be associated with distinguishing between the groups. Indeed, the late treatment samples seem to be well separated from other points if we carefully examine the rows for Topic 3 and Topic 4. Topic 2 shares five metabolites with Topic 3 so the separation of late treatment samples can be observed as well. For comparison, a similar analysis can be performed on the PCA model. Figure 3.7 presents projections of the spectra to the principal component space, also referred to as PCA scores. Component 2, when plotted against other components seems to separate late treatments from the rest of the samples.

Table 3.5: Two selected topics inferred with LDA on *S. mansoni* data. The previous study of the data (Li et al., 2011) found biomarkers (in red) differentiating between the experimental groups. The top 10 metabolites in the two selected topics include many of these red biomarkers. We investigated closer remaining (black) metabolites. Two of them N-acetyl-glycoprotein (N_AG) and 2-oxoglutarate (OGT) seem related to *S. mansoni* infection, we marked them with asterisk (more details in Section 3.4.3).

Topic 3	Topic 4
Pyr_2.36_s	OGT_3.01_t *
p-CG_2.3_s	Hip_7.55_t
p-CG_7.23_d	Hip_7.84_d
PAG_3.68_s	Hip_7.64_t
PAG_3.75_d	OGT_2.45_t *
p-CG_7.06_d	OAP_2.22_t
PAG_7.37_m	Hip_3.97_d
PAG_7.43_m	OAP_1.82_m
N-AG_2.06_s *	OIC_2.61_d
CRT_4.05_s	OAP_2.77_t

3.4.3 Potential new biomarkers

Topic 3 and 4 contain many metabolites found previously as the ones driving separation between the groups. Table 3.5 list the top metabolites in the two topics, red colour indicates previously identified metabolites. Let us look

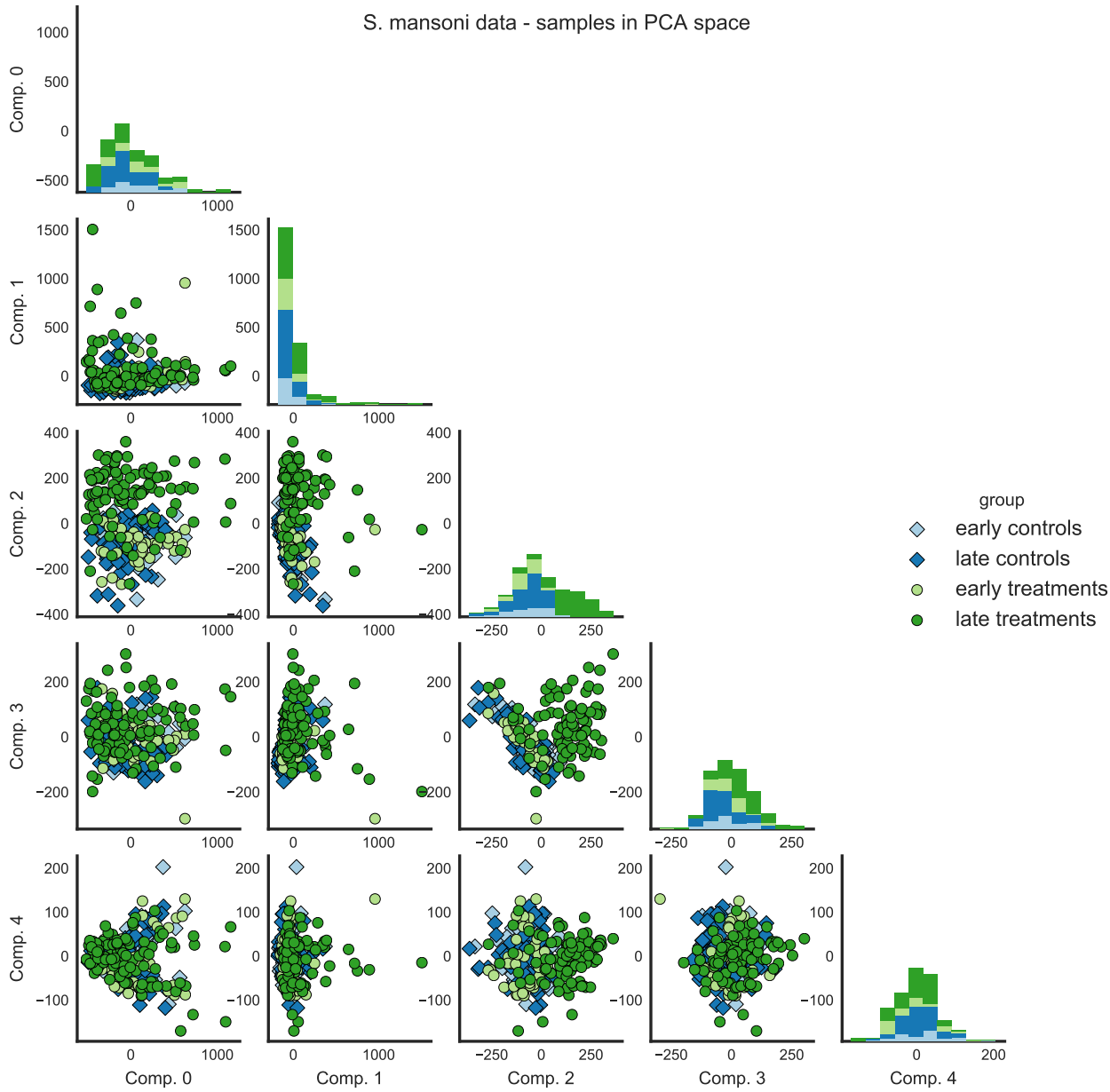

Figure 3.7: *S. mansoni*: NMR spectra in PCA space. Each animal has multiple points in a pair plot, corresponding to 14 time points. The samples were split into early and late with day 41 acting as a division line, so each individual has five early data points and nine late points. Component 2, when plotted against other components seems to separate the late treatments from the rest of the samples. T-test and Kolmogorov-Smirnov test for late treatments were run on the principal components, see Tables A.3 and A.4 in Appendix A.

closer at two metabolites which were not previously indicated: N-acetyl-glycoprotein (N_AG) and 2-oxoglutarate (OGT). N_AG was found by Li et al. (2011) as related to *S. mansoni* infection in faecal water, but our data is from urine. Wu et al. (2010) found elevated levels of N-acetyl-glycoprotein in plasma spectra in mice infected by *Schistosoma japonicum*, another species of *Schistosoma*. N-acetyl-glycoprotein seems to be a candidate biomarker. 2-oxoglutarate (OGT) was found by Wang et al. (2004) to be associated with *S. mansoni* infection in mice (urine samples). OGT is an intermediate of the tricarboxylic acid cycle. Disturbance of this cycle was found to be significant in the *S. japonicum* study by Wu et al. (2010). Figure 3.8 explores the relationships between four selected metabolites, two already identified as discriminatory for *S. mansoni* infection (p-CG and PAG) and two new ones N_AG and OGT. It can be determined from the scatter plots if the metabolites are higher or lower between the late treatments and the other samples. For example, PAG and p-CG shows late treatment samples to the right from the others, indicating that PAG and p-CG levels are higher than controls and early samples. This is consistent with the results from Li et al. (2011).

3.5 Discussion

In this chapter we investigated LDA inference on a variety of NMR metabolomics data sets in a meaningful way. LDA proved to be capable of modelling NMR data. We started with simulated data where LDA performed better than PCA in topic recovery. The comparison with PCA could be challenged as simulated data was produced with the LDA-based generative model. We recognise that this was an advantage for LDA. However, as shown in

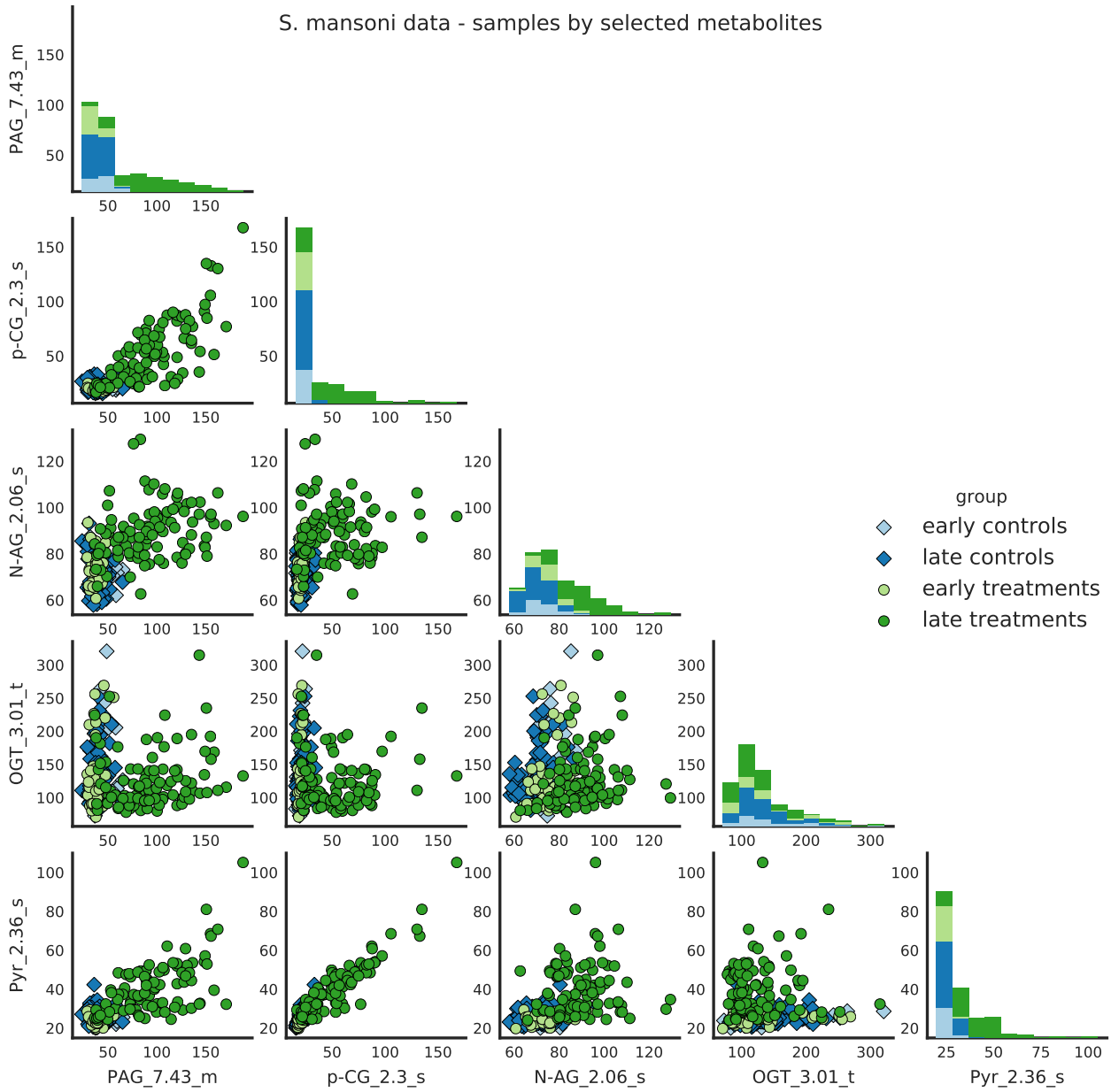

Figure 3.8: *S. mansoni*: a pairwise study of selected metabolites levels: two already identified as discriminatory for *S. mansoni* infection (p-CG and PAG) and two new ones N_AG and OGT. T-test and Kolmogorov-Smirnov test for late treatments were run on the metabolites levels, see Tables A.5 and A.6 in Appendix A.

Chapter 2 the simulated spectra are a reasonable approximation of real data sets. This could have been a fair point against LDA having a sort of

home advantage in experiments where only simulated concentrations were used (recall that LDA performed flawlessly), but once NMR spectra are simulated, translation to the NMR domain and associated loss of information makes it a levelled playing field for all models.[6]. It is noteworthy that that overlapping topics scenario, which is an attempt to mimic how topics would be in real biological systems, was not harder to infer compared to non-overlapping topics. Early on we speculated that perhaps the overlapping metabolites would be non-informative variables. It turned out that LDA picked correctly fourteen out of sixteen instances (a single square in Figure 3.4) which were shared between topics. Lastly, PCA did not perform as well but most notably it utterly failed to recover the last topic in the non-overlapping scenario.

In the case of the *S. mansoni* data set, LDA performed well, capturing all previously known metabolites associated with the disease. The model also suggested new metabolites to investigate. Two of the metabolites we selected for a follow-up analysis proved to be interesting from the biology point of view. PCA did not perform well, only component 2 seems to help to differentiate between the groups. Li et al. (2011) used PCA as well, and they did not find it helpful for differentiation between groups.

Limitations

Commonly in NMR metabolomics data analysis, multiple spectral bins may represent the same metabolite. Such variables will be correlated although not perfectly, as a single bin frequently contains signals from other metabolites. LDA attempts to group variables which change together within a sample

[6]The reader is kindly asked to excuse those sports metaphors, this section was written during FIFA World Cup 2018.

into topics. Therefore a topic may contain multiple bins related to the same metabolite. This is usually not a problem as observed in the results of *S. mansoni* data. What may be a potential problem for LDA in the context of NMR metabolomics is that metabolites with multiple bins can overshadow metabolites with only one bin, especially if the concentration in the latter is relatively low. Effectively, the multi-bin metabolites have a stronger signal and this can bias the analysis in their favour. We did not find a way to mitigate this bias with the application on LDA.

Our algorithm for identifying metabolites and estimating concentrations is not very sophisticated. Finding peaks in NMR spectra could easily grow into a separate project. Figuring out concentrations in peak overlap regions can be very difficult. Our algorithm finds only peaks present in a local database which is limited to normal urine metabolites. Real life NMR data will be subject to peak shifts which we do not address.

Future work

The method described above is only the first attempt at applying LDA to NMR data. We focused on achievable tasks to prove that this approach works. The scope of future work is outlined as follows:

1. Topics are sets of metabolites; high probability metabolites within a set tend to covary in a sample. Those metabolites could be mapped to a biological process or a metabolic pathway. Such mapping, although never perfect, could be a useful tool for providing an interesting angle to look at NMR spectra and connect them to the underlying biology of the sample.
2. Detailed analysis of metabolites could be extended and driven by

more than just LDA topics. Metabolites with high PCA loadings could be considered.

3. More latent variable methods could be investigated to compare with LDA, for example, Non-negative Matrix factorisation.
4. In simulated data experiments, we chose the number of topics to be the same as the original simulation. What would happen if we chose other K, i.e. if simulation K and inference K are different? Five topics were chosen for *S. mansoni* as the smallest K where previously known metabolites were captured coherently. However, for a data set without any prior information, how would we choose the number of topics?
5. Each topic is a probability distribution over all metabolites of interest, i.e. each metabolite has been assigned a probability. Different topics are the same list of all the metabolites of interest but with different probabilities assigned. In practice, we always focus on the highest probability metabolites, usually the top 10 in the list. However, 10 was an arbitrarily chosen number. Other options need to be investigated. Is there an optimal top N number?

Chapter 4

Supervised Latent Dirichlet Allocation for continuous response in ^{1}H NMR data

4.1 Introduction

This chapter will continue to focus on topic models but this time in a supervised learning context. We will apply a variant of the LDA model, called SLDA, which is extended to push topics composition according to a continuous variable representing a measured response variable, for example some clinical outcome like BMI. The continuous response variables are associated with each document or spectrum.

We are more interested in an interpretable model than prediction metrics but, of course, they should be satisfactory at least. The interpretability of the model is our primary objective. We want to find interesting patterns in data, not to predict response variables from unseen NMR spectra. We claim that a small number of topics consisting of spectral bins are interpretable

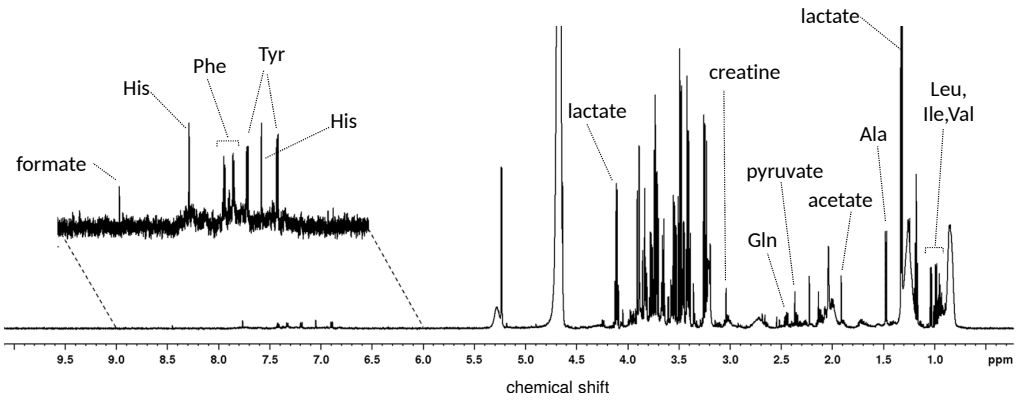

Figure 4.1: MESA study, sample spectrum of serum. Figure courtesy of Gonçalo Graça (private communication)

for a human researcher. In contrast, PCA, a popular method in NMR metabolomics, can be used as a latent variable model but its loadings do not easily lend themselves to interpretation. Although we could normalise the loadings and take their absolute value, so they fall in [0,1] range, they are not as easily interpretable as probabilities in topics which can be understood intuitively.

4.2 Data sets

4.2.1 MESA data set

We obtained untargeted one-dimensional (1D) serum 1H NMR metabolomics spectra of participants from the Multi-Ethnic Study of Atherosclerosis (MESA), see sample spectrum in Figure 4.1. The MESA study (Bild et al., 2002) set out to investigate the cardiovascular disease (CVD) in a sample of 6500 men and women aged over 40 years from the US. The proportions of ethnicity were as follows: 38% White, 28% African-American, 23% Hispanic, and 11% Asian (of Chinese descent). Each sample was associated with 86 measurements such as coronary calcium, ventricular mass, carotid

intimal-medial wall thickness, blood pressures, standard CVD risk factors, socio-demographic factors, and life habits. We provide more details on how we preprocessed the data in Section 4.3.3.

4.3 Methods

4.3.1 Supervised LDA model

A supervised Latent Dirichlet Allocation (SLDA) model builds on the LDA model to include a response variable which is associated with a document. SLDA model in plate notation is depicted in Figure 4.2. Y_d is a continuous response variable associated with a document or spectrum. The model was first introduced in Mcauliffe and Blei (2008) and refined in Blei and McAuliffe (2010). The later version uses the generalised linear model (GLM) by McCullagh and Nelder (1989) to model the Y_d response variable. In the original version, which we use here, the Y_d response variable is assumed to be distributed by the Gaussian distribution with mean η and variance σ^2.

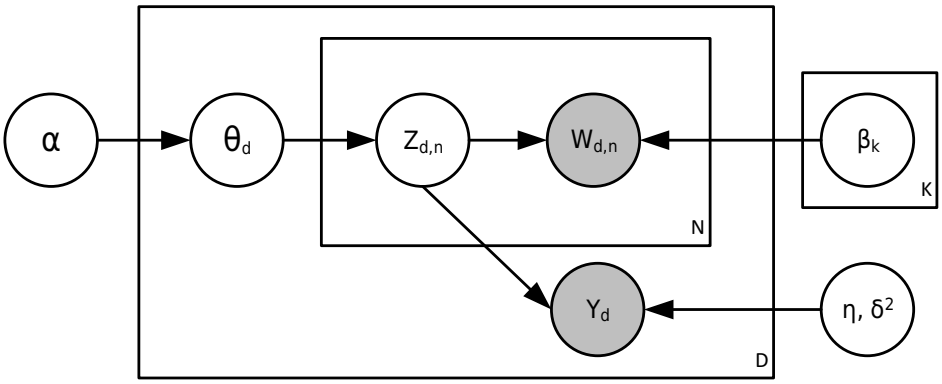

Figure 4.2: SLDA model plate diagram

4.3.2 Linear regression of latent variable models

Our primary model in this chapter is SLDA. However, we additionally fit three other models for comparison. They are a combination of latent variable models with linear regression which is a way of turning an unsupervised method into a supervised method.

Figure 4.3 gives an overview of the process in which we combine latent variable models with linear regression. At the top, we have a spectrum within a process of digitisation is converted into N bins (or columns in matrix X). N can be quite large for high-resolution spectra, in order of tens of thousands. At the same time, the number of samples M is usually in the order of hundreds or thousands. The conventional way of storing NMR spectra is in the form of a matrix (let us call it X) of size $M \times N$. Fitting linear models in a situation where the number of columns is much higher than the number of rows may be problematic. We reduce the number of columns by applying a latent variable model to the matrix X. Such models, LDA or PCA, will create new variables (K latent variables) and represent spectral data expressed in the new coordinates. The result is matrix T. Matrix T is much easier to work with because the number of latent variables is much smaller than N. So far we did not take into account the response variables related to NMR spectra. The linear regression can fit matrix T along with response variable vector b. The result is linear coefficient matrix B which is used to predict values of the response variable.

We use ElasticNet (Zou and Hastie, 2005) as our linear regression model of choice. We use ElasticNet because it works well for correlated features (Hastie, 2015). The latent variable models we combine with ElasticNet are SLDA, LDA and PCA. SLDA combined with ElasticNet means that the

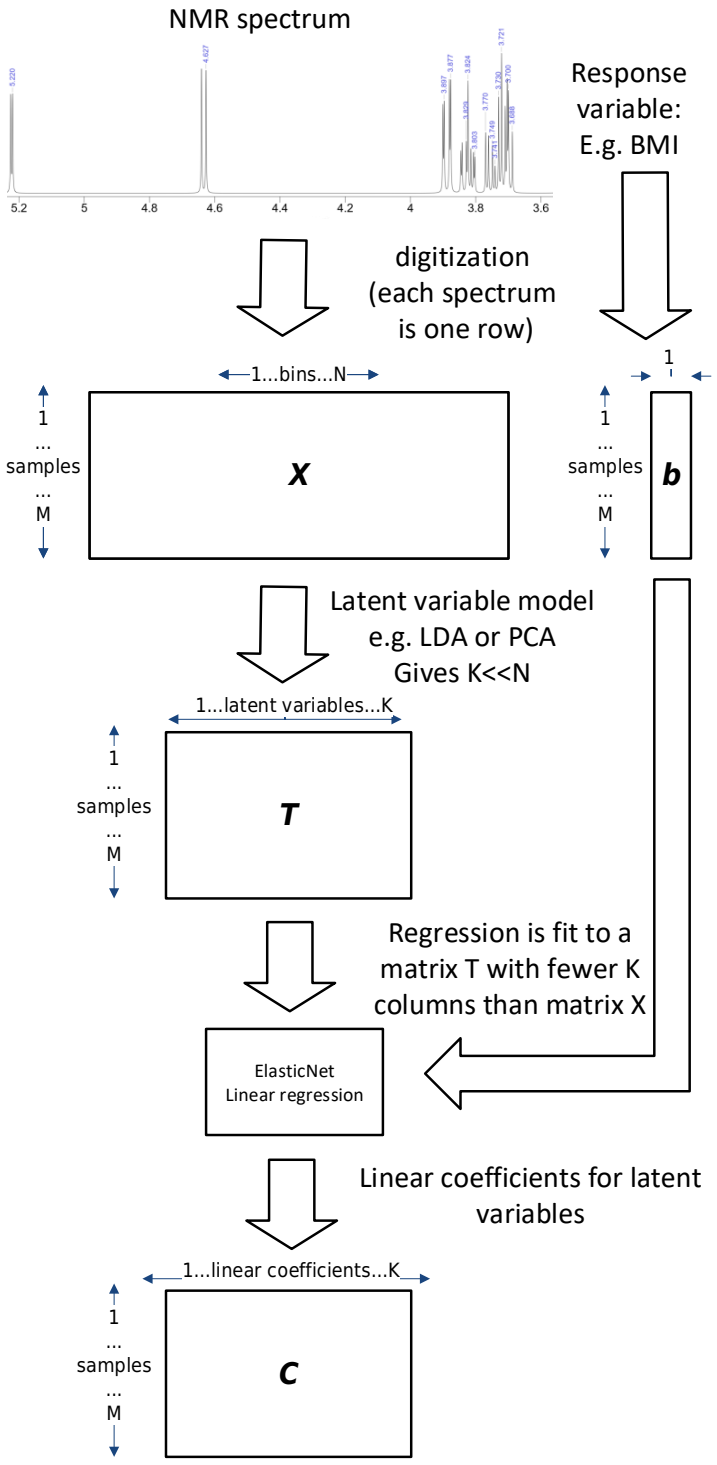

Figure 4.3: Linear regression of latent variable models

topics from sLDA (but not predictions) are used as features for ElasticNet. We expect that topics from sLDA will be different to unsupervised LDA topics. In the case of combined models, we first fit latent variable models; next, we fit ElasticNet to the spectra from the training set represented in the latent variable space. For example, if we take a topic model with $K = 10$ topics, a spectrum which originally was represented by 444 bins is transformed into 10 latent variables. ElasticNet is fitted on those 10 latent variables and will result in 10 linear coefficients, one per variable.

4.3.3 MESA data set preprocessing

Clinical variable selection

We obtained data files containing 3948 spectra. Each spectrum is associated with 86 clinical variables; some are continuous, some categorical and some binary. Here, we list a subset of the variables to give an idea what sort of information was recorded for each spectrum: age, sex, height, weight, BMI, ethnicity, smoking, education, blood pressure (SBP, DBP), glucose level, LDL and HDL cholesterol levels and many others. Due to the long running-time required to fit the sLDA model[1], we narrowed down the list to six clinically important variables: height, BMI, total cholesterol, LDL cholesterol, HDL cholesterol and glucose level. Height act as a negative control variable, we do not expect to find a metabolic signature related to height in the cohort of adults 40 years old and older. Glucose level acts as a positive control since the glucose signal in the spectra is directly present in the data. We expect glucose to be present in the inferred topic is some particular way. BMI and three cholesterol variables were chosen as they

[1]For K=20 topics and 10,000 iterations of Gibbs sampler, it takes about 48h to fit a single variable sLDA model using a virtual server on Google Cloud Platform.

play a role in atherosclerosis, which is central to the MESA study. Each of our variables of interest is represented in multiple spectral bins.

Spectral bin selection and annotation adjustment

The MESA spectrum data file was obtained as a matrix of 3948 rows and 710 columns. Each row represents an NMR spectrum of a serum sample from participants of the MESA study. The columns are manually created bins based on high-resolution spectra. An expert NMR spectroscopist manually annotated about 500 bins. Each annotation specifies one or more metabolites. We did not use the bins without annotations for modelling as we would not be able to interpret the topics. We also eliminated bins representing ethanol (possibly part of the diet), bins without a metabolite signal (noise only bins), and bins annotated as a baseline. Lastly, we eliminated two bins related to L-Lactic acid due to its very high peak, compared to the other bins; this is similar to the removal of common words in documents which would have a high word count. Also, L-Lactic acid is not important in the context of our selected clinical variables. In total, there were 444 bins left which we used for analysis. The bins annotations describe potential signal sources[2] and, of course, several bins could be associated with a single molecule. This leads to a significant repetition in the bin annotations; there are only 169 distinct annotations for 444 bins. Out of 169 annotations, 131 falls into only three broad categories: glucose related(55), cholesterol related(14) and fatty acyl chains related(62). We used this fact to simplify the annotations by using those three labels instead of the original lists of candidate source metabolites. The remaining 38 annotations

[2]As an example annotation, here is a list for bin #6: `CH3(CH2)n (fatty acyl chains); CH3 (C26 and C27 from Cholesterol); unknown 5.`

were changed to "other". This simplification allowed us to focus on the interpretability of the topic models, concretely, our inferred topics consist of bin numbers and simple annotations instead of using original lists which, although they may be precise from the chemistry point of view, are difficult to read when presented in a topic format.

Selection of spectra for train and test sets

Not all samples in the MESA data set were associated with clinical variables; some values were missing. In our analysis, we use only samples for which we have all the values of our selected six clinical variables. Because we operate in a supervised learning regime, we can evaluate the performance of our models. We split all the samples into two groups: training and test sets. The training set is used for inferring latent variables (topics and principal components). The test set is used to calculate the performance of the models. This is a standard approach in machine learning to ensure that models do not overfit to data. We decided on splitting the data set into 1000 spectra for training and 200 spectra for testing. This was primarily influenced by our experience of training the models and their run-times. Training on ~3,500 spectra was too slow. Special care was taken to select spectra for which the train and test clinical variables were balanced between classes. The train and test classes are balanced if 90% confidence intervals endpoints of both sets were within 10% from each other and also their means were within 10% from each other. For example, this is how this split works for Glucose. The 1000 train spectra have a mean Glucose of 98.66, and the endpoints of the range that contains 90% of that data are 48.31 and 149.02. The test set consists of 200 spectra for which the mean Glucose

was 96.38 and the endpoints of the range that contains 90% are 50.21 and 142.55. The final split train/test consists of 1000 train spectra and 200 test spectra for which class balance conditions are fulfilled for all six clinical variables of our choice thus we are able to use the same train/test sets for all MESA modelling in this chapter.

4.3.4 Model evaluation

The evaluation of models in this chapter is twofold: (a) strength of prediction and (b) interpretability of topics. Part (a) is straightforward as we can use some standard prediction metrics. Part (b) is not easily quantifiable; we evaluate topics in the context of the clinical variable in question. For example, when fitting models with Glucose levels, do we observe topics related to Glucose?

The two prediction metrics we use here are the Pearson product-moment correlation coefficient r, and the normalised root mean square error (NRMSE). The former is a standard measure for linear dependency between two variables. We use a correlation coefficient to assess linear dependency between values of clinical variables from the test set (200 values) versus predicted values from the models. We arbitrary chose a cut-off value $r > 0.75$ as a boundary to indicate a satisfactory relationship between test values and predictions.

NRMSE is used to asses errors in predictions from our models. A mean square error (MSE) is a standard measure of goodness of prediction. Its value is an average squared error: $\frac{1}{N}\sum_1^N (y_i - \hat{y})^2$ where y_i are data points and $\hat{y}$ is a mean value. The reason for squaring the errors is somewhat traditional and can serve, to our knowledge, three reasons: (a) squaring

makes all errors positive values so they can be summed up meaningfully, (b) squaring the errors makes larger errors more pronounced, (c) MSE can be used as an objective function in optimisation algorithms such as gradient descent which require differentiable objective functions in order to work[3]. Lastly, we take the square root of MSE (it becomes RMSE) so that the errors are on the same scale as the y_i values, and we normalise RMSE by the standard deviation of the response variable to be able to compare NRMSE between clinical variables. The full formula for NRMSE is:

$$NRMSE = \frac{\sqrt{\frac{1}{N}\sum_1^N (y_i - \hat{y}_i)^2}}{\sqrt{\frac{1}{N-1}\sum_1^N (y_i - \bar{y})^2}}$$

where y_i is the observed clinical variable for the i-th sample in the test set, $\hat{y}_i$ is i-th predicted value and $\bar{y}$ is the observed mean value.

$NRMSE < 1.0$ can be interpreted as: on average errors are less that one standard deviation. We deem the models with $NRMSE < 1.0$ and $r > 0.75$ to be satisfactory in the prediction task for the clinical variables.

We are primarily interested in the interpretability of our models, but of course, those models must fit the data, at least in a satisfactory manner. We use prediction metrics to select good models in order to investigate topics and try to interpret the patterns found (or lack of).

Choosing number of latent variables

One of the usual problems with any latent variable models is how to decide on a number of variables K. There are suggestions in the literature (Murphy, 2012), but they are usually of a theoretical nature and not always practical

[3]We potentially could use absolute value of errors $|y_i - \hat{y}_i|$ as we do not require differentiability ($|x|$ is not differentiable at zero).

to apply. For example, in the case of SLDA, running-time can take days for a large enough data set and a number of topics. Any strategy requiring fitting the model for a range of Ks and calculating some score to help to find optimal K is not feasible for us due to its high cost. Our solution to this problem was by fixing K for 2, 5, 10, 15 and 20 topics. This approach provides some insight as to what happens when a number of topics increases.

4.4 Results

4.4.1 MESA study data

In this section, we describe the results of fitting models to the MESA data. Our primary model is SLDA, but we also fit additional models: PCAEN, LDAEN and SLDAEN. The EN suffix indicates that those are a combination of latent variable models (PCA, LDA and SLDA) with ElasticNet linear regression. All models were fitted for each of our chosen clinical variables: height, BMI, total cholesterol, LDL cholesterol, HDL cholesterol and glucose. Each model for each clinical variable is fitted with five different topic numbers $K = 2, 5, 10, 15, 20$. This gives 20 variants per clinical variable, 120 variants in total.

Convergence in topic models

All our models other than the PCA use collapsed Gibbs sampling for inference. Gibbs sampler is a Markov Chain Monte Carlo (MCMC) type of algorithm. It is a standard procedure for MCMC to assess the convergence of the log-likelihood of the model to target distribution in the data. The

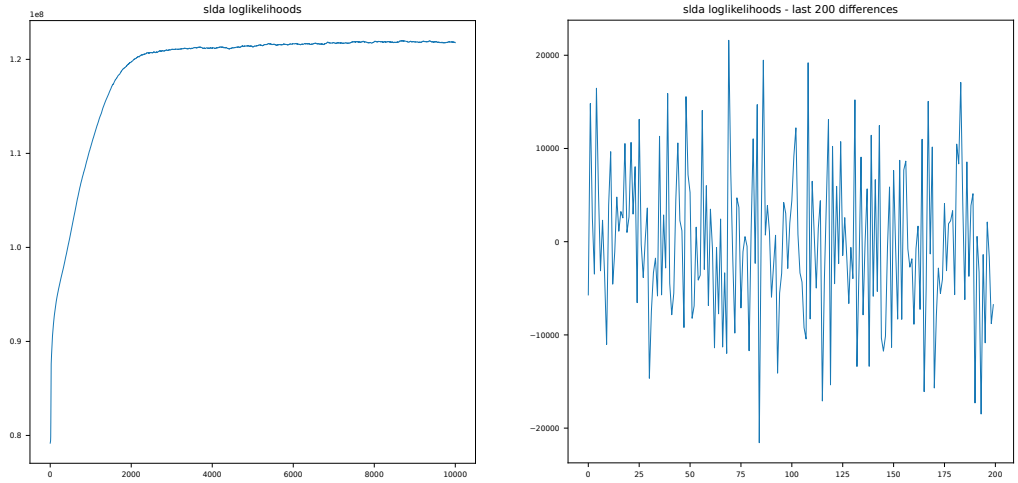

Figure 4.4: Assessing sLDA model convergence on Glucose example

key question is how many iterations of Gibbs sampling must be performed, so that the Markov Chain converges to its stationary distribution. This number of iterations must be executed in a reasonable amount of time, given available hardware. Within the timescale of our project, a reasonable amount of time translates to a couple of days at most. This depends on hardware; we used 24-core virtual machines on Google Cloud Platform. We run our experiments using an increasing number of iterations N_{iter}: starting with 2000, moving to 5000 and finally reaching 10,000 iterations where all our models are converged. Each time we plotted log-likelihood as in Figure 4.4. This Figure is specific to SLDA and Glucose as a clinical variable, but it gives a general idea of how those plots work. The left panel shows the value of not normalised log-likelihood over a number of iterations. In this kind of plot, we look for the likelihood to plateau to the right (high end of the number of iterations). The values of non-normalised log-likelihood might be high. With high values of this magnitude, it may be hard to assess if the log-likelihood plateaued or not. The solution to this

problem is in the right panel of Figure 4.4. In this plot, we show the first differences between the last $N_D = 200$ values of the likelihood. The first difference is defined as $D[n] = loglikelihood[n+1] - loglikelihood[n]$ where $n = N_{iter} - N_D, \ldots, N_{iter}$. The idea is that if the log-likelihood converged, the differences should oscillate around zero, as seen in this case on the right panel. To put a numeric measure on this visual intuition, we created a metric representing a log-likelihood difference standard score (LDSS) defined as the mean of the differences, $\bar{D}$, over the standard deviation of the differences:

$$LDSS = \frac{\bar{D}}{\sqrt{\frac{1}{N}\sum_1^{N-1}(D_i - \bar{D})^2}}$$

A value of LDSS less than 1.0 is an indicator that the model converged successfully. In fact, for all our models, for each of the six variables of interest, we observed $LDSS < 0.2$ with the number of iterations $N_{iter} = 10,000$ and the number of differences $N_D = 2000$, giving us confidence that the Gibbs sampling procedure converged successfully.

Prediction results

Six clinical variables were used for fitting the models. We defined the conditions for a model to be considered satisfactory in the prediction task to be $NRMSE < 1.1$ and $r > 0.7$. Models for two clinical variables: HDL cholesterol and Glucose, fulfilled these criteria. The models for other variables did not produce adequate results. Table 4.1 presents detailed results for the two clinical variables: Glucose and HDL cholesterol. We wish to highlight a number of observations. (A) the SLDA model prediction is satisfactory but using just topics from SLDA and combining them with

linear regression provides better results. (B) Unsupervised LDA topics combined with ElasticNet is as good, or slightly better than SLDAEN, suggesting that SLDA did not infer better-related topics with regards to the response variable. (C) PCA combined with ElasticNet outperforms topics models in the prediction task. This is perhaps unsurprising as the objective of PCA is to find a direction of greatest variability in the data thus providing excellent representation in the lower dimension. Topic models on the other hand focus on proving easy interpretability of the model which might not help the linear regression score highly on the prediction task.

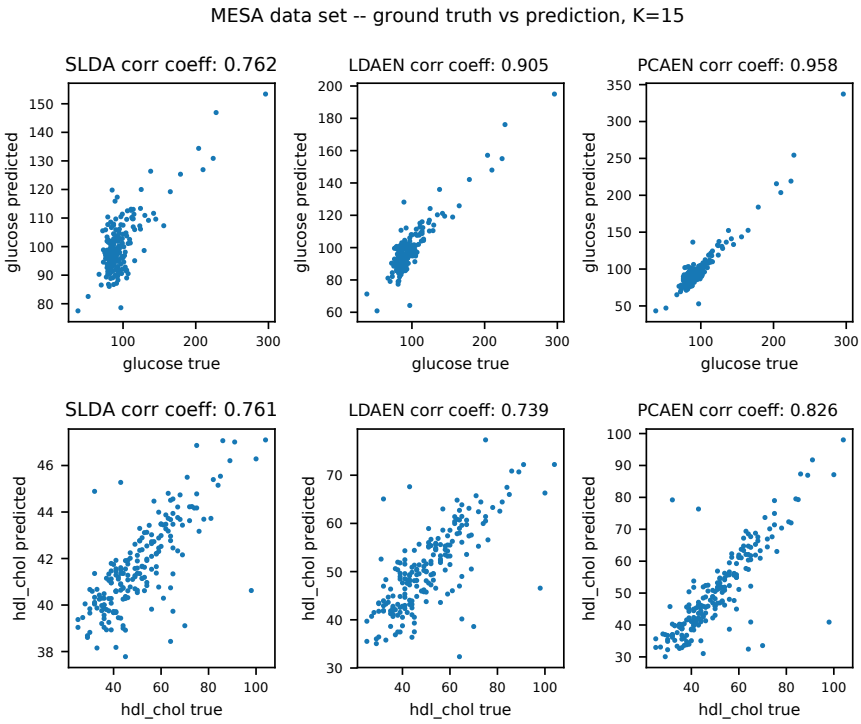

Figure 4.5: Scatter plots for K=15 topics

Nevertheless, in many cases, the performance of all the models is comparable suggesting that topic models have proven to come up with reasonable latent variables to represent the data. For illustration see Figure 4.5, showing the case with $K = 15$, top row panels represent models for Glucose and

bottom row panels for HDL cholesterol. The models are arranged from left to right in order of performance, on the left SLDA, LDAEN in the middle, and PCAEN on the right. The metrics from Table 4.1 show which models perform the best but by examining the scatter plots we note that although there are differences, all models capture the linear nature of the relationship between the predicted values and the actual values from the test set. Looking at the scale of the axes we note that SLDA makes more substantial errors in the prediction than PCA which has both axes on the same scale. Errors can also be visually assessed by looking at the spread of the "cloud" where the data points form. For example, for HDL cholesterol, the left panel for SLDA has NRMSE$\approx$1.0 while the right panel for PCAEN has NRMSE=0.56.

We investigate how to choose a number of topics by fixing K for 2, 5, 10, 15 and 20 topics and watching how it influences the prediction performance metrics. First, let us notice that very high values for K will defeat our goal of having interpretable models. In the most extreme case, K would approach a number of spectral bins. This would increase the prediction performance in the models with linear regression as ElasticNet is very effective in picking up variables in the data, so the prediction metrics are rather good. We confirmed this by running ElasticNet on all 444 spectral bins, i.e. without any latent variable models, and obtained correlations coefficients and NRMSE which were better than any of the models with latent variables. On the other hand, K=2 reduces the dimensionality of the spectra too much, resulting in poor prediction performance. This approach gives some insight as to what happens when a number of topics increases but it is possible only in supervised learning. In unsupervised learning it

is hard to construct such a metric because we do not make a prediction, an error cannot be calculated. A goodness of fit metric could help to make a judgement about choices of a number of topics. One such metric is perplexity. Its origin is the information theory and it is frequently used in natural language processing as a metric to evaluate how well the model predicts a set of documents $\boldsymbol{w}$. Perplexity is defined as:

$$perplexity(\boldsymbol{w}) = exp\left\{ -\frac{1}{D} \sum_{d=1}^{D} \log p(\boldsymbol{w}_d | \boldsymbol{\beta}, \boldsymbol{\theta}_d) \right\}$$

Lower perplexity indicates a better fit between the model and the data. The model is less surprised or perplexed by the sample. In practice, perplexity may not always agree which human intuition. Chang et al. (2009) show that for text documents, perplexity and human perceived topics may not be correlated, and sometimes may even be anti-correlated. There are alternatives to perplexity which were investigated by Wallach et al. (2009).

Despite the differences in the prediction performance, we consider that for K=10 and K=15 all the models performed, at least at a satisfactory level and this conclusion allows us to focus on the interpretation of the topics. Our primary goal was to find topics which may potentially lead to interesting biological insights, not a prediction of clinical variables from NMR spectra.

In Appendix B (page 134) we included a full listings of topics (from SLDA and LDA) and principal components inferred from MESA data. Before we present the selected findings, we describe the format we use for reporting topics. Here is an example topic:

```
LDA 10 topics
Topic 0
reg. coef. hdl_chol: 34.323
```

```
reg. coef. glucose: -26.383
53   fatty acyl chains
5    cholesterol
6    cholesterol
4    cholesterol
502  other
7    cholesterol
282  other
119  fatty acyl chains
120  fatty acyl chains
283  glucose
```

In the first line, a model name and a number of topics are stated. Here we show[4] `Topic 0`. Next, we have regression coefficients for this topic from the linear regression model. In the case of LDA and PCA, there will be two regression coefficients values for the two variables that we focus on: HDL cholesterol and Glucose. LDA and PCA models are unsupervised methods, so the clinical variables do not play any part in fitting the topics; there is just one set of topic irrespective of any clinical variables. Next, we have a list of the top ten spectral bins in the topics, in order of importance. Each row consists of a bin number[5] and a label. The labels are simplified bins annotations (see Section 4.3.3 for details). Notably, in this particular topic, we have four adjacent bins with contributions from cholesterol and a positive regression coefficient for HDL cholesterol indicating that this topic has a positive contribution in cholesterol prediction.

We now focus on the SLDA results for Glucose. We fitted the SLDA model for Glucose for K=2, 5, 10, 15 and 20 topics. For K=2 none of the two topics emphasises Glucose, which is consistent with the metrics in Table 4.1 where all the models for K=2 and Glucose yielded insufficient correlation coefficients. For K=5, the topic with the highest regression coefficient of 336.08 contains nine Glucose bins:

```
SLDA 5 topics
Topic 3, reg. coef. 336.08
4       cholesterol
```

[4]Topics are numbered starting with zero (Python convention).

[5]Full list of spectral bins with annotations is available in Appendix B (page 134).

```
309     glucose
325     glucose
367     glucose
311     glucose
388     glucose
500     glucose
323     glucose
501     glucose
365     glucose
```

We have a similar situation for K=10 and K=15. We present the topics with the highest regression coefficients.

```
SLDA 10 topics
Topic 7, reg. coef. 328.28
4       cholesterol
53      fatty acyl chains
309     glucose
502     other
325     glucose
367     glucose
311     glucose
388     glucose
500     glucose
323     glucose
```

In the case of K=15, we observe that topic 10 is almost identical to Topic 3 for K=5. The only difference is that the last two bins are in reverse order.

```
SLDA 15 topics
Topic 10, reg. coef. 352.843
4       cholesterol
309     glucose
325     glucose
367     glucose
311     glucose
388     glucose
500     glucose
323     glucose
365     glucose
501     glucose
```

It is also worth noting that for K=15 and K=20 there are topics which consist exclusively of Glucose bins, for example:

```
SLDA 15 topics
Topic 11, reg. coef. 90.959
367     glucose
309     glucose
325     glucose
283     glucose
289     glucose
311     glucose
388     glucose
368     glucose
500     glucose
399     glucose
```

The results above show that the SLDA models for Glucose not only give good prediction performance but also give topics which are clearly Glucose focused. This is an important confirmation that we can model clinical variables using SLDA. We expected this result as Glucose is directly present in the NMR spectra. Although a Glucose pattern might seem simple to find, its metabolic profile is complex, with many peaks which frequently overlap with other molecules. We must note that LDA and PCA also have Glucose only topics. LDA for all values of K other than K=2 (for tables with K=2 see Appendix B, page 134) produced all Glucose topics with high correlation coefficients.

LDA 5 topics	
Topic 3	
reg. coef. glucose:	
370.689	
367	glucose
309	glucose
325	glucose
311	glucose
388	glucose
500	glucose
501	glucose
365	glucose
323	glucose
310	glucose

LDA 15 topics	
Topic 13	
reg. coef. glucose:	
391.078	
309	glucose
367	glucose
325	glucose
311	glucose
388	glucose
500	glucose
365	glucose
501	glucose
323	glucose
310	glucose

LDA 10 topics	
Topic 7	
reg. coef. glucose:	
400.667	
309	glucose
325	glucose
367	glucose
311	glucose
388	glucose
500	glucose
501	glucose
365	glucose
323	glucose
310	glucose

LDA 20 topics	
Topic 14	
reg. coef. glucose:	
461.921	
309	glucose
325	glucose
367	glucose
311	glucose
388	glucose
323	glucose
500	glucose
501	glucose
365	glucose
310	glucose

The third principal component (PC3) from PCA is all Glucose. This is an important observation as PCA models the directions of the most significant variability in the data. PC3 inform us that, regardless of any clinical variables, the MESA data set strongly vary in the direction associated with

Glucose bins. Glucose bins are variables which tend to vary together in this data set.

```
Principal component 3 (PC3)
reg. coef. glucose: 0.129
309     glucose
325     glucose
367     glucose
311     glucose
388     glucose
323     glucose
500     glucose
501     glucose
365     glucose
310     glucose
```

This explains why SLDA and LDA inferred the Glucose topics. The SLDA model is perhaps not more useful in this case compared to LDA or PCA combined with linear regression which can provide as good or a better result.

We wow switch focus to the HDL cholesterol results. For K=2 only PCAEN gives a satisfactory prediction indicating that key spectral bins are present in the two first principal components (we only list top ten scores):

```
Principal component 1 (PC1)
53      fatty acyl chains
7       cholesterol
81      fatty acyl chains
502     other
8       cholesterol
6       cholesterol
57      fatty acyl chains
120     fatty acyl chains
119     fatty acyl chains
123     other

Principal component 2 (PC2)
4       cholesterol
281     other
5       cholesterol
282     other
52      fatty acyl chains
7       cholesterol
502     other
81      fatty acyl chains
51      fatty acyl chains
309     glucose
```

Out of all the bins in PC1 and PC2, let us focus on spectral bins 4 to 8 related to cholesterol and bins 51-53 and 81 related to fatty acyl chains. Those bins play a role in the topics which scored the highest correlation coefficients in the SLDA and LDA models. Let us start with SLDA:

SLDA, 5 topics	
Topic 1, reg. coef. 169.752	
53	fatty acyl chains
4	cholesterol
281	other
502	other
5	cholesterol
282	other
52	fatty acyl chains
6	cholesterol
51	fatty acyl chains
7	cholesterol

SLDA, 15 topics	
Topic 1, reg. coef. 196.7	
53	fatty acyl chains
4	cholesterol
281	other
5	cholesterol
282	other
502	other
52	fatty acyl chains
6	cholesterol
119	fatty acyl chains
118	fatty acyl chains

SLDA, 10 topics	
Topic 2, reg. coef. 187.643	
53	fatty acyl chains
4	cholesterol
281	other
5	cholesterol
282	other
502	other
52	fatty acyl chains
51	fatty acyl chains
6	cholesterol
119	fatty acyl chains

SLDA, 15 topics	
Topic 11, reg. coef. 181.526	
53	fatty acyl chains
4	cholesterol
5	cholesterol
502	other
282	other
6	cholesterol
281	other
7	cholesterol
120	fatty acyl chains
52	fatty acyl chains

Similar topic composition exists for the LDA topics. Here are LDA topics with the highest correlation coefficients for HDL cholesterol.

LDA, 5 topics	
Topic 1	
reg. coef. hdl_chol: 233.040	
53	fatty acyl chains
4	cholesterol
281	other
502	other
5	cholesterol
282	other
52	fatty acyl chains
6	cholesterol
51	fatty acyl chains
7	cholesterol

LDA, 15 topics	
Topic 14	
reg. coef. hdl_chol: 209.533	
4	cholesterol
281	other
52	fatty acyl chains
51	fatty acyl chains
50	fatty acyl chains
49	other
356	other
354	other
117	fatty acyl chains
53	fatty acyl chains

LDA, 10 topics	
Topic 2	
reg. coef. hdl_chol: 232.685	
53	fatty acyl chains
4	cholesterol
281	other
52	fatty acyl chains
282	other
5	cholesterol
502	other
51	fatty acyl chains
50	fatty acyl chains
118	fatty acyl chains

LDA, 20 topics	
Topic 19	
reg. coef. hdl_chol: 237.779	
4	cholesterol
281	other
52	fatty acyl chains
51	fatty acyl chains
50	fatty acyl chains
117	fatty acyl chains
356	other
53	fatty acyl chains
354	other
118	fatty acyl chains

4.5 Discussion

In this chapter, we evaluated SLDA, a probabilistic model for supervised learning, in the context of NMR metabolomics data with continuous response. We chose six response variables to test SLDA, and obtained good predictive metrics for two: Glucose and HDL Cholesterol. We compared the prediction results of SLDA with three other models which were a combination of the latent variable models and ElasticNet linear regression. We repeated all the modelling for K=2, 5, 10, 15 and 20 topics. We found that the predictive performance of most models was good with the exception of K=2. The poor performance of K=2 can be explained by the fact that NMR metabolomics data is inherently multivariate and extreme reduction of dimensionality to just 2D leads to too much loss of information. PCA combined with ElasticNet was slightly superior when the performance metrics were carefully scrutinised, but the interpretability of principal components was not as easy as it was for topics. SLDA and LDA topics were similar. Overall, on this data, LDA combined with ElasticNet is the best combination of interpretability and prediction.

Good prediction metrics are important but in the context of metabolomics (and omics in general) interpretability of the models is as important if not more. We analysed inferred topics from our models and found that they are informative in the context of the response variable. Especially good topics were inferred for Glucose. Glucose may seem to be a simple metabolite to predict in urine. However, its complex NMR signature (see Fig 4.6) consisting of 48 peaks could be seen as complex as an NMR profile of a biological process, in which a dozen of metabolites always respond together. This makes glucose a good model for a more complex set of metabolites. In

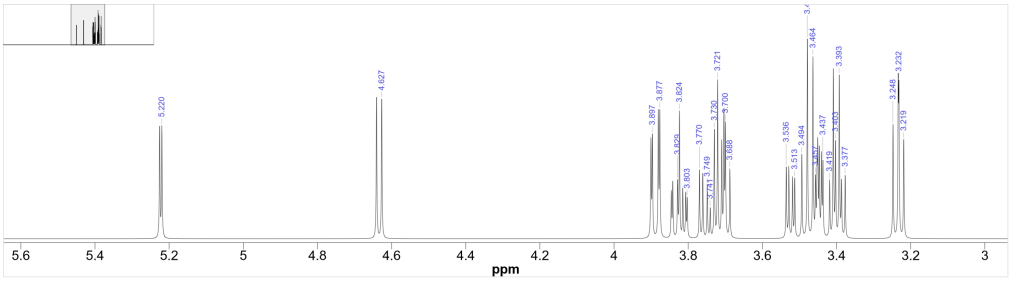

Figure 4.6: 1H NMR spectrum of D-Glucose is a complex profile consisting of 48 peaks. Image source: HMDB Wishart et al. (2018).

conclusion, topic models are an efficient way of finding a high-level pattern in data. They work well, both in supervised and unsupervised contexts. In supervised mode, they give good prediction results, but their main strength is interpretability.

4.5.1 Future work

Work presented in this chapter would benefit from the following extensions and further research:

1. Evaluation of the models on other NMR metabolomics data sets.
2. Evaluation of other approaches to deciding on the number of topics K.
3. Trying other inference methods such as variational inference because Gibbs sampling is an inherently slow procedure.
4. We tested our models only for continuous variables, thus other types of response variables, e.g. binary should be investigated.

Table 4.1: A Comparison of the supervised sLDA and unsupervised LDA/PCA combined with regression. The models with names ending in EN are combinations with ElasticNet (see section 4.3.2). The column 'prediction' summarises the two numerical metrics (corr. coeff. and NRMSE) as follows: 'good' means $NRMSE < 1.1$ and $corr.coeff. > 0.7$ and 'very good' means $NRMSE < 0.5$ and $corr.coeff. > 0.9$. sLDA and LDA models were inferred with Gibbs sampler. Each Gibbs sampler run for 10,000 iterations but the first 2,000 iterations were discarded (burn-in).

variable name	**K**	**model**	corr. coeff.	NRMSE	prediction	LDSS
glucose	**2**	**PCAEN**	0.284	0.967		
		LDAEN	0.161	0.996		0
		SLDAEN	0.162	0.996		0
		SLDA	0.162	0.994		0
	5	**PCAEN**	0.94	0.364	very good	
		LDAEN	0.914	0.476	very good	0.015
		SLDAEN	0.887	0.571	good	0.02
		SLDA	0.668	1.705		0.02
	10	**PCAEN**	0.945	0.356	very good	
		LDAEN	0.923	0.492	very good	0.021
		SLDAEN	0.876	0.612	good	-0.005
		SLDA	0.736	0.883	good	-0.005
	15	**PCAEN**	0.958	0.317	very good	
		LDAEN	0.905	0.556	good	0.177
		SLDAEN	0.858	0.647	good	0.016
		SLDA	0.762	0.774	good	0.016
	20	**PCAEN**	0.958	0.32	very good	
		LDAEN	0.909	0.585	good	0.09
		SLDAEN	0.864	0.622	good	0.005
		SLDA	0.794	1.333		0.005
hdl_chol	**2**	**PCAEN**	0.771	0.636	good	
		LDAEN	0.358	0.932		0
		SLDAEN	0.359	0.932		0
		SLDA	0.359	0.954		0
	5	**PCAEN**	0.776	0.632	good	
		LDAEN	0.754	0.671	good	0.015
		SLDAEN	0.757	0.666	good	0.021
		SLDA	0.733	0.782	good	0.021
	10	**PCAEN**	0.817	0.579	good	
		LDAEN	0.744	0.69	good	0.021
		SLDAEN	0.745	0.689	good	0
		SLDA	0.713	0.823	good	0
	15	**PCAEN**	0.826	0.564	good	
		LDAEN	0.739	0.696	good	0.177
		SLDAEN	0.719	0.721	good	0.035
		SLDA	0.761	1.091	good	0.035
	20	**PCAEN**	0.828	0.563	good	
		LDAEN	0.749	0.685	good	0.09
		SLDAEN	0.738	0.698	good	0.017
		SLDA	0.666	1.179		0.017

Chapter 5

Conclusions

In this chapter, we discuss the conclusions of the results chapter, some general conclusions, and long-term future work.

Simulations

In Chapter 2, we investigated if the generative process for Latent Dirichlet Allocation can successfully simulate NMR metabolomics spectra. It was shown by Blei (2012) that LDA can infer topics for various data types but the reversed process, simulation, was not investigated perhaps because the bag-of-words model for text is unsuitable for generating text documents. Our novel approach to simulate NMR spectra with specified underlying data patterns is reported in this thesis (see Chapter 2). We successfully demonstrated that the generative process for LDA can produce such realistic looking NMR spectra and tested that the predefined patterns can be found by standard tools like PCA. We have confidence that LDA is a valid model for NMR metabolomics spectra. This also gives us tools (Chapter 3) to assess the inferred topics in an unsupervised scenario where the evaluation of performance is difficult without data with known ground truth.

Unsupervised learning

Chapter 3 is an investigation of LDA inference on NMR metabolomics data sets. We began with simulated data from Chapter 2. The LDA inference of topics performed well, also in comparison with PCA. The most complex simulation with overlapping topics resembling a real biological system was a highlight of the LDA performance were it inferred correctly 87.5% of the metabolites which were shared between topics and missed only 10% of non-shared between topics metabolites. For the *S. mansoni* data set, LDA also performed well, finding all the metabolites associated with the disease indicated in the literature but also suggested new metabolites to investigate.

Supervised learning

Chapter 4 focused on evaluation SLDA, a supervised probabilistic model, in the context of NMR metabolomics data with continuous response. We modelled six response variables, two of them yielded good predictive metrics: Glucose and HDL Cholesterol. For comparison with SLDA, we trained three other models. Those models were a hybrid of latent variable models and ElasticNet linear regression. A decision about the ”right” number of topics was approached empirically by modelling for K=2, 5, 10, 15 and 20 topics. The predictive performance of most models was good except for K=2. K=2 represents reducing the dimensionality of multivariate data to points on a 2D plane. Such a process most likely removes too much information from the data. Overall, on MESA data, all models performed well, but LDA combined with ElasticNet struck a good balance between interpretability and prediction performance. Interpretability of models in omics may be even

more critical than their predictive performance. For example, informative topics were inferred for Glucose. Glucose has a complex metabolic profile that can be compared to a biological process where a dozen metabolites always respond together. In supervised mode, topic models give good prediction results, but their main strength is interpretability.

5.1 Future work

There is a number of possible additions to the results chapters. Our simulations work should be considered as a proof of principle. Creating a software tool for this work could be considered. For example, providing a user's interface to allow researchers to simulate data and iterate on the results. The simulation engine could be extended to include features of real NMR spectra such as baseline noise or peak shifts. More parameters could be added, for example, the width of a peak. A broader and more comprehensive database of metabolites could be extended to include more molecules available for simulation.

Considering the unsupervised LDA chapter, we note that metabolites with high probabilities vary together in a sample so that they could be mapped to a biological process. Such mapping could provide a new perspective to look at NMR metabolomics data and contribute to new biological insights. Another possible extension of the proposed analysis in Chapter 3 is to combine metabolites highlighted in topics of just one model, e.g. metabolites with high PCA loadings. Also, more latent variable methods could be investigated to be compared with LDA, for example, Non-negative Matrix factorisation (Lee and Seung, 2001).

Future work on the supervised learning from Chapter 4 could include

evaluation of the models on other NMR metabolomics data sets. Although the MESA data set was a large NMR study, further testing on other NMR metabolomics data sets would be a good idea. We tested our models only for continuous variables, other types of the response variable, for example, a binary response. Such a task would require to adjust the choice of our models. ElasticNet would need to be replaced with Logistic Regression. The SLDA model would need adjustment to cater for non-continuous variables by modelling the response variable with the Generalized Linear Model (GLM) instead of Normal distribution. In every dimensionality reduction method, choosing a number of lower dimensions K (in our case a number of topics) is always a challenge. It would be beneficial to survey other approaches and evaluate their usefulness to our problem. Lastly, Gibbs sampling in our experience was slow so perhaps other inference methods such as variational inference (VI) would bring improvement in running-time of models training.

Long-term future work, outside of what is related to the results chapters, could look at broader possibilities of using probabilistic models in NMR metabolomics. There are many other types of LDA models. It could be desirable to survey them and see what possibilities there are for our domain. For example, Dynamic Topic Model (Blei and Lafferty, 2006) can capture how topics evolve over time. This could have applications for metabolomics studies with time courses. The Dynamic Topic Model seems to work for short time series which is what we usually get in metabolomics. We could observe some topic representing a quiescent state, nothing much happening, maybe another couple of topics representing some aspects of infection or a topic is quiescent initially, then indicates infection and then it is quiescent again but differently than at the beginning.

Visualisation of topics is an exciting avenue to explore. In this thesis, we only presented topics by listing the top ten of their variables. There are sophisticated and interactive ways to visualise topic models, most notable is LDAVis (Sievert and Shirley, 2014) which would be interesting to explore. The apparent problem is the presentation of interactive methods in written material, but this should not be a reason not to pursue this line of investigation.

References

Bild, D. E., Bluemke, D. A., Burke, G. L., Detrano, R., Diez Roux, A. V., Folsom, A. R., Greenland, P., Jacob, D. R., Kronmal, R., Liu, K., Nelson, J. C., O'Leary, D., Saad, M. F., Shea, S., Szklo, M. and Tracy, R. P. (2002), Multi-Ethnic Study of Atherosclerosis: Objectives and design, *American Journal of Epidemiology* **156**(9): 871–881.

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

Blei, D. M. and Lafferty, J. D. (2006), Dynamic Topic Models, *in* 'Proceedings of the 23rd International Conference on Machine Learning', ACM, 113–120.

Blei, D. M. and McAuliffe, J. D. (2010), Supervised Topic Models, *Submitted to the Statistical Science. Preprint arXiv:1003.0783* .

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003), Latent Dirichlet Allocation, *The Journal of Machine Learning Research* **3**: 993–1022.

Bouatra, S., Aziat, F., Mandal, R., Guo, A. C., Wilson, M. R., Knox, C., Bjorndahl, T. C., Krishnamurthy, R., Saleem, F., Liu, P., Dame, Z. T., Poelzer, J., Huynh, J., Yallou, F. S., Psychogios, N., Dong, E., Bogumil, R., Roehring, C. and Wishart, D. S. (2013), The Human Urine Metabolome, *PLOS ONE* **8**(9): e73076.

Boyd-Graber, J., Hu, Y. and Mimno, D. (2017), Applications of Topic Models, *Foundations and Trends® in Information Retrieval* **11**(2-3): 143–296.

Carpenter, B. (2010), Integrating Out Multinomial Parameters in Latent Dirichlet Allocation and Naive Bayes for Collapsed Gibbs Sampling, Technical report, LingPipe, Inc.

Casella, G. and George, E. I. (1992), Explaining the Gibbs Sampler, *The American Statistician* **46**(3): 167–174.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L. and Blei, D. M. (2009), Reading tea leaves: How humans interpret topic models, *in* 'Advances in Neural Information Processing Systems 22', Curran Associates, Inc., 288–296.

Chong, W., Blei, D. and Li, F.-F. (2009), Simultaneous image classification and annotation, *in* 'Computer Vision and Pattern Recognition, 2009', IEEE, 1903–1910.

Chung, M.-H., Wang, Y., Tang, H., Zou, W., Basinger, J., Xu, X. and Tong, W. (2015), Asymmetric author-topic model for knowledge discovering of big data in toxicogenomics, *Frontiers in Pharmacology* **6**: 81.

Coen, M., Holmes, E., Lindon, J. C. and Nicholson, J. K. (2008), NMR-based metabolic profiling and metabonomic approaches to problems in molecular toxicology, *Chemical Research in Toxicology* **21**(1): 9–27.

Darling, W. M. (2011), A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling, Technical report, University of Guelph, Ontario, Canada.

Dieterle, F., Ross, A., Schlotterbeck, G. and Senn, H. (2006), Probabilistic quotient normalization as robust method to account for dilution of complex bi-

ological mixtures. Application in 1H NMR metabonomics, *Analytical Chemistry* **78**(13): 4281–4290.

Ebbels, T., Lindon, J. and Coen, M. (2011), Processing and Modeling of Nuclear Magnetic Resonance (NMR) Metabolic Profiles, *in* T. O. Metz, ed., 'Metabolic Profiling', Humana Press, 365–388.

Erosheva, E., Fienberg, S. and Lafferty, J. (2004), Mixed-membership models of scientific publications, *Proceedings of the National Academy of Sciences* **101**(suppl 1): 5220–5227.

Fiehn, O. (2002), Metabolomics – the link between genotypes and phenotypes, *Plant Molecular Biology* **48**(1-2): 155–171.

Fiehn, O., Kopka, J., Dörmann, P., Altmann, T., Trethewey, R. N. and Willmitzer, L. (2000), Metabolite profiling for plant functional genomics, *Nature biotechnology* **18**(11): 1157.

Ghahramani, Z. (2015), Probabilistic machine learning and artificial intelligence, *Nature* **521**(7553): 452–459.

Gilks, W., Richardson, S. and Spiegelhalter, D., eds (1996), *Markov Chain Monte Carlo in Practice*, Springer.

Gowda, G. N., Zhang, S., Gu, H., Asiago, V., Shanaiah, N. and Raftery, D. (2008), Metabolomics-based methods for early disease diagnostics, *Expert review of molecular diagnostics* **8**(5): 617–633.

Griffiths, T. (2002), Gibbs sampling in the generative model of LDA, Technical report, Stanford University.

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

Hao, J., Liebeke, M., Astle, W., De Iorio, M., Bundy, J. G. and Ebbels, T. M. D. (2014), Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN, *Nature Protocols* **9**(6): 1416–1427.

Hastie, T. (2015), *Statistical Learning with Sparsity: The Lasso and Generalizations*, CRC Press, Taylor & Francis Group.

Hausser, K. H. and Kalbitzer, H. R. (1991), *NMR in Medicine and Biology: Structure Determination, Tomography, in Vivo Spectroscopy*, Springer-Verlag.

Heinrich, G. (2008), Parameter estimation for text analysis, Technical report, University of Leipzig.

Holmes, E., Loo, R. L., Cloarec, O., Coen, M., Tang, H., Maibaum, E., Bruce, S., Chan, Q., Elliott, P., Stamler, J., Wilson, I. D., Lindon, J. C. and Nicholson, J. K. (2007), Detection of urinary drug metabolite (xenometabolome) signatures in molecular epidemiology studies via statistical total correlation (NMR) spectroscopy, *Analytical Chemistry* **79**(7): 2629–2640.

Holmes, E., Loo, R. L., Stamler, J., Bictash, M., Yap, I. K. S., Chan, Q., Ebbels, T., Iorio, M. D., Brown, I. J., Veselkov, K. A., Daviglus, M. L., Kesteloot, H., Ueshima, H., Zhao, L., Nicholson, J. K. and Elliott, P. (2008), Human metabolic phenotype diversity and its association with diet and blood pressure, *Nature* **453**(7193): 396–400.

Hore, P. (2015), *Nuclear Magnetic Resonance*, 2nd edn, Oxford University Press.

Jackson, J. E. (2003), *A User's Guide to Principal Components*, Wiley-Interscience, Hoboken, N.J.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. and Saul, L. K. (1999), An Introduction to Variational Methods for Graphical Models, *Machine Learning* **37**(2): 183–233.

Kanehisa, M. and Goto, S. (2000), KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Research* **28**(1): 27–30.

Keeler, J. (2011), *Understanding NMR Spectroscopy*, John Wiley & Sons.

Lee, D. D. and Seung, H. S. (2001), Algorithms for non-negative matrix factorization, *in* 'Advances in Neural Information Processing Systems 13', MIT Press, 556–562.

Lee, M., Liu, Z., Huang, R. and Tong, W. (2016), Application of dynamic topic models to toxicogenomics data, *BMC Bioinformatics* **17**(13): 368.

Li, J. V., Saric, J., Wang, Y., Keiser, J., Utzinger, J. and Holmes, E. (2011), Chemometric analysis of biofluids from mice experimentally infected with Schistosoma mansoni, *Parasites & Vectors* **4**: 179.

Liu, L., Tang, L., Dong, W., Yao, S. and Zhou, W. (2016), An overview of topic modeling and its current applications in bioinformatics, *SpringerPlus* **5**(1): 1608.

Lu, H.-M., Wei, C.-P. and Hsiao, F.-Y. (2016), Modeling healthcare data using multiple-channel latent Dirichlet allocation, *Journal of Biomedical Informatics* **60**: 210–223.

Mcauliffe, J. D. and Blei, D. M. (2008), Supervised Topic Models, *in* J. C. Platt, D. Koller, Y. Singer and S. T. Roweis, eds, 'Advances in Neural Information Processing Systems 20', Curran Associates, Inc., 121–128.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models, Second Edition*, CRC Press.

Muncey, H. J., Jones, R., De Iorio, M. and Ebbels, T. M. (2010), MetAssimulo:Simulation of Realistic NMR Metabolic Profiles, *BMC Bioinformatics* **11**: 496.

Murphy, K. P. (2012), *Machine Learning: A Probabilistic Perspective*, MIT Press.

Neal, R. M. (1993), Probabilistic Inference Using Markov Chain Monte Carlo Methods, Technical report, University of Toronto.

Nicholson, J. K. (2006), Global systems biology, personalized medicine and molecular epidemiology, *Molecular Systems Biology* **2**: 52.

Nicholson, J. K., Connelly, J., Lindon, J. C. and Holmes, E. (2002), Metabonomics: A platform for studying drug toxicity and gene function, *Nature Reviews Drug Discovery* **1**(2): 153–161.

Nicholson, J. K. and Lindon, J. C. (2008), Systems biology: Metabonomics, *Nature* **455**(7216): 1054–1056.

Nicholson, J. K., Lindon, J. C. and Holmes, E. (1999), Metabonomics: Understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data, *Xenobiotica* **29**(11): 1181–1189.

Pearce, J. T. M. (2010), Novel Computational Approaches to Characterising Metabolic Responses to Toxicity via an NMR-Based Metabonomic Database, PhD thesis, Imperial College, London.

Pearson, K. (1901), LIII. On lines and planes of closest fit to systems of points in space, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**(11): 559–572.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000), Inference of population structure using multilocus genotype data, *Genetics* **155**(2): 945–959.

Rabi, I. I., Zacharias, J. R., Millman, S. and Kusch, P. (1938), A New Method of Measuring Nuclear Magnetic Moment, *Physical Review* **53**(4): 318–318.

Resnik, P. and Hardisty, E. (2010), Gibbs sampling for the uninitiated, Technical report, University of Maryland Institute for Advanced Computer Studies.

Roessner, U., Luedemann, A., Brust, D., Fiehn, O., Linke, T., Willmitzer, L. and Fernie, A. R. (2001), Metabolic Profiling Allows Comprehensive Phenotyping of Genetically or Environmentally Modified Plant Systems, *The Plant Cell* **13**(1): 11–29.

Sievert, C. and Shirley, K. (2014), LDAvis: A method for visualizing and interpreting topics, *in* 'Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces', Association for Computational Linguistics, 63–70.

Stamler, J., Elliott, P., Dennis, B., Dyer, A. R., Kesteloot, H., Liu, K., Ueshima, H. and Zhou, B. F. (2003), INTERMAP: Background, aims, design, methods, and descriptive statistics (nondietary), *Journal of Human Hypertension* **17**(9): 591–608.

Steyvers, M. and Griffiths, T. (2007), Probabilistic topic models, *in* 'Latent Semantic Analysis: A Road to Meaning', Laurence Erlbaum.

Takis, P. G., Schäfer, H., Spraul, M. and Luchinat, C. (2017), Deconvoluting interrelationships between concentrations and chemical shifts in urine provides a powerful analysis tool, *Nature Communications* **8**(1): 1662.

Veselkov, K. A., Lindon, J. C., Ebbels, T. M. D., Crockford, D., Volynkin, V. V., Holmes, E., Davies, D. B. and Nicholson, J. K. (2009), Recursive Segment-Wise Peak Alignment of Biological 1H NMR Spectra for Improved Metabolic Biomarker Recovery, *Analytical Chemistry* **81**(1): 56–66.

Wainwright, M. J. and Jordan, M. I. (2008), Graphical Models, Exponential Families, and Variational Inference, *Foundations and Trends® in Machine Learning* **1**(1–2): 1–305.

Wallach, H. M., Murray, I., Salakhutdinov, R. and Mimno, D. (2009), Evaluation methods for topic models, *in* 'Proceedings of the 26th Annual International Conference on Machine Learning', ACM, 1105–1112.

Wang, Y., Holmes, E., Nicholson, J. K., Cloarec, O., Chollet, J., Tanner, M., Singer, B. H. and Utzinger, J. (2004), Metabonomic investigations in mice infected with Schistosoma mansoni: An approach for biomarker identification, *Proceedings of the National Academy of Sciences* **101**(34): 12676–12681.

Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., Sayeeda, Z., Lo, E., Assempour, N., Berjanskii, M., Singhal, S., Arndt, D., Liang, Y., Badran, H., Grant, J., Serra-Cayuela, A., Liu, Y., Mandal, R., Neveu, V., Pon, A., Knox, C., Wilson, M., Manach, C. and Scalbert, A. (2018), HMDB 4.0: The human metabolome database for 2018, *Nucleic Acids Research* **46**(D1): D608–D617.

Wohlgemuth, G., Haldiya, P. K., Willighagen, E., Kind, T. and Fiehn, O. (2010), The Chemical Translation Service—a web-based tool to improve standardization of metabolomic reports, *Bioinformatics* **26**(20): 2647–2648.

Wold, S., Esbensen, K. and Geladi, P. (1987), Principal component analysis, *Chemometrics and Intelligent Laboratory Systems* **2**(1): 37–52.

Wold, S., Sjöström, M. and Eriksson, L. (2001), PLS-regression: A basic tool of chemometrics, *Chemometrics and Intelligent Laboratory Systems* **58**(2): 109–130.

Wu, J., Xu, W., Ming, Z., Dong, H., Tang, H. and Wang, Y. (2010), Metabolic Changes Reveal the Development of Schistosomiasis in Mice, *PLoS Neglected Tropical Diseases* **4**(8): 1–11.

Zou, H. and Hastie, T. (2005), Regularization and variable selection via the Elastic Net, *Journal of the Royal Statistical Society, Series B* **67**(2): 301–320.

Appendix A

Unsupervised learning: Statistical tests for the *S. mansoni* data

Table A.1: Two-tailed p-values for Kolmogorov-Smirnov test applied pairwise on *S. mansoni* NMR spectra in LDA topic space for late treatments group (see Figure 3.6). The p-value smaller 1% (marked red) indicates that we can reject the null hypothesis of equal averages for the given pair of topics.

	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
Topic 0					
Topic 1	0.0875				
Topic 2	3.06e-05	0.00391			
Topic 3	1.51e-05	1.57e-09	8.5e-17		
Topic 4	0.0401	0.529	0.0065	1.02e-11	

Table A.2: Two-tailed p-values for t-test applied pairwise on *S. mansoni* NMR spectra in LDA topic space for late treatments group (see Figure 3.6). The p-value smaller 1% (marked red) indicates that we can reject the null hypothesis of equal averages for the given pair of topics.

	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
Topic 0					
Topic 1	0.15				
Topic 2	4.59e-06	0.0025			
Topic 3	0.000979	4.08e-07	3.62e-15		
Topic 4	0.0024	0.206	0.0057	4.19e-11	

Table A.3: Two-tailed p-values for Kolmogorov-Smirnov test applied pairwise on *S. mansoni* NMR spectra in principal component space for late treatments group (see Figure 3.7). The p-value smaller 1% (all in this case) indicates that we can reject the null hypothesis of equal averages for the given pair of principal components.

	Comp. 0	Comp. 1	Comp. 2	Comp. 3	Comp. 4
Comp. 0					
Comp. 1	5.99e-10				
Comp. 2	5.99e-10	2.92e-11			
Comp. 3	1.01e-08	0.00132	2.24e-10		
Comp. 4	5.99e-10	0.00391	3.92e-24	7.32e-06	

Table A.4: Two-tailed p-values for t-test applied pairwise on *S. mansoni* NMR spectra in principal component space for late treatments group (see Figure 3.7). The p-value smaller 1% (marked red) indicates that we can reject the null hypothesis of equal averages for the given pair of principal components.

	Comp. 0	Comp. 1	Comp. 2	Comp. 3	Comp. 4
Comp. 0					
Comp. 1	0.115				
Comp. 2	0.00117	0.0484			
Comp. 3	0.211	0.326	1.25e-07		
Comp. 4	0.9	0.0054	4.4e-17	2.01e-05	

Table A.5: Two-tailed p-values for Kolmogorov-Smirnov test applied pairwise to study the selected metabolites levels for late treatments group (see Figure 3.8). The p-value smaller 1% (all in this case) indicates that we can reject the null hypothesis of equal averages for the given pair of metabolites levels.

	PAG_7.43_m	p-CG_2.3_s	N-AG_2.06_s	OGT_3.01_t	Pyr_2.36_s
PAG_7.43_m					
p-CG_2.3_s	3.49e-12				
N-AG_2.06_s	0.000746	1.8e-23			
OGT_3.01_t	1.61e-06	1.59e-30	3.86e-14		
Pyr_2.36_s	1.73e-25	1.51e-05	2.26e-39	5.55e-42	

Table A.6: Two-tailed p-values for t-test applied pairwise to study the selected metabolites levels for late treatments group (see Figure 3.8). The p-value smaller 1% (marked red) indicates that we can reject the null hypothesis of equal averages for the given pair of metabolites levels.

	PAG_7.43_m	p-CG_2.3_s	N-AG_2.06_s	OGT_3.01_t	Pyr_2.36_s
PAG_7.43_m					
p-CG_2.3_s	4.49e-31				
N-AG_2.06_s	0.116	1.78e-19			
OGT_3.01_t	2.93e-08	2.03e-24	7.03e-14		
Pyr_2.36_s	4.13e-33	1.16e-10	1.27e-46	3e-34	

Appendix B

Supervised learning: MESA data full results

Listing B.1: Glucose-SLDA, 2 topics

```
----------------
glucose SLDA, 2 topics
---------------
Topic 0, reg. coef. 90.92
  53  fatty acyl chains
  7 cholesterol
  502 other
  81  fatty acyl chains
  6 cholesterol
  8 cholesterol
  57  fatty acyl chains
  120 fatty acyl chains
  123 other
  119 fatty acyl chains

Topic 1, reg. coef. 100.728
  53  fatty acyl chains
  4 cholesterol
  502 other
  281 other
  5 cholesterol
  123 other
  6 cholesterol
  7 cholesterol
  52  fatty acyl chains
  367 glucose
```

Listing B.2: Glucose-SLDA, 5 topics

```
----------------
glucose SLDA, 5 topics
---------------
Topic 0, reg. coef. 104.469
  53  fatty acyl chains
  7 cholesterol
  502 other
  81  fatty acyl chains
  8 cholesterol
  6 cholesterol
  4 cholesterol
  57  fatty acyl chains
  120 fatty acyl chains
  123 other

Topic 1, reg. coef. 105.045
  53  fatty acyl chains
  7 cholesterol
  502 other
  81  fatty acyl chains
  6 cholesterol
  8 cholesterol
  4 cholesterol
  144 fatty acyl chains
  57  fatty acyl chains
  120 fatty acyl chains

Topic 2, reg. coef. 72.096
  53  fatty acyl chains
  4 cholesterol
  281 other
  502 other
  5 cholesterol
  7 cholesterol
  52  fatty acyl chains
  282 other
  6 cholesterol
  51  fatty acyl chains

Topic 3, reg. coef. 336.08
  4 cholesterol
  309 glucose
  325 glucose
  367 glucose
  311 glucose
  388 glucose
  500 glucose
  323 glucose
  501 glucose
  365 glucose

Topic 4, reg. coef. 54.971
```

```
  4 cholesterol
  123 other
  281 other
  502 other
  5 cholesterol
  367 glucose
  309 glucose
  325 glucose
  283 glucose
  399 glucose
```

Listing B.3: Glucose-SLDA, 10 topics

```
---------------
glucose SLDA, 10 topics
---------------
Topic 0, reg. coef. 21.43
  53  fatty acyl chains
  5 cholesterol
  6 cholesterol
  7 cholesterol
  282 other
  119 fatty acyl chains
  283 glucose
  120 fatty acyl chains
  118 fatty acyl chains
  333 other

Topic 1, reg. coef. 109.676
  53  fatty acyl chains
  7 cholesterol
  81  fatty acyl chains
  8 cholesterol
  6 cholesterol
  502 other
  57  fatty acyl chains
  120 fatty acyl chains
  119 fatty acyl chains
  123 other

Topic 2, reg. coef. 44.824
  53  fatty acyl chains
  4 cholesterol
  281 other
  282 other
  5 cholesterol
  52  fatty acyl chains
  502 other
  51  fatty acyl chains
  50  fatty acyl chains
  118 fatty acyl chains

Topic 3, reg. coef. 10.586
  53  fatty acyl chains
  6 cholesterol
  5 cholesterol
  4 cholesterol
  333 other
  372 glucose
  123 other
  7 cholesterol
  526 other
  289 glucose

Topic 4, reg. coef. 6.898
  53  fatty acyl chains
  4 cholesterol
  50  fatty acyl chains
  49  other
  281 other
  5 cholesterol
  282 other
  52  fatty acyl chains
  154 other
  6 cholesterol

Topic 5, reg. coef. 30.368
  502 other
  53  fatty acyl chains
  4 cholesterol
  5 cholesterol
  6 cholesterol
  123 other
  7 cholesterol
  220 fatty acyl chains
  225 fatty acyl chains
  226 fatty acyl chains

Topic 6, reg. coef. 110.82
  53  fatty acyl chains
  7 cholesterol
  81  fatty acyl chains
  6 cholesterol
  8 cholesterol
  502 other
  144 fatty acyl chains
  57  fatty acyl chains
  120 fatty acyl chains
  119 fatty acyl chains

Topic 7, reg. coef. 328.28
  4 cholesterol
  53  fatty acyl chains
  309 glucose
  502 other
  325 glucose
  367 glucose
  311 glucose
  388 glucose
  500 glucose
  323 glucose

Topic 8, reg. coef. 77.762
  4 cholesterol
  502 other
  281 other
  123 other
  367 glucose
  309 glucose
  325 glucose
  399 glucose
  368 glucose
  289 glucose

Topic 9, reg. coef. -2.341
  53  fatty acyl chains
  6 cholesterol
  5 cholesterol
  7 cholesterol
  282 other
  119 fatty acyl chains
  120 fatty acyl chains
  283 glucose
  502 other
  118 fatty acyl chains
```

Listing B.4: Glucose-SLDA,15 topics

```
----------------
glucose SLDA, 15 topics
---------------
Topic 0, reg. coef. 8.385
  53  fatty acyl chains
  445 other
  444 other
  72  other
  73  other
  443 other
  446 other
  168 other
  165 other
  167 other

Topic 1, reg. coef. 43.483
  53  fatty acyl chains
  4 cholesterol
  282 other
  5 cholesterol
  281 other
  502 other
  6 cholesterol
  52  fatty acyl chains
  119 fatty acyl chains
  120 fatty acyl chains

Topic 2, reg. coef. 14.565
  53  fatty acyl chains
  8 cholesterol
  502 other
  81  fatty acyl chains
  57  fatty acyl chains
  52  fatty acyl chains
  7 cholesterol
  9 cholesterol
  147 fatty acyl chains
  16  other

Topic 3, reg. coef. 20.161
  123 other
  5 cholesterol
  333 other
  126 other
  6 cholesterol
  124 other
  119 fatty acyl chains
  120 fatty acyl chains
  356 other
  357 other

Topic 4, reg. coef. 119.323
  53  fatty acyl chains
  7 cholesterol
  81  fatty acyl chains
  502 other
  6 cholesterol
  8 cholesterol
  120 fatty acyl chains
  57  fatty acyl chains
  123 other
  119 fatty acyl chains

Topic 5, reg. coef. 11.013
  123 other
  333 other
  5 cholesterol
  126 other
  4 cholesterol
  502 other
  6 cholesterol
  124 other
  356 other
  120 fatty acyl chains

Topic 6, reg. coef. 35.044
  53  fatty acyl chains
  5 cholesterol
  6 cholesterol
  7 cholesterol
  4 cholesterol
  502 other
  282 other
  119 fatty acyl chains
  120 fatty acyl chains
  283 glucose

Topic 7, reg. coef. 7.25
  53  fatty acyl chains
  123 other
  5 cholesterol
  6 cholesterol
  7 cholesterol
  126 other
  332 glucose
  367 glucose
  502 other
  119 fatty acyl chains

Topic 8, reg. coef. 43.73
  502 other
  5 cholesterol
  4 cholesterol
  6 cholesterol
  123 other
  220 fatty acyl chains
  226 fatty acyl chains
  225 fatty acyl chains
  224 fatty acyl chains
  221 fatty acyl chains

Topic 9, reg. coef. 120.851
  53  fatty acyl chains
  7 cholesterol
  502 other
  81  fatty acyl chains
  6 cholesterol
  8 cholesterol
  144 fatty acyl chains
  123 other
  5 cholesterol
  120 fatty acyl chains

Topic 10, reg. coef. 352.843
  4 cholesterol
  309 glucose
  325 glucose
  367 glucose
  311 glucose
  388 glucose
  500 glucose
  323 glucose
  365 glucose
  501 glucose

Topic 11, reg. coef. 90.959
  367 glucose
  309 glucose
  325 glucose
  283 glucose
  289 glucose
  311 glucose
  388 glucose
  368 glucose
  500 glucose
  399 glucose

Topic 12, reg. coef. 5.367
  4 cholesterol
  50  fatty acyl chains
  49  other
  281 other
  52  fatty acyl chains
  154 other
  51  fatty acyl chains
  53  fatty acyl chains
  160 other
  282 other

Topic 13, reg. coef. 8.721
  4 cholesterol
  281 other
  52  fatty acyl chains
  51  fatty acyl chains
  50  fatty acyl chains
  117 fatty acyl chains
  49  other
  372 glucose
  118 fatty acyl chains
  354 other

Topic 14, reg. coef. 34.584
  4 cholesterol
```

```
281 other
52  fatty acyl chains
51  fatty acyl chains
50  fatty acyl chains
117 fatty acyl chains
```

```
49  other
118 fatty acyl chains
356 other
53  fatty acyl chains
```

Listing B.5: Glucose-SLDA, 20 topics

```
----------------
glucose SLDA, 20 topics
---------------
Topic 0, reg. coef. 7.513
  8 cholesterol
  7 cholesterol
  81  fatty acyl chains
  57  fatty acyl chains
  9 cholesterol
  147 fatty acyl chains
  10  cholesterol
  56  fatty acyl chains
  11  cholesterol
  148 fatty acyl chains

Topic 1, reg. coef. 3.453
  123 other
  372 glucose
  126 other
  81  fatty acyl chains
  213 fatty acyl chains
  375 glucose
  356 other
  374 glucose
  399 glucose
  357 other

Topic 2, reg. coef. 18.062
  53  fatty acyl chains
  4 cholesterol
  281 other
  52  fatty acyl chains
  51  fatty acyl chains
  50  fatty acyl chains
  117 fatty acyl chains
  49  other
  118 fatty acyl chains
  81  fatty acyl chains

Topic 3, reg. coef. 23.368
  53  fatty acyl chains
  4 cholesterol
  7 cholesterol
  144 fatty acyl chains
  81  fatty acyl chains
  275 other
  8 cholesterol
  145 fatty acyl chains
  57  fatty acyl chains
  499 other

Topic 4, reg. coef. 10.337
  289 glucose
  309 glucose
  325 glucose
  367 glucose
  388 glucose
  311 glucose
  283 glucose
  365 glucose
  500 glucose
  501 glucose

Topic 5, reg. coef. 29.572
  53  fatty acyl chains
  6 cholesterol
  7 cholesterol
  5 cholesterol
  4 cholesterol
  502 other
  120 fatty acyl chains
  119 fatty acyl chains
  52  fatty acyl chains
  118 fatty acyl chains

Topic 6, reg. coef. 14.934
  53  fatty acyl chains
  502 other
  6 cholesterol
  7 cholesterol
  5 cholesterol
  4 cholesterol
  120 fatty acyl chains
  119 fatty acyl chains
  232 fatty acyl chains
  231 fatty acyl chains

Topic 7, reg. coef. 6.019
  53  fatty acyl chains
  17  other
  22  other
  16  other
  21  other
  27  other
  26  other
  7 cholesterol
  72  other
  18  other

Topic 8, reg. coef. 114.534
  53  fatty acyl chains
  7 cholesterol
  81  fatty acyl chains
  8 cholesterol
  502 other
  6 cholesterol
  120 fatty acyl chains
  57  fatty acyl chains
  119 fatty acyl chains
  123 other

Topic 9, reg. coef. 13.55
  53  fatty acyl chains
  4 cholesterol
  49  other
  50  fatty acyl chains
  281 other
  5 cholesterol
  282 other
  154 other
  6 cholesterol
  52  fatty acyl chains

Topic 10, reg. coef. 9.46
  123 other
  126 other
  356 other
  400 glucose
  399 glucose
  368 glucose
  367 glucose
  124 other
  357 other
  7 cholesterol

Topic 11, reg. coef. 14.224
  502 other
  5 cholesterol
  6 cholesterol
  123 other
  4 cholesterol
  220 fatty acyl chains
  226 fatty acyl chains
  225 fatty acyl chains
  16  other
  224 fatty acyl chains

Topic 12, reg. coef. 115.548
  53  fatty acyl chains
  7 cholesterol
  81  fatty acyl chains
  502 other
  8 cholesterol
  144 fatty acyl chains
  6 cholesterol
  120 fatty acyl chains
  57  fatty acyl chains
  119 fatty acyl chains

Topic 13, reg. coef. 5.774
  502 other
  6 cholesterol
  5 cholesterol
  445 other
  123 other
  444 other
  7 cholesterol
  120 fatty acyl chains
  72  other
  73  other

Topic 14, reg. coef. 340.676
  4 cholesterol
```

```
  502 other
  309 glucose
  325 glucose
  367 glucose
  311 glucose
  388 glucose
  5 cholesterol
  500 glucose
  501 glucose

Topic 15, reg. coef. 128.103
  502 other
  5 cholesterol
  282 other
  123 other
  6 cholesterol
  367 glucose
  309 glucose
  283 glucose
  325 glucose
  501 glucose

Topic 16, reg. coef. 20.483
  17  other
  16  other
  22  other
  72  other
  21  other
  26  other
  27  other
  73  other
  445 other
  18  other

Topic 17, reg. coef. 10.922
  4 cholesterol
  281 other
  52  fatty acyl chains
  51  fatty acyl chains
  280 other
  50  fatty acyl chains
  333 other
  220 fatty acyl chains
  218 fatty acyl chains
  225 fatty acyl chains

Topic 18, reg. coef. 39.455
  4 cholesterol
  281 other
  52  fatty acyl chains
  51  fatty acyl chains
  50  fatty acyl chains
  333 other
  280 other
  220 fatty acyl chains
  225 fatty acyl chains
  218 fatty acyl chains

Topic 19, reg. coef. 12.665
  4 cholesterol
  281 other
  52  fatty acyl chains
  51  fatty acyl chains
  280 other
  50  fatty acyl chains
  220 fatty acyl chains
  332 glucose
  225 fatty acyl chains
  218 fatty acyl chains
```

Listing B.6: hdl-chol-SLDA-2-topics.txt

```
----------------
hdl_chol SLDA, 2 topics
---------------
Topic 0, reg. coef. 17.958
  53  fatty acyl chains
  7 cholesterol
  502 other
  81  fatty acyl chains
  6 cholesterol
  8 cholesterol
  57  fatty acyl chains
  120 fatty acyl chains
  123 other
  144 fatty acyl chains

Topic 1, reg. coef. 59.685
  53  fatty acyl chains
  4 cholesterol
  502 other
  281 other
  5 cholesterol
  123 other
  6 cholesterol
  7 cholesterol
  52  fatty acyl chains
  367 glucose
```

Listing B.7: hdl-chol-SLDA-5-topics.txt

```
----------------
hdl_chol SLDA, 5 topics
---------------
Topic 0, reg. coef. 23.974
  53  fatty acyl chains
  7 cholesterol
  81  fatty acyl chains
  502 other
  8 cholesterol
  6 cholesterol
  4 cholesterol
  57  fatty acyl chains
  120 fatty acyl chains
  123 other

Topic 1, reg. coef. 169.752
  53  fatty acyl chains
  4 cholesterol
  281 other
  502 other
  5 cholesterol
  282 other
  52  fatty acyl chains
  6 cholesterol
  51  fatty acyl chains
  7 cholesterol

Topic 2, reg. coef. 54.86
  53  fatty acyl chains
  4 cholesterol
  7 cholesterol
  502 other
  81  fatty acyl chains
  6 cholesterol
  5 cholesterol
  281 other
  8 cholesterol
  144 fatty acyl chains

Topic 3, reg. coef. 16.538
  367 glucose
  309 glucose
  325 glucose
  311 glucose
  388 glucose
  365 glucose
  500 glucose
  501 glucose
  323 glucose
  310 glucose

Topic 4, reg. coef. -8.608
  123 other
  280 other
  72  other
```

```
  126 other
  333 other
  16  other
  50  fatty acyl chains
```

Listing B.8: hdl-chol-SLDA-10-topics.txt

```
---------------
hdl_chol SLDA, 10 topics
---------------
Topic 0, reg. coef. 18.698
  53  fatty acyl chains
  5 cholesterol
  6 cholesterol
  7 cholesterol
  4 cholesterol
  502 other
  282 other
  120 fatty acyl chains
  119 fatty acyl chains
  283 glucose

Topic 1, reg. coef. 18.353
  53  fatty acyl chains
  7 cholesterol
  81  fatty acyl chains
  8 cholesterol
  502 other
  6 cholesterol
  57  fatty acyl chains
  120 fatty acyl chains
  119 fatty acyl chains
  123 other

Topic 2, reg. coef. 187.643
  53  fatty acyl chains
  4 cholesterol
  281 other
  5 cholesterol
  282 other
  502 other
  52  fatty acyl chains
  51  fatty acyl chains
  6 cholesterol
  119 fatty acyl chains

Topic 3, reg. coef. 12.036
  502 other
  4 cholesterol
  5 cholesterol
  6 cholesterol
  123 other
  333 other
  372 glucose
  309 glucose
  388 glucose
  289 glucose

Topic 4, reg. coef. 23.858
  53  fatty acyl chains
  50  fatty acyl chains
  49  other
  5 cholesterol
  4 cholesterol
  282 other
  6 cholesterol
  73  other
  17  other
  445 other

  154 other
  160 other
  448 other

Topic 5, reg. coef. 27.394
  502 other
  5 cholesterol
  4 cholesterol
  6 cholesterol
  123 other
  220 fatty acyl chains
  226 fatty acyl chains
  225 fatty acyl chains
  224 fatty acyl chains
  221 fatty acyl chains

Topic 6, reg. coef. 25.01
  53  fatty acyl chains
  7 cholesterol
  502 other
  81  fatty acyl chains
  6 cholesterol
  8 cholesterol
  144 fatty acyl chains
  5 cholesterol
  120 fatty acyl chains
  57  fatty acyl chains

Topic 7, reg. coef. 38.684
  309 glucose
  325 glucose
  367 glucose
  311 glucose
  388 glucose
  500 glucose
  501 glucose
  365 glucose
  323 glucose
  310 glucose

Topic 8, reg. coef. 71.53
  4 cholesterol
  281 other
  123 other
  52  fatty acyl chains
  51  fatty acyl chains
  280 other
  72  other
  73  other
  126 other
  282 other

Topic 9, reg. coef. 1.474
  50  fatty acyl chains
  49  other
  53  fatty acyl chains
  526 other
  4 cholesterol
  275 other
  154 other
  160 other
  51  fatty acyl chains
  499 other
```

Listing B.9: hdl-chol-SLDA-15-topics.txt

```
----------------
hdl_chol SLDA, 15 topics
---------------
Topic 0, reg. coef. 6.188
  7 cholesterol
  6 cholesterol
  5 cholesterol
  8 cholesterol
  57  fatty acyl chains
  120 fatty acyl chains
  283 glucose
  121 fatty acyl chains
  9 cholesterol
  56  fatty acyl chains

Topic 1, reg. coef. 196.7
  53  fatty acyl chains
  4 cholesterol
  281 other
  5 cholesterol
  282 other
  502 other
  52  fatty acyl chains
  6 cholesterol
  119 fatty acyl chains
  118 fatty acyl chains

Topic 2, reg. coef. -0.119
  53  fatty acyl chains
  502 other
  4 cholesterol
  5 cholesterol
  444 other
  445 other
  72  other
  52  fatty acyl chains
  73  other
  282 other

Topic 3, reg. coef. 11.453
  53  fatty acyl chains
  4 cholesterol
  333 other
  502 other
  289 glucose
  309 glucose
  368 glucose
  283 glucose
  282 other
  5 cholesterol

Topic 4, reg. coef. 27.436
  53  fatty acyl chains
  7 cholesterol
  81  fatty acyl chains
  502 other
  8 cholesterol
  6 cholesterol
  57  fatty acyl chains
  120 fatty acyl chains
  119 fatty acyl chains
  123 other

Topic 5, reg. coef. 5.333
  5 cholesterol
  6 cholesterol
  502 other
  4 cholesterol
  281 other
  7 cholesterol
  289 glucose
  50  fatty acyl chains
  49  other
  526 other

Topic 6, reg. coef. -3.139
  53  fatty acyl chains
  5 cholesterol
  6 cholesterol
  4 cholesterol
  7 cholesterol
  502 other
  119 fatty acyl chains
  52  fatty acyl chains
  120 fatty acyl chains
  118 fatty acyl chains

Topic 7, reg. coef. 10.051
  5 cholesterol
  6 cholesterol
  7 cholesterol
  289 glucose
  281 other
  435 other
  119 fatty acyl chains
  262 other
  372 glucose
  304 other

Topic 8, reg. coef. 33.821
  502 other
  4 cholesterol
  5 cholesterol
  6 cholesterol
  123 other
  220 fatty acyl chains
  226 fatty acyl chains
  7 cholesterol
  225 fatty acyl chains
  224 fatty acyl chains

Topic 9, reg. coef. 42.684
  53  fatty acyl chains
  7 cholesterol
  81  fatty acyl chains
  502 other
  8 cholesterol
  144 fatty acyl chains
  6 cholesterol
  120 fatty acyl chains
  119 fatty acyl chains
  57  fatty acyl chains

Topic 10, reg. coef. 69.639
  4 cholesterol
  123 other
  281 other
  502 other
  52  fatty acyl chains
  51  fatty acyl chains
  280 other
  282 other
  50  fatty acyl chains
  126 other

Topic 11, reg. coef. 59.159
  4 cholesterol
  123 other
  281 other
  367 glucose
  309 glucose
  325 glucose
  399 glucose
  311 glucose
  283 glucose
  388 glucose

Topic 12, reg. coef. 21.133
  50  fatty acyl chains
  49  other
  4 cholesterol
  281 other
  154 other
  160 other
  282 other
  448 other
  52  fatty acyl chains
  170 other

Topic 13, reg. coef. 33.546
  309 glucose
  325 glucose
  367 glucose
  311 glucose
  388 glucose
  500 glucose
  323 glucose
  501 glucose
  365 glucose
  310 glucose
```

```
Topic 14, reg. coef. 41.126
  4 cholesterol
  281 other
  52  fatty acyl chains
  51  fatty acyl chains
  53  fatty acyl chains
  50  fatty acyl chains
  49  other
  117 fatty acyl chains
  8 cholesterol
  118 fatty acyl chains
```

Listing B.10: hdl-chol-SLDA-20-topics.txt

```
----------------
hdl_chol SLDA, 20 topics
---------------
Topic 0, reg. coef. 6.654
  7 cholesterol
  8 cholesterol
  57  fatty acyl chains
  9 cholesterol
  81  fatty acyl chains
  56  fatty acyl chains
  147 fatty acyl chains
  10  cholesterol
  124 other
  125 other

Topic 1, reg. coef. 8.582
  49  other
  50  fatty acyl chains
  154 other
  160 other
  448 other
  170 other
  449 other
  173 other
  446 other
  313 other

Topic 2, reg. coef. 4.781
  53  fatty acyl chains
  444 other
  445 other
  7 cholesterol
  8 cholesterol
  72  other
  73  other
  443 other
  446 other
  81  fatty acyl chains

Topic 3, reg. coef. 28.157
  53  fatty acyl chains
  144 fatty acyl chains
  81  fatty acyl chains
  7 cholesterol
  145 fatty acyl chains
  6 cholesterol
  8 cholesterol
  120 fatty acyl chains
  146 fatty acyl chains
  119 fatty acyl chains

Topic 4, reg. coef. 9.218
  282 other
  280 other
  72  other
  73  other
  289 glucose
  17  other
  22  other
  27  other
  21  other
  26  other

Topic 5, reg. coef. -7.811
  53  fatty acyl chains
  5 cholesterol
  6 cholesterol
  7 cholesterol
  4 cholesterol
  120 fatty acyl chains
  283 glucose
  502 other
  119 fatty acyl chains
  8 cholesterol

Topic 6, reg. coef. 10.49
  53  fatty acyl chains
  6 cholesterol
  5 cholesterol
  7 cholesterol
  4 cholesterol
  502 other
  81  fatty acyl chains
  275 other
  120 fatty acyl chains
  144 fatty acyl chains

Topic 7, reg. coef. 13.419
  53  fatty acyl chains
  502 other
  5 cholesterol
  4 cholesterol
  6 cholesterol
  52  fatty acyl chains
  220 fatty acyl chains
  226 fatty acyl chains
  118 fatty acyl chains
  16  other

Topic 8, reg. coef. 15.302
  53  fatty acyl chains
  502 other
  7 cholesterol
  81  fatty acyl chains
  123 other
  4 cholesterol
  6 cholesterol
  224 fatty acyl chains
  221 fatty acyl chains
  122 fatty acyl chains

Topic 9, reg. coef. 11.497
  53  fatty acyl chains
  5 cholesterol
  289 glucose
  372 glucose
  6 cholesterol
  282 other
  73  other
  72  other
  4 cholesterol
  22  other

Topic 10, reg. coef. 29.685
  53  fatty acyl chains
  7 cholesterol
  81  fatty acyl chains
  6 cholesterol
  8 cholesterol
  120 fatty acyl chains
  502 other
  5 cholesterol
  119 fatty acyl chains
  57  fatty acyl chains

Topic 11, reg. coef. 181.526
  53  fatty acyl chains
  4 cholesterol
  5 cholesterol
  502 other
  282 other
  6 cholesterol
  281 other
  7 cholesterol
  120 fatty acyl chains
  52  fatty acyl chains

Topic 12, reg. coef. 39.08
  53  fatty acyl chains
  7 cholesterol
  81  fatty acyl chains
  6 cholesterol
  8 cholesterol
  144 fatty acyl chains
  5 cholesterol
  502 other
  120 fatty acyl chains
  119 fatty acyl chains

Topic 13, reg. coef. 29.046
  502 other
  4 cholesterol
  5 cholesterol
  123 other
  220 fatty acyl chains
  224 fatty acyl chains
  221 fatty acyl chains
  225 fatty acyl chains
  226 fatty acyl chains
  121 fatty acyl chains
```

```
Topic 14, reg. coef. 24.607
  309 glucose
  325 glucose
  367 glucose
  311 glucose
  388 glucose
  323 glucose
  500 glucose
  365 glucose
  501 glucose
  310 glucose

Topic 15, reg. coef. 61.639
  123 other
  309 glucose
  325 glucose
  399 glucose
  500 glucose
  311 glucose
  388 glucose
  501 glucose
  400 glucose
  365 glucose

Topic 16, reg. coef. 71.425
  123 other
  309 glucose
  325 glucose
  399 glucose
  311 glucose
  388 glucose
  500 glucose
  400 glucose
  501 glucose
  365 glucose

Topic 17, reg. coef. 44.984
  289 glucose
  282 other
  367 glucose
  368 glucose
  72  other
  73  other
  17  other
  16  other
  332 glucose
  280 other

Topic 18, reg. coef. 45.428
  4 cholesterol
  281 other
  52  fatty acyl chains
  51  fatty acyl chains
  50  fatty acyl chains
  333 other
  117 fatty acyl chains
  368 glucose
  280 other
  116 fatty acyl chains

Topic 19, reg. coef. 50.631
  4 cholesterol
  281 other
  52  fatty acyl chains
  51  fatty acyl chains
  50  fatty acyl chains
  117 fatty acyl chains
  53  fatty acyl chains
  280 other
  116 fatty acyl chains
  115 fatty acyl chains
```

Listing B.11: LDA-2-topics.txt

```
----------------
LDA, 2 topics
---------------
Topic 0, reg. coef. hdl_chol: -25.737
     reg. coef. glucose: -5.289
  53  fatty acyl chains
  7 cholesterol
  502 other
  81  fatty acyl chains
  6 cholesterol
  8 cholesterol
  57  fatty acyl chains
  120 fatty acyl chains
  123 other
  119 fatty acyl chains

Topic 1, reg. coef. hdl_chol: 25.694
    reg. coef. glucose: 5.289
  53  fatty acyl chains
  4 cholesterol
  502 other
  281 other
  5 cholesterol
  123 other
  6 cholesterol
  7 cholesterol
  52  fatty acyl chains
  367 glucose
```

Listing B.12: LDA-5-topics.txt

```
----------------
LDA, 5 topics
---------------
Topic 0, reg. coef. hdl_chol: -75.742
     reg. coef. glucose: -5.642
  53  fatty acyl chains
  7 cholesterol
  81  fatty acyl chains
  502 other
  8 cholesterol
  6 cholesterol
  57  fatty acyl chains
  4 cholesterol
  120 fatty acyl chains
  123 other

Topic 1, reg. coef. hdl_chol: 233.040
     reg. coef. glucose: -68.511
  53  fatty acyl chains
  4 cholesterol
  281 other
  502 other
  5 cholesterol
  282 other
  52  fatty acyl chains
  6 cholesterol
  51  fatty acyl chains
  7 cholesterol

Topic 2, reg. coef. hdl_chol: -78.972
     reg. coef. glucose: -30.121
  53  fatty acyl chains
  7 cholesterol
  502 other
  81  fatty acyl chains
  6 cholesterol
  8 cholesterol
  4 cholesterol
  144 fatty acyl chains
  57  fatty acyl chains
  120 fatty acyl chains

Topic 3, reg. coef. hdl_chol: -39.823
     reg. coef. glucose: 370.689
  367 glucose
  309 glucose
  325 glucose
  311 glucose
  388 glucose
  500 glucose
  501 glucose
```

```
  365 glucose
  323 glucose
  310 glucose

Topic 4, reg. coef. hdl_chol: -35.474
     reg. coef. glucose: -263.401
  4 cholesterol
  123 other
  281 other
  502 other
  5 cholesterol
  6 cholesterol
  50  fatty acyl chains
  282 other
  51  fatty acyl chains
  280 other
```

Listing B.13: LDA-10-topics.txt

```
----------------
LDA, 10 topics
---------------
Topic 0, reg. coef. hdl_chol: 34.323
    reg. coef. glucose: -26.383
  53  fatty acyl chains
  5 cholesterol
  6 cholesterol
  4 cholesterol
  502 other
  7 cholesterol
  282 other
  119 fatty acyl chains
  120 fatty acyl chains
  283 glucose

Topic 1, reg. coef. hdl_chol: -97.851
     reg. coef. glucose: 36.482
  53  fatty acyl chains
  7 cholesterol
  81  fatty acyl chains
  8 cholesterol
  502 other
  6 cholesterol
  57  fatty acyl chains
  120 fatty acyl chains
  119 fatty acyl chains
  123 other

Topic 2, reg. coef. hdl_chol: 232.685
     reg. coef. glucose: -60.518
  53  fatty acyl chains
  4 cholesterol
  281 other
  52  fatty acyl chains
  282 other
  5 cholesterol
  502 other
  51  fatty acyl chains
  50  fatty acyl chains
  118 fatty acyl chains

Topic 3, reg. coef. hdl_chol: -46.869
     reg. coef. glucose: -16.594
  502 other
  4 cholesterol
  5 cholesterol
  6 cholesterol
  53  fatty acyl chains
  123 other
  333 other
  372 glucose
  309 glucose
  388 glucose

Topic 4, reg. coef. hdl_chol: 63.374
    reg. coef. glucose: -27.574
  53  fatty acyl chains
  4 cholesterol
  50  fatty acyl chains
  49  other
  5 cholesterol
  282 other
  281 other
  6 cholesterol
  154 other
  160 other

Topic 5, reg. coef. hdl_chol: 12.913
    reg. coef. glucose: -40.995
  502 other
  5 cholesterol
  4 cholesterol
  6 cholesterol
  123 other
  220 fatty acyl chains
  226 fatty acyl chains
  225 fatty acyl chains
  224 fatty acyl chains
  221 fatty acyl chains

Topic 6, reg. coef. hdl_chol: -95.996
     reg. coef. glucose: 4.169
  53  fatty acyl chains
  7 cholesterol
  502 other
  81  fatty acyl chains
  6 cholesterol
  8 cholesterol
  144 fatty acyl chains
  5 cholesterol
  120 fatty acyl chains
  57  fatty acyl chains

Topic 7, reg. coef. hdl_chol: -52.856
     reg. coef. glucose: 400.667
  309 glucose
  325 glucose
  367 glucose
  311 glucose
  388 glucose
  500 glucose
  501 glucose
  365 glucose
  323 glucose
  310 glucose

Topic 8, reg. coef. hdl_chol: -49.870
     reg. coef. glucose: -250.550
  4 cholesterol
  281 other
  123 other
  52  fatty acyl chains
  51  fatty acyl chains
  280 other
  72  other
  73  other
  126 other
  17  other

Topic 9, reg. coef. hdl_chol: 0.160
    reg. coef. glucose: -14.692
  50  fatty acyl chains
  49  other
  53  fatty acyl chains
  526 other
  275 other
  4 cholesterol
  154 other
  160 other
  280 other
  500 glucose
```

Listing B.14: LDA-15-topics.txt

```
---------------
LDA, 15 topics
---------------
Topic 0, reg. coef. hdl_chol: -38.598
     reg. coef. glucose: 6.480
  7 cholesterol
  6 cholesterol
  5 cholesterol
  8 cholesterol
  283 glucose
  120 fatty acyl chains
  57  fatty acyl chains
  121 fatty acyl chains
  9 cholesterol
  285 other

Topic 1, reg. coef. hdl_chol: 165.257
     reg. coef. glucose: -32.584
  53  fatty acyl chains
  4 cholesterol
  282 other
  281 other
  502 other
  5 cholesterol
  52  fatty acyl chains
  119 fatty acyl chains
  283 glucose
  120 fatty acyl chains

Topic 2, reg. coef. hdl_chol: -52.674
     reg. coef. glucose: 43.449
  53  fatty acyl chains
  8 cholesterol
  81  fatty acyl chains
  4 cholesterol
  52  fatty acyl chains
  57  fatty acyl chains
  51  fatty acyl chains
  281 other
  9 cholesterol
  147 fatty acyl chains

Topic 3, reg. coef. hdl_chol: 14.700
    reg. coef. glucose: -22.942
  53  fatty acyl chains
  4 cholesterol
  5 cholesterol
  6 cholesterol
  333 other
  502 other
  282 other
  119 fatty acyl chains
  368 glucose
  283 glucose

Topic 4, reg. coef. hdl_chol: -92.198
     reg. coef. glucose: 83.763
  53  fatty acyl chains
  7 cholesterol
  81  fatty acyl chains
  502 other
  8 cholesterol
  6 cholesterol
  120 fatty acyl chains
  123 other
  57  fatty acyl chains
  119 fatty acyl chains

Topic 5, reg. coef. hdl_chol: -8.966
     reg. coef. glucose: -55.404
  4 cholesterol
  5 cholesterol
  333 other
  502 other
  6 cholesterol
  50  fatty acyl chains
  289 glucose
  283 glucose
  49  other
  368 glucose

Topic 6, reg. coef. hdl_chol: 30.507
    reg. coef. glucose: 10.789
  53  fatty acyl chains
  5 cholesterol
  4 cholesterol
  6 cholesterol
  502 other
  52  fatty acyl chains
  282 other
  119 fatty acyl chains
  7 cholesterol
  118 fatty acyl chains

Topic 7, reg. coef. hdl_chol: -8.065
    reg. coef. glucose: -65.208
  5 cholesterol
  4 cholesterol
  6 cholesterol
  333 other
  372 glucose
  123 other
  289 glucose
  119 fatty acyl chains
  52  fatty acyl chains
  136 other

Topic 8, reg. coef. hdl_chol: 35.813
    reg. coef. glucose: -29.580
  502 other
  5 cholesterol
  6 cholesterol
  4 cholesterol
  123 other
  220 fatty acyl chains
  226 fatty acyl chains
  225 fatty acyl chains
  16  other
  224 fatty acyl chains

Topic 9, reg. coef. hdl_chol: -98.220
     reg. coef. glucose: 67.366
  53  fatty acyl chains
  7 cholesterol
  502 other
  81  fatty acyl chains
  144 fatty acyl chains
  6 cholesterol
  8 cholesterol
  120 fatty acyl chains
  119 fatty acyl chains
  123 other

Topic 10, reg. coef. hdl_chol:
    -74.844  reg. coef. glucose:
    -124.932
  123 other
  502 other
  72  other
  73  other
  282 other
  126 other
  16  other
  280 other
  17  other
  445 other

Topic 11, reg. coef. hdl_chol:
    -93.756  reg. coef. glucose:
    -139.211
  123 other
  72  other
  73  other
  126 other
  280 other
  17  other
  445 other
  16  other
  444 other
  282 other

Topic 12, reg. coef. hdl_chol: 70.155
     reg. coef. glucose: -34.005
  4 cholesterol
  50  fatty acyl chains
  49  other
  281 other
  154 other
  282 other
  51  fatty acyl chains
  52  fatty acyl chains
  160 other
  448 other
```

```
Topic 13, reg. coef. hdl_chol:
    -55.664  reg. coef. glucose:
    391.078
  309 glucose
  367 glucose
  325 glucose
  311 glucose
  388 glucose
  500 glucose
  365 glucose
  501 glucose
  323 glucose
  310 glucose
```

```
Topic 14, reg. coef. hdl_chol:
    209.533  reg. coef. glucose:
    -96.007
  4 cholesterol
  281 other
  52  fatty acyl chains
  51  fatty acyl chains
  50  fatty acyl chains
  49  other
  356 other
  354 other
  117 fatty acyl chains
  53  fatty acyl chains
```

Listing B.15: LDA-20-topics.txt

```
----------------
LDA, 20 topics
---------------
Topic 0, reg. coef. hdl_chol: -76.920
     reg. coef. glucose: 33.682
  7 cholesterol
  8 cholesterol
  81  fatty acyl chains
  57  fatty acyl chains
  9 cholesterol
  147 fatty acyl chains
  56  fatty acyl chains
  10  cholesterol
  148 fatty acyl chains
  121 fatty acyl chains

Topic 1, reg. coef. hdl_chol: 43.983
    reg. coef. glucose: -29.485
  49  other
  50  fatty acyl chains
  282 other
  5 cholesterol
  154 other
  160 other
  448 other
  445 other
  449 other
  446 other

Topic 2, reg. coef. hdl_chol: 5.909
    reg. coef. glucose: -25.327
  7 cholesterol
  5 cholesterol
  445 other
  444 other
  6 cholesterol
  282 other
  73  other
  72  other
  8 cholesterol
  446 other

Topic 3, reg. coef. hdl_chol: -64.913
     reg. coef. glucose: 26.756
  53  fatty acyl chains
  144 fatty acyl chains
  7 cholesterol
  81  fatty acyl chains
  145 fatty acyl chains
  8 cholesterol
  6 cholesterol
  146 fatty acyl chains
  120 fatty acyl chains
  57  fatty acyl chains

Topic 4, reg. coef. hdl_chol: -74.723
     reg. coef. glucose: -85.615
  333 other
  280 other
  368 glucose
  289 glucose
  72  other
  73  other
  283 glucose
  17  other
  262 other
  6 cholesterol

Topic 5, reg. coef. hdl_chol: -19.221
     reg. coef. glucose: 11.734
  53  fatty acyl chains
  7 cholesterol
  6 cholesterol
  5 cholesterol
  4 cholesterol
  120 fatty acyl chains
  8 cholesterol
  119 fatty acyl chains
  283 glucose
  118 fatty acyl chains

Topic 6, reg. coef. hdl_chol: 92.975
    reg. coef. glucose: -44.033
  4 cholesterol
  502 other
  5 cholesterol
  282 other
  6 cholesterol
  7 cholesterol
  120 fatty acyl chains
  232 fatty acyl chains
  231 fatty acyl chains
  220 fatty acyl chains

Topic 7, reg. coef. hdl_chol: -30.068
     reg. coef. glucose: 13.606
  53  fatty acyl chains
  49  other
  50  fatty acyl chains
  7 cholesterol
  6 cholesterol
  526 other
  81  fatty acyl chains
  8 cholesterol
  120 fatty acyl chains
  119 fatty acyl chains

Topic 8, reg. coef. hdl_chol: -30.279
     reg. coef. glucose: 39.684
  53  fatty acyl chains
  7 cholesterol
  6 cholesterol
  275 other
  81  fatty acyl chains
  5 cholesterol
  120 fatty acyl chains
  119 fatty acyl chains
  8 cholesterol
  269 other

Topic 9, reg. coef. hdl_chol: -68.851
     reg. coef. glucose: -113.482
  53  fatty acyl chains
  333 other
  289 glucose
  372 glucose
  72  other
  368 glucose
  73  other
  262 other
  22  other
  526 other

Topic 10, reg. coef. hdl_chol:
    -95.181  reg. coef. glucose:
```

```
    41.523
  53  fatty acyl chains
  7 cholesterol
  81  fatty acyl chains
  502 other
  8 cholesterol
  6 cholesterol
  123 other
  57  fatty acyl chains
  120 fatty acyl chains
  147 fatty acyl chains

Topic 11, reg. coef. hdl_chol:
    -24.160  reg. coef. glucose:
    59.120
  53  fatty acyl chains
  81  fatty acyl chains
  6 cholesterol
  120 fatty acyl chains
  119 fatty acyl chains
  52  fatty acyl chains
  5 cholesterol
  468 other
  51  fatty acyl chains
  118 fatty acyl chains

Topic 12, reg. coef. hdl_chol:
    -55.403  reg. coef. glucose:
    44.024
  53  fatty acyl chains
  7 cholesterol
  81  fatty acyl chains
  6 cholesterol
  502 other
  144 fatty acyl chains
  8 cholesterol
  120 fatty acyl chains
  119 fatty acyl chains
  57  fatty acyl chains

Topic 13, reg. coef. hdl_chol:
    -31.531  reg. coef. glucose:
    -51.261
  502 other
  5 cholesterol
  4 cholesterol
  6 cholesterol
  123 other
  220 fatty acyl chains
  225 fatty acyl chains
  226 fatty acyl chains
  16  other
  224 fatty acyl chains

Topic 14, reg. coef. hdl_chol:
    -12.222  reg. coef. glucose:
    461.921
  309 glucose
  325 glucose
  367 glucose
  311 glucose
  388 glucose
  323 glucose
  500 glucose
  501 glucose
  365 glucose
  310 glucose

Topic 15, reg. coef. hdl_chol:
    171.528  reg. coef. glucose:
    -67.669
  502 other
  4 cholesterol
  5 cholesterol
  282 other
  281 other
  6 cholesterol
  220 fatty acyl chains
  119 fatty acyl chains
  225 fatty acyl chains
  120 fatty acyl chains

Topic 16, reg. coef. hdl_chol:
    -94.683  reg. coef. glucose:
    52.063
  123 other
  309 glucose
  325 glucose
  399 glucose
  311 glucose
  400 glucose
  388 glucose
  365 glucose
  500 glucose
  501 glucose

Topic 17, reg. coef. hdl_chol:
    -80.819  reg. coef. glucose:
    -143.089
  289 glucose
  367 glucose
  368 glucose
  280 other
  332 glucose
  72  other
  17  other
  16  other
  73  other
  262 other

Topic 18, reg. coef. hdl_chol:
    214.828  reg. coef. glucose:
    -111.346
  4 cholesterol
  281 other
  52  fatty acyl chains
  51  fatty acyl chains
  50  fatty acyl chains
  333 other
  117 fatty acyl chains
  356 other
  354 other
  118 fatty acyl chains

Topic 19, reg. coef. hdl_chol:
    237.779  reg. coef. glucose:
    -112.838
  4 cholesterol
  281 other
  52  fatty acyl chains
  51  fatty acyl chains
  50  fatty acyl chains
  117 fatty acyl chains
  356 other
  53  fatty acyl chains
  354 other
  118 fatty acyl chains
```

Listing B.16: PCA-20-topics.txt

```
----------------
PCA, 20 topics
---------------
Topic 0, reg. coef. hdl_chol: -0.002
    reg. coef. glucose: 0.001
  53  fatty acyl chains
  7 cholesterol
  81  fatty acyl chains
  502 other
  8 cholesterol
  6 cholesterol
  57  fatty acyl chains
  120 fatty acyl chains
  119 fatty acyl chains
  123 other

Topic 1, reg. coef. hdl_chol: 0.031
    reg. coef. glucose: -0.029
  4 cholesterol
  281 other
  5 cholesterol
  282 other
  52  fatty acyl chains
  7 cholesterol
  502 other
  81  fatty acyl chains
  51  fatty acyl chains
  309 glucose

Topic 2, reg. coef. hdl_chol: 0.008
    reg. coef. glucose: 0.129
  309 glucose
  325 glucose
  367 glucose
  311 glucose
  388 glucose
  323 glucose
  500 glucose
  501 glucose
  365 glucose
  310 glucose

Topic 3, reg. coef. hdl_chol: 0.005
    reg. coef. glucose: -0.018
  502 other
  6 cholesterol
  5 cholesterol
  7 cholesterol
  281 other
  144 fatty acyl chains
  53  fatty acyl chains
  8 cholesterol
  123 other
  220 fatty acyl chains

Topic 4, reg. coef. hdl_chol: -0.005
    reg. coef. glucose: 0.000
  502 other
  144 fatty acyl chains
  5 cholesterol
  6 cholesterol
  81  fatty acyl chains
  7 cholesterol
  145 fatty acyl chains
  8 cholesterol
  57  fatty acyl chains
  123 other

Topic 5, reg. coef. hdl_chol: -0.002
    reg. coef. glucose: -0.028
  144 fatty acyl chains
  145 fatty acyl chains
  5 cholesterol
  6 cholesterol
  281 other
  502 other
  146 fatty acyl chains
  282 other
  7 cholesterol
  51  fatty acyl chains

Topic 6, reg. coef. hdl_chol: -0.001
    reg. coef. glucose: 0.001
  49  other
  50  fatty acyl chains
  160 other
  150 fatty acyl chains
  449 other
  170 other
  448 other
  168 other
  154 other
  281 other

Topic 7, reg. coef. hdl_chol: -0.005
    reg. coef. glucose: -0.002
  7 cholesterol
  281 other
  8 cholesterol
  282 other
  5 cholesterol
  6 cholesterol
  4 cholesterol
  53  fatty acyl chains
  57  fatty acyl chains
  123 other

Topic 8, reg. coef. hdl_chol: 0.063
    reg. coef. glucose: -0.043
  282 other
  123 other
  52  fatty acyl chains
  51  fatty acyl chains
  81  fatty acyl chains
  144 fatty acyl chains
  8 cholesterol
  57  fatty acyl chains
  6 cholesterol
  145 fatty acyl chains

Topic 9, reg. coef. hdl_chol: -0.013
    reg. coef. glucose: -0.005
  123 other
  282 other
  126 other
  7 cholesterol
  52  fatty acyl chains
  502 other
  283 glucose
  399 glucose
  400 glucose
  4 cholesterol

Topic 10, reg. coef. hdl_chol: 0.000
    reg. coef. glucose: 0.008
  333 other
  372 glucose
  332 glucose
  367 glucose
  81  fatty acyl chains
  8 cholesterol
  366 glucose
  289 glucose
  213 fatty acyl chains
  4 cholesterol

Topic 11, reg. coef. hdl_chol: 0.003
    reg. coef. glucose: -0.011
  281 other
  4 cholesterol
  8 cholesterol
  280 other
  57  fatty acyl chains
  119 fatty acyl chains
  282 other
  444 other
  72  other
  445 other

Topic 12, reg. coef. hdl_chol: -0.024
     reg. coef. glucose: -0.041
  445 other
  444 other
  281 other
  72  other
  73  other
  4 cholesterol
  446 other
  443 other
  52  fatty acyl chains
  168 other

Topic 13, reg. coef. hdl_chol: -0.000
```

```
     reg. coef. glucose: 0.063
  333 other
  8 cholesterol
  81  fatty acyl chains
  57  fatty acyl chains
  280 other
  52  fatty acyl chains
  4 cholesterol
  5 cholesterol
  9 cholesterol
  332 glucose

Topic 14, reg. coef. hdl_chol: 0.055
    reg. coef. glucose: 0.110
  444 other
  445 other
  333 other
  17  other
  280 other
  368 glucose
  289 glucose
  52  fatty acyl chains
  51  fatty acyl chains
  262 other

Topic 15, reg. coef. hdl_chol: -0.000
     reg. coef. glucose: -0.016
  333 other
  372 glucose
  368 glucose
  57  fatty acyl chains
  444 other
  445 other
  5 cholesterol
  22  other
  8 cholesterol
  21  other

Topic 16, reg. coef. hdl_chol: 0.024
    reg. coef. glucose: -0.000
  282 other
  51  fatty acyl chains
  52  fatty acyl chains
  5 cholesterol
  7 cholesterol
  57  fatty acyl chains
  81  fatty acyl chains
  275 other
  368 glucose
  367 glucose

Topic 17, reg. coef. hdl_chol: -0.000
     reg. coef. glucose: -0.000
  275 other
  51  fatty acyl chains
  52  fatty acyl chains
  282 other
  72  other
  73  other
  435 other
  262 other
  332 glucose
  57  fatty acyl chains

Topic 18, reg. coef. hdl_chol: -0.027
     reg. coef. glucose: -0.001
  57  fatty acyl chains
  275 other
  16  other
  52  fatty acyl chains
  17  other
  51  fatty acyl chains
  22  other
  21  other
  15  cholesterol
  26  other

Topic 19, reg. coef. hdl_chol: -0.000
     reg. coef. glucose: 0.000
  289 glucose
  72  other
  73  other
  444 other
  445 other
  332 glucose
  275 other
  375 glucose
  57  fatty acyl chains
  262 other
```

Appendix C

Software

Software for this thesis was developed and run on Ubuntu Linux 18.04. All the code was written with Python v3.5.4 (using Anaconda Distribution v5.0.1).

Our software

Our Python modules are availavle on Github, see `https://github.com/dataoverflow/ezNMR` for details.

3rd party software

The important packages along with the versions are listed in table C.1.

From modelling point of view the two most important libraries are `scikit-learn` and `slda`. `scikit-learn` is the most popular Python library for Machine Learning. We used `scikit-learn` for PCA and ElasticNet. `slda` is implementations of Gibbs sampling for supervised and unsupervised LDA. This package was used for all the LDA related work. It can be installed as a binary in Anaconda Distribution, alternatively the sources are available at `https://github.com/dataoverflow/slda`.

Table C.1: Used Python libraries

Name	Version
bioservices	1.5.2
cython	0.26.1
cythongsl	0.2.2
gsl	2.2.1
matplotlib	2.1.0
numpy	1.13.3
pandas	0.20.3
PeakUtils	1.3.0
scikit-learn	0.19.1
scipy	0.19.1
seaborn	0.8.0
slda	0.1.6
statsmodels	0.8.0