MANCHESTER
1824
The University of Manchester

MODELLING OF LONGITUDINAL DIGITAL HEALTH DATA TO UNDERSTAND UNDERLYING PHENOTYPES

By

Rajenki Das

School of Natural Sciences
Department of Mathematics

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF SCIENCE AND ENGINEERING

2023

Contents

List of Tables

List of Figures

Outline

This thesis focuses on achieving two main objectives:

1. Novel application of cutting-edge statistical learning methods to longitudinal health data;
2. Development of a Bayesian approach to model-based clustering of time series of categorical variables.

In Chapter 1, we introduce topics in mental health, which is a relevant issue in modern healthcare. In Chapter 2, we provide brief introductions to few of the methods and other tools which have been used throughout the thesis. Chapters 3, 4 and 5 can be read independently as they are written in academic paper format, and each of these chapters includes a separate abstract at the beginning of the paper. In Chapter 3, we perform residual analysis and clustering of mood-pain trajectories on the basis of transitions taken from a longitudinal study. In Chapter 4, we perform Bayesian inference on the same data considered before in the previous chapter. We assume the data to be distributed multinomially and taking Dirichlet distribution as a conjugate prior, we use Hamiltonian Monte Carlo method to sample estimates of the model parameters. In doing so, we also address the problem of label-switching in the mixture model. In Chapter 5, we consider all the self-reported symptoms, not just mood and pain, and implement dimensionality reduction to investigate the relationships amongst those. Chapters 6 and 7 include further prospects of the thesis, and conclusion respectively.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright statement

Acknowledgements

Amongst many . . . ,

I would like to express my deepest gratitude towards my supervisors Thomas House, Mark Muldoon and Mark Lunt for their continuous support and guidance throughout my PhD journey. Thomas has been very inclusive and always optimistic which has helped me maintain faith in myself and my research. Mark M. has given his valuable time throughout and motivated me to work on new concepts. Working alongside him has been a privilege, thus allowing me to grow and learn. Mark L. has provided valuable input and advice as needed. These examples represent only a fraction of the reasons for my gratitude, as they have consistently shown kindness, support and encouragement.

I am indebted to the Engineering and Physical Sciences Research Council (EPSRC), a part of the UKRI research council, for funding my research.

Many thanks to Xiaoxi and my office mates Jacob, Xiaodong, Bindu, Zubier, Wang, Lubo, Peej, Martyn, QS, Steve and Anthony for their company and support. I would also like to acknowledge the Mathematics Epidemiology group with whom I had a great time attending seminars and conferences, and engaging in social activities. Furthermore, I would like to acknowledge my collaborators, the University of Manchester Mathematics Department, IT services and the Mathematics Teaching and Learning Office, particularly Gemma, for her patience in addressing my numerous queries as a postgraduate researcher. I extend my thanks to other members of the group, including Tracey, Laura, and James. I am sincerely grateful to my examiners Korbinian Strimmer and Robert Goudie for their valuable time and the insightful feedback on my thesis.

Thanks to my friends Sampurna, Rishiraj, Shabnam, Abhimanyu, Miruna, Rhea and Esther for being pillars of emotional strength in this journey.

Last but not the least, thank you to mamma and baba who have been by my side through the maxima and many minima of this journey with me.

Thank you all for reading this!

Rajenki

Chapter 1

Introduction

> If a man has lost a leg or an eye, he knows he has lost a leg or an eye; but if he has lost a self—himself—he cannot know it, because he is no longer there to know it.
>
> Oliver Sacks

Health is an integral part of human well being, and a holistic perspective on the same involves physical, mental and social factors (World Health Organization et al., 1948). However, "health" itself is a complex concept where many aspects play a role in building a healthy person and meanings of the same can vary across individuals, as per the capability of tackling an illness (Leonardi, 2018). Most of us are aware of physical illness and associated concerns but mental health often gets neglected in the larger discourse on health and wellbeing, and it remains important to remember the adage "no health without mental health" as used used by e.g. Prince et al. (2007) in the context of overall health. Even though recently there has been an increase in awareness towards mental health, the stigma around it continues to exist (Bharadwaj et al., 2017; Gold et al., 2016). Additionally this social stigma around mental health as well as a self-perceived notion of underestimating an issue (Andrade et al., 2014) often leads to lack of treatments. Proper treatments or diagnoses are still unavailable or inaccessible for many individuals (Moreno et al., 2020; Camm-Crosbie et al., 2019; MacDonald et al., 2018). In 2019, the World Health Organisation (WHO) estimated 970 million people in the world to be living with a mental disorder (Organization, 2022). Keeping all of this in mind, it needs to be re-iterated that mental ailments are widespread and can affect anyone, just like a physical illness. There are many aspects of this topic that need to be dealt with carefully, but in this thesis, the emphasis has been on finding mental health traits using digital health records and how it can facilitate the process of developing treatments for the same.

Mental health can be affected by numerous factors, with examples being: physical health problems, socio-economic conditions, nutrition, genetics, and the environment around us (Kola et al., 2021; Adan et al., 2019; Bhugra et al., 2013; Tew et al., 2012; Rutter, 2005; Morris, 2003; Tsuang, 2000). Even though identifying causes remains an extremely challenging problem, symptoms associated with a decline in mental health can aid diagnoses. It can be found that withdrawal in life, many times shown by lack of motivation, is quite prominent amongst those severely affected by mental affliction and can serve as a vital warning. But, a more comprehensive understanding of mental health requires an interdisciplinary approach (Fried, 2021) which can benefit from insights from various fields including neurology, psychology, sociology, biology etc. and, importantly in the current context, mathematical sciences. The advent of COVID-19 presented a global health crisis which affected lives across the world quite disproportionately (Gibson et al., 2021) and resulted in further widening of inequalities. A rise in mental health problems, as an associated result of worsening physical health or in this case, a physical health calamity, has become a major concern (Vigo et al., 2020) and may have long-term implications (Bourmistrova et al., 2022; Kola et al., 2021). The motivation behind this thesis was to help quantify mental health, model it and see underlying behaviour using mathematical, statistical and computational tools as these would help in comparing and providing better tools for understanding the differences and commonalities in behaviour patterns.

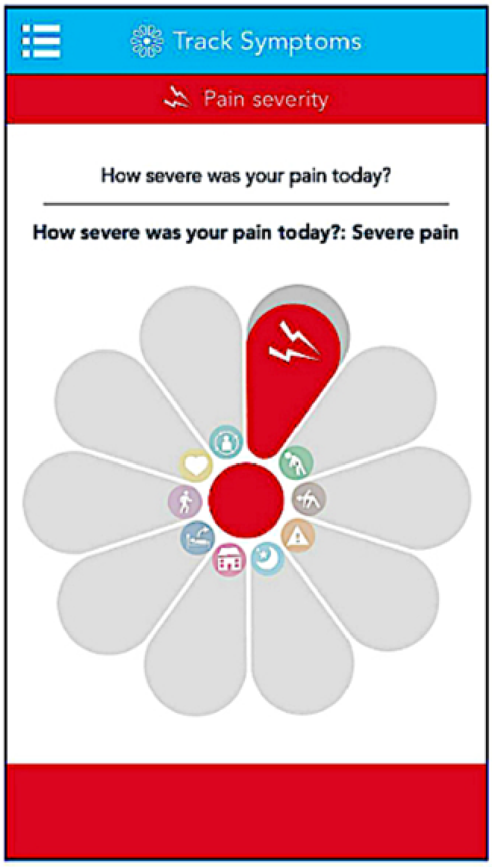
While there are pros and cons to the advancement of technology and electronic health (eHealth) (Vitacca et al., 2009), in this context, it has provided us with digital health tools which have facilitated the collection of information on health. Mobile health (mHealth) is potentially revolutionary and opening up doors to opportunities for exploring research in healthcare (Fiordelli et al., 2013). Especially in the sphere of mental health, mhealth helps in overcoming many barriers related to accessibility (Price et al., 2014). Such apps or platforms can be beneficial to those who are unable or are reluctant to be available for an in person appointment, as well as those who want to keep their information completely private or anonymous. This is particularly true in the context of mental health, where many prefer not to be identified while reporting their issues as a result of the stigma attached, although it is difficult to tell if a mHealth based solution or therapy is indeed better than an in-person one (Olff, 2015) but nevertheless, these apps can be useful as they maintain the confidentiality of the patient. Not only this, digital health apps also enable a person to track and share their health behaviour easily. Especially with the rise in smartwatches and other devices, it gets easier to monitor one's health as one can be notified if there are abrupt changes in their patterns of health. These technological interventions are not just beneficial for the participants, but also for the clinicians and doctors, who can now access patient behaviour easily and take necessary actions and intervene accordingly. For researchers, mHealth enables them to study a wider range of problems with the data collected – understanding mental health has rather been a long, time-consuming process and mHealth can offer to speed up the research by easier collection of data. Other drawbacks of mHealth are related to its authentication (Mathews et al., 2019). Digitisation is not so pervasive in many countries, so we could be excluding significant proportions of people before coming to any conclusion based on an mHealth based analysis. At the same time, technology in general can have a negative effect on mental health (Haidt and Allen, 2020; Scott et al., 2017) so it can sound ironic to rely on such devices. Thus, ethics in digital health and its widespread impact still need to be actively discussed. Regardless, the potential benefits of digital health are clear (Triantafyllidis and Tsanas, 2019), and it is hoped that new technology can benefit wider population with correct implementation. In the context of mental health, for example, we may see apps that allow individuals to share their problems with ease and seek help quickly.

In this thesis, efforts have been put into understanding mental health by taking a quantitative approach to mHealth data. The dataset used in this research is provided by the Cloudy with a Chance of Pain study (Sergeant et al., 2015; Dixon et al., 2019) led by the University of Manchester. Details of the dataset specific to the corresponding study are provided in Chapters 3, 4 and 5. These are the three main chapters pertinent to the work carried out as part of the doctoral research presented here. These chapters are also arranged chronologically which helps in connecting the motivations of moving from one study to another. In Chapters 3 and 4, the primary focus is on the combined trajectories of mood and pain which were taken as the real data for the analyses. Chapter 5 includes mood and pain as part of a unified analysis of many other self-reported symptoms that were recorded in the Cloud with a Chance of Pain study. Each of these three chapters is written in the form of a journal paper, but additionally contains a motivation section at the beginning of the main write-up. As a technical introduction, in Chapter 2 the methods and tools used throughout this thesis are briefly discussed. More specific methods are elaborated in the respective chapters themselves. Chapter 3 talks about clustering the participants of the study on the basis of their self reported mood and pain trajectories. We discovered four digital phenotypes on the basis of the mood-pain behaviour over a period of time, and emphasised the need for personalised treatment. In chapter 4, we take a Bayesian inference approach in further analysing these mood-pain trajectories. We develop a Dirichlet-multinomial distribution based on the Markov chains derived from the given data. We consider existence of clusters, and while doing this kind of mixture modelling, we had to deal with the inherent problem of label switching, all of which is talked about in the chapter itself. Chapter 5 gives an overview of the data and shows us how the features in the dataset are related and can be grouped. It gives insights on what parameters to take care of when studying mental health. Chapter 6 talks about the unaccomplished goals and what else we would have liked to do. More importantly, it includes pointers on potential further research based on what has been carried out in this thesis. Finally in Chapter 7, we make some concluding remarks regarding the thesis. Please note that notations within a chapter are kept consistent, unless stated otherwise.

The results presented in this thesis highlight the applicability of the findings to real-world problems such as mental health, which is an extremely relevant subject of concern and further discussion. In addition to that, the overall objective of the research project was to be able to forecast mood and build a mobile phone application or a software, keeping in mind the structure of forecast tools such as weather applications, which could facilitate greater understanding of mental health and build predictive technologies that ensure better treatment of the same. As this was my rather larger goal, I believe the work in this thesis will contribute along those lines and that some day in the future, we will be to talk about mental health without any shame whatsoever, as well as assess and predict (which is also a term with quite a broad meaning) the behaviour patterns pertaining to mental health in a more confident and scientific way.

Methodology motivation

We are presented with data taken from Cloudy with a Chance of Pain study (`https://www.cloudywithachanceofpain.com/`) in which a mobile phone application collected details on self-reported variables like mood, pain, sleep quality, weather parameters like humidity, dew point and some baseline information like age and sex mainly to investigate the relationship between weather and pain (Dixon et al., 2019). The cohort contained residents from the UK aged 17 or above who were already experiencing chronic pain for at least three months preceding the study. More details of data pertaining to the studies have been provided in later chapters. Data obtained from such mobile health surveys are often in the form of longitudinal data which means same data are collected repeatedly over a period of time. Such kind of data collection is very common in the fields of health where health parameters of several people are tracked over a period of time. This helps in recognising patterns or trends of behaviour across a span of time. Commonly implemented methods include linear models (Garcia and Marder, 2017; Diggle et al., 2002) which capture correlation for e.g. mixed effect models are used in estimating random and fixed effects allowing to analyse behaviour inter and intra subjects. For identification of latent classes in longitudinal trajectories, Herle et al. (2020) compared mixed effect models with growth mixture models and latent class growth analysis and discussed the possibility of complexities in the case of multivariate trajectories instead of univariate. Proust-Lima et al. (2015); Komárek and Komárková (2013) extended mixed models to identify latent



Track Symptoms

Pain severity

How severe was your pain today?

How severe was your pain today?: Severe pain

Figure 1.1: Screenshot of the mobile application for daily symptom collection (Source: Druce et al. (2017))

classes in longitudinal data. Different methods may allow to relax the Markov chain assumption that we have taken, allowing for e.g. consideration of patterns in longer sequences of data, but that would be at the cost of the ability to model out of sample behaviour as Markov chains allow. Defining a distance measure in longitudinal data is difficult Liao (2005). Identifying markers of a progression of a condition and finding the sub-conditions which are called as endotypes or phenotypes are often discovered with the help of model-based clustering where a mixture model framework is considered e.g.: Gaussian Mixture Models are commonly used for mixture modelling (McNicholas and Murphy, 2010) while another approach is demonstrated by De la Cruz-Mesía et al. (2008) where a mixture of non-linear hierarchical models are considered. Hidden Markov Models (Eddy, 2004) are usually applied in order to model longitudinal data trajectories which are assumed to be Markov chains. They are especially helpful in estimating the underlying trajectories. However, in our case, we implemented the Expectation-Maximisation (EM) (Dempster et al., 1977) algorithm, similar to what is done in a Hidden Markov Model, to cluster the participants of the study on the basis of their self-reported trajectories of data and thereby, recognise the patterns in the longitudinal data– this resulted in the discovery of the distinct digital phenotypes that we talk about later. The algorithm has been elaborated in the next chapter.

Recalling that the overall goal is to set up a system which could help predicting mental health of an individual, we moved on to do Bayesian inference in studying given mood-pain trajectories of participants of the study. Once we classified the participants according to their mood-pain trajectories in Chapter 3, we wanted to investigate how considering a multinomial distribution, which is a natural assumption for such data (Tu, 2014), will help in estimating parameters of Dirichlet distribution which is taken to be the conjugate prior thereby, giving potential to predict a state of an individual given the history of their severities of mood-pain. To do so, we built a Bayesian inference model for this data and reported the findings for real-data. We carried out Bayesian Inference on the same dataset and compared the results. Similar methods are demonstrated by Holmes et al. (2012) and Cadez et al. (2003) where Dirichlet-multinomial modelling is done for genome data and web navigation data respectively. Li et al. (2019) implements a variation of Dirichlet-multinomial mixture model for topic modelling over short texts. But in our case, we work with digital health data and model transition matrices instead of vectors. Grimshaw and Alexander (2011) considers transition matrices to denote monthly movement of loans between delinquency states and then model them using Dirichlet-multinomial distribution to make forecasts. Frühwirth-Schnatter and Pamminger (2010) have done model based clustering on transition matrices to model the deviation of each row from the mean of a group-specific transition matrix and applied to wage data. We assume each row of transition matrix derived from the observational data to be sampled from Dirichlet distribution and the group is specified by Dirichlet parameters and mixture weights. Using these we introduce more parameters later which have been utilised to address the problem of label-switching in mixture models. The method is elaborated in Chapter 4.

We primarily analysed behaviour based on self-reported daily symptoms and developed methods to recognise underlying phenotypes which would also help in predicting behaviour. In the next chapter, we provide a background of the methods to familiarise with the methodology sections of the following chapters.

Chapter 2

Methods

With help of the mobile health study, the very ultimate goal is to improve mental health. To achieve so, we need to analyse the data, model it for further predictions and evaluate the outcomes to emphasise on interventions and treatments, policy making and subsequent research. We start by dealing with the following sub-problems first which are very much related to each other and cannot be solved independently. In view of the problem, at present, only mood and pain are taken into account. But other symptoms like sleep quality and physical activity may be taken into account to reinforce the model and considering more features will increase the dimensions which will need to be dealt with simultaneously. Missing values exist in the dataset when a participant did not enter a value for a symptom– such values in our studies in the thesis have been ignored. For e.g. value for mood is not recorded on a day, then the entire row from the table is removed. We begin our analysis with the consideration of the following sub-problems:

i) Fitting a model:
We have already got a dataset of observations. The observations on their own are meaningless, statistical methods are what give meaning to the observations by detecting patterns. Statistical modelling is the technique to encapsulate an entire dataset with the help of equation(s).

We assume the trajectories of the symptoms to be Markov chains and take a maximum-likelihood approach for the statistical modelling. The steps to build a suitable model for the given dataset is discussed in the next chapter.

ii) Identification of endotypes or phenotypes:
Understanding the aetilogy of a medical condition is very important to be able to understand a health condition. An endotype is “a subtype of a condition, which is defined by a distinct functional or pathophysiological mechanism (Lötvall et al., 2011)” whereas a phenotype is an observable trait. Identification of phenotypes can contribute towards

understanding the underlying mechanisms therefore helping with finding endotypes. For e.g. McInnes et al. (2016) proposed that with the help of characterising endotypes, it would be possible to select RA (Rheumatoid Arthritis) patients who are most likely to benefit from a certain anti-cytokine therapy. In the similar way, if we are able to identify endotypes for our problem, it will be helpful in advancing therapies targeted to specific mechanisms. Additionally, endotypes can help in prescribing medications or treatments accordingly or/ and predicting responses to given treatments. All in all, identifying endotypes is a crucial step which can answer many questions thereby, enhancing the overall management of a health issue.
To summarise, there are two key questions: i) are there any endotypes present?, and ii) if yes, what are they?

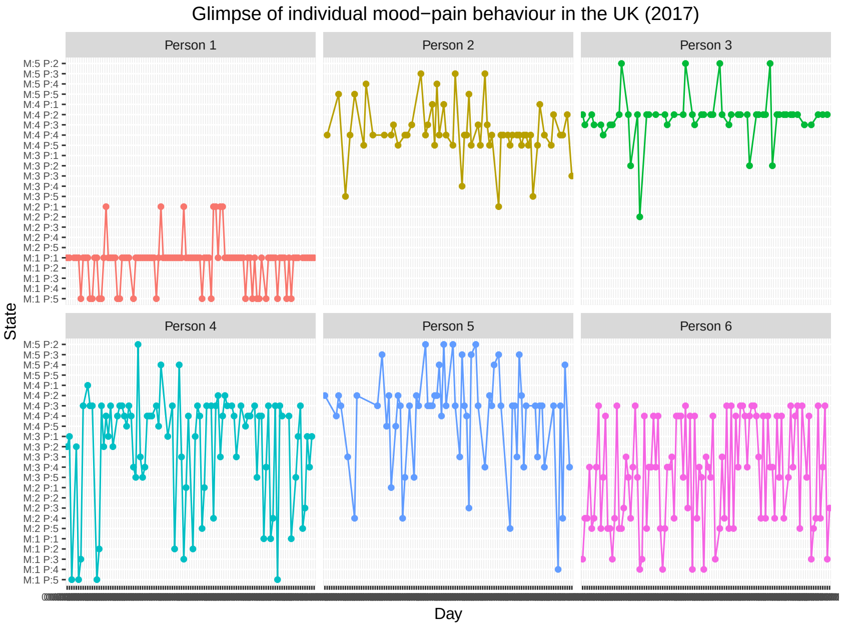


Figure 2.1: Glimpse of mood-pain behaviour of a random sample over a period of time

To address this, we begin with residual analysis which tells us about the possibility of clusters (thereby endotypes or phenotypes) and then perform clustering with the help of an EM algorithm based model in an attempt to spot the endotypes or phenotypes.

In the Chapter 3, we clustered the trajectories with the initial assumption that all participants belong to one group, and then we gradually distributed them to four clusters therefore, all participants of a cluster have the same transition probability matrix. In other words, the initial set up of our previous analysis considered all the observed transitions altogether while, in Chapter 4, we focus on the observed transition probability matrix per participant. The clusters are then defined by a distribution over transition matrices and it is these that lead to the Dirichlet-multinomial distributions which this chapter emphasises. Thus in this chapter, we:

1. Invent a distribution over transition matrices. If the number of states in n, then the n-rows of a transition matrix are drawn from n separate, n-dimensional Dirichlet distributions. This means that a component of the model is specified by an $n \times n$ matrix of Dirichlet shape parameters whose i-th row gives the shape parameters for the i-th row of the transition matrix.

2. Formulate a k-component mixture model whose components are specified by the sort of distribution over transition matrices defined above.

3. Marginalise-out the individual subjects' transition matrices: this ultimately leads to the Dirichlet-multinomial distribution for the counts in a given row which is discussed in the chapter.

4. Address the label-switching problem by (i) imposing an ordering constraint on the sums of all the shape parameters and (ii) getting a good starting guess via EM and then fitting Dirichlet distributions to the rows of those participants assigned to the same cluster.

We carry out the above steps on a test dataset by synthesising trajectories such that the rows of transition matrices are sampled from a Dirichlet distribution. Then, we run it on the real data made available to us. It must not be forgotten that this method is not restricted to the dataset considered throughout this PhD thesis, but can be applied to any data with similar data structure.

In Chapters 3 and 4, we considered how complex trajectories of mood and pain may, through unsupervised learning, be indicative of the presence of multiple disease phenotypes. Here, we consider a complementary unsupervised learning approach in which all ten variables (mood, pain severity, impact of pain on daily activity, physical activity, time spent outside, fatigue, sleep quality, morning stiffness, waking up tired) measured in the Cloudy With a Chance of Pain data `https://www.cloudywithachanceofpain.com/` may co-occur in participants in ways that indicate they arise from a number of underlying factors that is significantly less than 10. Such a dimensionality reduction is in fact often used as part of determination of clusters in data (Hastie et al., 2009) although here we will focus on the insights gained from it as a standalone analysis. So finally, we applied dimensionality reduction techniques to the given dataset and analysed the results. This has been included in Chapter 5.

Although the methods along with the applications are included in the following three chapters, in this chapter we aim to provide background of few topics that could help form a basis of the methods used later. Additionally, in the Section 2.2, we elaborate an expectation-maximisation algorithm specific to Markov chains which has been implemented in Chapter 3.

2.1 Mixture modelling

A mixture model is a probabilistic model that assumes the presence of sub-populations. Throughout the thesis, we concern ourselves only with finite mixtures which means there are finitely many components within the population considered for modelling. In Figure 2.2, we can see mixture population density where the sample of the population is composed of two Gaussian distributions.

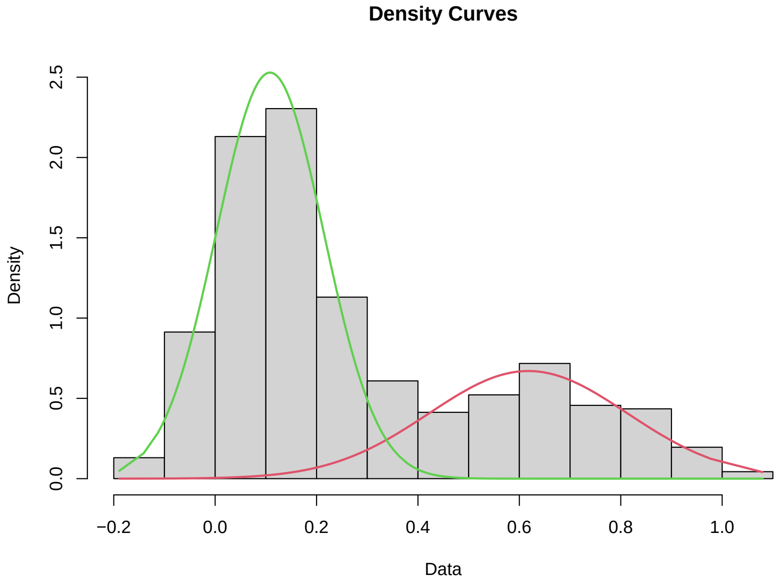


Figure 2.2: Mixture density of two Gaussian distributions (green and red curves) together with a grey histogram showing a finite sample from this density.

More formally, let $\mathbf{y}$ be a random vector of n observations $y_1, \ldots, y_n$ which are sampled from one of the K *mixture components*. We use the word 'component' interchangeably with 'cluster' and 'group'. Then each observation y_i is associated with a component label $k \in \{1, \ldots, K\}$. Let z_i, which is sometimes referred to as *latent variable*, denote the unknown or unobserved component label for an observation y_i. The marginal probability (density) of y_i is given by:

$$P(y_i) = \sum_{k=1}^{K} P(y_i \mid z_i = k) \underbrace{P(z_i = k)}_{\omega_k} = \sum_{k=1}^{K} \omega_k P(y_i \mid z_i = k) \tag{2.1}$$

Here the ω_k are the *mixture weights* or *mixture proportions* that represent probability of observation y_i belonging to the k-th component where $k \in \mathbb{Z}^+$. Thus, assuming $\omega_k \neq 0 \forall k$ and $\sum\limits_{k=1}^{K} \omega_k = 1$. Also the mixture component $P(y_i \mid z_i = k)$ gives the distribution of y_i given it was taken from component k.

We can then write a likelihood for a mixture model by taking products of the individual terms in Equation (2.1):

$$P(\mathbf{y}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \omega_k P(y_i \mid z_i = k)\,.$$

Note that for continuous random variables, P given above is a probability density function and for discrete variables, it is a probability mass function.

A major challenge for mixture modelling is the inherent issue of *label switching*, where the likelihood of the model is invariant under relabelling of the mixture components. In simpler language, the labels of mixture components can be swapped without changing the likelihood, which raises issues with identifying the component associated with a set of parameters. This occurs particularly while estimating parameters of a mixture model by taking a Bayesian approach. Hence this gives rise to $K!$ (*i.e.* K factorial) modes, where K is the total number of mixture components. Many methods (Papastamoulis, 2015) have been suggested in the past to address this problem, the primary one being the imposition of ordering constraints to the set of hyperparameters associated with the mixture components. This is discussed further in Chapter 4.

2.1.1 EM algorithm

Now we introduce the *Expectation-Maximisation* (EM) algorithm that describes the method of computing a maximum likelihood estimate of the parameters with an underlying distribution for a given dataset, especially when the data is incomplete or contains missing values. The following outline is based on the exposition of Bilmes et al. (1998).

Let $\mathbf{y}$ be the observed data generated by some distribution which represents the incomplete data. We assume a *complete* dataset that contains the observation set $\mathbf{y}$ and latent/ missing/ unobserved variable set $\mathbf{z}$, *i.e.* the complete dataset is $\mathbf{x} = (\mathbf{y}, \mathbf{z})$. Let Θ be the set of parameters, then the joint density distribution can be given by:

$$P(\mathbf{x} \mid \Theta) = P(\mathbf{y}, \mathbf{z} \mid \Theta) = P(\mathbf{z} \mid \mathbf{y}, \Theta) P(\mathbf{y} \mid \Theta)\,.$$

With this new density function of observed and missing/ latent variables, the likelihood function can be written as $L(\Theta \mid \mathbf{x}) = L(\Theta \mid \mathbf{y}, \mathbf{z}) = P(\mathbf{y}, \mathbf{z} \mid \Theta)$, which is the complete-data likelihood. Note that this likelihood function is a random variable as it contains missing information $\mathbf{z}$ which is unknown, random and usually, assumed to be sampled from some underlying distribution. The likelihood $L(\Theta \mid \mathbf{y})$ of the observed data $\mathbf{y}$ is known as the incomplete-data likelihood function.

As the name of the EM algorithm suggests, we first find the expected value of the complete-data log-likelihood $\log P(\mathbf{y}, \mathbf{z} \mid \Theta)$ with respect to the unknown set $\mathbf{z}$ given observation data $\mathbf{y}$ and current parameter estimates. Therefore, the expectation is given by:

$$Q(\Theta, \Theta^{(i-1)}) = \mathrm{E}\left[\log P(\{\mathbf{y}, \mathbf{z} \mid \Theta\} \mid \{\mathbf{y}, \Theta^{(i-1)}\})\right] \tag{2.2}$$

where, $\Theta^{(i-1)}$ is the set of current parameter estimates which is used to calculate expectation and Θ is the set of new parameters that is to be optimised to increase Q. Here, $\mathbf{z}$ is the random vector and its variables are sampled from the distribution $f(\zeta \mid \mathbf{y}, \Theta^{(i-1)})$. Therefore, the right side of (2.2) can be further written as:

$$\mathrm{E}\left[\log P(\{\mathbf{y}, \mathbf{z} \mid \Theta\} \mid \{\mathbf{y}, \Theta^{(i-1)}\})\right] = \int_{\zeta \in \mathcal{Z}} \log P(\mathbf{y}, \zeta \mid \Theta) f(\zeta \mid \mathbf{y}, \Theta^{(i-1)})\, d\zeta \tag{2.3}$$

where $f(\zeta \mid \mathbf{y}, \Theta^{(i-1)})$ is the marginal distribution of the unobserved data and is a function of observed data and current parameters. $\mathcal{Z}$ is the domain space of ζ.

In the second step of EM algorithm, we maximise the expectation computed in the previous step by finding

$$\Theta^{(i)} = \arg\max_{\Theta} Q(\Theta, \Theta^{(i-1)}) \tag{2.4}$$

Both these steps of EM are iterated in order to increase the log-likelihood, thereby converging to a local maximum of the likelihood function.

2.1.2 Maximum likelihood estimation

Let $\Theta = (\omega_1, \ldots, \omega_K, \theta_i, \ldots, \theta_K)$ be the set of parameters describing a mixture model with K components; then the probabilistic model is written as:

$$P(y \mid \Theta) = \sum_{k=1}^{K} \omega_k P_k(y \mid \theta_k)$$

for mixture weights ω_k and conditional probability density function $P_k = P(y_i \mid z_i = k)$ parameterised by θ_k for each component k.

Due to missing observations, the data remain incomplete. The incomplete data log-likelihood expression for the density for data with observations $\mathbf{y} = y_1, \ldots, y_n$ is given by

$$\log L(\Theta \mid \mathbf{y}) = \log P(\mathbf{y} \mid \Theta) = \log \prod_{i=1}^{n} P(y_i \mid \Theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \omega_k P_k(y_i \mid \theta_k) \right)$$

which can't be optimised easily since the integrand contains the log of a sum. Now considering $\mathbf{y}$ to be incomplete and assuming $\mathbf{z}$ to be the latent variable set that indicates which mixture component is associated with each data point *i.e.* $z_i \in \{1, \ldots, K\}$ for each i, so $z_i = k$ if i-th observation is generated by the k-th component of the mixture model. If the values of $\mathbf{z}$ are known, then the log-likelihood can be simplified as

$$\log L(\Theta \mid \mathbf{y}, \mathbf{z}) = \log P(\mathbf{y}, \mathbf{z} \mid \Theta) = \sum_{i=1}^{n} \log(P(y_i \mid z_i)P(z_i)) = \sum_{i=1}^{n} \log(\omega_{z_i} P_{z_i}(y_i \mid \theta_{z_i}))\,,$$

which gets rid of the log-sum expression and gives a form that can be optimised using appropriate techniques. As the values of $\mathbf{z}$ are unknown, if we simply consider $\mathbf{z}$ to be a random vector, we can proceed with the method.

We start by deriving a general expression for the distribution of the unobserved data $\mathbf{z}$. Let's assume a new set of parameters for the mixture model as $\Theta' = ({\omega_1}', \ldots, {\omega_K}', {\theta_1}', \ldots, {\theta_K}')$ with likelihood $L(\Theta' \mid \mathbf{y}, \mathbf{z})$. Given Θ', we can compute $P_k(y_i \mid {\theta_k}')$ for each i and j indices for numbers of observations and components respectively. Additionally, the mixture weights ω_k can be thought of as prior probabilities of each of the mixture components, *i.e.* $\omega_k = P(\text{component } k)$.

Using Bayes's rule (2.30), we get

$$P(z_i \mid y_i, \Theta') = \frac{{\omega_{z_i}}' P_{z_i}(y_i \mid {\theta_{z_i}}')}{P(y_i \mid \Theta')} = \frac{{\omega_{z_i}}' P_{z_i}(y_i \mid {\theta_{z_i}}')}{\sum_{k=1}^{K} {\omega_k}' P_k(y_i \mid {\theta_k}')}$$

and

$$P(\zeta \mid \mathbf{y}, \Theta') = \prod_{i=1}^{n} P(z_i \mid y_i, \Theta')$$

where $\zeta = (z_1, \ldots, z_n)$ is an instance of the unobserved data, which is independently drawn from a multinomial distribution with parameters equal to the mixture weights ω.

Now, substituting in Equation (2.3), we get

$$\begin{aligned}
Q(\Theta,\Theta') &= \sum_{\zeta\in\mathcal{Z}} \log(L(\Theta \mid \mathbf{y},\zeta))p(\zeta \mid \mathbf{y},\Theta') && (2.5)\\
&= \sum_{\zeta\in\mathcal{Z}} \sum_{i=1}^{n} \log(\omega_{z_i} p_{z_i}(y_i \mid \theta_{z_i})) \prod_{j=1}^{n} p(z_j \mid y_j,\Theta') && (2.6)\\
&= \sum_{z_1=1}^{K}\sum_{z_2=1}^{K}\cdots\sum_{z_n=1}^{K}\sum_{i=1}^{n} \log(\omega_{z_i} p_{z_i}(y_i \mid \theta_{z_i})) \prod_{j=1}^{n} p(z_j \mid y_j,\Theta') && (2.7)\\
&= \sum_{z_1=1}^{K}\sum_{z_2=1}^{K}\cdots\sum_{z_n=1}^{K}\sum_{i=1}^{n}\sum_{l=1}^{K} \delta_{l,z_i} \log(\omega_{z_l} p_l(y_i \mid \theta_l)) \prod_{j=1}^{n} p(z_j \mid y_j,\Theta') && (2.8)\\
&= \sum_{l=1}^{K}\sum_{i=1}^{n} \log(\omega_{z_l} p_l(y_i\theta_l)) \sum_{z_1=1}^{K}\sum_{z_2=1}^{K}\cdots\sum_{z_n=1}^{K} \delta_{l,z_i} \prod_{j=1}^{n} p(z_j \mid y_j,\Theta') && (2.9)
\end{aligned}$$

For $l \in \{1,\dots,K\}$,

$$\begin{aligned}
&\sum_{z_1=1}^{K}\sum_{z_2=1}^{K}\cdots\sum_{z_n=1}^{K} \delta_{l,z_i} \prod_{j=1}^{n} p(z_j \mid y_j,\Theta') && (2.10)\\
&= \left(\sum_{z_1=1}^{K}\cdots\sum_{z_{i-1}=1}^{K}\sum_{z_{i+1}=1}^{K}\cdots\sum_{z_n=1}^{K}\prod_{j=1,j\neq i}^{n} p(z_j \mid y_j,\Theta')\right) p(l \mid y_i,\Theta') && (2.11)\\
&= \prod_{j=1,j\neq i}^{n}\left(\sum_{z_j=1}^{K} p(z_j \mid y_j,\Theta')\right) p(l \mid y_i,\Theta') && (2.12)\\
&= p(l \mid y_i,\Theta'), && (2.13)
\end{aligned}$$

where $\sum_{i=1}^{K} p(i \mid y_j,\Theta') = 1$, so using Equations (2.9) and (2.13), we get

$$\begin{aligned}
Q(\Theta,\Theta') &= \sum_{l=1}^{K}\sum_{i=1}^{n} \log(\omega_l p_l(y_i \mid \theta_l))p(l \mid y_i,\Theta') && (2.14)\\
&= \sum_{l=1}^{K}\sum_{i=1}^{n} \log(\omega_l)p(l \mid y_i,\Theta') + \sum_{l=1}^{K}\sum_{i=1}^{n} log(p_l(y_i \mid \theta_l))p(l \mid y_i,\Theta') && (2.15)
\end{aligned}$$

Now to maximise the expectation, the two additive terms in Equation (2.15) are maximised separately.

2.2 An EM algorithm for a mixture of Markov chains

We have sequences of some data which we have assumed to be discrete-state and discrete-time Markov chains. These can be described as $x_{s,0}, x_{s,1}, \ldots, x_{s,t}, \ldots x_{s,n_s}$ where $x_{s,t} \in \{1, \ldots, M\}$ is a state, the first subscript, $s \in \{1, \ldots, S\}$, which we refer to by subject, indicates which of the S observed chains is considered and the second subscript, $t \in \{0, \ldots, n_s\}$, is a discrete-time.

Hence, $x_{s,t} =$ the state reported by subject s at time t. Therefore, a single Markov chain with M states gets characterised by an initial distribution $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)$ over the states that governs the first entry in the sequence, $\alpha_j = P(x_0 = j)$ and $M \times M$ *transition matrix* $\boldsymbol{T}$ whose entries are $T_{i,j} = P(x_t = j \mid x_{t-1} = i)$ such that $\sum_{j=1}^{M} T_{i,j} = 1$.

Therefore, the probability of observing a single sequence of states $\boldsymbol{x} = x_0, \ldots, x_n$ can be computed as:

$$\begin{aligned} P(\boldsymbol{x} \mid \boldsymbol{\alpha}, \boldsymbol{T}) &= P(x_0 \mid \boldsymbol{\alpha}, \boldsymbol{T}) \prod_{t=1}^{n} P(x_t \mid x_{t-1}, \boldsymbol{\alpha}, \boldsymbol{T}) \\ &= \alpha_{x_0} \prod_{t=1}^{n} T_{x_{t-1}, x_t} \\ &= \alpha_{x_0} \prod_{i,j \,\in \{1,\ldots,M\}} T_{i,j}^{N_{i,j}} \end{aligned}$$

where $N_{i,j} \in \mathbb{N}$ is the number of times a transition $i \to j$ appears in the state sequence $x_0, \ldots, x_n$.

Our goal is to model in a way such that each subject is a member of one of the classes (or ‘components’) and within each class, all the subjects report state sequences drawn from the same Markov chain.

2.2.1 Finite mixtures of Markov chains

Now consider a K-component mixture of Markov Chains specified by pairs of parameters $(\boldsymbol{\alpha}_k, \boldsymbol{T}_k)$ with $k \in \{1, \ldots, K\}$ defined so that

- $\boldsymbol{\alpha}_k$ is the distribution over initial states for sequences drawn from component k, so that

$$\alpha_{k,j} = P(x_0 = j \mid \text{sequence is drawn from chain component } k). \tag{2.16}$$

- T_k is a transition matrix for the k-th component, so that

$$T_{k,i,j} = P(x_t = j \mid x_{t-1} = i \text{ and sequence is drawn from chain component } k). \tag{2.17}$$

- ω_k is the discrete distribution of mixture weights ω over the K components such that

$$\omega_k = P(\text{component } k) \tag{2.18}$$

Given a mixture of Markov Chains, we can generate samples that look like many realisations of a discrete- state, discrete-time Markov Chain by performing the following steps for each subject s.

1. Choose a component number k_s by sampling from the discrete distribution $\boldsymbol{\omega}$ in Equation (2.18).

2. Choose an initial state by sampling from the discrete distribution $\boldsymbol{\alpha}_k$:

$$P(x_{s,0} = j \mid k) = \alpha_{k,j}. \tag{2.19}$$

3. Extend the sequence up to the appropriate length, n_s, by sampling successive states from rows of T_k:

$$P(x_{s,t} = j \mid x_{s,t-1} = i, k_s = k) = T_{k,i,j}. \tag{2.20}$$

If we know that a given observed sequence $\boldsymbol{x} = x_0, \ldots x_n$ was drawn from component k, the likelihood can be computed as:

$$P(\boldsymbol{x} \mid k) \;=\; \alpha_{k,x_0} \prod_{t=1}^{n} T_{k,x_{t-1},x_t} \;=\; \alpha_{k,x_0} \prod_{i,j\,\in\,\{1,\ldots,M\}} T_{k,i,j}^{N_{i,j}}$$

where, as above, $N_{i,j} \in$ is the number of times a transition $i \to j$ appears in the state sequence $\boldsymbol{x}$. Using this, we can also compute

$$P(\boldsymbol{x}) \;=\; \sum_{k=1}^{K} P(\boldsymbol{x}\mid k)\, P(k) \;=\; \sum_{k=1}^{K} \left(\alpha_{k,x_0} \prod_{i,j\,\in\,\{1,\ldots,M\}} T_{k,i,j}^{N_{i,j}} \right) \omega_k \tag{2.21}$$

Using Bayes' theorem (2.30), a posterior probability for the class assignments:

$$P(k \mid \boldsymbol{x}) = \frac{P(\boldsymbol{x} \mid k)P(k)}{P(\boldsymbol{x})} = \frac{\left(\alpha_{k,x_0} \prod_{i,j\,\in\,\{1,\ldots,M\}} T_{k,i,j}^{N_{i,j}}\right)\omega_k}{\sum_{k'=1}^{K}\left(\alpha_{k',x_0} \prod_{i,j\,\in\,\{1,\ldots,M\}} T_{k',i,j}^{N_{i,j}}\right)\omega_{k'}}$$

2.2.2 Maximum-likelihood estimation for mixture of chains

We want to estimate the parameters of a mixture of Markov chains but a problem that arises is that we do not know the class assignments of the each of the subject's chains *i.e.* we do not know which Markov chain belongs to which component. Therefore we treat the class assignments as latent variables and introduce a vector $\boldsymbol{k} \in \{1, \ldots, K\}^S$ with $\boldsymbol{k} = (k_1, \ldots, k_S)$, to hold the class assignments, and we'll write $k_s = k$ to indicate that subject s belongs to class k. The contribution to the likelihood from subject s with data $\boldsymbol{x}_s$ is then

$$\begin{aligned} P(\boldsymbol{x}_s \mid \Theta) &= \sum_{k=1}^{K} P(\boldsymbol{x}_s \mid \Theta, k_s = k) P(k_s = k \mid \Theta) \\ &= \sum_{k=1}^{K} \left(\alpha_{k_s, x_{s,0}} \prod_{i,j \,\in\, \{1,\ldots,M\}} T_{k_s,i,j}^{N_{s,i,j}} \right) \omega_k \,, \end{aligned}$$

where Θ represents all the parameters of the mixture of chains. The likelihood for the full dataset is then a product over subjects:

$$\begin{aligned} L &= \prod_{s=1}^{S} P(\boldsymbol{x}_s \mid \Theta) \\ &= \prod_{s=1}^{S} \left[\sum_{k=1}^{K} P(\boldsymbol{x}_s \mid \Theta, k_s = k)\, \omega_k \right] \\ &= \prod_{s=1}^{S} \left[\sum_{k=1}^{K} \omega_k \left(\alpha_{k, x_{s,0}} \prod_{i,j \,\in\, \{1,\ldots,M\}} T_{k,i,j}^{N_{s,i,j}} \right) \right] \end{aligned} \tag{2.22}$$

with the constraints for each component $k \in \{1, \ldots, K\}$ as:

$$\sum_{j=1}^{M} \alpha_{k,j} = 1, \qquad \text{and} \qquad \sum_{j=1}^{M} T_{k,i,j} = 1 \ \text{ for each } i \in \{1, \ldots, M\} \qquad \text{and} \qquad \sum_{k=1}^{K} \omega_k = 1. \tag{2.23}$$

Applying Lagrange multipliers, the expression for the constrained log-likelihood is

$$\begin{aligned} L &= \log(L) - \beta \left(\sum_{k=1}^{K} \omega_k \right) - \sum_{k=1}^{K} \left(\mu_k \sum_{j=1}^{M} \alpha_{k,j} + \sum_{i=1}^{M} \lambda_{k,i} \sum_{j=1}^{M} T_{k,i,j} \right) \\ &= \sum_{s=1}^{S} \log \left[\sum_{k=1}^{K} \omega_k \left(\alpha_{k, x_{s,0}} \prod_{i,j \,\in\, \{1,\ldots,M\}} T_{k,i,j}^{N_{s,i,j}} \right) \right] - \beta \left(\sum_{k=1}^{K} \omega_k \right) \\ &\quad - \sum_{k=1}^{K} \left(\mu_k \sum_{j=1}^{M} \alpha_{k,j} + \sum_{i=1}^{M} \lambda_{k,i} \sum_{j=1}^{M} T_{k,i,j} \right) \end{aligned} \tag{2.24}$$

and to maximise the log likelihood,we get the following equations for all $k \in \{1, \dots, K\}$,

$$\left.\frac{\partial L}{\partial \omega_k}\right|_{\widehat{\omega},\widehat{\boldsymbol{\alpha}},\widehat{T}} = 0, \qquad \left.\frac{\partial L}{\partial \alpha_{k,j}}\right|_{\widehat{\omega},\widehat{\boldsymbol{\alpha}},\widehat{T}} = 0 \qquad \text{and} \qquad \left.\frac{\partial L}{\partial T_{k,i,j}}\right|_{\widehat{\omega},\widehat{\boldsymbol{\alpha}},\widehat{T}} = 0,$$

where the last two equations hold, respectively, for all $j \in \{1, \dots, M\}$ and for all $i, j \in \{1, \dots, M\}$.

Considering the derivative of the optimisation target with respect to some transition probability $T_{k,p,q}$, where $k \in \{1, \dots, K\}$ and $p, q \in \{1, \dots, M\}$ are fixed, we get:

$$\begin{aligned}
\frac{\partial L}{\partial T_{k,p,q}} &= \left[\sum_{s=1}^{S} \left(\frac{N_{s,p,q}}{T_{k,p,q}}\right) \frac{\omega_k \left(\alpha_{k,x_{s,0}} T_{k,p,q}^{N_{s,p,q}}\right)}{\sum_{k'=1}^{K} \omega_{k'} \left(\alpha_{k',x_{s,0}} \prod_{i,j \in \{1,\dots,M\}} T_{k',i,j}^{N_{s,i,j}}\right)}\right] - \lambda_{k,p} \\
&= \frac{1}{T_{k,p,q}} \left[\sum_{s=1}^{S} N_{s,p,q} \left(\frac{P(k_s = k) P(\boldsymbol{x}_s \mid \Theta, k_s = k)}{P(\boldsymbol{x}_s \mid \Theta)}\right)\right] - \lambda_{k,p} \\
&= \frac{1}{T_{k,p,q}} \left[\sum_{s=1}^{S} \Gamma_{s,k} \, N_{s,p,q}\right] - \lambda_{k,p} \qquad (2.25)
\end{aligned}$$

where,

$$\Gamma_{s,k} = \frac{\omega_k \left(\alpha_{k,x_{s,0}} T_{k,p,q}^{N_{s,p,q}}\right)}{\sum_{k'=1}^{K} \omega_{k'} \left(\alpha_{k',x_{s,0}} \prod_{i,j \in \{1,\dots,M\}} T_{k',i,j}^{N_{s,i,j}}\right)} = P(k_s = k \mid \boldsymbol{x}_s, \Theta). \qquad (2.26)$$

So to compute maximum likelihood, we equate Equation (2.25) to 0, but $\Gamma_{s,k}$ includes the parameters of the mixture. If we ignore this dependence, we can solve the equation above to get an estimate $\widehat{T}_{k,p,q}$, which is given by

$$\widehat{T}_{k,p,q} = \frac{\sum_{s=1}^{S} \Gamma_{s,k} \, N_{s,p,q}}{\lambda_{k,p}}.$$

Since $\sum_{q=1}^{M} \widehat{T}_{k,p,q} = 1$, summing over the left-hand and right-hand sides gives:

$$\lambda_{k,p} = \sum_{q=1}^{M} \sum_{s=1}^{S} \Gamma_{s,k} \, N_{s,p,q}.$$

Therefore,

$$\widehat{T}_{k,p,q} = \frac{\widetilde{N}_{k,p,q}}{\sum_{r=1}^{M} \widetilde{N}_{k,p,r}}, \quad \text{where} \ \widetilde{N}_{k,p,q} = \sum_{s=1}^{S} \Gamma_{s,k} \, N_{s,p,q}. \qquad (2.27)$$

Here $N_{p,q}$ is the total number of transitions $p \to q$ observed in the data and $\widetilde{N}_{k,p,q}$ is a sum over the data in which the transition count for a subject s, $N_{s,p,q}$, is weighted by $\Gamma_{s,k}$

which is the posterior probability that subject s belongs to class k. This probability is a crucial element in our EM-algorithm outlined in Chapter 3.

Taking derivative with respect to $\alpha_{k,j}$, we get

$$\frac{\partial L}{\partial \alpha_{k,j}} = \left[\sum_{s\,|\,x_{s,0}=j} \left(\frac{1}{\alpha_{k,j}}\right) \frac{\omega_k \left(\alpha_{k,j} \prod_{p,q\in\{1,\ldots,M\}} T_{k,p,q}^{N_{s,p,q}}\right)}{\sum_{k'=1}^{K} \omega_{k'} \left(\alpha_{k',x_{s,0}} \prod_{p,q\in\{1,\ldots,M\}} T_{k',p,q}^{N_{s,p,q}}\right)} \right] - \mu_k,$$

and performing similar calculations as before, we get: and thus that

$$\widehat{\alpha}_{k,j} = \frac{\sum_{s\,|\,x_{s,0}=j} \Gamma_{s,k}}{\sum_{s=1}^{S} \Gamma_{s,k}}. \tag{2.28}$$

Similarly for ω_k, the derivative is

$$\frac{\partial L}{\partial \omega_k} = \left[\sum_{s=1}^{S} \left(\frac{1}{\omega_k}\right) \frac{\omega_k \left(\alpha_{k,x_{s,0}} \prod_{p,q\in\{1,\ldots,M\}} T_{k,p,q}^{N_{s,p,q}}\right)}{\sum_{k'=1}^{K} \omega_{k'} \left(\alpha_{k',x_{s,0}} \prod_{p,q\in\{1,\ldots,M\}} T_{k',p,q}^{N_{s,p,q}}\right)} \right] - \beta,$$

and the final estimate is

$$\widehat{\omega}_k = \frac{\sum_{s=1}^{S} \Gamma_{s,k}}{S}. \tag{2.29}$$

Thus we have the following algorithm:

Algorithm (The EM algorithm for a mixture of K Markov chains)**.** *Given a collection of state sequences, find maximum-likelihood estimates of the parameters $\boldsymbol{\omega}$, $\boldsymbol{\alpha}_k$ and $\boldsymbol{T}_k$ for a mixture of Markov chains.*

1. *Make initial assignments of the S subjects to K classes in the set $\Gamma_{s,k}$.*

2. *Given the $\Gamma_{s,k}$, Use Equations.* (2.29), (2.28) *and* (2.27) *to estimate the parameters of the mixture, $\widehat{\boldsymbol{\omega}}$, $\widehat{\boldsymbol{\alpha}}_k$ and $\widehat{\boldsymbol{T}}_k$.*

3. *Use Equation* (2.26) *to re-estimate the $\Gamma_{s,k}$, the posterior probabilities of class-membership.*

4. *Repeat steps 2 and 3 until the estimated parameters converge.*

The results from Borman (2009) ensure that this algorithm increases the likelihood (2.22) in each cycle.

2.3 Bayesian inference

Suppose I walk into my office in the university and find a chocolate cake at my desk. There could be numerous reasons why it is there, two obvious causes that would come to mind are – it is my birthday, it is somebody else's birthday, or some celebration in the office or university. Each of these possibilities has a prior probability. If it was my birthday yesterday and I had already received a cake from my friends then it is unlikely that it the same reason for this event. Another question in my mind is whose birthday it is then. As I consider more information, I will narrow down to a few possibilities. This scenario simply helps us in understanding how Bayesian statistics works where we take into consideration other information before arriving to a conclusion of how likely an event is.

Throughout the thesis we consider longitudinal health data, so we will give another example to form an intuition of Bayesian statistics. The case of missing data is a common challenge when collecting information from participants of a study over a period of time, which can lead to biases and inefficient inferences if not dealt with appropriately (Mason et al., 2010; Ma and Chen, 2018). It can arise due to several reasons: malfunctioning of the survey method (*e.g.* a mobile application had a bug and stopped working), unavailability of a participant on a specific day and so on. In fact, Rubin (1976) says these missing data can be classified into missing completely at random, missing at random and not missing at random. To address this, the missing values can be treated as random variables and a Bayesian model can be built considering different circumstances based on the importance of the missing values.

A one line explanation of Bayesian inference is that a subjective probability (density) is updated as we receive more information about the event. The following set of steps provides a more detailed description of carrying out Bayesian Data Analysis:

1. Find data based on the area of research. Identify the variables to be predictors and those that are predicted.
2. Specify a prior distribution.
3. Compute the posterior distribution, and analyse by making inferences about the parameters.

Bayesian inference is based on the Bayes' theorem which can be defined as: For data D and hypothesis H,

$$P(H \mid D) = \frac{P(D \mid H)P(H)}{P(D)} \tag{2.30}$$

where $P(H \mid D)$ is the posterior probability,
$P(D \mid H)$ is the likelihood,

$P(H)$ is prior and
$P(D)$ is the marginal likelihood.
In the above expression, it is the marginal likelihood which is often hard to compute.

Writing it even more explicitly, for observable random variables $x_1 \dots x_n$ the posterior density for parameter θ can be expressed as a parametric model of $x_1 \dots x_n$ given θ and the prior probability density for θ,

$$P(\theta \mid x_1, \dots, x_n) = \frac{\prod_{i=1}^{n} P(x_i \mid \theta) P(\theta)}{\int \prod_{i=1}^{n} P(x_i \mid \theta) P(\theta) \mathrm{d}\theta},$$

and $P(x_1, \dots, x_n \mid \theta) = L(\theta \mid x_1, \dots, x_n)$ is the likelihood function.

2.4 Selected distributions

We now show our preferred parameterisation of some distributions used throughout the thesis.

2.4.1 Bernoulli and Binomial distribution

Let $y_i \in \{1, \dots, n\}$ be a random variable with n possible outcomes of an event. When $n = 2$, we call it a Bernoulli distribution which gets extended to Binomial distribution for multiple trials.

Let y_i be a binary random variable with two possible outcomes 0 and 1, so $y_i \in \{0, 1\}$. Suppose for one trial- which we call Bernoulli trial in this scenario, probability $P(y_i = 1) = \theta$, then $P(y_i = 0) = 1 - \theta$. Therefore, the probability mass function $P(y_i \mid \theta)$ of **Bernoulli** distribution is written as $\text{Bernoulli}(y_i \mid \theta) = \theta^{y_i}(1-\theta)^{1-y_i}$. The expected value is given as:

$$\mathrm{E}[y] = \sum_{y_i \in 0,1} y_i P(y_i) = P(y_i = 1) = \theta$$

Variance is:

$$\mathrm{Var}[y] = \mathrm{E}[y^2] - (\mathrm{E}[y])^2 = \theta(1 - \theta)$$

Likelihood for a sequence of data $\mathcal{S} = \{y_1, \dots, y_N\}$ with 2 possible outcomes is given by:

$$\begin{aligned} L(\theta) &= P(\mathcal{S} \mid \theta) \\ &= \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{1-y_i} \\ &= \theta^{\sum\limits_{i=1}^{N} y_i}(1-\theta)^{\sum\limits_{i=1}^{N}(1-y_i)} \\ &= \theta^{N_1}(1-\theta)^{N-N_1} \end{aligned}$$

where $N_1 = \sum\limits_{i=1}^{N} y_i$ is the number of 1's *i.e.* it is the number of times when $y = 1$. Now to find the likelihood of N_1 outcomes out of N total trials instead of a sequence, we include a combinatorial factor to the likelihood found above. Thus, this gives the probability mass function for **Binomial** distribution which is written as:

$$P(N_1 \mid N, \theta) = \binom{N}{N_1}\theta^{N_1}(1-\theta)^{N-N_1} = \mathrm{Binom}(\theta, N)$$

where the Binomial coefficient $\binom{N}{N_1} = \frac{N!}{(N-N_1)!N_1!}$ represents the number of ways of selecting N_1 outcomes out of N trials of an event which is the Binomial distribution on the counts of possible outcomes. It is a generalisation of Bernoulli distribution where number of trials is more than 1.

2.4.2 Multinomial distribution

Multinomial distribution extends Binomial distribution by taking the possible number of outcomes to be more than 2.

Let $y_i \in \{1, \ldots, n\}$ be a categorical random variable with n possible outcomes and $P(y_i) = \theta_i$, then the probability for one trial:

$$P(y_i \mid \theta) = \mathrm{Mult}(y_i \mid \theta) = \prod_{i=1}^{n} \theta_i^{\mathbb{I}(y_i=1)}$$

where $\mathbb{I}$ is the Indicator function so $\mathbb{I}(y = i) = 1$, if $y = i$ or it is 0, otherwise. Therefore, we get:

$$P(y_i \mid \theta) = \prod_{i=1}^{n} \theta_i^{y_i}$$

Likelihood for the sequence of data $\mathcal{S} = \{y_1, \ldots, y_N\}$ with n possible outcomes is given by:

$$L(\theta) = P(\mathcal{S} \mid \theta) = \prod_{j=1}^{N}\prod_{i=1}^{n} \theta_i^{\mathbb{I}(y_j=i)} = \prod_{i=1}^{n} \theta_i^{N_i}$$

where $N_i = \sum\limits_{j=1}^{N} \mathbb{I}(y_j = i)$ is the number of times $y = i$.

Now to write the likelihood for a given number of counts $N_1, \ldots, N_n$ out of total N trials, we include Multinomial coefficient, which gives the probability mass function of **Multinomial** distribution on counts of data as :

$$P(N_1, \ldots, N_n \mid N) = \text{Mult}(\theta, N) = \binom{N}{N_1 \ldots N_n} \prod_{i=1}^{n} \theta_i^{N_i}$$

where $\binom{N}{N_1 \ldots N_n} = \frac{N!}{N_1! \ldots N_n!} = \binom{N}{N_1}\binom{N-N_1}{N_2} \ldots \binom{N-N_1-\cdots-N_{n-1}}{N_n}$ is the Multinomial coefficient.

2.4.3 Gamma and Beta distributions

A random variable y has **Gamma** distribution with shape and scale parameters α and β respectively if the probability density function is given by:

$$P(y) = \text{Gam}(\alpha, \beta) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^{\alpha}}\, y^{\alpha-1} \mathrm{e}^{-y/\beta}, & \text{if } 0 < y < \infty \\ 0, & \text{otherwise} \end{cases}$$

where $\alpha, \beta > 0$ and $\Gamma(\alpha) = \int\limits_{0}^{\infty} t^{\alpha-1}\mathrm{e}^{-t}\mathrm{d}t$ is the Gamma function.

Let us take another random variable x and the same parameters α and β, the probability density function of **Beta** distribution is given as:

$$P(y) = \text{Beta}(\alpha, \beta) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}, & \text{if } 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

Beta function is given by: $\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x$

2.4.4 Dirichlet distribution

Here, we derive an expression for the Dirichlet distribution by implementing a sampling strategy of generating random variables from Gamma distribution. Let y_i be a random variable generated from Gamma distribution $\text{Gam}(\alpha_i, 1)$ for $i = 1, \ldots, n$. Let $y_1, \ldots, y_n$ be independent samples, then the joint probability distribution function is given as:

$$P(y_1, \ldots, y_n) = \begin{cases} \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha_i)\beta_i^{\alpha}}\, y_i^{\alpha_i-1}\mathrm{e}^{-y_i}, & \text{if } 0 < y_i < \infty \\ 0, & \text{otherwise} \end{cases}$$

We perform the following transformation on the random variables y_i.

$$\begin{aligned} p_i &= \frac{y_i}{y_1 + \cdots + y_n},\ i \in \{1, \ldots, n-1\} \\ p_n &= y_1 + \cdots + y_n \end{aligned} \tag{2.31}$$

This means $0 \leq p_i \leq 1$ for $i \in \{1, \ldots, n-1\}$ and $0 \leq p_n < \infty$. Equating the transformation Equations (2.31), we get:

$$\begin{aligned} y_1 &= p_1 p_n, \\ &\vdots \\ y_{n-1} &= p_{n-1} p_n, \\ y_n &= p_n(1 - p_1 - \cdots - p_{n-1}) \end{aligned}$$

Now, the Jacobian of the transformation is given by:

$$J = \begin{bmatrix} p_n & 0 & \ldots & 0 & p_1 \\ 0 & p_n & \ldots & 0 & p_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \ldots & p_n & p_{n-1} \\ -p_n & -p_n & \ldots & -p_n & 1 - p_1 - \cdots - p_{n-1} \end{bmatrix}_{n \times n}$$

and its determinant is:

$$\mid J \mid \ = p_n^{n-1}$$

Therefore, the joint probability density function is re-written as:

$$\begin{aligned} &f(p_1, \ldots, p_{n-1}, p_n) = f(y_1, \ldots, y_{n-1}, y_n) \mid J \mid \\ &= \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha_i)} \prod_{i=1}^{n-1} (p_i p_n) \mathrm{e}^{-(p_i p_n)} p_n^{\alpha_n - 1} (1 - p_1 - \cdots - p_{n-1})^{\alpha_n - 1} \mathrm{e}^{-p_n(1 - p_1 - \cdots - p_{n-1})} p_n^{n-1} \\ &= \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha_i)} \prod_{i=1}^{n-1} p_i^{\alpha_i - 1} (1 - p_1 - \cdots - p_{n-1})^{\alpha_n - 1} p_n^{\alpha_1 + \cdots + \alpha_n - 1} \mathrm{e}^{-p_n} \end{aligned}$$

Integrating out the n-th term of the transformed variables to get the marginal density as:

$$\begin{aligned} &f(p_1, \ldots, p_{n-1}) = \int_0^\infty f(p_1, \ldots, p_{n-1}, p_n) \mathrm{d}p_n \\ &= \int_0^\infty \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha_i)} \prod_{i=1}^{n-1} p_i^{\alpha_i - 1} (1 - p_1 - \cdots - p_{n-1})^{\alpha_n - 1} p_n^{\alpha_1 + \cdots + \alpha_n - 1} \mathrm{e}^{-p_n} \mathrm{d}p_n \\ &= \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha_i)} \prod_{i=1}^{n-1} p_i^{\alpha_i - 1} (1 - p_1 - \cdots - p_{n-1})^{\alpha_n - 1} \Gamma(\alpha_1 + \cdots + \alpha_n) \int_0^\infty \frac{p_n^{\alpha_1 + \cdots + \alpha_n - 1} \mathrm{e}^{-p_n}}{\Gamma(\alpha_1 + \cdots + \alpha_n)} \mathrm{d}p_n \end{aligned}$$

where,

$$\int_0^\infty \frac{p_n^{\alpha_1+\cdots+\alpha_n-1}\mathrm{e}^{-p_n}}{\Gamma(\alpha_1+\cdots+\alpha_n)}\mathrm{d}p_n = 1$$

since the integrand is probability density function of Gamma distribution Gam$(\alpha_1+\cdots+\alpha_n, 1)$. This gives us the probability density function of the Dirichlet distribution with parameters $\{\alpha_1, \ldots \alpha_n\}$ as:

$$f(p_1,\ldots,p_{n-1}) = \frac{\Gamma(\alpha_1+\cdots+\alpha_n)}{\prod\limits_{i=1}^{n}\Gamma(\alpha_i)}\prod_{i=1}^{n-1}p_i^{\alpha_i-1}(1-p_1-\cdots-p_{n-1})^{\alpha_n-1} \tag{2.32}$$

On simplification of the Equation (2.32), we rewrite the probability density function in the following form. Let $\mathbf{p} = \{p_1, \ldots, p_n\}$ be a probability vector of n components such that $\sum\limits_{i=1}^{n} p_i = 1$ and $p_i \geq 0$ for $i \in \{1, \ldots, n\}$. Then the probability density function of **Dirichlet distribution** over simplex Δ_{n-1} of dimension $n-1$ is given as:

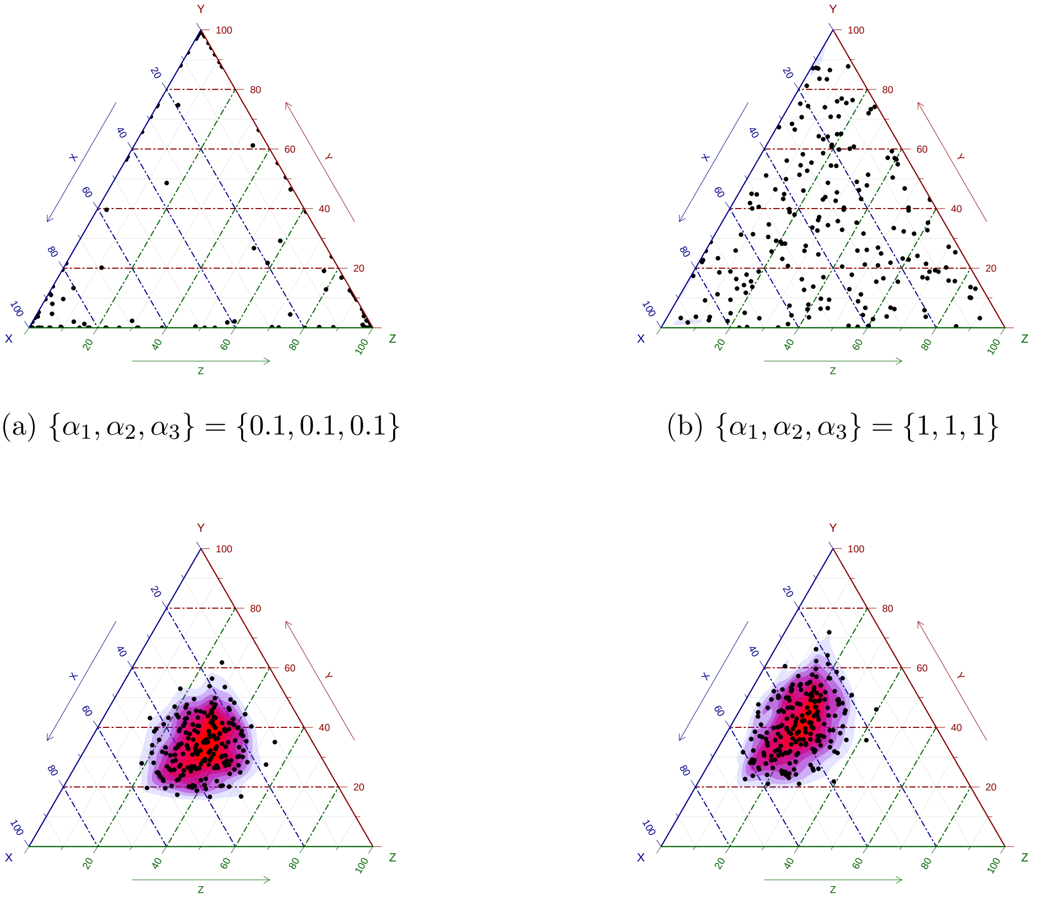
$$f(p_1,\ldots,p_n) = \mathrm{Dir}(\alpha_1,\ldots,\alpha_n) = \frac{\Gamma(\boldsymbol{\alpha})}{\prod\limits_{i=1}^{n}\Gamma(\alpha_i)}\prod_{i=1}^{n}p_i^{\alpha_i-1} \tag{2.33}$$

where $\boldsymbol{\alpha} = \sum\limits_{i=1}^{n} \alpha_i$ and $\{\alpha_1, \ldots, \alpha_n\}$ are the Dirichlet shape parameters such that $\alpha_i > 0$ for $i \in \{1, \ldots, n\}$.

An $(n-1)$-*dimensional simplex* represented by Δ_{n-1}, is a vector of length n which has been defined by the following set:

$$\Delta_n = \left\{ \{p_1, \ldots, p_n\} \in \mathbb{R}^n \mid p_i \geq 0 \text{ and } \sum_{i=1}^{n} p_i = 1 \right\}.$$

Points lying in the interior of the simplex Δ_{n-1} are probability distributions over the numbers $\{1, \ldots, n\}$. In the case of Dirichlet distribution, we talk about the shape parameters α over the simplex. Figure 2.3 shows us how the shape parameters distribute the weight of the random variables generated from a Dirichlet distribution.



(a) $\{\alpha_1, \alpha_2, \alpha_3\} = \{0.1, 0.1, 0.1\}$

(b) $\{\alpha_1, \alpha_2, \alpha_3\} = \{1, 1, 1\}$

(c) $\{\alpha_1, \alpha_2, \alpha_3\} = \{10, 10, 10\}$

(d) $\{\alpha_1, \alpha_2, \alpha_3\} = \{10, 10, 5\}$

Figure 2.3: Ternary diagrams for different sets of Dirichlet shape parameters

Now we compute the mean E$[y]$ and variance Var$[y]$ (Lin, 2016).

$$\begin{aligned}
\mathrm{E}[y_i] &= \int_0^1 \cdots \int_0^1 y_i \frac{\Gamma(\alpha)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} p_k^{\alpha_k - 1} \mathrm{d}y_1 \ldots \mathrm{d}y_K \\
&= \frac{\Gamma(\alpha)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{\substack{k=1 \\ i \neq k}}^{K} \frac{\Gamma(\alpha_k)\Gamma(\alpha_i + 1)}{\Gamma(\alpha + 1)} \int_0^1 \cdots \int_0^1 \prod_{\substack{k=1 \\ i \neq k}}^{K} \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha_k)\Gamma(\alpha_i + 1)} \prod_{\substack{k=1 \\ i \neq k}}^{K} y_k^{\alpha_k - 1} y_i^{\alpha_i + 1 - 1} \mathrm{d}y_1 \ldots \mathrm{d}y_K \\
&= \frac{\Gamma(\alpha)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{\substack{k=1 \\ i \neq k}}^{K} \frac{\Gamma(\alpha_k)\Gamma(\alpha_i + 1)}{\Gamma(\alpha + 1)} \\
&= \frac{\Gamma(\alpha)\Gamma(\alpha_i + 1)}{\Gamma(\alpha_i)\Gamma(\alpha + 1)} \\
&= \frac{\alpha_i}{\alpha}
\end{aligned}$$

Using similar steps as that for E$[y]$, we get the second moment E$[y^2]$ as:

$$\mathrm{E}[y_i^2] = \frac{\Gamma(\alpha)\Gamma(\alpha_i + 2)}{\Gamma(\alpha_i)\Gamma(\alpha + 2)} = \frac{\alpha_i(\alpha_i + 1)}{\alpha(\alpha + 1)}$$

and therefore, substituting the values from above, we get the variance to be:

$$\mathrm{Var}[y_i] = \mathrm{E}[y_i^2] - (\mathrm{E}[y_i])^2 = \frac{\alpha_i(\alpha - \alpha_i)}{\alpha^2(\alpha + 1)}$$

Now to compute the covariance, we begin by writing the product moment E$[y_i y_j]$ as:

$$\mathrm{E}[y_i y_j] = \frac{\Gamma(\alpha)\Gamma(\alpha_i + 1)\Gamma(\alpha_j + 1)}{\Gamma(\alpha + 2)\Gamma(\alpha_i)\Gamma(\alpha_j)} = \frac{\alpha_i \alpha_j}{\alpha(\alpha + 1)}$$

where $i \neq j$.

Therefore, the covariance Cov$[y_i, y_j]$:

$$\mathrm{Cov}[y_i, y_j] = \mathrm{E}[y_i y_j] - \mathrm{E}[y_j]\mathrm{E}[y_j] = \frac{\alpha_i \alpha_j}{\alpha^2(\alpha + 1)}, \ i \neq j$$

2.4.5 Historical note

The Dirichlet distribution is named after Johann Peter Gustav Lejeune Dirichlet who was a German mathematician born with a French last name. He was born in a town which was under was the First French Empire at that time. “Lejeune Dirichlet” can be translated to the boy from Richelet. The point of mentioning this is that it is unknown what pronunciation Dirichlet himself would have preferred and the correct pronunciation can be debatable. On this note, an interesting fact to add is that Dirichlet's primary three advisors were known to be Siméon Poisson, Joseph Fourier and Carl Gauss.

2.5 Markov chain Monte Carlo

Markov Chain Monte Carlo methods have been known about for some time, but it is only in the last 30 years that they have gained widespread adoption due to availability of computational resources. They are based on adapting Monte Carlo methods for Markov chains, as the name suggests.

A Markov chain is a sequence of events where the probability of occurrence of the next event is determined by the previous one and it is independent of all the other events before that. This characteristic of retaining no memory of the past is also known as *memorylessness* and often referred to as Markov property in the context of Markov chains or processes. In this thesis, we consider only finite-dimensional, discrete-time Markov chains.

Monte Carlo refers to computational methods that simulate a probability model. Monte Carlo is a gambling casino in Monaco which is said to have given name to the method implying simulation through (pseudo-) random number generation.

Let $X_1, \dots, X_n$ be an n- sequence of independent, identically distributed (i.i.d) simulations of a probability model. Let X be a generic realisation of the probability model such that all the X_i have the same distribution $\pi(x)$ as that of X. As we have simulated a random process, now we can compute probabilities and expectations by taking average over these simulations (Geyer, 1998). To calculate expectation of a random variable $g(X)$ analytically, $\mathrm{E}(g(X)) = \int g(x)\pi(x)\mathrm{d}x$ can be tricky sometimes, so we can use Monte Carlo integration instead (Metropolis et al., 1953). Therefore, we get $\mu = \mathrm{E}(g(X))$. Monte Carlo approximation gives:

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i).$$

Here, $\hat{\mu}_n$ is the sample mean of i.i.d random variables $g(X_1), \dots, g(X_n)$ with expectation μ. The Strong Law of Large Numbers tells us that $\hat{\mu}_n$ converges to μ almost surely as the number of simulations n tends to infinity. Furthermore, if the variance $\mathrm{Var}(g(X))$ is finite say σ^2, then by Central Limit Theorem says that $\hat{\mu}_n$ is asymptotically normal with mean μ and variance σ^2/n. But sampling those sequences from model still remains a problem.

Now combining the concept of Monte Carlo and Markov chains, we talk about Markov chain Monte Carlo, mostly referred to as MCMC. MCMC is a tool for generating samples of a probability model while exploring the state space through a Markov chain mechanism so that the chain spends more time in the regions with high probability mass. In short, Markov chain samples are generated from a probability distribution and then Monte Carlo approximation gets implemented.

Let $x^{(i)}$ be a Markov chain drawn from a target distribution $\pi(x)$, then

$$\pi(x^{(i)} \mid x^{(i-1)}, \ldots, x^{(1)}) = T(x^{(i)} \mid x^{(i-1)})$$

where T denotes transition probability kernel and the chain is homogeneous (Andrieu et al., 2003) if it remains invariant *i.e.* $\pi T = \pi$ with $\sum_{x^{(i)}} T(x^{(i)} \mid x^{(i-1)}) = 1$ for any i. For any starting point, a Markov chain will converge to its invariant, also known as *stationary distribution* $\pi(x)$ if T has the following properties:

1. *Irreducibility*: A given Markov chain is irreducible if it is possible to visit all other states from any given state *i.e.* transition probability from one state to another is positive for the complete state space. So the Markov chain should not be reducible. In Figure 2.4, we see a reducible Markov chain as we cannot visit state C if we start from state A or B.

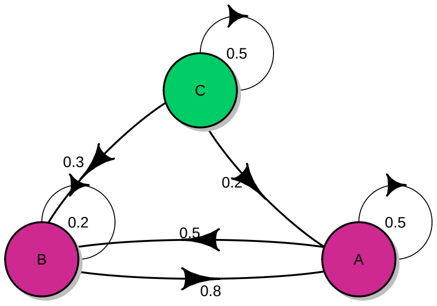


Figure 2.4: Reducible Markov chain

2. *Aperiodicity*: A Markov chain is aperiodic if it does not get trapped in any of the loops *i.e.* the greatest common divisor (g.c.f) of number of times of possible return of state to itself is 1 holds true for all states of the Markov process. Hence the Markov chain has no periodic states. In Figure 2.5, we see that all the states are periodic since a state can be visited back only in even number of jumps thereby the

g.c.f is greater than 1. However, this Markov chain is reducible while the previous Markov chain shown in Figure 2.4 was aperiodic.

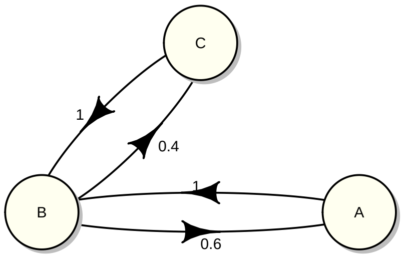


Figure 2.5: Periodic Markov chain

There are many MCMC algorithms which describe how a sampling is carried out. One of the widely used algorithms is the Metropolis-Hastings algorithm which is introduced in the next section.

2.5.1 Metropolis-Hastings

The Metropolis-Hastings (MH) algorithm is an MCMC method which involves sampling a new value x^* given x according to proposal distribution $q(x^* \mid x)$ and the invariant distribution $\pi(x)$. The Markov chain selects x^* with the acceptance probability (Hastings, 1970) given by:

$$\min \left\{ 1, \frac{\pi(x^*)q(x \mid x^*)}{\pi(x)q(x^* \mid x)} \right\}, \tag{2.34}$$

otherwise it remains at x. More details can be found in Chib and Greenberg (1995) amongst many other sources.

2.5.2 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) previously known as Hybrid Monte Carlo is another MCMC technique which is based on Hamiltonian dynamics. HMC modifies the MH algorithm of sampling from a proposal distribution $q(y \mid x)$ by adding two steps of proposal on the basis of the Hamiltonian system and improving the acceptance probability accordingly.

HMC was developed by Duane et al. (1987) as a numerical simulation method for lattice field theory which was then known as Hybrid Monte Carlo, in which MCMC and deterministic simulation methods got combined. Later Neal et al. (2011) and others implemented it in the context of statistics thus, soon the method got started to be known as Hamiltonian Monte Carlo.

Now we briefly describe the Hamiltonian dynamics before formulating HMC. A way of visualising the Hamiltonian system can be by imagining someone on a roller-coaster ride whose state is determined by their position r and momentum at a point ρ in the ride. Potential energy $U(r)$ is directly proportional to the height of the position of the person on the roller coaster with respect to the ground, and Kinetic energy $K(\rho)$ is given by $|\rho|^2/2m$ where m is the mass of the object which is the rider in this context. Thus, the Hamiltonian of the rider will be determined by the sum of their potential and kinetic energies which remains conserved by the Law of Conservation of Energy. So we write the following equation of the Hamiltonian system in a generic form:

$$H(r, \rho) = U(r) + K(\rho)$$

with

$$\frac{\mathrm{d}r_i}{\mathrm{d}t} = \frac{\partial H}{\partial \rho_i} \quad \text{and} \quad \frac{\mathrm{d}\rho_i}{\mathrm{d}t} = -\frac{\partial H}{\partial r_i} \tag{2.35}$$

for $i = 1, \ldots, d$ where d is the dimension of position and momentum vectors r and ρ respectively.

Alternatively, the vectors r and ρ can be combined to $z = (r, \rho)$ with $2d$ dimensions, so the Equations (2.35) can be rewritten as:

$$\frac{\mathrm{d}z}{\mathrm{d}t} = J\nabla H(z) \tag{2.36}$$

where ∇H is the gradient of H and

$$J = \begin{bmatrix} \mathbf{0}_{d\times d} & \mathbf{I}_{d\times d} \\ -\mathbf{I}_{d\times d} & \mathbf{0}_{d\times d} \end{bmatrix}_{2d\times 2d} \tag{2.37}$$

where $\mathbf{0}$ and $\mathbf{I}$ are the Zero and Identity matrices respectively. There are several properties specified by Neal et al. (2011) which are essential for Markov Chain Monte Carlo (MCMC) updates to happen. An outline of the properties is given below:

- Reversibility: For time $t \to t + s$, the mapping T_s: $H(r_t, \rho_t) \to H(r_{t+s}, \rho_{t+s})$ is one-one, therefore, T_{-s} exists.

- Conservation of Hamiltonian *i.e* $\frac{\mathrm{d}H}{\mathrm{d}t} = 0$.

- Preservation of volume: This is also given by Liouville's theorem (Liouville). If the mapping T_s is applied to points in some region R of (r, ρ) space with volume V then the image of R under T_s will have same volume V.

- Symplecticness: Preservation of volume is also a result of the symplecticness. Let $z = (r, \rho)$ and J is given by (2.37), the symplecticness condition is Jacobian matrix B_s of T_s satisfying $B_s^T J^{-1} B_s = J^{-1}$.

For computational implementation of HMC, it is essential to discretise the time in Hamiltonian equations. Time is broken down into small step-sizes ϵ as $\epsilon, 2\epsilon, 3\epsilon, \ldots$, and then I solve the Hamiltonian system of equations using an appropriate numerical method. One of such numerical methods is Leapfrog integrator which updates the r and ρ according to the following set of equations (Ziegler, 2019):

$$\begin{aligned}
\rho_i(t + \frac{\epsilon}{2}) &= \rho_i(t) - \frac{\epsilon}{2}\nabla U(r(t)) \\
r_i(t + \epsilon) &= r_i t + \epsilon \rho_i(t + \frac{\epsilon}{2}) \\
\rho_i(t + \epsilon) &= \rho_i(t + \frac{\epsilon}{2}) - (\frac{\epsilon}{2})\nabla U(r(t + \epsilon))
\end{aligned}$$

To summarise, with the help of leapfrog integrator, the Hamiltonian Monte Carlo algorithm uses a Markov chain containing alternate stochastic updates of momentum ρ and Hamiltonian updates, and the resulting state is accepted or rejected on the basis of Metropolis-Hasting criterion on Hamiltonian H (Bishop and Nasrabadi, 2006). The probability of accepting a candidate state is given by:

$$\min\left(1, \mathrm{e}^{H(r,\rho) - H(r^*,\rho^*)}\right) \tag{2.38}$$

where (r, ρ) is the initial state and (r^*, ρ^*) is the new state after leapfrog integration.

Statistical interpretation

In the context of probability and statistics, a distribution $p(\phi)$ on the momenta augments the posterior probability density $p(\theta \mid y)$, then the joint distribution $p(\theta, \phi \mid y) = p(\phi)p(\theta \mid y)$ where ϕ is an auxiliary variable. In addition to the posterior probability density, gradient of the log-posterior probability density $\frac{\mathrm{d}\log p(\theta \mid y)}{\mathrm{d}\theta}$ is also supplied to HMC. Here θ and ϕ play the roles of position and momentum vectors simultaneously satisfying the Hamiltonian system of equations (Gelman et al., 2014; Betancourt, 2017) where the joint density $p(\theta, \phi \mid y)$ defines the Hamiltonian as:

$$\begin{aligned}
H(\theta, \phi) &= -\log p(\theta, \phi) \\
&= -\log p(\theta) - \log p(\phi \mid \theta) \\
&= U(\theta) + K(\phi \mid \theta)
\end{aligned}$$

where, $U(\theta) = -\log p(\theta)$ is the Potential energy and

$K(\phi \mid \theta) = -\log p(\phi \mid \theta)$ is the Kinetic energy.

Chapter 3

Modelling and classifying joint trajectories of self-reported mood and pain in a large cohort study

Rajenki Das[1], Mark Muldoon[1], Mark Lunt[2], John McBeth[2], Belay Birlie Yimer[2], Thomas House[1]

1 – Department of Mathematics, University of Manchester, Manchester, UK

2 – Centre for Epidemiology Versus Arthritis, University of Manchester, Manchester, UK

Publication link

https://doi.org/10.1371/journal.pdig.0000204

Abstract

It is well-known that mood and pain interact with each other, however individual-level variability in this relationship has been less well quantified than overall associations between low mood and pain. Here, we leverage the possibilities presented by mobile health data, in particular the “Cloudy with a Chance of Pain” study, which collected longitudinal data from the residents of the UK with chronic pain conditions. Participants used an App to record self-reported measures of factors including mood, pain and sleep quality. The richness of these data allows us to perform model-based clustering of the data as a mixture of Markov processes. Through this analysis we discover four endotypes with distinct patterns of co-evolution of mood and pain over time. The differences between endotypes are sufficiently large

to play a role in clinical hypothesis generation for personalised treatments of co-morbid pain and low mood.

3.1 Introduction

Mental disorder has been associated with a substantial excess in all-cause mortality risk (Prince et al., 2007). It is often accompanied by mood disorders which, according to the World Health Organisation (WHO) (Organization and Others, 2017), are one of the leading causes of disability. Mental health can suffer due to many social, physical and other factors, and mathematical approaches are uniquely placed to disentangle these complex issues. In view of the difficulty in clearly defining “mental illness” itself, simply linking its absence with positive mental health is not enough (Jahoda, 1958; Galderisi et al., 2015). One may not suffer from any “mental illness”, yet not be mentally fit. So, identifying markers of mental health disorders remains a vital challenge.

Chronic pain is a persistent or intermittent pain that lasts for more than 3 months (Sheng et al., 2017), and approximately one fifth of the population in the USA and Europe are affected by it (Breivik et al., 2006). Chronic pain can cause a lot of emotional distress and affect lifestyle by interrupting activities (van den Berg-Emons et al., 2007) thereby it can potentially lower a person’s mood. Low mood and low self esteem often give birth to mental disorders like depression. (Fordyce, 1976) and (Sternbach, 1974) noted that depression is a frequent accompaniment to chronic pain while, (Von Knorring et al., 1983) observed that those who suffer from depression often complain of pain. Depression, which is commonly associated with chronic pain (Fishbain et al., 1997; Zis et al., 2017), is one of the leading contributors to global disease burden (Whiteford et al., 2013; Collins et al., 2011). It has been seen that chronic pain and depression tend to coexist (Romano and Turner, 1985) and the relationship between the two is widely studied. (Tang et al., 2008) showed that when a depressed mood was induced in patients with chronic back pain, their pain ratings increased, while participants with a happy mood had lower pain ratings. (Fishbain et al., 1997) observed evidence against the hypothesis of depression preceding the development of pain and indicated that pain may play a causal role for depression. Chronic pain could be due to presence of inflammatory diseases (Ji et al., 2016), which cause inflammation in the body that can produce cytokines which can lower mood (Wright et al., 2005), and according to Irwin (2002), higher levels of biomarkers associated with inflammation are linked with depression. So today, the causal relationship of these associations between inflammation and mood disorders is said to be bi-directional (Rosenblat et al., 2014; Jones et al., 2020; Lwin et al., 2020).

It is widely recognised that healthcare increasingly involves dealing with comorbidities (Gijsen et al., 2001), and also personalisation of treatment plans (Vicente et al., 2020).

Health issues such as mood disorders and conditions associated with chronic pain are often comorbid (Tunks et al., 2008; Agüera et al., 2010), but the manner in which these conditions influence each other varying from person to person is still considerably uncertain. Mood disorders or depression can be treated in three ways: antidepressants, psychotherapy and electro-convulsive therapy (ECT) (Nemeroff and Owens, 2002). Chronic pain treatments can be based on multiple aspects of pain experience like the intensity and quality of pain, and use of rescue analgesic medications (Patel et al., 2021). For certain types of chronic pain, drug therapy including intake of analgesics like non-steroidal anti-inflammatory drugs (NSAIDs) could be the option, while for others, a multimodal approach may be required (Portenoy, 2000) e.g.: a pharmacotherapy consisting analgesics and Cognitive–behavioural therapy (CBT) together can be effective when chronic pain and anxiety disorders co-occur (Asmundson and Katz, 2009). But when dealing with both mood disorders and chronic pain, especially when considering only pharmaceutical interventions, it must be noted that the combined usage of anti-depressants and NSAIDs can have negative effect, as shown in (Shin et al., 2015; Hou et al., 2021) where there observed a risk of intracranial haemorrhage although there was no such association found in independent use.

Nowadays, technology is making its presence felt in several sectors, one of which is the health sector. It is only in the early 21st century that eHealth, a broad term for the combined usage of electronic and communication technologies in the health sector, emerged (Harrison and Lee, 2006). Many novel ways have developed to tackle healthcare issues and provide support. From wearable accessories to smartphone applications, all of these are aiding healthcare. From a global perspective, e-health is useful in dissemination of health information as well as ensuring that the most updated information is used to improve the health (Kwankam, 2004; Kendall et al., 2020). The WHO's Global Observatory for eHealth defines mobile health (mHealth) as "medical and public health practice supported by mobile devices, such as mobile phones, patient monitoring devices, personal digital assistants (PDAs), and other wireless devices". mHealth is a powerful way to cater to individual requirements. Few of the benefits of the mHealth tools, especially for the purpose of research, are: (i) cost-effectiveness while collecting voluminous amount of data; (ii) more honesty in answers received as there is no direct human intervention in collection of data; and (iii) convenience of easily linking mHealth apps to other link to other sensing tools. More than one in four people are affected by mental health disorders like depression, anxiety etc. worldwide (Ginn and Horder, 2012), and digital technology interventions show the potential to extend support to those who suffer from mental health problems. There is a growing need to make digital based mental health care aid accessible to as many people as possible (Naslund et al., 2017), and in this study, we make use of digital health data to analyse mood-pain patterns in a cohort of residents of the UK with chronic pain conditions.

We explore the association between pain and mood by analysing long records of self-reported, daily data collected using a mobile phone application. We perform clustering on the basis of the transitions of mood-pain and show how an intervention to improve low mood or high pain symptoms can affect the clusters differently.

3.2 Methods

3.2.1 Data

We use data from the Cloudy with a Chance of Pain study (Reade et al., 2017; Dixon et al., 2019), which was conducted to investigate the relationship between weather and pain, but in doing so created an extremely rich dataset suitable to answer a diversity of research questions. Data were collected from January 2016 to April 2017 from participants resident in the UK who were aged 17 or above and had experienced chronic pain for at least 3 months preceding the survey (Druce et al., 2017).

The cohort had 10,584 survey participants, each of whom was asked to rate their symptoms and other variables on a mobile application in five ordinal categories (e.g. pain scores ranged from 1 for no pain to 5 for very severe pain). Data were recorded for pain interference, sleep quality, time spent outside, tiredness, activity, mood, well-being, pain severity, fatigue severity and stiffness on a daily basis. However, participants did not always report all the data daily so we considered only those (Mood, Pain) states where both the values are available, leaving us with $N = 9990$ participants for our analysis.

In this paper we analyse trajectories of pairs of self-reported pain severity and mood scores. Participants were asked to provide information on these on a five-point Likert scale, with accompanying text for each of the ordinal levels. For mood, a score of 1 represents worst mood and 5 represents best, whereas for pain a score of 1 represents least pain and 5 represents most.

For easier analysis of the data and interpretation of results, we regrouped the severity of mood and pain into two categories each on the basis of the descriptions associated with each ordinal value. Mood scores of 1–3 and 4–5 were labelled Bad (B) and Good (G) respectively, while pain levels of 1–2 and 3–5 were, respectively, labelled Low (L) and High (H). Thus, at a given time, a participant's mood and pain scores fall into one of four states: GL; GH; BL; and BH. Full details are shown in Table 3.1

Mood			**Pain**		
Score	*Description*	*Binary*	*Score*	*Description*	*Binary*
1	Depressed	Bad	5	Very severe pain	High
2	Feeling low	Bad	4	Severe pain	High
3	Not very happy	Bad	3	Moderate pain	High
4	Quite happy	Good	2	Low pain	Low
5	Very happy	Good	1	No pain	Low

Table 3.1: Mood and pain scores, descriptions, and binary classifications. *Score* is the value on a Likert scale available to participants, *Text* is the description presented to them when recording these data, and *Binary* is our binary classification into 'Good' (G) or 'Bad' (B) for Mood, and 'Low' (L) and 'High' (H) for Pain.

Participants self-reported diagnoses, and also provided information on age, sex, pain condition diagnosed and the site of pain. They might have more than one condition and site of pain. The list of conditions includes Rheumatoid arthritis, Osteoarthritis, Spondyloarthropathy, Gout, Unspecific arthritis, Fibromyalgia, Chronic headache and Neuropathic pain. The list of sites of pain taken in this analysis includes mouth or jaw, neck or shoulder, back pain, stomach or abdominal, hip pain, knee pain, and hands.

Code for this study is made available at: `https://github.com/rajenkidas/EM-clustering-on-Markov-Chains`. The data is scheduled to be made available to the wider research community via a trusted research environment in 2023.

3.2.2 Residual analysis

We performed an initial data analysis based on Pearson residuals, looking for notable patterns in the co-evolution of mood and pain over time using standard methodology as outlined by e.g. Bishop et al. (1975). Such an analysis particularly helps to visualise the ways in which observed patterns deviate from a simple 'null' model.

We begin by visualising a matrix of transitions observed in the data. Let $\mathbf{Y}$ be the count matrix whose element Y_{ij} denotes the total number of observed transitions—across all participants—from state i one reporting day to state j the next reporting day. We then perform Pearson residual analysis to compare observed transition probabilities with the expected values given a specified 'null' model assumption, which we fit by maximum likelihood estimation. Throughout this work we will use the standard result that the maximum likelihood estimator for a probability of an outcome is the observed number of such outcomes divided by the number of observations under binomial and Poisson sampling (which we also assume throughout as appropriate).

We have seen that participants are most likely to remain in their current state rather

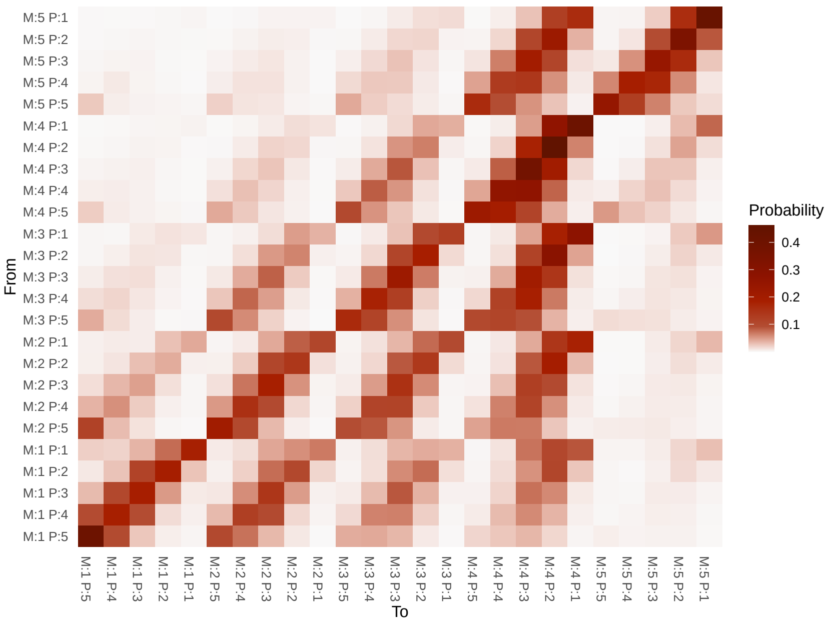


Figure 3.1: Transition probability matrix: Heatmap showing probabilities of transitions from one state to another.

than move to another one. That is, their mood and pain scores do not usually change from one day to the next, as shown in Figure 3.1. These observations allow us to define a simple first model for their behaviour and perform residual analyses as described below. In this exploratory analysis we work with the original data, so there are $n = 5 \times 5 = 25$ states.

We therefore define a null model in which the number of participants starting in state i is N_i, the probability of staying in state i is π_i and when a person does change state, the probabilities P_{ij} of a transition from state i to state $j \neq i$ are uniform. The model parameters can then have maximum likelihood estimators (indicated with hats) as follows. For $i, j \in \{1, 2, \ldots, n\}$,

$$\hat{N}_i = \sum_{k=1}^{n} Y_{ik}, \quad \hat{\pi}_i = \frac{Y_{ii}}{\sum_{k=1}^{n} Y_{ik}}, \quad \hat{P}_{ij} = \begin{cases} \hat{\pi}_i & \text{if } i = j, \\ \dfrac{1-\hat{\pi}_i}{n-1} & \text{otherwise,} \end{cases} \quad E_{ij} = \hat{N}_i \hat{P}_{ij}, \tag{3.1}$$

where E_{ij} is the (i, j)-th element of the matrix of expected counts, $\mathbf{E}$. The associated entry in the Pearson residual matrix $\mathbf{R}$ is then given by

$$R_{ij} = \frac{Y_{ij} - E_{ij}}{\sqrt{E_{ij}}}. \tag{3.2}$$

Since we expect such residuals to be asymptotically standard normal under the null (Bishop et al., 1975), we will interpret these as values over 2 indicating significantly more events than expected under the null, and values under -2 indicating significantly fewer.

3.2.3 Clustering analysis

In this section we outline methods used to classify the participants using unsupervised learning, organising the participants into clusters on the basis of their sequences of reduced (Mood, Pain) states: GL, GH, BL, BH.

Model setup

We assume the sequence of self-reported mood-pain states $X = (X_t; t \geq 0)$ is generated by a Markov chain:

$$\Pr(X_{t+1} = j \mid X_0 = k_0, ..., X_t = i) = \Pr(X_{t+1} = j \mid X_t = i) =: P_{ij}, \tag{3.3}$$

where the P_{ij} are called the chain's *transition probabilities*.

Our data consists of trajectories of mood-pain pairs that we reduce to matrices tabulating numbers of transitions observed for each participant individually. We then cluster these count-matrices by using the EM algorithm to fit a mixture of Markov chains with a distinct matrix of transition probabilities for each component of the cluster.

Let the number of states be n and the number of participants be S. We write $\mathbf{C}$ for the matrix of total count of transitions from one state to another, and use $\mathbf{C}_s$ for the matrix of counts of transitions that appear in the trajectory of states of mood-pain of participant s. We note that $\mathbf{C}$ is distinguished from the count matrix $\mathbf{Y}$ introduced in Section 3.2.2 since it involves only the four reduced states.

The expectation-maximisation algorithm

The classical Expectation-Maximisation (EM) algorithm (Dempster et al., 1977) provides a way to do maximum-likelihood estimation of parameters in a setting where some variables are unobserved or unknown. In our case, the latent variables are the classes to which the participants belong. The algorithm involves iteration of two alternating steps: the E, or *expectation* step, during which one computes the expected value of the log likelihood for the observed data, given the current estimates of the parameters, and the M, or *maximisation*, step during which one re-estimates the parameters is maximising the expected value as calculated in the E-step.

The details of this algorithm are given in Supplementary Material §3.5.3. It gives an $S \times K$ matrix $\mathbf{\Gamma}$ such that its (s, c)-th element Γ_{sc} is the probability that participant s belongs to cluster c. Finally, cluster assignments are then made on the basis of the class membership probabilities: participants are assigned to whichever cluster they have the highest probability of belonging to.

Associated stationary distribution

The stationary distribution for a Markov chain with $n \times n$ transition matrix $\mathbf{M}$ has probability x_i associated with state i, where $\mathbf{x} = (x_i)$ solves the left Eigenvalue equation

$$x_k = \sum_{i=1}^{n} x_i \, M_{ik}, \tag{3.4}$$

where $k \in \{1, \ldots, n\}$, and we impose conditions ensuring that $\mathbf{x}$ is a probability vector: $x_i \geq 0$ and $\sum_{i=1}^{n} x_i = 1$.

The solution to Eqn. (3.4) need not be unique, but as the transition matrices of our problem are regular, we do get a unique stationary distribution for each component of the mixture (Stirzaker, 2003). That is, for each cluster, we get a distribution over the states BH, BL, GH and GL. Further, as the Markov chains are ergodic, the modelled expected fraction of time an individual participant spends in state k is given by x_k.

3.2.4 Intervention

In this section, we explore the prospect of alleviating low mood or high pain, which can be done by taking the appropriate treatment targeting mood or pain. We naïvely examine how the interventions could work by altering the transition probabilities associated with the clusters and see what effect this has on the cluster's stationary distribution. Throughout, we will let the transition probability matrix before intervention be represented as:

$$\begin{array}{c} \\ \text{BH} \\ \text{BL} \\ \text{GH} \\ \text{GL} \end{array} \begin{array}{c} \begin{array}{cccc} \text{GL} & \text{GH} & \text{BL} & \text{BH} \end{array} \\ \begin{bmatrix} M_{11} & M_{12} & M_{13} & M_{14} \\ M_{21} & M_{22} & M_{23} & M_{24} \\ M_{31} & M_{32} & M_{33} & M_{34} \\ M_{41} & M_{42} & M_{43} & M_{44} \end{bmatrix} \end{array}. \tag{3.5}$$

Improving mood

To model an improvement in mood, we increase the probabilities of transitions from states of bad mood to those with good mood. We get an updated transition matrix $\mathbf{M}'_c$

for every cluster c in the following way:

$$\begin{array}{c|cccc|} & \text{GL} & \text{GH} & \text{BL} & \text{BH} \\ \text{BH} & M_{11}+\beta_M & M_{12}+\beta_M & 0.8\times(M_{13}+M_{14}-2\beta_M) & 0.2\times(M_{13}+M_{14}-2\beta_M) \\ \text{BL} & M_{21}+\beta_M & M_{22}+\beta_M & 0.8\times(M_{23}+M_{24}-2\beta_M) & 0.2\times(M_{23}+M_{24}-2\beta_M) \\ \text{GH} & M_{31} & M_{32} & M_{33} & M_{34} \\ \text{GL} & M_{41} & M_{42} & M_{43} & M_{44} \end{array}, \tag{3.6}$$

where the rows are labelled by the (Mood, Pain) states from which the transition starts, while the columns are labelled by the states to which it goes. Here β_M must be chosen so that all transition probabilities remain in the range $0 \leq M'_{cij} \leq 1$. For our fitted transition matrices, these constraints mean that $0 \leq \beta_M \leq 0.15$.

One can see that we distribute the probabilities disproportionately between transitions to BH and BL from BH and BL. This has been done to reduce the probability of moving to BL, which we wish to model as less likely under an intervention assumed to be beneficial. In fact, in general the probability of moving to good mood from bad mood could have been achieved in numerous other ways through changes to the full matrices. The choice used here permits a more substantial increase in the probabilities of improved mood than simpler formulæ, many of which are strongly constrained by the necessity of keeping all probabilities to the laws of probability.

Improving pain

Similar to improvement of mood, we considered altering the transition probabilities to improve pain, which means increasing probability of transitioning to low pain through adding and subtracting β_P as shown below for the updated transition probability matrix $\mathbf{M}'_c$ for every cluster c in the following way:

$$\begin{array}{c|cccc|} & \text{GL} & \text{GH} & \text{BL} & \text{BH} \\ \text{BH} & M_{11}+\beta_P & 0.8\times(M_{12}+M_{14}-2\beta_P) & M_{13}+\beta_P & 0.2\times(M_{12}+M_{14}-2\beta_M) \\ \text{BL} & M_{21} & M_{22} & M_{23} & M_{24} \\ \text{GH} & M_{31}+\beta_P & 0.8\times(M_{32}+M_{34}-2\beta_M) & M_{33}+\beta_P & 0.8\times(M_{32}+M_{34}-2\beta_M) \\ \text{GL} & M_{41} & M_{42} & M_{43} & M_{44} \end{array}. \tag{3.7}$$

Here $0 \leq \beta_P \leq 0.2$. In both cases, we then examine the resulting changes in the stationary distributions to see the consequences of the intervention for each cluster individually.

3.3 Results

3.3.1 Residual analysis

The resulting transition probability matrix is illustrated in Figure 3.1, which is a heatmap illustrating the probabilities with which participants switch from one pair of mood-pain scores to another. It is based on the original data and so has $5 \times 5 = 25$ possible states and $25 \times 25 = 625$ possible transitions. It has rows labelled by a current mood-pain pair and columns labelled by the mood-pain pair on the following day.

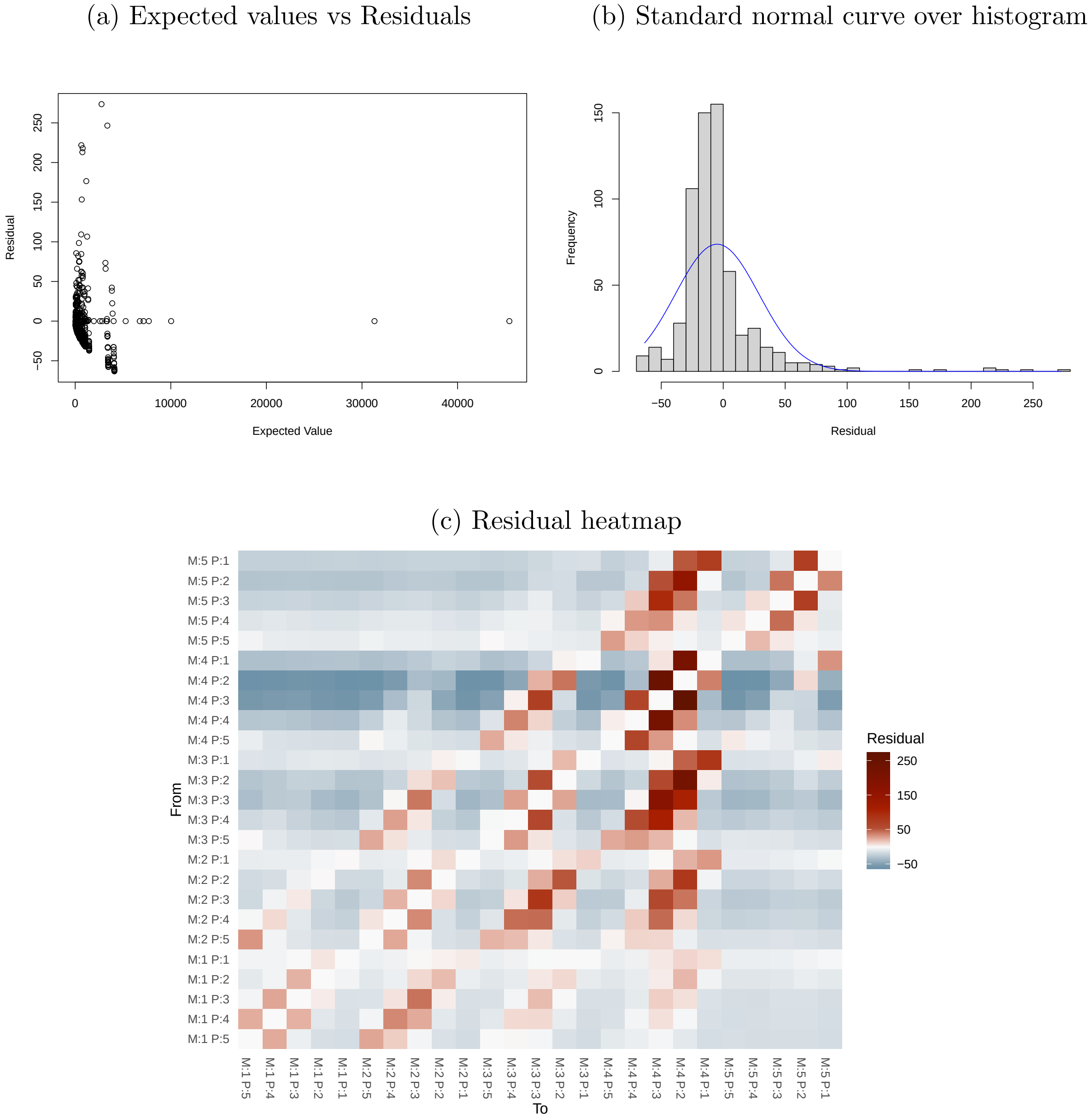
Note that the diagonal elements—those that correspond to remaining in the same state on successive days—have high probabilities. The entries at upper right and lower left, which correspond, respectively, to the worst and best mood-pain scores, are especially large (near their maximum value, 1) indicating that participants at the extremes of the scale have a strong tendency to remain there.

Figures 3.2a and 3.2b, which illustrate the distribution of residuals for this model as computed with Eqn. (3.2), clearly show that the residuals do not appear to be normally distributed. Looking at the residual heatmap in 3.2c, we can say that the naïve model specified by Eqn. (3.1) does not describe the data well.

This suggests we try another model or check for latent variables or clusters. We try another model in the Supplementary (3.12) which showed an improvement in fitting since the residual range decreases in 3.15, but it still did not fit the data well as we see in 3.16. So we move on to clustering the data, as explained in the next section.

3.3.2 Clustering

We found four clusters using the EM algorithm to do model-based clustering using a mixture of Markov chains, as illustrated in Figure 3.3, where the clusters are represented by heatmaps of their transition matrices.


(a) Expected values vs Residuals

(b) Standard normal curve over histogram

(c) Residual heatmap

Figure 3.2: **A** is the scatter plot of expected values and the residuals. **B** shows a histogram of the residuals as well as a blue curve giving the probability density function of a normal distribution having the same mean and variance as the residuals. **C** is a heatmap of the matrix of residuals based on the model specified by Eqn. (3.1).

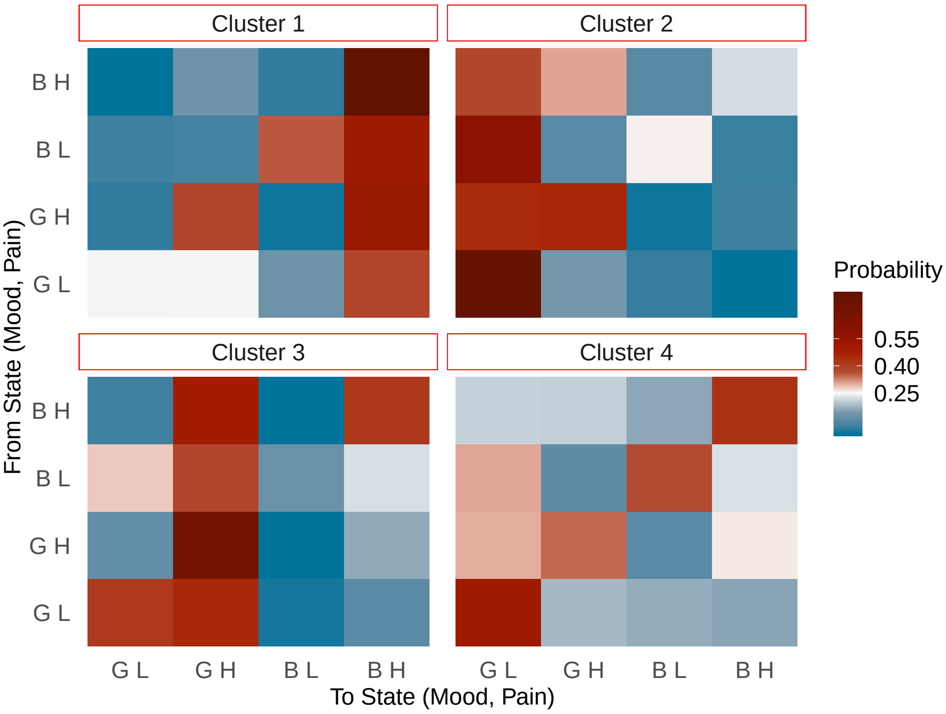


Figure 3.3: Heatmaps of the transition probability matrices for the four clusters where G, B, L and H imply good mood, bad mood, low pain and high pain respectively.

Before describing the clusters, it should be noted that GL is the best state as both mood and pain are good, while BH is the least preferable state to be in as both mood and pain are bad here. Based on the transition probabilities, the four clusters for mood-pain dynamics can be broadly characterised as:

Cluster 1: Movement to the least preferable state. 1783 members.

Here, we see that there are high probabilities of moving to the state where there is bad mood and high pain.

Cluster 2: Movement to the ideal state. 1558 members.

In this cluster, we observe, irrespective of the current state, a participant is most likely to be in good mood and low pain the next day.

Cluster 3: Good mood, high pain. 2019 members.

In this cluster, the dominant movement is to the state with good mood and high pain.

Cluster 4: Remain in the same state. 4630 members.

Most of the participants tend to stay in the same state.

Given the total of 9990 participants, we see that it is most common for participants (46%) to be members of Cluster 4 involving staying in the same state, which is consistent with our exploratory analysis of transitions. The smallest cluster (number 2) with 16%

of participants, consists of those who tend to the ideal state, but at the same time, not many (18%) are in Cluster number 1 that tends to the worst state. The remainder (20%) belong to the third cluster: good mood, high pain.

In Figure 3.6, we present a set of comparisons of properties of the clusters. The stationary distributions as defined by Eqn. (3.4) are shown in Figure 3.4, and are as would be expected from the full estimated transition probability estimates they are derived from: Cluster 1 has most probability mass on BH; Cluster 2 has most probability mass on GL; Cluster 3 has most probability mass on GH; and Cluster 4 has evenly distributed probability masses.

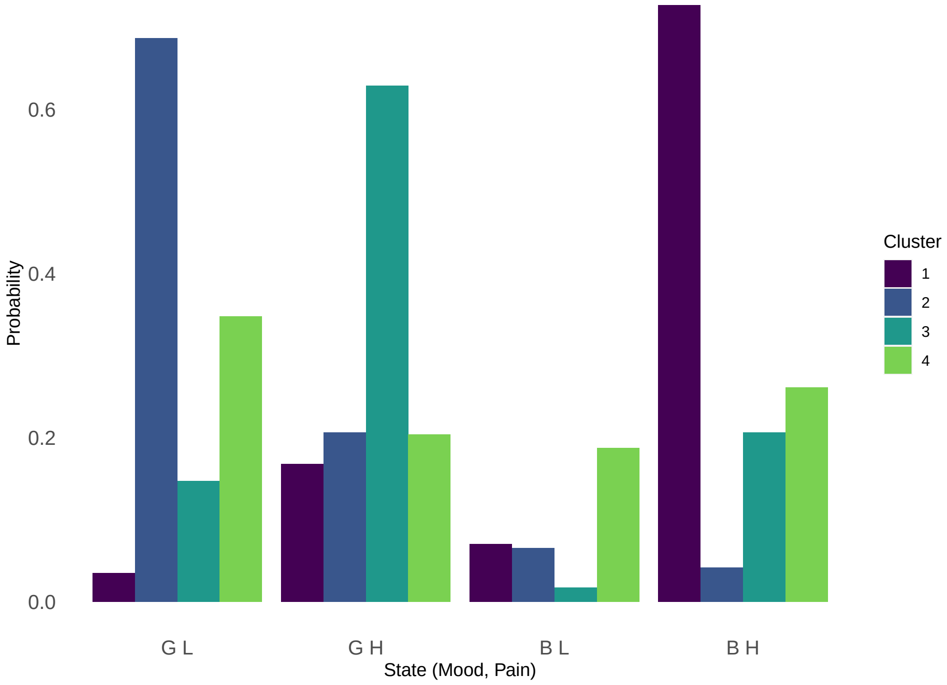


Figure 3.4: Stationary distribution for the four clusters

In Figure 3.5, we compare age distributions by sex and cluster, seeing that Clusters 1 and 4 have comparable age distributions, but Cluster 3 is associated with older ages than these two and Cluster 2 is associated with older ages than all three other clusters. Males are typically older than females in all clusters.

Participants had one or more conditions and sites of pain, with log odds ratios for these by cluster shown in Figures 3.6a and 3.6b. These show that while some conditions and sites such as gout and hands are not strongly associated with any cluster, for others this is not the case. Fibromyalgia and stomach pain are particularly strongly associated with Cluster 2, for example.

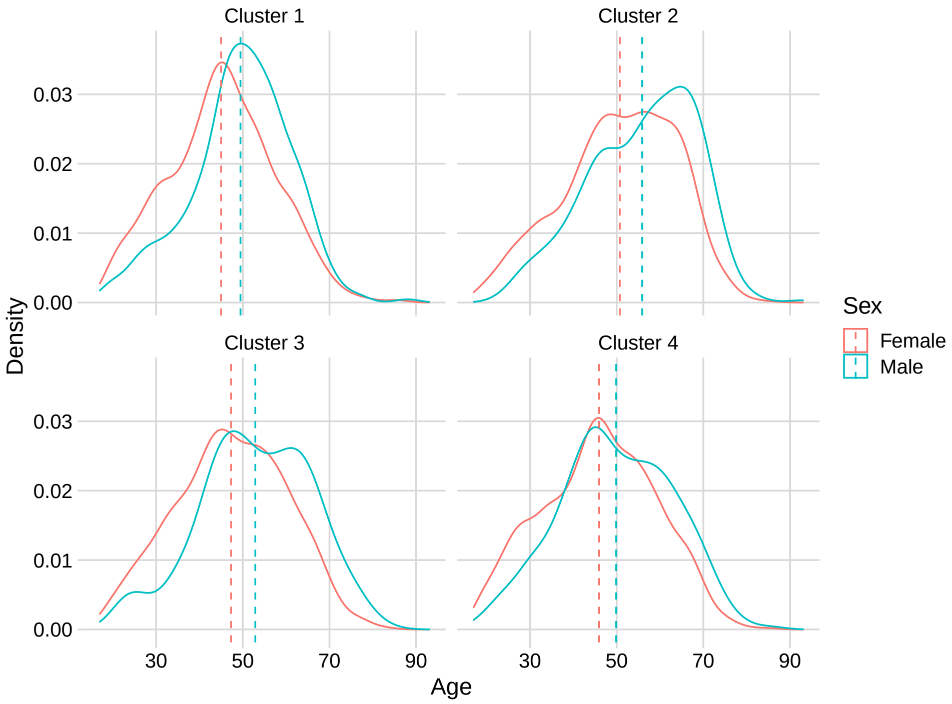


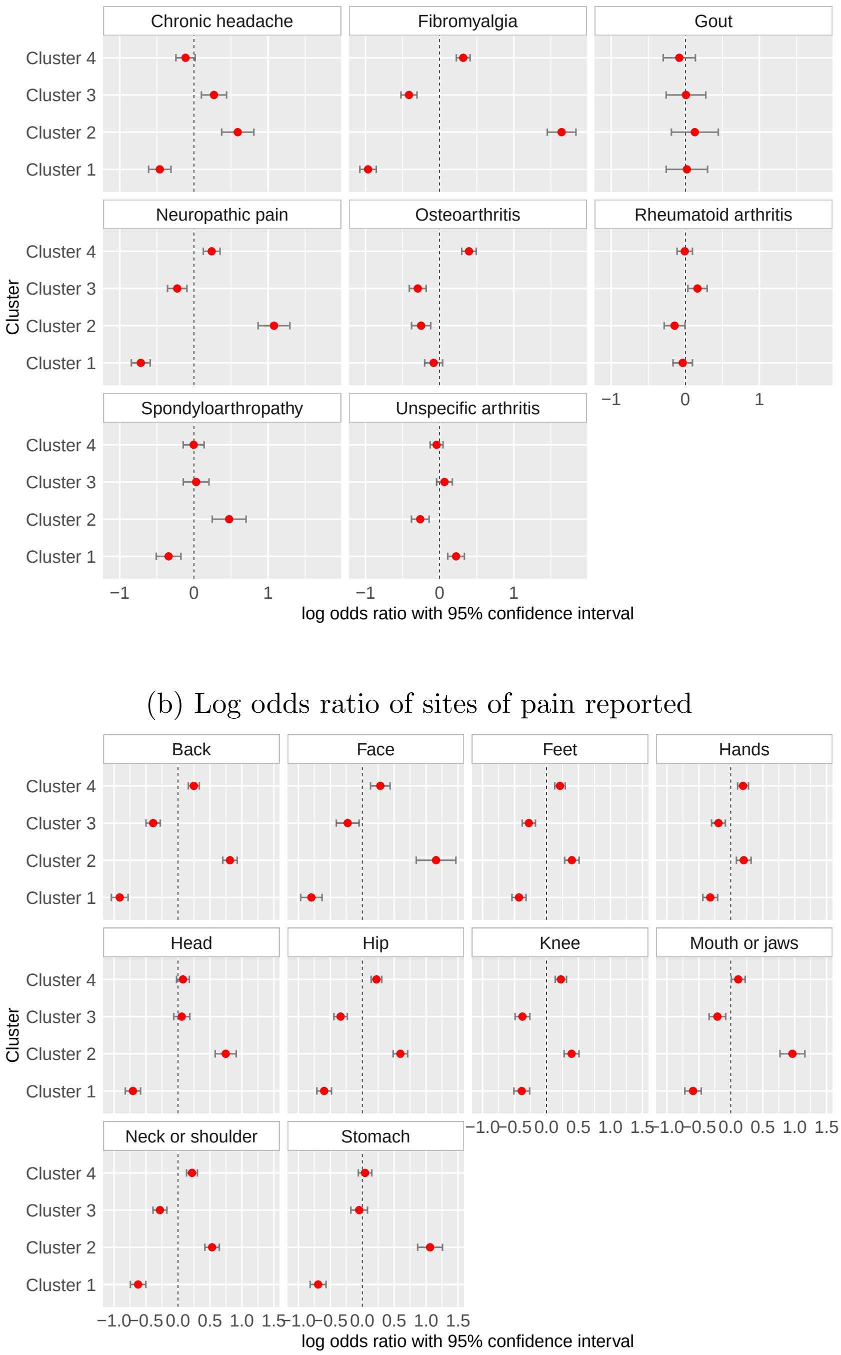
Figure 3.5: Age distribution amongst the clusters

3.3.3 Intervention

We look at how interventions could work help alleviate the symptoms of bad mood and high pain.

In Figure 3.7a, Cluster 2 shows least improvement in mood, while Cluster 1 shows the most followed by Cluster 4. Decrease in state BH is the highest for Cluster 1, followed by Cluster 4 and least for Cluster 2. Overall, Cluster 1 shoes the most drastic changes in probability distribution while Cluster 2 is the least. We also note that in the case of improving mood from bad mood, state BL probability drops for Clusters 2 and 4, while it increases for 1 and 3.

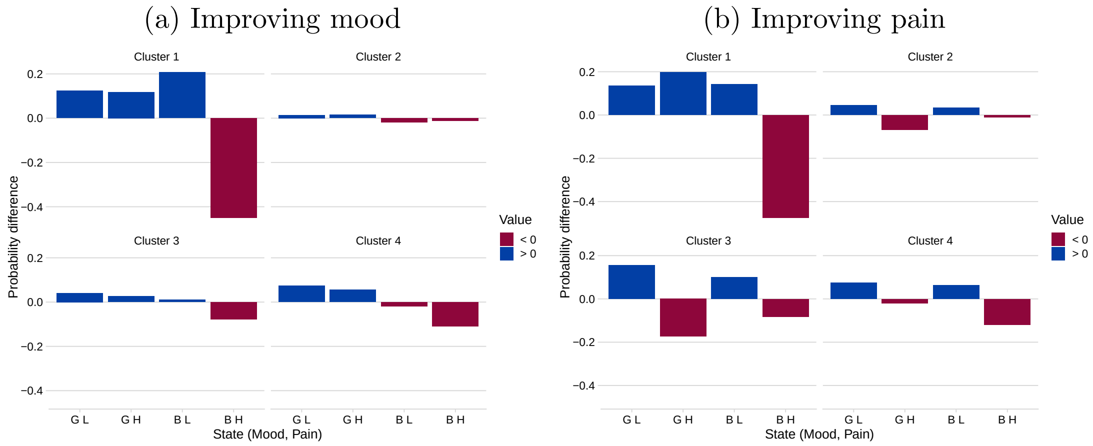
In Figure 3.7b, we again find Cluster 1 with maximum changes. When intervened to improve pain by lessening the intensity of pain, probability of GH state improves only for Cluster 1.



(a) Log odds ratio of conditions reported

(b) Log odds ratio of sites of pain reported

Figure 3.6: **A** and **B** represent log odds ratio of condition and site of pain respectively, per cluster.


(a) Improving mood

(b) Improving pain

3.4 Discussion

In this work, we have performed an analysis of joint trajectories of mood and pain of participants in the large mobile health cohort, “Cloudy with a Chance of Pain”. In addition to analysis of the full set of transitions using residuals, we performed clustering on transitions between a simplified set of variables and in doing so found four digital behavioural phenotypes on the basis of people’s past trajectories of their mood-pain states. This suggests that even though mood and pain have been known to be correlated, the association may not be generalised in one single way for an entire population.

Previous studies on mood-pain relationships have tended to reach the conclusions on universal associations between mood and pain – i.e. generalising the result for everyone. The clusters found in this study emphasise that mood-pain relationships may differ between (groups of) individuals. The varying relationships between mood and pain, as shown by the clusters, highlights that such variability should be taken into account when considering expected future associations - for example, in a clinical prediction model, an approach of personalising forecasts could be taken.

Going beyond association to look at mechanism and causation, we stress that we have not performed causal inference and so results should all be interpreted as indicative of (potential) association magnitudes rather than as causal statements. Nevertheless, the interpretability of the observed clusters and their diversity in terms of e.g. conditions and sites of pain represented suggests that there may be associated endotypes – i.e. clusters representing distinct mechanisms of disease. If such causally distinct groups exist, then our hypothetical investigation of interventions that target either mood or pain individually suggests that we might expect clinically significant differences from different treatment depending on an individual’s endotype.

Our study has some limitations that should be borne in mind when interpreting results. The first of these is, as discussed above, that we consider associations rather than causation. Furthermore, we have assumed missing values – primarily arising when participants did not enter data on one day – can be ignored and so have removed them; although this is not a major component of the data an alternative would be to model non-response as a separate value. Along related lines, the simplification of the state space, while necessary for the EM algorithm to produce plausible transition matrices for each cluster, involves some information loss and this leaves open the possibility of more sophisticated methodology to perform the clustering. Also, factors common to all observational studies such as this one are important to bear in mind, particularly that individuals are selected from the general rather than a clinical population.

Extension of the work presented here could include applying the same methodology to more datasets to check if the phenotypes found are reproducible. This would further strengthen the likelihood of different causal relationships holding within clusters. To make a fuller assessment of likely causation, however, expected relationships between all observed and unobserved variables would need to be specified, and ideally intervention studies run. Additionally, this work can be extended by including socio-economic factors, extra latent variables like sleep quality, environment etc. Another direction would be to apply different techniques to this dataset, such as linear model based approaches that can identify latent classes (Proust-Lima et al., 2015; Komárek and Komárková, 2013). Different methods may allow the Markovian assumption made in our work to be relaxed, allowing for e.g. consideration of patterns in longer sequences of data, but at the cost of the ability to model out of sample behaviour as Markov chains allow.

Ultimately, our hope is that work on observational data such as that presented here can aid with hypothesis generation for future clinical studies of more personalised interventions for common problems such as low mood and chronic pain.

Funding

RD and TH are supported by the Engineering and Physical Sciences Research Council. ML, JMcB and BBY are supported by Centre for Epidemiology Versus Arthritis. TH is also supported by the Royal Society, the Medical Research Council and the Alan Turing Institute for Data Science and Artificial Intelligence.

Data availability statement

Our work involved secondary analysis of data from the project Cloudy with a Chance of Pain (see public website at `https://www.cloudywithachanceofpain.com`). The data is in the process of being made available to researchers outside the University of Manchester via a trusted research environment. Please contact Professor Will Dixon [will.dixon@manchester.ac.uk] to enquire about access.

Acknowledgement

We thank Dr Elaine Mackey and Prof Will Dixon for granting access to the data.

Ethics statement

This is a secondary analysis. Ethical approval of the primary data was obtained from the University of Manchester Research Ethics Committee (ref: ethics/15522) and from the NHS IRAS (ref: 23/NW/0716).

Conflicts of interest

None declared.

Author contributions

RD performed the analysis. All authors contributed to the formulation of research questions and the writing of the paper.

Supplementary Material

3.5 Supplementary text

3.5.1 Summary statistics of data and results

Main data

Here we provide an overview of the data. Both mood and pain were rated on a five point scale where 1 for mood is the worst score while 1 for pain is the best. Similarly 5 means the mood is the best and pain is at its worst. Mean values of mood and pain are 3.6 and 2.7 respectively. Number of NA's: 344784 in mood and 349760 in pain. After removing these NA's from the data, we are left with 9990 participants instead of 10584. More details about the data can be found in `www.cloudywithachanceofpain.com`.

With five possibilities of each of mood and pain, there are 5 × 5 *i.e.* 25 possible (mood, pain) pairs, which would become the states of our Markov processes, leading to transition matrices with $25 \times 25 = 625$ entries. But instead, we reduce the total number of states by regrouping the scores of mood and pain into good and bad, and low and high categories respectively and then taking pairs of these regrouped scores. For Mood, the Bad (B) scores are {1,2,3} while the Good (G) ones are {4,5}. For Pain, Low (L) is {1,2} while High (H) is {3,4,5}.

For Mood, number of B's and G's are, respectively, 153922 and 288145. For Pain the number of H's and L's are, respectively, 245344 and 196723. The frequencies with which the four (Mood, Pain) states are observed is shown in Figure 3.8. The frequencies for BH, BL, GH and GL are 113632, 40290, 131712 and 156433, respectively.

Figure 3.9 gives the overall age distribution of the cohort. It shows that the mean age for the women and men are approximately 47 years and 52 years respectively.

Tables 3.2 and 3.3 give the characteristics of the study participants. Total number of participants is 9990.

Clustered data

We have also included the clustered heatmaps without any regrouping of the categorical variables in Figure 3.19. Note that the probability of moving to a state gets reduced from 0.25 (in case of grouped data) to 0.04 which brings problems of interpretability and generalisations which have been talked about before. But it can help in understanding the contribution of a transition probability to the grouped clusters, which even 3.18 helps. We have chosen not to discuss the granular details with 25-states and focus on the clustering with respect to the reduced states which can help in easy comparison with similar studies performed or yet to be performed.

We see that within cluster composition of the conditions and sites of pain are both more or less the same across the clusters as shown in Figures 3.10a and 3.10b. The biggest difference which can be immediately noted is that in the composition of conditions in cluster 2, Fibromyalgia and Neuropathic pain are less while unspecified arthritis is comparatively more. Next, when we look at how much of a cluster constitutes a condition, we find the proportions are in sync with the in general sizes of the clusters. Noticeable difference in 3.10c is that cluster 2 amounts to very little proportion of Fibromyalgia, while for site of pain, we find in 3.10d, cluster 2 constitutes very little of site of pain as the face, while cluster 1 takes up a high proportion compared to the rest of its contributing proportions.

Table 3.4 gives the mean age in years and the response rate in percentage per cluster. Mean age was calculated by taking the average of age in a group, after removing all the NA values. Response rate is the percentage of participants in a cluster who gave their date of birth details.

3.5.2 Definition of the log odds ratio

Tables 3.5 and 3.6 give the log odds ratio of a condition and site of pain respectively in a cluster compared to the other clusters. To calculate log odds ratio, we use Multinomial Logistic Regression and build the following contingency table first:

	Cluster	Remaining clusters
Condition	n_{11}	n_{12}
Remaining conditions	n_{21}	n_{22}

Explicitly:

n_{11} is the number of participants with a specific condition in a cluster.
n_{12} is the number of participants with the specific condition not in the cluster.
n_{21} is the number of participants without the specific condition in the cluster.
n_{22} is the number of participants without the specific condition not in the cluster.

The log odds ratio is then given by:

$$L = \log\left(\frac{n_{11}n_{22}}{n_{12}n_{21}}\right) = \log(n_{11}) + \log(n_{22}) - \log(n_{12}) - \log(n_{21})$$

The standard error of this quantity is asymptotically equal to

$$\sigma = \sqrt{n_{11}{}^{-1} + n_{12}{}^{-1} + n_{21}{}^{-1} + n_{22}{}^{-1})},$$

as shown in Bishop et al. (1975). Therefore, the 95% Confidence Interval is approximately $L \pm 1.96\,\sigma$.

Similarly we calculate for site of pain by building the following contingency table:

	Cluster	Remaining clusters
Site of pain	n_{11}	n_{12}
Remaining sites of pain	n_{21}	n_{22}

Explicitly:

n_{11} is the number of participants with a specific site of pain in a cluster.
n_{12} is the number of participants with the specific site of pain not in the cluster.
n_{21} is the number of participants without the specific site of pain in the cluster.
n_{22} is the number of participants without the specific site of pain not in the cluster.

And we can then calculate a log odds ratio as for conditions. In general, a positive log odds ratio indicates that the site or condition is more commonly in a cluster, and a negative that it is less commonly so.

3.5.3 Description of the EM algorithm

The matrix $\mathbf{\Gamma}$ is initialised randomly with probabilities chosen such that every row sums to 1. The mixture of Markov chains is then specified by a weight vector $\boldsymbol{\omega}$ of length K in which

$$\hat{\omega}_k = \frac{\sum_{s=1}^{S} \Gamma_{sk}}{S} \tag{3.8}$$

Using the mixture weights and count matrices, we then estimate the parameters of the per-cluster Markov chains which are transition probability matrices $\mathbf{M}$. For cluster k, the estimate for i to j transition is given by

$$\hat{M}_{kij} = \frac{\sum_{s=1}^{S} \Gamma_{sk}\, C_{sij}}{\sum_{k=1}^{K} \sum_{s=1}^{S} \Gamma_{sk} C_{sij}}. \tag{3.9}$$

The rows of the a participant's count matrix are taken to follow a Multinomial distribution and so define an $S \times K$ matrix of expected likelihoods whose entry Λ_{sk} gives the likelihood of observing participant s's trajectory given that they participant s belongs to cluster k. It is given by

$$\Lambda_{sk} = \prod_{i,j=1}^{n} M_{kij}^{C_{sij}}$$

where we have suppressed a multinomial coefficient that does not depend on the parameters of the mixture model and so does not affect maximum-likelihood estimates. The log-likelihood for participant s and cluster k is thus given by,

$$\log \Lambda_{sk} = \sum_{i,j=1}^{n} C_{sij} \log\left(M_{kij}\right). \tag{3.10}$$

Using Eqn. (3.10), we can specify the algorithm steps as follows:

- **Expectation step:** The expected values of the matrix of class membership probabilities are computed using

$$\widehat{\Gamma}_{sk} = \frac{\omega_k \, \Lambda_{sk}}{\sum_{c=1}^{K} \omega_c \, \Lambda_{sc}}$$

- **Maximisation step:** This involves re-estimating the parameters of the mixture using Eqns. (3.8) and (3.9).

One performs the steps in alternation until the matrix $\mathbf{\Gamma}$ converges. That is, one keeps track of the two most recent estimates of $\mathbf{\Gamma}$ — call them $\widehat{\mathbf{\Gamma}}$ and $\widehat{\mathbf{\Gamma}}'$ — and continues iterating until a convergence criterion such as

$$|\widehat{\mathbf{\Gamma}} - \widehat{\mathbf{\Gamma}}'| < \epsilon,$$

for some sufficiently small ϵ is met. Straightforward arguments by Borman (2009) establish that every cycle of this algorithm increases the likelihood thereby the log likelihood as defined in Eqn. (3.11).

3.5.4 Choosing the number of clusters

To find the total log-likelihood of the observed data, we made use of the log-likelihood per participant per cluster $\log(\Lambda_{sk})$ as found in Eqn. (3.10):

$$\sum_{s=1}^{S} \log\left(\sum_{k=1}^{K} \omega_k \Lambda_{sk}\right), \tag{3.11}$$

where s denotes the participant and k ranges over the clusters.

Figure 3.11 shows the negative log-likelihood as a function of the number of clusters. We see a massive drop between $K = 1$ and $K = 2$ which suggests that the participants may fall into two or more clusters . Further we see relatively big decreases as we increase the number of components to $K = 3$ and $K = 4$. The curve continues falling after $K = 4$, but as the decreases in negative log-likelihood are modest, we decided to work with 4 clusters. Additionally, we notice that the curve seems to flatten between 4 and 5. We have also plotted differences between consecutive negative Log-Likelihoods in Figure 3.12 to show how the negative Log-Likelihoods vary across number of clusters. We can see that the differences show a dropping trend with more number of clusters. Taking number of clusters to be 4 seems to be sufficient to describe the data.

Now we talk about another model selection criterion which was considered during this study. Bayesian Information Criterion (BIC) is a method to compare statistical models by calculating the information loss between the true and evaluated model by penalising the sample size to address the problem of overestimating the number of parameters (Dorea et al., 2014).

First we compute the likelihood L for the model in consideration. Then, we write BIC as

$$BIC = -2\log L + k\log(n),$$

where $\log L$ same as that given in 3.11. k is the total number of parameters and n is the number of observations.
For the problem in question, $k = K \times \text{size}(\mathbf{T}) \times (\text{size}(\mathbf{T}) - 1) + K - 1$ where K is the total number of clusters and $\mathbf{T}$ is the transition probability matrix therefore, for the 4-state Markov mixture models, we get $k = K \times 4 \times (4-1) + K - 1 = 13K - 1$. Number of observations is the total number of transitions which is 432077. We select the optimal number of clusters in the similar way as that done for negative log-likelihood above. Its plot is given in Figure 3.14. However, we don't see much difference from the previous model selection plot in Figure 3.11. It is because in comparison to the large size of the dataset, the penalty in BIC is too small and does not give criteria value much different from negative Log-Likelihood. It is common to have BIC and other criteria to keep on decreasing in case of large datasets and not be penalised much. We can see that Yin et al. (2016) used similar reasoning for model selection.

Additionally, we performed clustering into 5 to 8 components as shown in Figures 3.20, 3.21, 3.22 and 3.23, which refine the clusters but since cluster 4, identified by high probabilities along the diagonal bottom left to top right, pattern seems to exist consistently and it is the cluster with highest number of participants (46 % of population belong to this group), further breakdown of other clusters will lead to much fewer people to the other groups which we might want to avoid. So, we decided 4 to be a good selection as the clusters from the 4-component mixture had certain natural interpretations (see Section 3.4) while those from the other component mixtures did not. Taking $K = 4$ gives us quite distinct patterns while higher-cluster models have repeated patterns amongst the clusters.

3.5.5 Residual analysis: a second model

In order to fit a better model, let's us go back to the observed transition matrix plotted in Figure 3.1. The observation that the diagonal elements are high has already been incorporated into the model described by Eqns. (3.1). If we look at the heatmap more carefully, we can see that there are more dark bands indicating high probabilities in certain off-diagonal regions as well.

Let m and p denote the original mood and pain scores respectively. The states are in the pairs of form (m, p) where, $m, p \in \{1, 2, 3, 4, 5\}$. Since people tend not only to remain in the same state, but also to move a single step up or down in either mood or pain, it is interesting to extend the simple model of Eqn. (3.1) to capture these features. Let the probability that a person remains in the same state (m, p) be $\pi_{m,p}$ and the probabilities that they move to a state with $p \pm 1$ be $\pi_{m,p\pm 1}$ and probability of moving to a state with $m \pm 1$ be $\pi_{m\pm 1,p}$.

Assuming independence holds, the model is re-defined using the following distribution. $P_{(m,p),(m',p')}$ is the probability of people moving from state (m, p) on a day to (m', p') the next day. For $m, p \in \{1, 2, 3, 4, 5\}$,

$$P_{(m,p),(m',p')} = \begin{cases} \pi_{m,p} & \text{if } (m', p') = (m, p) \\ \pi_{m,p\pm 1} & \text{if } (m' = m) \text{ and } (p' = p \pm 1) \\ \pi_{m\pm 1,p} & \text{if } (m' = m \pm 1) \text{ and } (p' = p) \\ \text{uniform} & \text{otherwise} \end{cases} \tag{3.12}$$

The new model is thus that the probabilities for staying at the same state or moving to states whose mode or pain scores differ by 1 agree with those implicit, but transitions to all other states are equally likely. When we overlay a standard normal curve on the histogram of standardised residuals, as shown in Figure 3.16, we find once again that the residuals do not appear to be normally distributed.

Using the same standardised residual formula as in Eqn. (3.2), the heatmap of the residuals shown in Figure 3.15 is obtained. Comparing it with Figure 3.2, the first noticeable difference is that the range of residuals for the new model has decreased, indicating a better fit. Also, the diagonal region connecting the top left to bottom right has smoothed out a bit.

We could in principle carry on, constructing models of increasing complexity and reducing the largest residuals until those that remain have the expected, near-normal distribution. But this modelling effort was only meant to be exploratory: our main goal was the clustering analysis as discussed before.

Figure 3.16a compares the expected values and residuals obtained from Figure 3.12, and Figure 3.16b shows how the normal distribution curve fits the histogram of residuals.

3.5.6 Clusters

Transition probability matrix based on the regrouped states is given in Figure 3.17.

Before clustering, we take a look at the transition probability matrix again, but with the new states where we have regrouped the states into two categories, Good (G) and Bad (B) for mood, and Low (L) and High (H) for pain. We see trends similar to those in Figure 3.1, where the probability to remain in any given state is high. Additionally, here we can also see that probability of moving from (Mood, Pain) state (B, L) to (G, L) is high.

Once clustering is done, in Figure 3.18 we note the distribution of transitions amongst the clusters. Here, the sum of probabilities for a particular transition across clusters add up to 1.

3.5.7 Computing the shift for interventions

Here, we show how we have shifted the transition matrices to intervene with either improving mood or improving pain. Please note that this is an arbitrarily developed method to find a shift in the given scenario, as it helps in following the laws of probability. This need not be the optimal solution *i.e.* the maximum possible shift.

Step 1: We first calculate an intermediate result α by taking the maximum of the maximum of the probabilities of transitioning from bad mood to good mood over the clusters:

$$\alpha = \max\{\max_{1\leq k\leq K} \Pr_k(\text{mood tomorrow} = G \mid \text{mood today} = B)\}.$$

Step 2: Given α, we calculate $\beta = (1/2)*(1-\alpha)$. Then our new probabilities become

$$\Pr'(\text{mood} = \text{good tomorrow} \mid \text{mood} = \text{bad today}) = \\ \Pr(\text{mood} = \text{good tomorrow} \mid \text{mood} = \text{bad today}) + \beta.$$

To ensure that the probabilities to add up to 1,
Pr'(mood = bad & pain = low | mood = bad) = 0.8 × (Pr'(mood = good | mood = bad)
and
Pr'(mood = bad & pain = high | mood = bad) = 0.2 × (Pr'(mood = good | mood = bad).
We split in the ratio of 4:1 sto give lesser probability to the least-ideal state BH.

For our data, we get the approximate maximum value of β as 0.15. So we shift by 0.15

and compare with the clusters.

In a similar way, we calculated β_P.

3.6 Supplementary tables

Diagnosis	N (% approx.)
Chronic headache	1040 (10.4)
Fybromalgia	2668 (26.7)
Gout	340 (3.4)
Neuropathic pain	1519 (15.2)
Osteoarthiritis	2283 (22.8)
Rheumotoid arhiritis	1838 (18.3)
Spondyloarthropathy	865 (86.5)
Unspecified arthiritis	3418 (34.2)

Table 3.2: Conditions reported by the participants of the study.

Site of pain	N (% approx.)
Head	1963 (19.6)
Face	740 (74)
Mouth or jaws	1606 (16)
Neck or shoulder	5692 (57)
Back	5910 (59.1)
Stomach	1713 (17.1)
Hip	5160 (51.6)
Knee	6260 (62.6)
Hands	5778 (57.8)
Feet	4749 (47.5)

Table 3.3: Sites of chronic pain reported by the participants of the study

Age mean (% response rate)					
Sex	Overall	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Female	47	46 (96)	51 (97)	47 (97)	46 (96)
Male	52	50 (96)	56 (96)	53 (95)	50 (93)

Table 3.4: Rounded off values of mean age and response rate

	Cluster	Condition	Log OR	Std. Error	CI low	CI high
1	Cluster 1	Rheumatoid arthritis	-0.04	0.07	-0.17	0.10
2	Cluster 2	Rheumatoid arthritis	-0.15	0.07	-0.29	-0.00
3	Cluster 3	Rheumatoid arthritis	0.16	0.07	0.03	0.29
4	Cluster 4	Rheumatoid arthritis	-0.01	0.05	-0.11	0.10
5	Cluster 1	Osteoarthritis	-0.08	0.06	-0.20	0.04
6	Cluster 2	Osteoarthritis	-0.25	0.07	-0.38	-0.12
7	Cluster 3	Osteoarthritis	-0.30	0.06	-0.41	-0.18
8	Cluster 4	Osteoarthritis	0.40	0.05	0.30	0.49
9	Cluster 1	Spondyloarthropathy	-0.34	0.08	-0.51	-0.18
10	Cluster 2	Spondyloarthropathy	0.47	0.12	0.25	0.70
11	Cluster 3	Spondyloarthropathy	0.03	0.09	-0.14	0.20
12	Cluster 4	Spondyloarthropathy	-0.00	0.07	-0.15	0.14
13	Cluster 1	Gout	0.02	0.14	-0.26	0.30
14	Cluster 2	Gout	0.13	0.16	-0.19	0.44
15	Cluster 3	Gout	0.01	0.14	-0.26	0.28
16	Cluster 4	Gout	-0.08	0.11	-0.30	0.14
17	Cluster 1	Unspecific arthritis	0.22	0.06	0.11	0.33
18	Cluster 2	Unspecific arthritis	-0.26	0.06	-0.38	-0.14
19	Cluster 3	Unspecific arthritis	0.07	0.05	-0.04	0.17
20	Cluster 4	Unspecific arthritis	-0.04	0.04	-0.13	0.05
21	Cluster 1	Fibromyalgia	-0.97	0.06	-1.08	-0.85
22	Cluster 2	Fibromyalgia	1.64	0.10	1.45	1.84
23	Cluster 3	Fibromyalgia	-0.41	0.06	-0.52	-0.31
24	Cluster 4	Fibromyalgia	0.32	0.05	0.23	0.41
25	Cluster 1	Chronic headache	-0.46	0.08	-0.61	-0.31
26	Cluster 2	Chronic headache	0.59	0.11	0.37	0.81
27	Cluster 3	Chronic headache	0.27	0.09	0.10	0.44
28	Cluster 4	Chronic headache	-0.11	0.07	-0.24	0.02
29	Cluster 1	Neuropathic pain	-0.72	0.06	-0.84	-0.59
30	Cluster 2	Neuropathic pain	1.08	0.11	0.87	1.29
31	Cluster 3	Neuropathic pain	-0.23	0.07	-0.36	-0.09
32	Cluster 4	Neuropathic pain	0.24	0.06	0.13	0.35

Table 3.5: 8618 out of 9990 participants of the study, reported their chronic pain condition. Log odds ratio of a condition in a cluster with 95% Confidence Interval

	Cluster	Condition	Log OR	Std. Error	CI low	CI high
1	Cluster 1	Head	-0.71	0.06	-0.82	-0.59
2	Cluster 2	Head	0.75	0.08	0.58	0.91
3	Cluster 3	Head	0.06	0.06	-0.07	0.18
4	Cluster 4	Head	0.08	0.05	-0.02	0.18
5	Cluster 1	Face	-0.80	0.09	-0.96	-0.63
6	Cluster 2	Face	1.16	0.16	0.85	1.47
7	Cluster 3	Face	-0.23	0.09	-0.41	-0.05
8	Cluster 4	Face	0.28	0.08	0.13	0.44
9	Cluster 1	Mouth or jaws	-0.59	0.07	-0.72	-0.46
10	Cluster 2	Mouth or jaws	0.96	0.10	0.77	1.16
11	Cluster 3	Mouth or jaws	-0.21	0.07	-0.34	-0.08
12	Cluster 4	Mouth or jaws	0.12	0.06	0.01	0.22
13	Cluster 1	Neck or shoulder	-0.62	0.06	-0.74	-0.50
14	Cluster 2	Neck or shoulder	0.53	0.06	0.42	0.65
15	Cluster 3	Neck or shoulder	-0.28	0.06	-0.39	-0.17
16	Cluster 4	Neck or shoulder	0.22	0.04	0.13	0.30
17	Cluster 1	Back	-0.91	0.07	-1.04	-0.78
18	Cluster 2	Back	0.81	0.06	0.70	0.93
19	Cluster 3	Back	-0.39	0.06	-0.50	-0.28
20	Cluster 4	Back	0.25	0.04	0.16	0.33
21	Cluster 1	Stomach	-0.69	0.06	-0.82	-0.57
22	Cluster 2	Stomach	1.06	0.10	0.87	1.26
23	Cluster 3	Stomach	-0.05	0.07	-0.18	0.08
24	Cluster 4	Stomach	0.04	0.05	-0.06	0.15
25	Cluster 1	Hip	-0.60	0.06	-0.71	-0.48
26	Cluster 2	Hip	0.60	0.06	0.48	0.71
27	Cluster 3	Hip	-0.34	0.05	-0.44	-0.23
28	Cluster 4	Hip	0.22	0.04	0.14	0.31
29	Cluster 1	Knee	-0.39	0.06	-0.51	-0.26
30	Cluster 2	Knee	0.39	0.06	0.28	0.51
31	Cluster 3	Knee	-0.38	0.06	-0.49	-0.26
32	Cluster 4	Knee	0.22	0.05	0.14	0.31
33	Cluster 1	Hands	-0.32	0.06	-0.44	-0.21
34	Cluster 2	Hands	0.20	0.06	0.09	0.32
35	Cluster 3	Hands	-0.19	0.06	-0.30	-0.09
36	Cluster 4	Hands	0.19	0.04	0.11	0.28
37	Cluster 1	Feet	-0.43	0.06	-0.54	-0.32
38	Cluster 2	Feet	0.40	0.06	0.28	0.51
39	Cluster 3	Feet	-0.28	0.05	-0.38	-0.17
40	Cluster 4	Feet	0.21	0.04	0.13	0.29

Table 3.6: 9146 out of 9990 participants of the study reported their site of pain. Log odds ratio of a site of pain in a cluster with 95% Confidence Interval

3.7 Supplementary figures

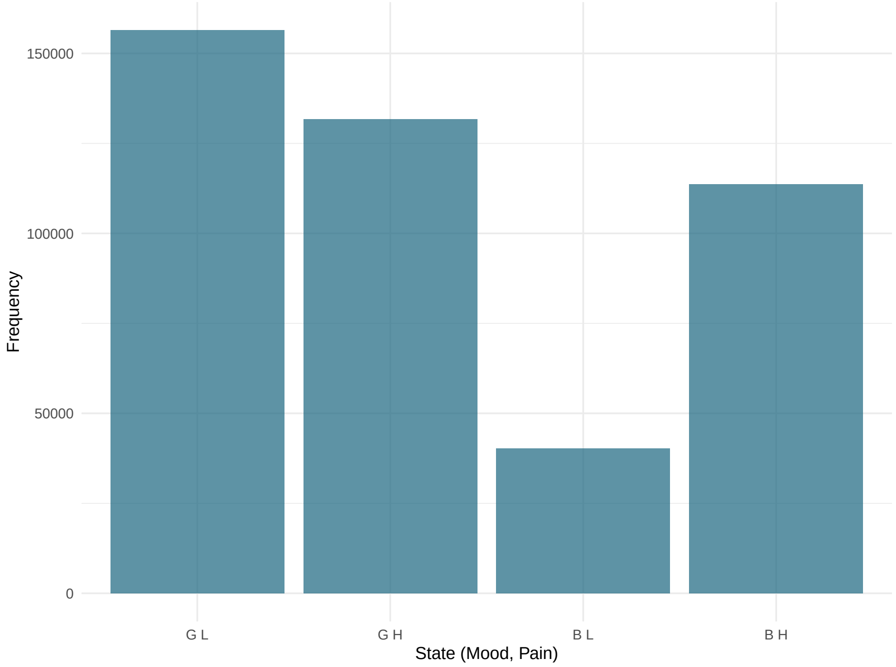


Figure 3.8: Frequency of states

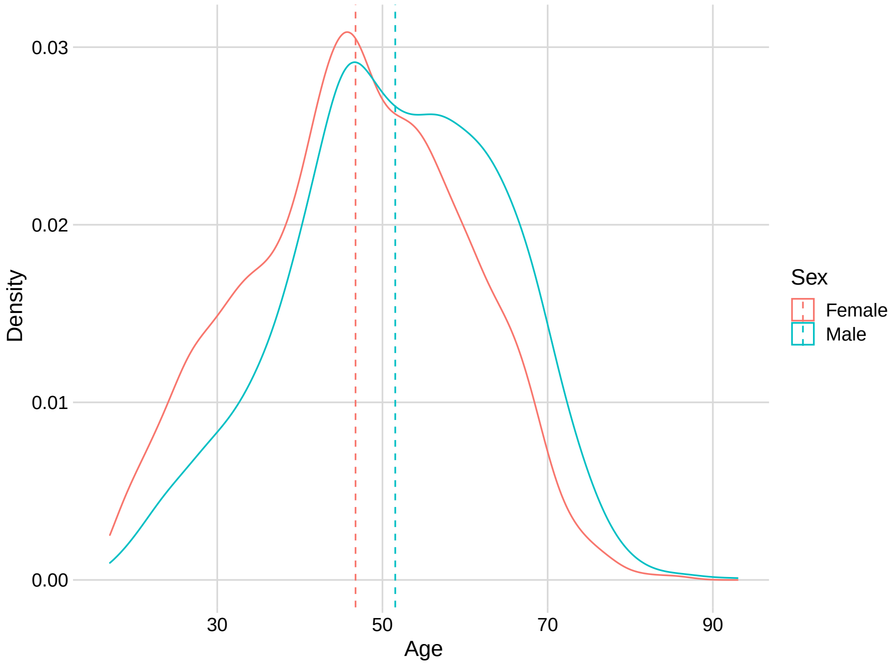
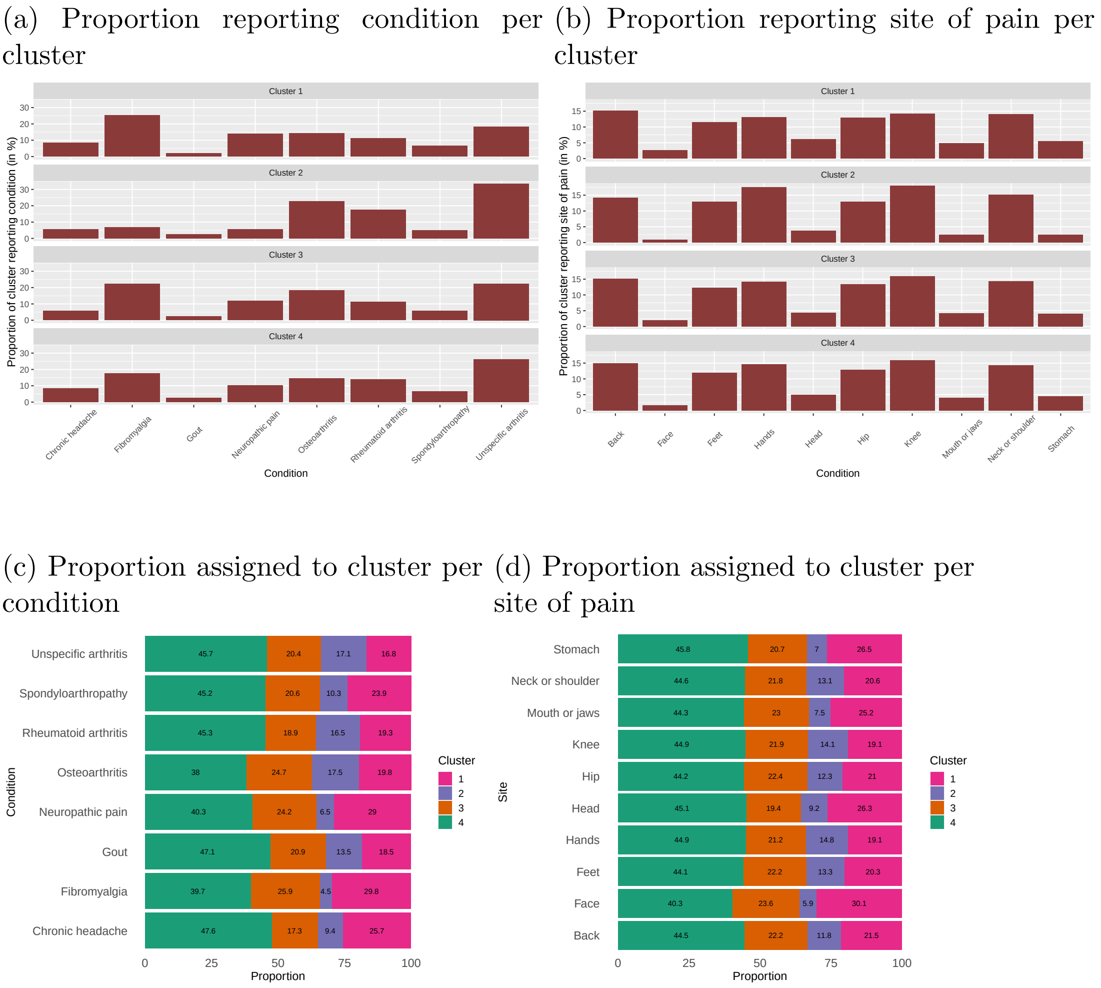


Figure 3.9: Overall age distribution, included for comparison with Figure 3.5.


(a) Proportion reporting condition per cluster

(b) Proportion reporting site of pain per cluster

(c) Proportion assigned to cluster per condition

(d) Proportion assigned to cluster per site of pain

Figure 3.10: **A** and **B** indicate the proportion of participants in a cluster reporting, respectively, a given condition and site of pain. **C** and **D** show the proportions of participants with, respectively, a given condition or site of pain who fall into each cluster.

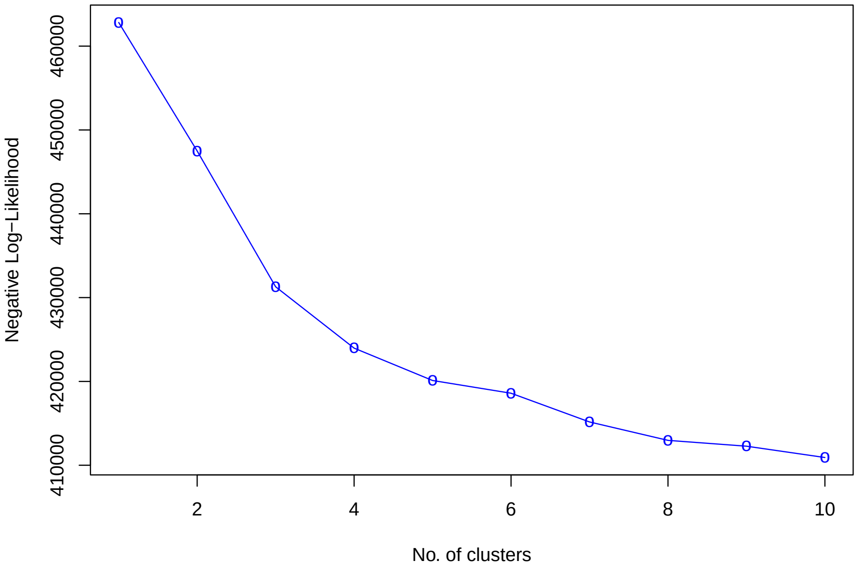


Figure 3.11: Negative Log Likelihood as a function of the number of components

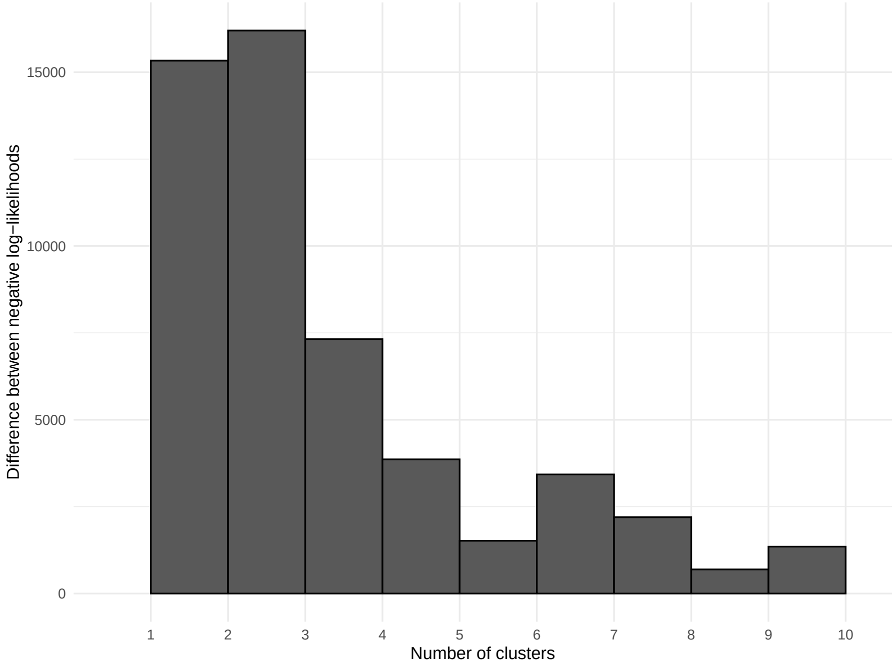


Figure 3.12: Difference of negative Log-Likelihoods of models with number of clusters k+1 and k

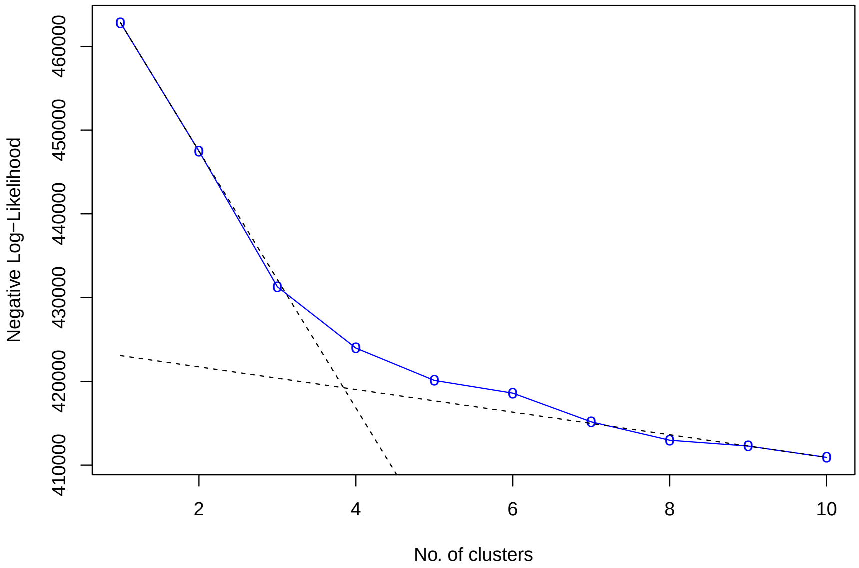


Figure 3.13: Dotted lines represent negative Log Likelihood gradients extrapolated from the difference between clusters 1 and 2, and clusters 9 and 10.

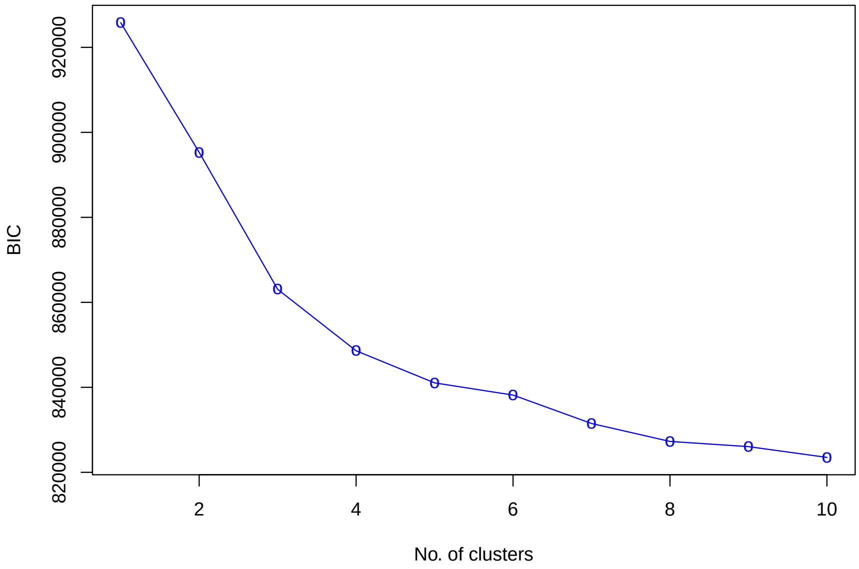


Figure 3.14: Bayesian Inference Criterion (BIC) per model

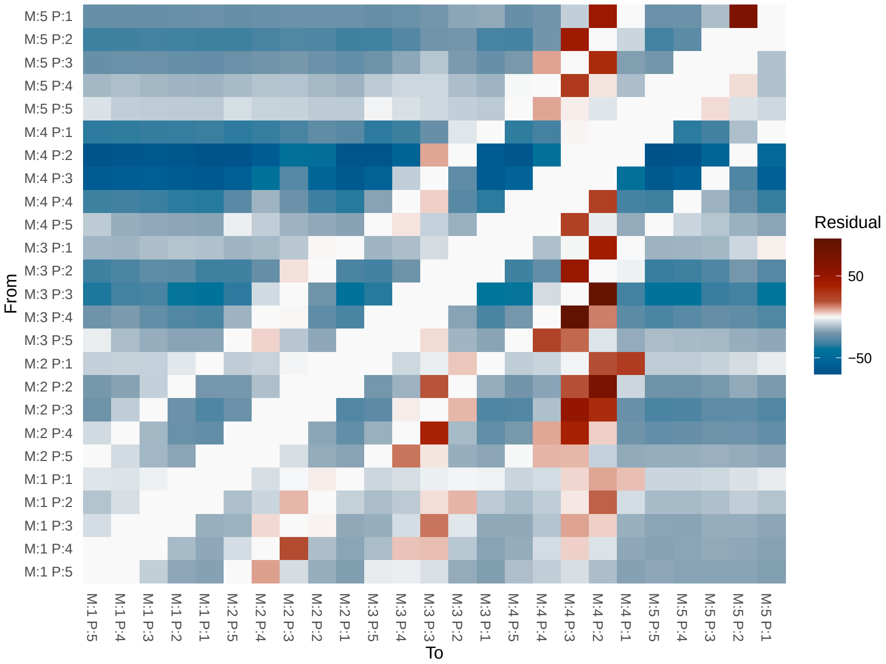
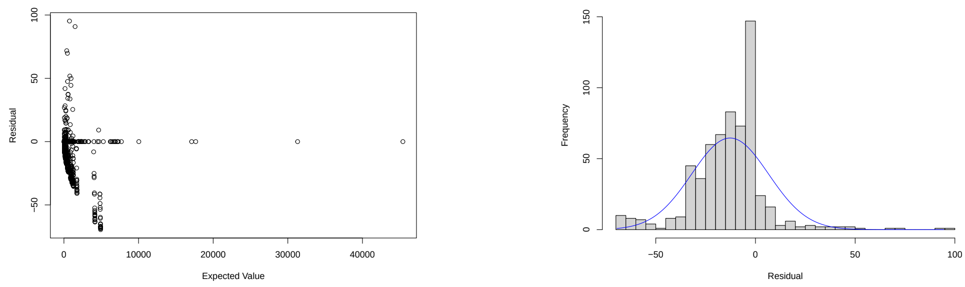


Figure 3.15: Residual heatmap of the model given by Eqn. (3.12)


(a) Expected vs residual values

(b) Normal curve over histogram Model 2

Figure 3.16: **A** is the scatter plot of expected values and the residuals. **B** shows a histogram of the residuals as well as a blue curve giving the probability density function of a normal distribution having the same mean and variance as the residuals.

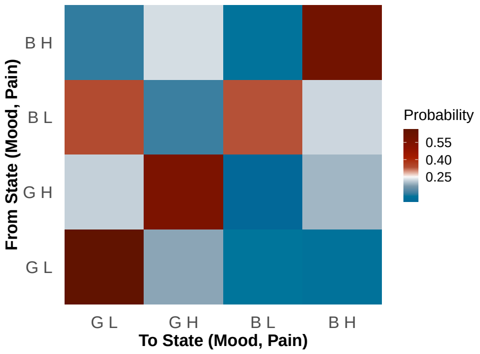


Figure 3.17: Transition Matrix based on the regroup scales

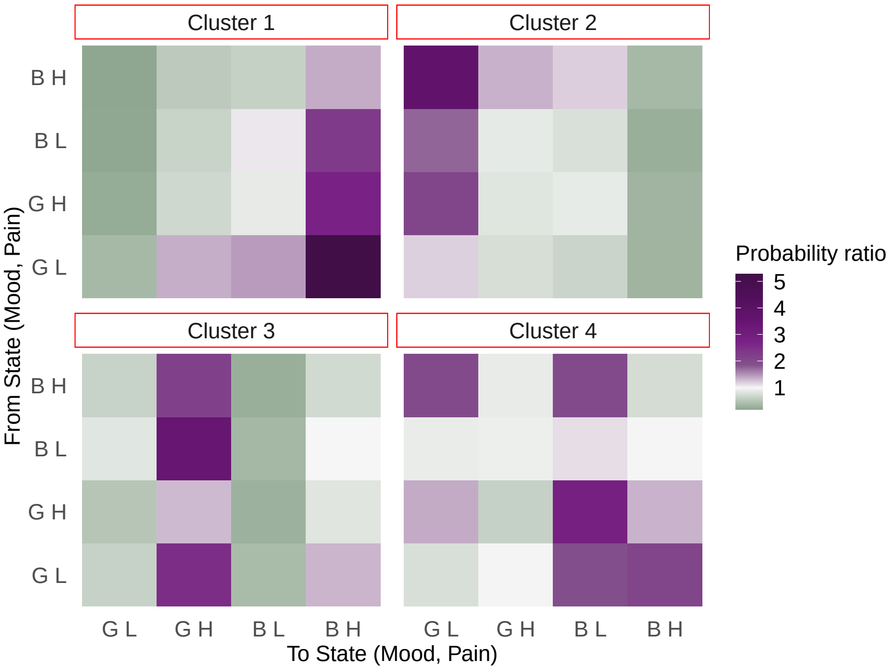


Figure 3.18: The ratio of the entries in the transition probability matrices for the clusters to the transition probabilities estimated from the whole sample without clustering.

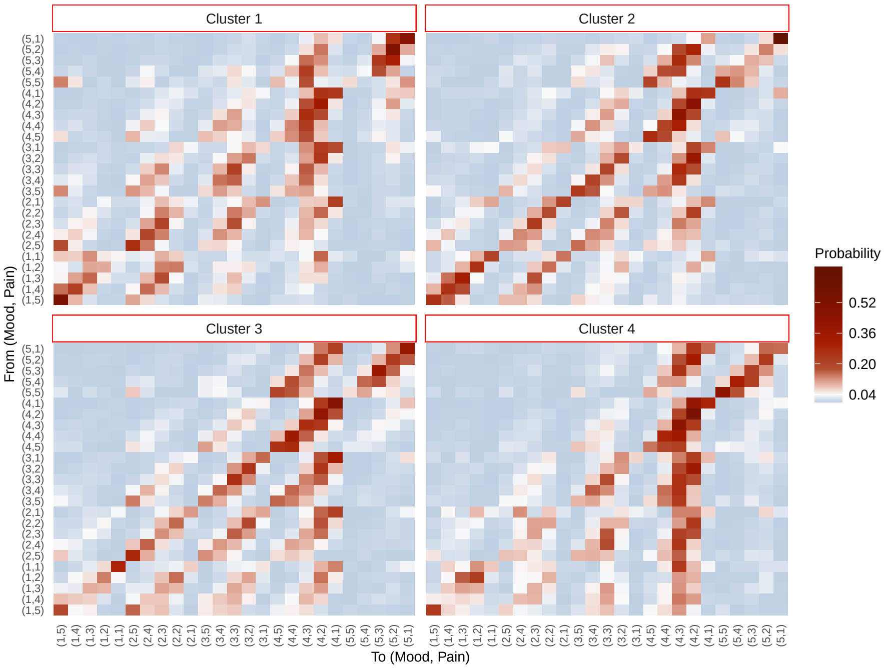


Figure 3.19: Four clusters without regrouping (Mood, Pain) states

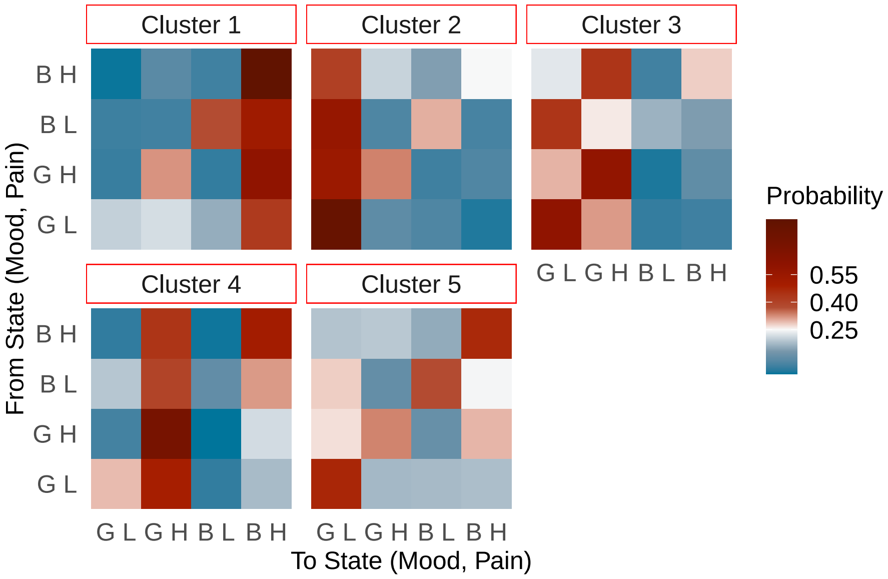


Figure 3.20: Heamtap of transition probability matrices when number of clusters is 5

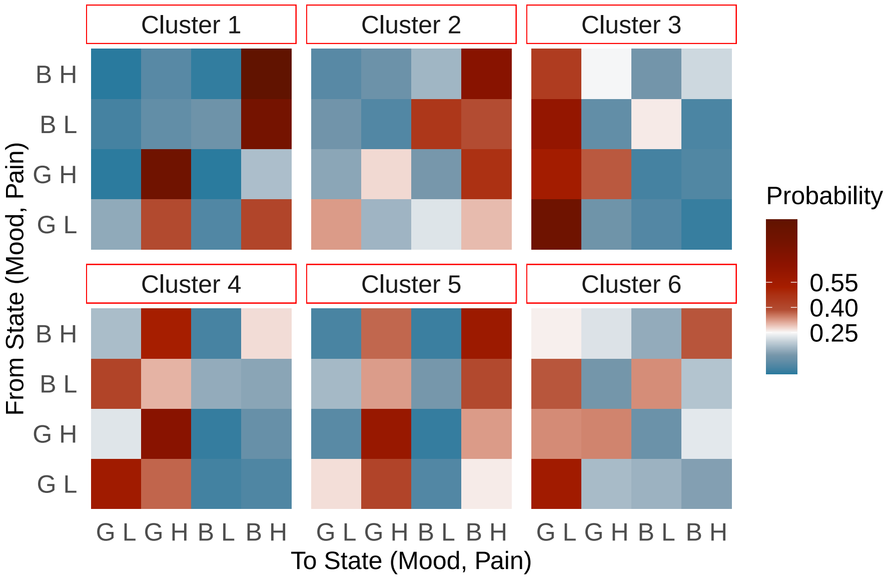


Figure 3.21: Heamtap of transition probability matrices when number of clusters is 6

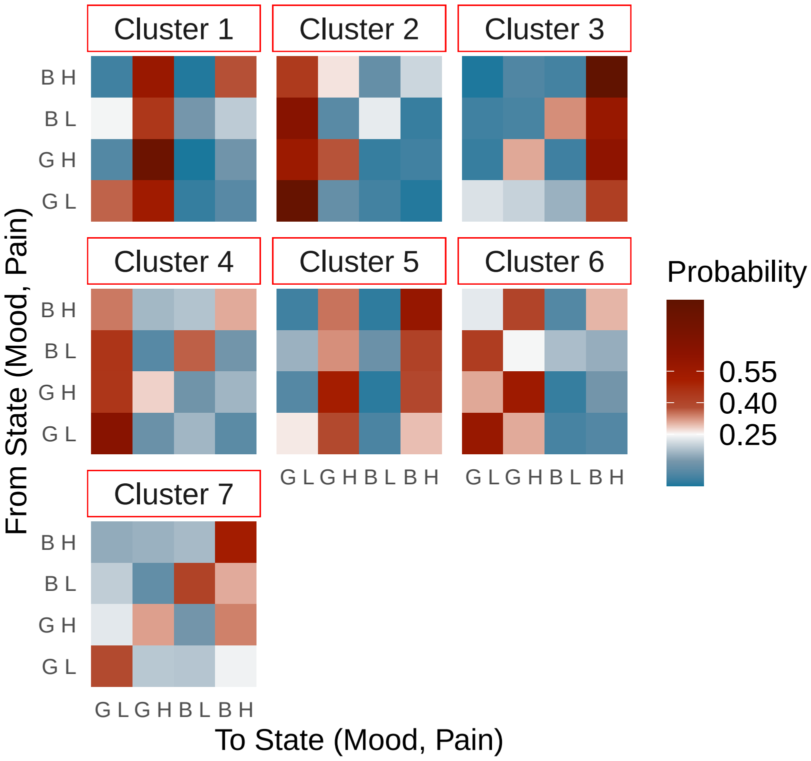


Figure 3.22: Heamtap of transition probability matrices when number of clusters is 7

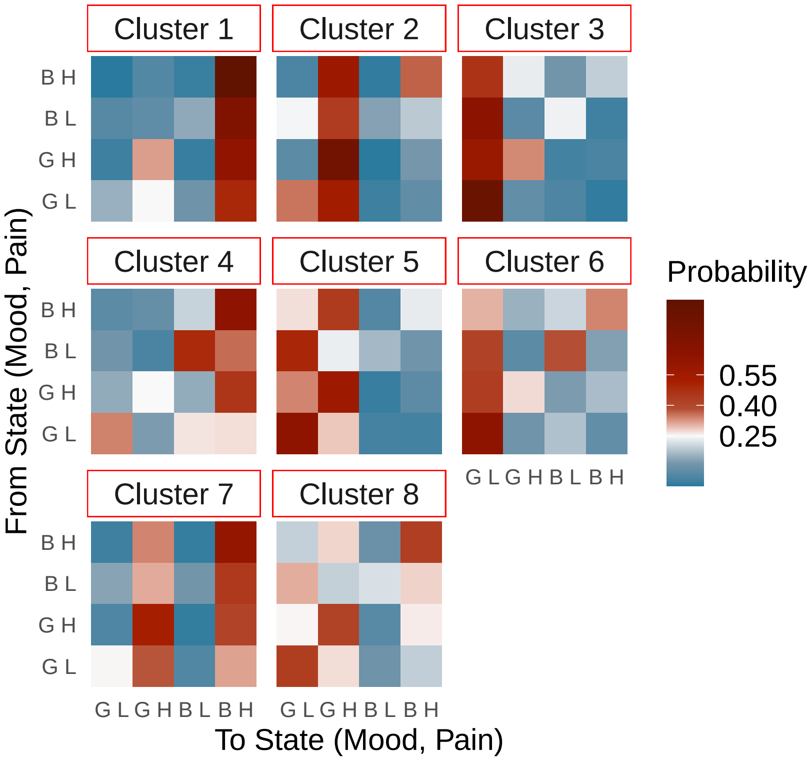


Figure 3.23: Heamtap of transition probability matrices when number of clusters is 8

Chapter 4

Fitting Dirichlet distribution to trajectories of self-reported data

Rajenki Das[1], Mark Muldoon[1], Thomas House[1]

1 – Department of Mathematics, University of Manchester, Manchester, UK

Abstract

We are given trajectories of data which we model to build a mixture of Markov chains parameterised by a vector of matrices each representing a class of Markov chains. We sample a finite mixture of Markov chains fitting Dirichlet distribution and implement Hamiltonian Monte Carlo imposing constraints on the parameters of the model to address the problem of label-switching. This method is then applied to a set of real data consisting of longitudinal trajectories of self reported mood and pain severities, and then the findings are reported.

4.1 Introduction

Data in real world are often heterogeneous in nature and can't be described by only one probability distribution. In statistics, mixture modelling is an approach to address such heterogeneity where sub-populations of the data are identified. It is useful in estimating unobserved variables, finding patterns and clustering into what are often known as *mixture* components. The main problem becomes estimating the parameters of the components of the mixtures. There are several applications of mixture modelling in day to day life, especially in healthcare such as finding endotypes of a given disease.

The multinomial distribution is used in the modelling of counts and is often useful when categorical data is presented. A classic approach taken for multinomial estimation is introduction of the Dirichlet distribution as a prior to the multinomial (Bouguila, 2008;

Minka, 2000). We can find applications of this Dirichlet-multinomial distribution in many fields like stock assessment (Bouguila, 2008), detection of protein sequence homology (Sjölander et al., 1996) and language modelling (MacKay and Peto, 1995).

In this paper, we are given trajectories of self-reported data which we have assumed to be sampled from a mixture of Dirichlet distribution on the transition matrices. Given these finite mixtures, we then build a Bayesian model to estimate the parameters by sampling each row of the transition matrix of a component from a Dirichlet distribution. For parameter estimation of this model, we implement Markov Chain Monte Carlo (MCMC) using the method of Hamiltonian Monte Carlo (HMC). In short, we show how HMC can be performed on mixture of count matrices parameterised by matrices of Dirichlet shape parameters by taking care of *label-switching* which is an inherent computational challenge in mixture modelling. Additionally, we run the model on real data of the trajectories of self-reported mood and pain, and report the inference on it.

4.2 Data and code

This is a secondary analysis where the data is taken from the Cloudy with a Chance of Pain study (Reade et al., 2017; Dixon et al., 2019) which was conducted in order to investigate the relationship between weather and pain, but in doing so a rich dataset was created which could be used to answer a diversity of research questions. Data were collected for 1 year 3 months which was from January 2016 to April 2017. The participants were residents in the UK who were aged 17 or above and had experienced chronic pain for at least 3 months preceding the survey (Druce et al., 2017).

The cohort had 10,584 survey participants, each of whom was asked to rate their symptoms and other variables on a mobile phone application in five ordinal categories of 1 to 5. Data were recorded for 10 variables, two of which were pain severity and mood. Participants were asked to provide information on these on a five-point Likert scale, with accompanying text for each of the ordinal levels. For mood, a score of 1 represents worst mood and 5 represents best, whereas for pain a score of 1 represents least pain and 5 represents most.

In this study, we analyse trajectories of self-reported pain severity and mood scores — individual mood and pain trajectories, and also a trajectory with pairs of mood and pain. In case of a state not reported *i.e.* recorded as NA, we remove that data point *i.e.* the entire row in the dataset entirely. Also, while performing this analysis, we considered only those participants who had at least 3 weeks *i.e.* 21 days of entry — participants outside this criterion were removed.

As we were considering these trajectories to be drawn from Markov chains, the transition probability of moving to another state is 1/(no. of states). If we regroup the data, where possible, in such a way that the number of states decreases therefore, our transition probability increases which helps in better understanding of the transitions. So for easier understanding of results, we regrouped the severity of mood and pain into two categories each on the basis of the descriptions associated with each ordinal value. Mood scores of 1–3 and 4–5 were labelled Bad (B) and Good (G) respectively, while pain levels of 1–2 and 3–5 were, respectively, labelled Low (L) and High (H). When combining mood and pain states, we get four states: GL; GH; BL; and BH.

Code will be made available at: `https://github.com/rajenkidas/`.

4.3 Bayesian framework

In this section, we talk about the steps taken in developing the distribution over transition matrices and establishing a mixture model whose components are represented by Dirichlet distributions over row-vectors of a transition matrix per component.

A similar method has been described by Frühwirth-Schnatter and Pamminger (2010) in studying wage mobility in the Austrian labour market where they model the deviations of each row of transition matrix using Dirichlet-multinomial and take the Dirichlet parameters as the group specific parameters. In our case, we have assumed each row of transition matrix to be sampled from Dirichlet and the group is specified by Dirichlet parameters and mixture weights. Using these we introduce few more parameters later which have been utilised to address the problem of label-switching in mixture models.

4.3.1 What data do we have?

We are given trajectories of ordinal data. We assume each trajectory is drawn from an n-state Markov chain *i.e.* the trajectories follow the Markov property in which a state depends only on the previous state. Using these trajectories, we derive count matrices $\mathbf{C_s}$ for each of the subject $s \in \{1, \ldots, S\}$. Count matrices contain frequencies of transitions observed from one state i to another state j. These can be reduced to matrices of transition probabilities $\mathbf{T_s}$ per subject s where each of its elements represents the proportion of counts in a row for a subject. The trajectories are further classified into K components based on the transitions observed in the data. We assume that each row of the transition count matrix $\mathbf{C}$ is sampled from a Dirichlet-multinomial distribution for a component. Therefore, each component gets defined by its mixture weight and a matrix of Dirichlet parameters. Our goal is to model the count data by implementing Bayesian inference in order to estimate the Dirichlet parameters of the model, which can thereby also help in predicting the probability of a state for a subject given the trajectory.

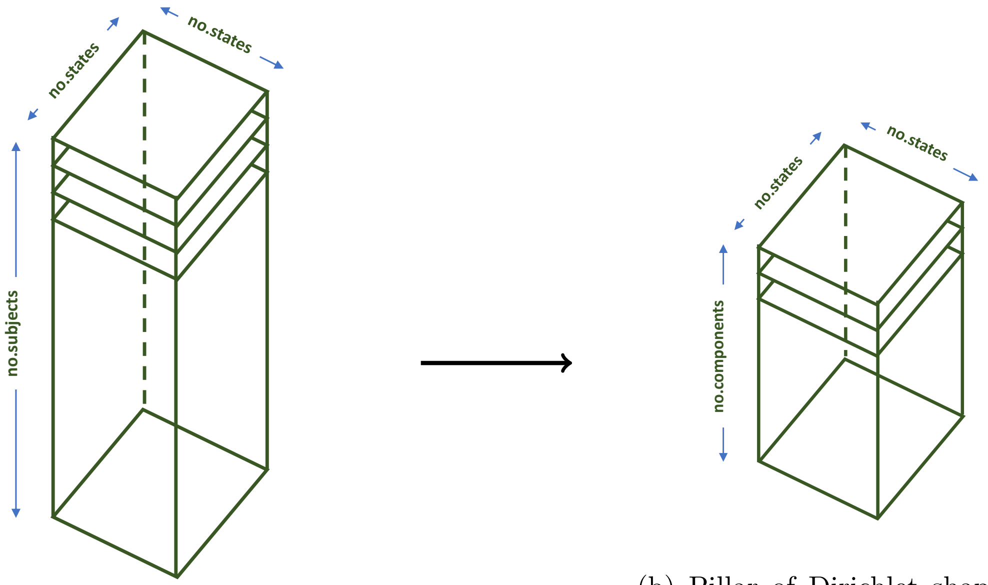
So, we have the following data for the model:

- Total number of states: n
- Total number of subjects: S
- Total number of components: K
- Count matrices: $\mathbf{C}_s$ for each subject s.

The real data specifications are given in the Supplementary in Table 4.1.

4.3.2 What is the model?

To summarise the gist of the model pictorially, we represent it as Figure 4.1a getting reduced to Figure 4.1b where each row of a count matrix is sampled from a Dirichlet distribution. To achieve this, we begin by defining some intermediate parameters for which we carry out transformations as stated in the equations: (4.2) to (4.5). Using these parameters, we can retrieve each of the individual Dirichlet shape parameters as shown in the Equation (4.6). The Dirichlet parameters are later used in computing likelihood and posterior distribution.



(a) Pillar of count matrices

(b) Pillar of Dirichlet shape parameters matrices

Let the matrix of Dirichlet shape parameters α associated with a single component be written as:

$$\begin{bmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \vdots & & & \\ \alpha_{n1} & \alpha_{n2} & \dots & \alpha_{nn} \end{bmatrix}, \tag{4.1}$$

then each row of the above matrix specifies a Dirichlet distribution.

Summing the elements of the matrix (4.1) over columns across rows, we get the row-sum for each row:

$$\hat{\alpha}_i = \sum_{j=1}^{n} \alpha_{ij} \tag{4.2}$$

Dividing every element of the matrix (4.1) by the row-sum defined in (4.2), we get the following parameter that adds up to 1 for each row:

$$\theta_{ij} = \frac{\alpha_{ij}}{\hat{\alpha}_i} \tag{4.3}$$

Now let $\alpha_{\text{GrandTotal}}$ be defined as the grand total of all the Dirichlet shape parameters of a component. This parameter helps in identifying a mixture-component, and used to impose an ordering constraint on the model. We sum all the elements of the matrix (4.1):

$$\alpha_{\text{GrandTotal}} = \sum_{i=1}^{n} \hat{\alpha}_i \tag{4.4}$$

The fraction of row-sum (4.2) in grand total is given by:

$$\phi_i = \frac{\hat{\alpha}_i}{\alpha_{\text{GrandTotal}}} \tag{4.5}$$

Therefore, ϕ_i for each component sums to 1.

We create the intermediate parameters so that it gets easier to reconstruct the Dirichlet shape parameters in the end using:

$$\alpha_{ij} = \phi_i \theta_{ij} \alpha_{\text{GrandTotal}} \tag{4.6}$$

So the parameters used in defining the model are:

- Mixture weights per component: ω, vector of length K
- Proportion of Dirichlet parameter per row: θ_i for row i, simplex vector of length n per state per component
- Ordering constraint parameter defined per component: $\alpha_{\text{GrandTotal}}$ per component
- Proportion of row-sum of Dirichlet parameters in $\alpha_{\text{GrandTotal}}$: ϕ, a simplex vector of length n per component
- Sum of Dirichlet parameters per row: $\hat{\alpha}$, a vector of length n
- Dirichlet shape parameter vector: $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$, a vector of length n per state per component

4.3.3 What are the priors taken for the model parameters?

Here we list the prior distributions considered for some of the parameters of the model. Since some parameters are related to other parameters so that their values can be computed once other parameter values are known, all parameters do not require a prior to be given.

Prior taken for:

- Mixture weights ω: Dirichlet distribution
- Parameter θ_i for row i: Dirichlet distribution
- Ordering constraint parameter, $\alpha_{\text{GrandTotal}}$: Gamma distribution

Exact prior values are provided in Table 4.2 in the Supplementary of this paper.

4.3.4 How do we initialise the MCMC chains?

This is not a necessary step for the MCMC sampling, but we do so for better and faster convergence of the chains. We specify some initial values for some of the parameters of the model so that the starting point of the MCMC chains are close to the same local maximum of the likelihood thus, avoiding label-switching. To obtain the point of initialisation, we implement Expectation-Maximisation (EM) algorithm by fitting mixture model to count matrices $\mathbf{C_s}$ per subject to estimate matrices of Dirichlet shape parameters of size $n \times n$ for K components. Additionally, the EM-algorithm also returns mixture weights of the components of the model which are obtained from class memberships of the subjects. Using step (4.5), we get the ϕ parameter of the model. Using the generated matrices, θ vectors per row per component are sampled from Dirichlet distribution with the shape parameters corresponding to the row and component. In the end, we generate a point of initialisation for the MCMC chains by providing estimates of parameters ω, ϕ, θ and $\alpha_{\text{GrandTotal}}$.

Initial values given for parameters: ω, ϕ, θ and $\alpha_{\text{GrandTotal}}$ (ordered set).

More details on the starting guess and the ordering of the parameters are given in the section 4.4.

4.3.5 Bayesian inference

Here we finally formulate the Bayesian inference framework for estimating the Dirichlet shape parameters. It remains to calculate the likelihood which is used to estimate the posterior probability in the STAN model.

We model vectors of count data $\boldsymbol{y}_s \in \mathbb{N}^n$ as arising from a multinomial distribution whose per sample probability vector $\boldsymbol{p}_s$ is drawn from a K-component mixture *i.e.* $\boldsymbol{y}_s \in \text{Mult}(\boldsymbol{p}_s)$ and $\boldsymbol{p}_s \in \text{Dir}(\boldsymbol{\alpha})$ where $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_N\}$ is vector of Dirichlet shape parameters in simplex space Δ_{n-1} of dimension $n-1$. Let $\boldsymbol{\omega}$ be the vector of K mixture weights. Given this scenario, the likelihood can be computed as:

$$\begin{aligned} P(\boldsymbol{y}_s, \boldsymbol{p}_s \mid \boldsymbol{\omega}, \boldsymbol{\alpha}) &= P(\boldsymbol{y}_s \mid \boldsymbol{p}_s) P(\boldsymbol{p}_s \mid \boldsymbol{\omega}, \boldsymbol{\alpha}) \\ &= \text{Mult}(\boldsymbol{y}_s \mid \boldsymbol{p}_s) \left(\sum_{k=1}^{K} \omega_k \text{Dir}(\boldsymbol{p}_s \mid \boldsymbol{\alpha}_k) \right) \end{aligned}$$

But $\boldsymbol{p}_s$ values are unknown to us, and therefore we would like to marginalise these out in the following way:

$$\begin{aligned} P(\boldsymbol{y}_s \mid \boldsymbol{\omega}, \boldsymbol{\alpha}) &= \int_{\Delta_{n-1}} P(\boldsymbol{y}_s \mid \boldsymbol{p}_s) P(\boldsymbol{p}_s \mid \boldsymbol{\omega}, \boldsymbol{\alpha}) \mathrm{d}\boldsymbol{p}_s \\ &= \int_{\Delta_{n-1}} \text{Mult}(\boldsymbol{y}_s \mid \boldsymbol{p}_s) (\sum_{k=1}^{K} \omega_k \text{Dir}(\boldsymbol{p}_s \mid \boldsymbol{\alpha}_k)) \mathrm{d}\boldsymbol{p}_s \\ &= \sum_{k=1}^{K} \omega_k \left(\int_{\Delta_{n-1}} \text{Mult}(\boldsymbol{y}_s \mid \boldsymbol{p}_s) \text{Dir}(\boldsymbol{p}_s \mid \boldsymbol{\alpha}_k) \mathrm{d}\boldsymbol{p}_s \right) \end{aligned} \tag{4.7}$$

Now we focus on the single integral included in the sum in the Equation (4.7) and drop the subscripts s and k. Therefore, the problem reduces to:

$$P(\boldsymbol{y} \mid \boldsymbol{\alpha}) = \int_{\Delta_{n-1}} \text{Mult}(\boldsymbol{y} \mid \boldsymbol{p}) \text{Dir}(\boldsymbol{p} \mid \boldsymbol{\alpha}) \mathrm{d}\boldsymbol{p} \tag{4.8}$$

where $\boldsymbol{y} \in \mathbb{N}^n, \boldsymbol{p} \in \Delta_{n-1}$ and $\boldsymbol{\alpha} \in \mathbb{R}_{+}^{N}$.

We have,

$$\text{Mult}(\boldsymbol{y} \mid \boldsymbol{p}) = \frac{N!}{\prod_{i=1}^{n} y_i} \left(\prod_{i=1}^{n} {p_i}^{y_i} \right) \text{ and } \text{Dir}(\boldsymbol{p} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha)}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \left(\prod_{i=1}^{n} {p_i}^{\alpha_i - 1} \right).$$

where, $N = \sum_{i=1}^{n} y_i$ and $\alpha = \sum_{i=1}^{n} \alpha_i$.

Substituting the probability density functions of multinomial and Dirichlet distributions in the integral (4.8), we get:

$$\begin{aligned}
&P(\boldsymbol{y} \mid \boldsymbol{\alpha}) \\
&= \int_{\Delta_{n-1}} \frac{N!}{\prod\limits_{i=1}^{n} y_i} \frac{\Gamma(\alpha)}{\prod\limits_{i=1}^{n} \Gamma(\alpha_i)} \prod_{i=1}^{n} {p_i}^{\alpha_i+y_i-1} \mathrm{d}\boldsymbol{p} \\
&= \frac{N!}{\prod\limits_{i=1}^{n} y_i} \frac{\Gamma(\alpha)}{\prod\limits_{i=1}^{n} \Gamma(\alpha_i)} \int_{\Delta_{n-1}} \prod_{i=1}^{n} {p_i}^{\alpha_i+y_i-1} \mathrm{d}\boldsymbol{p} \\
&= \frac{N!}{\prod\limits_{i=1}^{n} y_i} \frac{\Gamma(\alpha)}{\prod\limits_{i=1}^{n} \Gamma(\alpha_i)} \frac{\prod\limits_{i=1}^{n} \Gamma(\alpha_i + y_i)}{\Gamma(\boldsymbol{\alpha} + y)} \\
&= \frac{N!}{\prod\limits_{i=1}^{n} y_i} \frac{\mathcal{B}(\boldsymbol{\alpha} + \boldsymbol{y})}{\mathcal{B}(\boldsymbol{\alpha})} \\
&= \mathrm{DirMult}(\boldsymbol{y} \mid \boldsymbol{\alpha})
\end{aligned} \tag{4.9}$$

where

$$\mathcal{B}(\boldsymbol{\alpha}) = \frac{\prod \Gamma(\alpha_i)}{\prod(\sum \alpha_i)} = \int_{\Delta_{n-1}} \left(\prod {p_i}^{\alpha_i-1}\right) \mathrm{d}p$$

Now finally substituting in Equation (4.7), we find that the marginalising $\boldsymbol{p}$ out of $P(\boldsymbol{y}, \boldsymbol{p} \mid \boldsymbol{\omega}, \boldsymbol{\alpha})$ gives us **Dirichlet-Multinomial** distribution.

Thus, the likelihood of our model is written as :

$$\mathrm{DirMult}(\boldsymbol{y}_s \mid \boldsymbol{\omega}, \boldsymbol{\alpha}) = \sum_{k=1}^{K} \omega_k \mathrm{DirMult}(\boldsymbol{y}_s \mid \boldsymbol{\alpha}_k) \tag{4.10}$$

4.4 MCMC constraints

In this part of the section, we discuss the additional problems as part of the modelling. Since this is a problem of mixture modelling *i.e.* modelling using a distribution with more than one set of parameters, we encounter the inherent problem of label-switching while computing.

To address this issue, we focus on the following two steps of:
1) selecting a starting guess for running the MCMC sampling, and
2) imposing constraints on the parameters of the model.

4.4.1 Initialisation outline

To ensure good performance of the MCMC model, we initialise the Markov chains by finding a starting point via EM-algorithm. With MCMC starting from the same local maximum likelihood, it helps to address the problem of label-switching in mixture models. Broadly we take the following two steps to determine the point of initialisation for the MCMC: 1. Perform EM in order to find a matrix of class assignments for the subjects. 2. Using the class assignments, estimate Dirichlet shape parameters.

In step (1), we initiate EM by performing k-means in Centred-Log-Ratio space.

So to generate a starting guess for the MCMC inference, we fit a mixture model to the count matrices using EM algorithm. This organises the subjects' information to the specified number of clusters as previously shown in Chapter 3. Once, we have the class assignments for the participants, then using count data and cluster-membership probabilities, we estimate matrices of Dirichlet shape parameters where each row represents a Dirichlet distribution. These parameters are further fed into the MCMC model by fitting a Dirichlet distribution to the rows of proportions of the counts matrices. Using an R package (Heck et al., 2019), we estimate the shape parameters for samples of each component which give us the initial values.

4.4.2 Ordering of parameters

A common strategy is to impose an ordering constraint on the parameters to identify the components of the mixture model. So we imposed a constraint on the parameters by ordering the grand total of Dirichlet shape parameters, as derived in equation (4.4), per component to associate with the corresponding component.

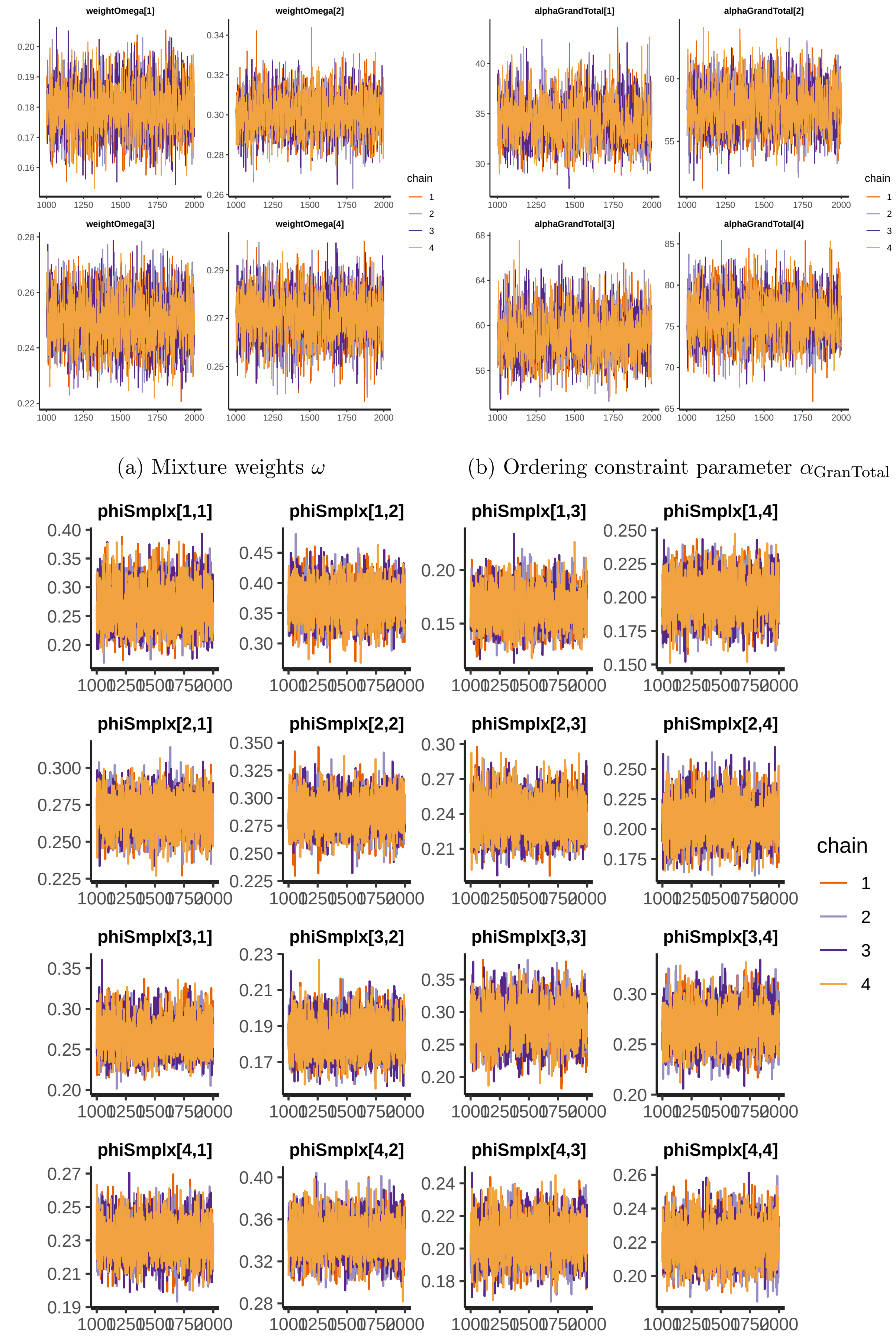
Hence, for a component $k \in \{1, \ldots, K\}$, associated with an $n \times n$ matrix of Dirichlet shape parameters $[\alpha]^k_{n\times n}$, we compute the total sum of all the parameters for a component k as $\alpha^k_{\text{GrandTotal}} = \sum_{1\leq i,j\leq n} \alpha^k_{ij}$. This helps the HMC-MCMC chains in identifying the mixture components. We do assume that these alpha grand totals are all different for the components, and by imposing the ordering constraint, we found that our Monte Carlo Markov chains of the model converged. This constraint confines HMC-MCMCM to a region of parameter space that will contain only a single example of the $K!$ symmetrically-related local maxima that label-switching produces.

Before proceeding to the next section, we would like to point out that we had initially run our method on synthetic data which was instrumental in assessing the robustness and limitations of our technique, *i.e.* the synthetic data analysis was mainly a debugging tool. Once we achieved decent results– primarily indicated by convergence of the sampling per chain around the expected values, we moved on to implementing our method to the real dataset which was introduced in Section 4.2. By doing so, we get results showing how our model worked which, at this point, we mainly check by the convergence of the Monte Carlo Markov chains. Additionally, we also learn about the mood and pain associations in this data and compare with those found by Das et al. (2023). These results based on the observed mood-pain trajectories bring the possibility of predicting mood-pain which is a pertinent question in the health sphere.

In the next section, we share the results from implementation of the Bayesian inference described in this Chapter on the real data which contain self-reported values of mood and pain over a period of time.

4.5 Results on real data

Trace plots show the movement of iterations in MCMC. It helps to assess the convergence and mixing of the chains. We show the trace plots of some parameters representing our model. We have taken 4 chains for sampling, so these are denoted by 4 trace plots of different colour in a figure.



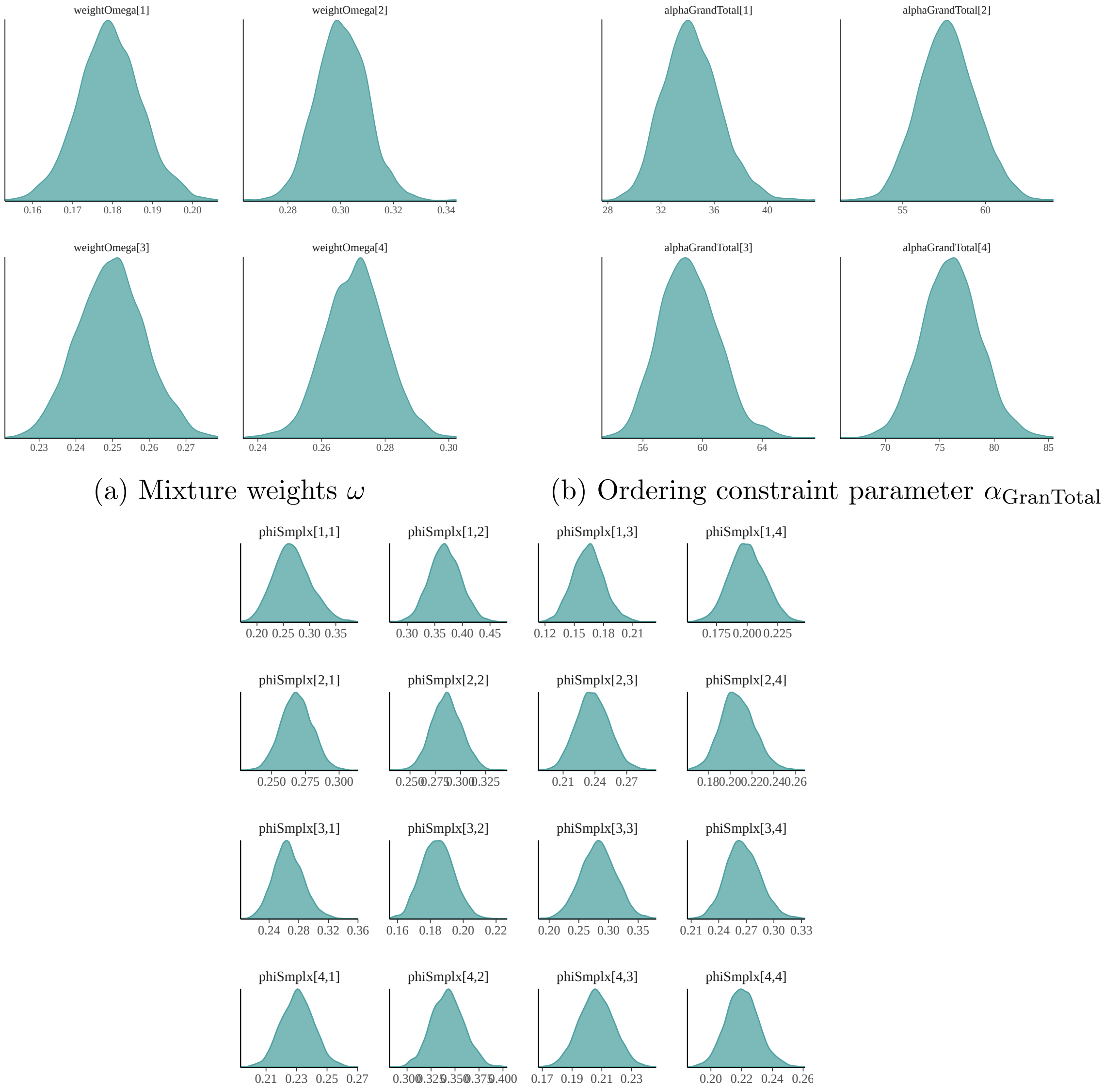
(a) Mixture weights ω

(b) Ordering constraint parameter $\alpha_{\text{GranTotal}}$

(c) Parameter ϕ

Figure 4.2: Parameters for modelling Mood-Pain trajectories with 4 states

We display density plots for the same parameters in Figure 4.3.



(a) Mixture weights ω

(b) Ordering constraint parameter $\alpha_{\text{GranTotal}}$

(c) Parameter ϕ

Figure 4.3: Parameters for modelling Mood-Pain trajectories with 4 states

In Figure 4.4, we show the trace plots for all the estimated Dirichlet shape parameters of the model. We see each MCMC chain gives almost same estimate for every parameter therefore, appear to converge to the same value.

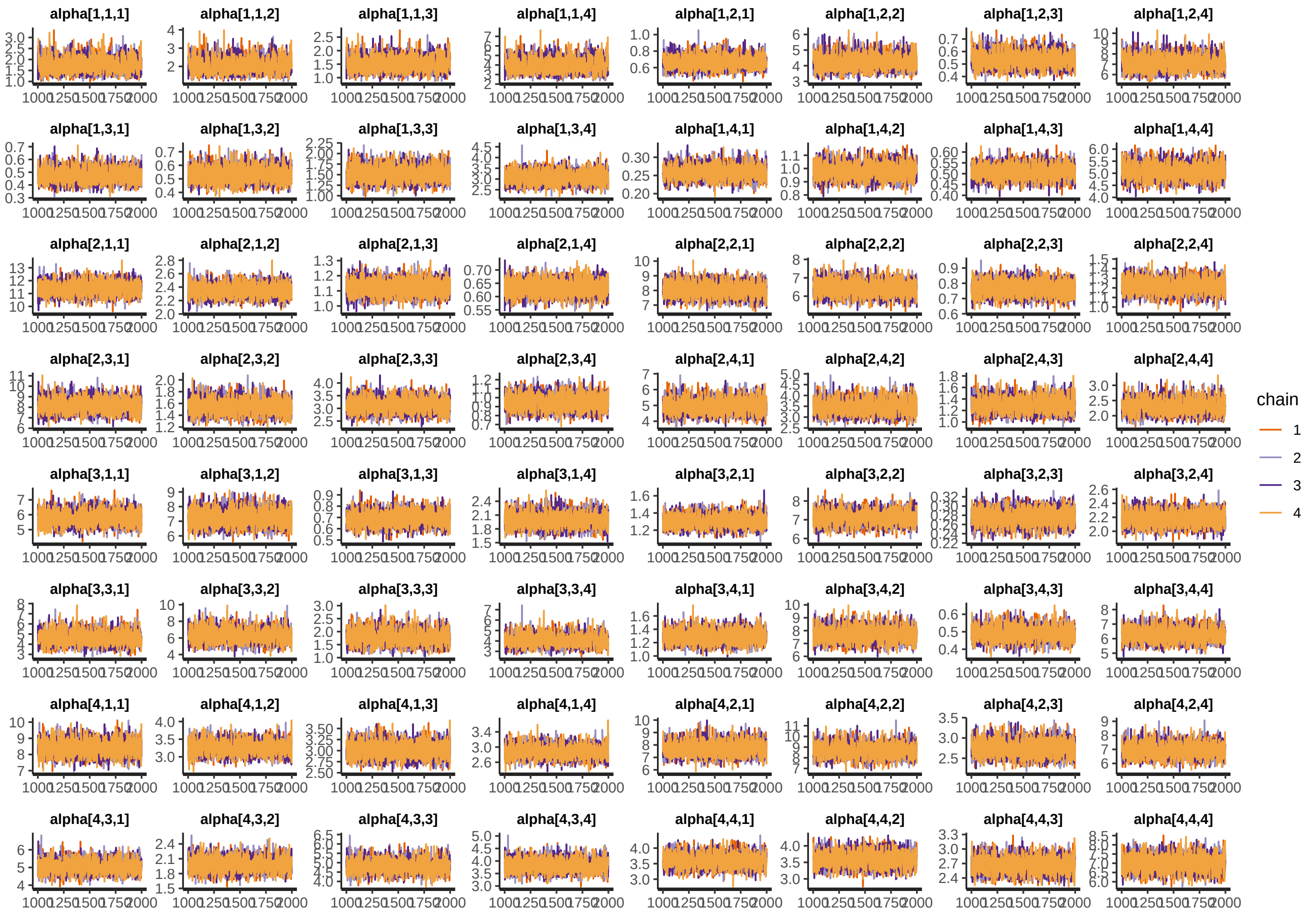


Figure 4.4: Dirichlet shape parameters of the model

Next, if we were to withdraw random variables from the Dirichlet distributions with the estimated Dirichlet shape parameters, which are the αs, then we get the density plots as shown in Figure 4.5. Each of the states is an ordered pair of mood and pain scores where G and B imply good mood and bad mood respectively, while L and H imply low pain and high pain respectively.

While studying these densities individually, we focus on the mean and standard deviation to understand how strong the transition probabilities are. We write about few of the observations here but provide a discussion by giving meaning to the states in the next section. In component 1, we see for transition from state 1 to any other state, the density plots representing the probabilities have more or less the same variance. Probability from state 1 to states 1 and 3 are almost same. From 2 and 3, it seems unlikely moving to state 1. From 4, there is a good probability of moving to 4, even though the density noted is low as compared to others.

In component 2, there is low probability of moving from 1 and 2 to 4. Movement to state 1 from any of the states is quite probable with probabilities above 0.5.

In component 3, probabilities of moving from any of the states to state 3 are low. Out of all the values displayed, probability from 2 to 2 is the highest.

In component 4, movement to state 1 from 1, 2, 3 are more than 0.3. Movement from 3 to 2 has low probability as the mean is slightly above zero.

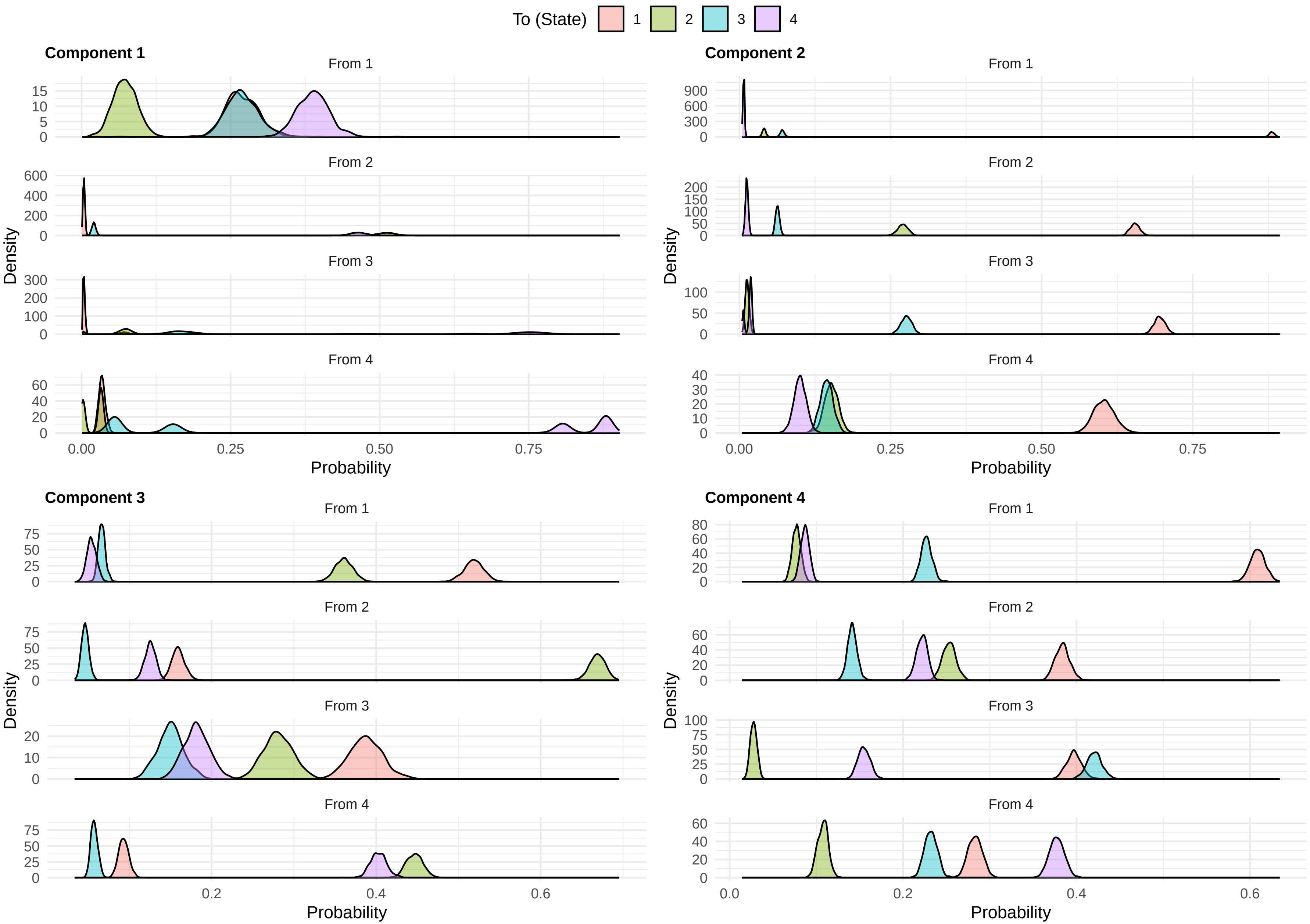


Figure 4.5: Probabilities sampled from Dirichlet distributions, where State 1: (G,L); State 2: (G,H); State 3: (B,L); State 4: (B,H)

4.6 Discussion

We developed a distribution over transition matrices. We set up a routine of steps that associates transition count matrices of participants of a longitudinal data survey to a matrix where the rows are vectors of Dirichlet shape parameters. We extend this by considering a mixture model, therefore the count matrix of every participant is linked to the set of Dirichlet parameters belonging to only one of the components. To elaborate on this, let's say there are people who have recorded some data containing 3 states. Then each person has a personal 3×3 transition count matrix. Now let there be 2 components, so each of person's transition matrices has rows which are assumed to be sampled from a Dirichlet distribution associated with one of the two components. In short, a personal matrix is sampled from a Dirichlet distribution (defined by a matrix of parameters) belonging to a component.

To achieve the mixture-modelling and its Bayesian inference, we had also introduced a new parameter $\alpha_{\text{GrandTotal}}$ which was used to impose ordering constraint on the cluster-specific parameters.

Now we discuss a few of the defining characteristics of the clusters. Cluster 1 shows high probability of movement from any of the states to the one with Bad Mood, High Pain. Cluster 2 has high probability of moving from any of the states to Good Mood, Low Pain. Cluster 3 shows high probability of staying in the same state, other than in the case of already being in Bad Mood, Low Pain when the probability of moving to Good Mood, Low Pain is the highest. In Cluster 4, we see high probability of movement to Good Mood and Low Pain from the same state, and the one with same mood but High Pain. For the other two states, there is a tendency of remaining in the state in this cluster. In a nutshell, in comparison to the results found in Chapter 3, we also find distinct characteristics of the clusters, few of which are quite similar to what we had observed before. In this chapter, we again notice that two of the clusters are distinguished by high probabilities of transitioning to the best state and the least ideal state.

A limitation is that this method may fail for scenarios which have not been explored yet. However, problems might be easy to deal with mathematically but computational challenges may arise. For example, if the sums of the shape parameters are equal for at least two of the components, then our ordering constraint will fail as at least two of the components will be linked by the same hyper-parameter giving rise to the label-switching problem. There already exists many methods concerning the label-switching problem, and those can also be implemented in this scenario where MCMC was performed using Hamiltonian Monte Carlo method. But another problem that could occur is poor convergence in a higher dimension data and it could require more attention. We have not yet checked the performance of our method in a situation with much more dimensions than currently taken in this paper. Even though higher dimensions may present new issues, we think the routine as set forth in this paper helps in setting up a basis for Dirichlet-multinomial parameter estimation of this kind of longitudinal data by performing Hamiltonian Monte Carlo. An improvement at any of the steps is a scope for further research. We did not perform sensitivity analysis of the parameters which can also be studied further.

The methodology provided at this paper can be extended to more similar types of longitudinal data, and utilised in prediction modelling. Another potential extension of this study is to consider the geometry of the Dirichlet shape parameters which have not been discussed in this paper.

Acknowledgement

We would like to thank Prof Korbinian Strimmer, Department of Mathematics, University of Manchester for granting access to Minerva which is a high-memory compute server. We would also acknowledge the IT services (University of Manchester) for providing support with Computational Shared Facility for running High Performance Computing.

Supplementary Material

4.7 Outline of steps taken

All the steps taken to develop our codes are:

- Generate synthetic data: First, we generated synthetic data of Markov chain mixtures drawn from Dirichlet distribution in the following manner:

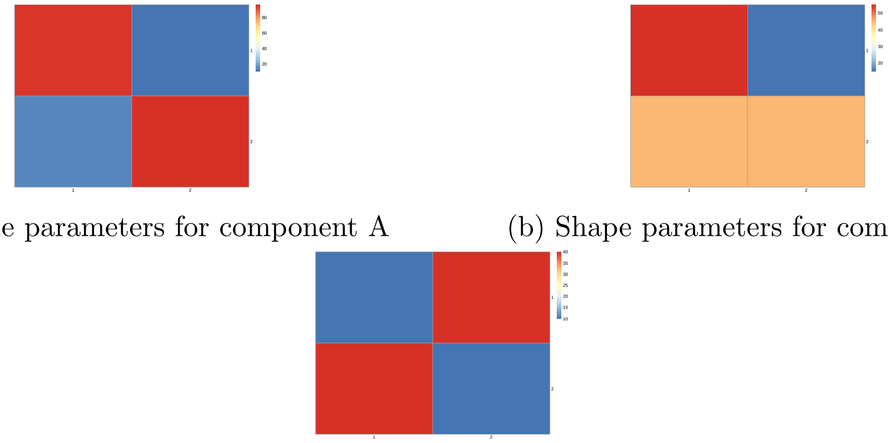
    1. Fix number of components K, number of states n and number of subjects S.
    2. Setup: Create a transition probability matrix for each component where each row of the transition matrix is sampled from a Dirichlet distribution.
    3. Setup: Create a set of K mixture weights.
    4. Sample: Draw S Markov chains of varying lengths from the complete model according to the transition probability matrix per component and the mixture weights.

- Construct the Bayesian inference model: Next, we formulate Bayesian inference by calculating the likelihood. This has been covered in the methods section.
- Sample from the posterior probability distribution: Once we derived the likelihood, we have the posterior distribution from which we sample data with the help of MCMC. This has been covered in the methods section.
- Check the results: We check if the results converged, and since we know the parameter values of the synthetic data, it is easy to compare with the parameter estimates. In case of low convergence or bad estimates, we make changes to the prior and take relevant steps to fix the issue.
- Run the model on real data in hand: Once we have decent results, we fit the model to longitudinal data of self-reported data trajectories.

We have already discussed the problem in the context of real data in the main paper.

4.8 Experiments on synthetic data

We generated samples of two-state Markov chains from a mixture of Dirichlet distribution with three components. In Figure 4.6, we can see how the shape parameters were varied for synthesis of 3-component mixture of Markov chains.



(a) Shape parameters for component A

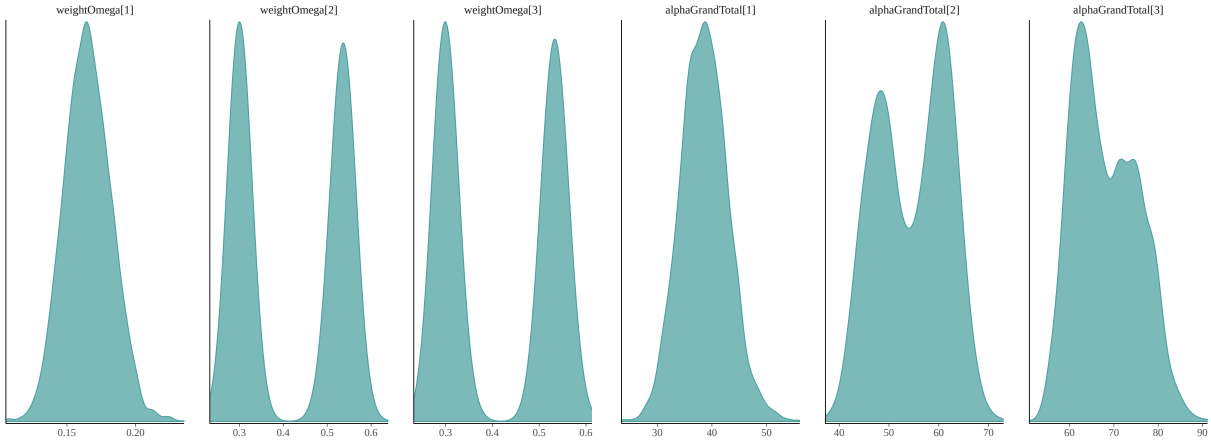
(b) Shape parameters for component B

(c) Shape parameters for component C

Figure 4.6: Synthetic data- shape parameters of Dirichlet distribution

4.8.1 Without specifying any point of initialisation

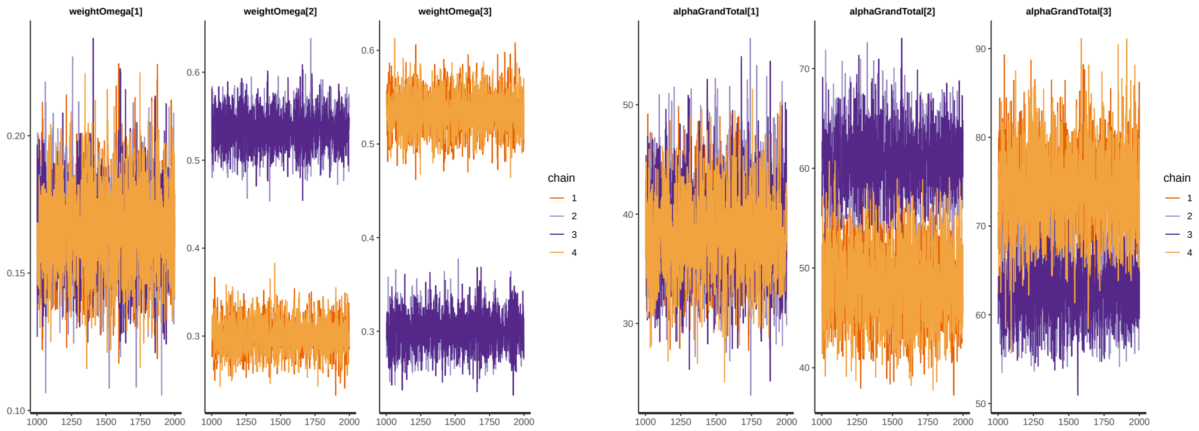
Here, we show how not assigning an initial point for the MCMC lead to no convergence of the Markov chains despite the ordering constraint on parameters. We can notice that the components 2 and 3 get swapped when third and fourth chains are compared.



(a) Mixture weight per component

(b) Alpha grand total per component

Figure 4.7: Density plots of hyper-parameters



(a) Mixture weight per component

(b) Alpha grand total per component

Figure 4.8: Trace plots of hyper-parameters

4.8.2 With an EM based point of initialisation

Here we show how feeding an initial point, which was computed using EM-algorithm, to the model for the MCMC to run on the same synthetic data resulted in convergence of MCMC for parameter estimation. This along with ordering of parameters addressed the problem of label-switching in mixture models as well.

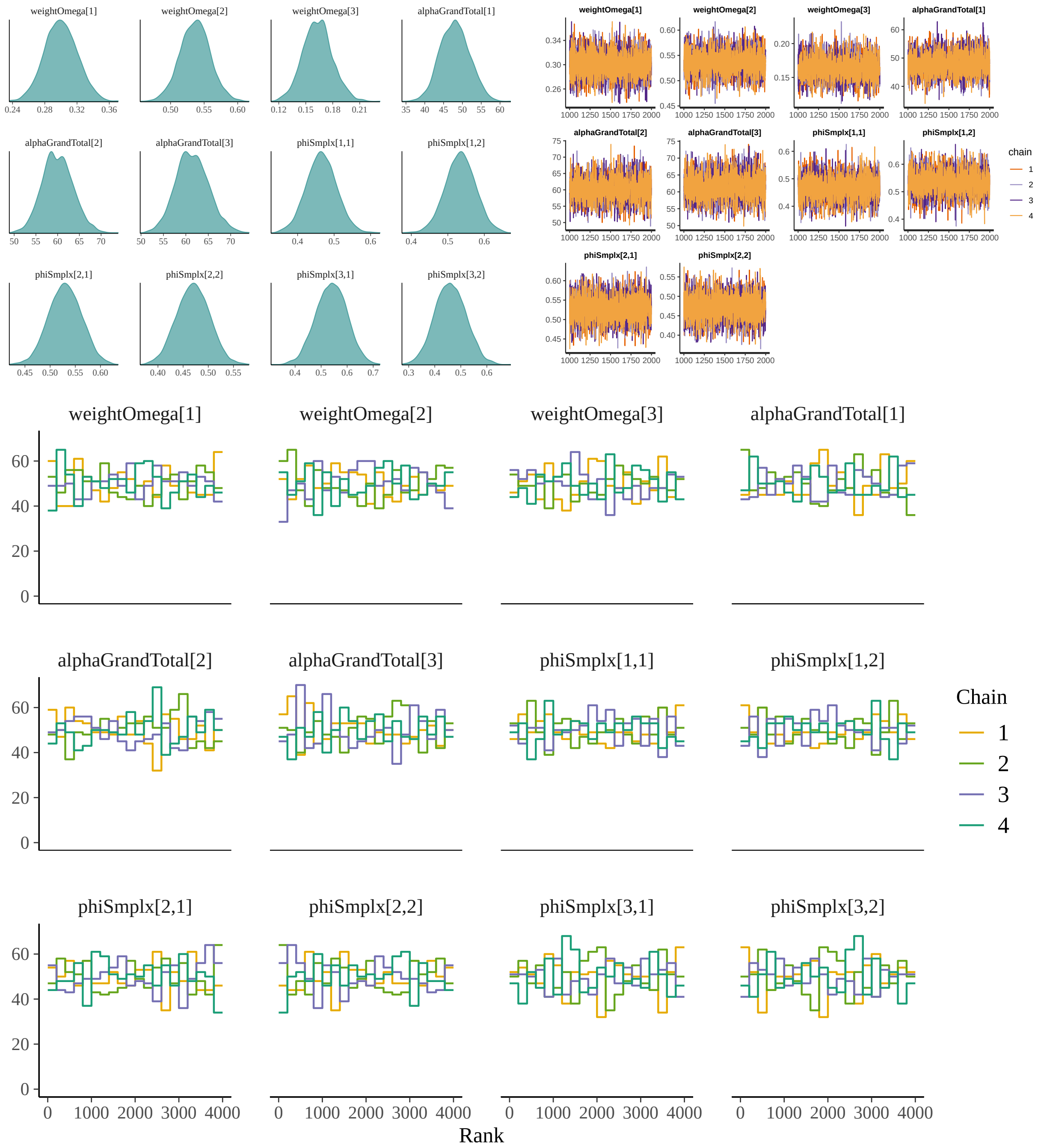


Figure 4.9: MCMC samples of parameter estimates

4.9 Real data specifications

Real data information required for the model is given in the Table 4.1, and uninformative priors taken for the parameters are in the Table 4.2. For $\alpha_{\text{GrandTotal}}$, 4^2 (4 being the size of the transition matrices) is multiplied arbitrarily and any other value could have been chosen.

Data	Value
No. of states n	4
No. of components K	4
No. of subjects S	3720

Table 4.1: Data

Parameter	Prior
ω	$\mathrm{Dir}(1,1,1,1)$
$\alpha_{\mathrm{GrandTotal}}$	$\mathrm{Gamma}(4^2 \times 1.5, 1.5)$
θ_i	$\mathrm{Dir}(1,1,1,1)$

Table 4.2: Priors

4.10 Background of methods

4.10.1 Multinomial distribution and Dirichlet distribution

The multinomial distribution is a distribution for a vector of counts and is parameterised by total number of trials and the probabilities per outcome.

Let y_i be the count of an outcome group or category i and $\boldsymbol{p}$ be a vector of probabilities associated with each outcome level such that p_i is the probability of category i being realised in a single trial and $\sum\limits_{i=1}^{n} p_i = 1$. Let N be the total number of trials and n be the total number of outcomes or categories which is same as the length of the vectors $\boldsymbol{y}$ and $\boldsymbol{p}$. Then $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ follows a multinomial distribution with parameters N and probabilities $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$. Note that vectors are denoted with a bar on them. So, for $\sum\limits_{i=1}^{n} y_i = N$, we get:

$$\begin{aligned}\mathrm{Mult}(\boldsymbol{y} \mid N, \boldsymbol{p}) &= \frac{N!}{y_1! y_2! \ldots y_n!} p_1^{y_1} p_2^{y_2} \ldots p_n^{y_n} \\ &= \frac{\Gamma(\sum_i y_i + 1)}{\prod_i \Gamma(y_i + 1)} \prod_{i=1}^{n} p_i^{y_i}\end{aligned}$$

when $\boldsymbol{y} \sim \mathrm{Mult}(N, \boldsymbol{p})$, then $y_i \sim \mathrm{Binom}(N, p_i)$. Therefore, the multinomial distribution is also known as multivariate binomial distribution.

The Dirichlet distribution is a distribution over the probability vectors $\boldsymbol{p}$. It is commonly used for compositional data. The Dirichlet distribution has a simplex as its sample space that takes the dimension of the data into account, and the distribution is parameterised by α which are known as Dirichlet shape/ concentration parameters. Let $\boldsymbol{p} \in \Delta_{n-1}$ (a simplex of dimension n-1) be a random vector whose elements represent the proportions of items in it therefore, summing up to be 1, then we get the probability distribution to be:

$$\boldsymbol{p} \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_n) = \frac{\Gamma(\sum\limits_{i=1}^{n} \alpha_i)}{\prod\limits_{i=1}^{n} \Gamma(\alpha_i)} \prod_{i=1}^{n} p_i^{\alpha_i - 1} \tag{4.11}$$

where $\sum\limits_{i=1}^{n} p_i = 1$ and $p_i > 0$ for i $\in \{1, \ldots, n\}$.

We know that the integration of any probability distribution over the sample space is 1. So we now integrate the Eqn. (4.11) over the simplex Δ_{n-1} which gives:

$$1 = \int_{\Delta_{n-1}} \frac{\Gamma \sum\limits_{i=1}^{n} \alpha_i}{\prod\limits_{i=1}^{n} \Gamma(\alpha_i)} \prod_{i=1}^{n} p_i^{\alpha_i - 1} \mathrm{d}p$$

$$\Longrightarrow 1 = \frac{\Gamma \sum\limits_{i=1}^{n} \alpha_i}{\prod\limits_{i=1}^{n} \Gamma(\alpha_i)} \int_{\Delta_{n-1}} \prod_{i=1}^{n} p_i^{\alpha_i - 1} \mathrm{d}p$$

Thus, we get the normalising constant of the Dirichlet distribution $\mathcal{B}(\boldsymbol{\alpha})$ as:

$$\mathcal{B}(\boldsymbol{\alpha}) = \int_{\Delta_{n-1}} \prod_{i=1}^{n} p_i^{\alpha_i - 1} \mathrm{d}p = \frac{\prod\limits_{i=1}^{n} \Gamma(\alpha_i)}{\Gamma(\sum\limits_{i=1}^{n} \alpha_i)} \tag{4.12}$$

Now, let us consider we are given a training set of data $\mathcal{D} = \{\boldsymbol{p_1}, \boldsymbol{p_2}, \ldots, \boldsymbol{p_N}\}$ containing N samples of probability vectors each of length n such that $p_{ij} > 0$ and $\sum\limits_{j=1}^{n} p_{ij} = 1$ for $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, n\}$. Then the likelihood is given as:

$$P(\mathcal{D} \mid \alpha) = \log \prod_{k=1}^{N} \mathrm{Dir}(p_i \mid \alpha) = \prod_{k=1}^{N} \frac{\Gamma(\sum\limits_{i=1}^{n} \alpha_i)}{\prod\limits_{i=1}^{n} \Gamma(\alpha_i)} \prod_{i=1}^{n} p_{ki}^{\alpha_i - 1}$$

Here we compute Maximum Likelihood Estimate (MLE) to get estimates for the Dirichlet shape parameters α (Minka, 2000). Therefore, we get the log likelihood as:

$$\log P(\mathcal{D} \mid \alpha) = N \log \Gamma(\sum_{i=1}^{n} \alpha_i) - N \sum_{i=1}^{n} \log \Gamma(\alpha_i) + N \sum_{i=1}^{n} (\alpha_i - 1) \log \tilde{p}_i$$

where $\log \tilde{p}_i = \frac{\sum\limits_{k=1}^{N} p_{ki}}{N}$.

Now we take the derivative with respect to a shape parameter α_i which gives us the gradient as:

$$\frac{\mathrm{d} \log p(\mathcal{D} \mid \alpha)}{\mathrm{d}\alpha_i} = N\Psi(\sum_{i=1}^{n} \alpha_i) - N\Psi(\alpha_i) + N \log \tilde{p}_i$$

where $\Psi_x = \frac{\mathrm{d} \log \Gamma(x)}{\mathrm{d}x}$ is the digamma function.

In an exponential family, when gradient is zero then expected sufficient statistic is equal to observed sufficient statistic which is $\mathrm{E}(\log(p_i)) = \Psi(\alpha_i) - \Psi(\sum \alpha_i)$. Next, an appropriate numerical method can be applied to maximise the likelihood. Example, in the case of fixed-point iteration, we use $\Psi(\alpha_i^{new}) = \Psi(\sum\limits_i \alpha_i^{old}) + \log \tilde{p}_i$. This finally results in obtaining the estimates of the Dirichlet shape parameters α by taking an MLE approach.

We will take a Bayesian approach to estimate the Dirichlet shape parameters in the modelling described in this paper. An important property to remember is that Dirichlet distribution is the prior conjugate of multinomial distribution which means the prior distribution of the parameters of a multinomial distribution is taken to be a Dirichlet distribution, then the posterior distribution is also a Dirichlet distribution with an updated set of parameters.

4.10.2 Dirichlet-multinomial distribution

The Dirichlet-Multinomial distribution is a probability distribution that arises naturally when doing Bayesian inference for Dirichlet models of count data. It is a probability distribution over vectors of counts $\boldsymbol{y} \in \mathbb{N}^n$, but parameterised by the total number of counts $N = \sum_{i=1}^n y_i$ and a vector of shape parameters $\boldsymbol{\alpha} \in \mathbb{R}_+^n$. Its probability mass function is

$$\begin{aligned} P(\boldsymbol{y} \mid N, \boldsymbol{\alpha}) &\equiv \mathrm{DirMult}(\boldsymbol{y} \mid \boldsymbol{\alpha}) \\ &= \left(\frac{N!}{\prod_{i=1}^n y_i!}\right) \left(\frac{\Gamma(\boldsymbol{\alpha})}{\prod_{i=1}^n \Gamma(\alpha_i)}\right) \left(\frac{\prod_{i=1}^n \Gamma(\alpha_i + y_i)}{\Gamma(\boldsymbol{\alpha} + N)}\right) \\ &= \left(\frac{N!}{\prod_{i=1}^n y_i!}\right) \frac{\mathcal{B}(\boldsymbol{\alpha} + \boldsymbol{y})}{\mathcal{B}(\boldsymbol{\alpha})} \end{aligned} \tag{4.13}$$

where, as above, $\boldsymbol{\alpha} = \sum\limits_{i=1}^{n} \alpha_i$ is the sum of the shape parameters and $\mathcal{B}(\boldsymbol{\alpha})$ is the normalising constant as defined in Equation (4.12).

4.10.3 Identifiability, exchangeability and label-switching

Random variables $x_1, \dots x_n$ are *exchangeable* if $\forall$ permutations σ on $1, \dots, n$, the corresponding probability mass or density function follows: $p(x_1, \dots, x_2) = p(x_{\sigma(1)}, \dots, x_{\sigma(n)})$. Exchangeability captures the notion of symmetry amongst random variables of the model. The concept is said to be introduced by de Finetti whose theorem talked about independence and exchangeability (Diaconis and Freedman, 1980).

To explain label-switching, we start by defining a parametric family of finite mixture densities with the following probability density function for random variables $x \in \mathbb{R}^K$ and mixture weights ω such that $\sum_{k=1}^{K} \omega_k = 1$:

$$P(x \mid \theta) = \sum_{k=1}^{K} \omega_k P_k(x \mid \theta_k) \tag{4.14}$$

where P_k is the probability density functions corresponding to parameter θ_k, and let $\phi = (\omega_{1 \leq k \leq K}, \theta_{1 \leq k \leq K})$. This parametric family of probability density functions is *identifiable* if for distinct parameters we get distinct probabilities. In view of (4.14), *identifiability* means that for two sets of parameter pair $\phi^{'} = \{\omega^{'}_{1 \leq k \leq K}, \theta^{'}_{1 \leq k \leq k}\}$ and $\phi^{''} = \{\omega^{''}_{1 \leq k \leq K}, \theta^{''}_{1 \leq k \leq K}\}$, probability densities $p(x \mid \theta^{'})$ and $p(x \mid \theta^{''})$ for almost all $x \in \mathbb{R}^n$ only if $\exists$ permutation $\pi_{1 \leq k \leq K}$ such that $\omega^{'}_k = \omega^{''}_{\pi(k)}$, and if $\omega^{'}_k \neq 0$ then $\theta^{'}_k = \theta^{''}_{\pi(k)}$ for $1 \leq k \leq K$. More details can be found in Redner and Walker (1984).

It often happens in such mixture models that the log of likelihood $L(\phi \mid x)$, given by $\log L(\phi \mid x)= \log \prod_{k=1}^{K} \sum_{k=1}^{K} p(x_k \mid \theta_k) = \sum_{k=1}^{K} \log\Big(\sum_{k=1}^{K} p(x_k \mid \theta_k)\Big)$ can be maximised at multiple different values of ϕ. In such a case, if the component parameters (ω_k, θ_k) and $(\omega_{k'}, \theta_{k'})$ for some $k \neq k'$ are interchanged, the log-likelihood still remains the same. This results in the problem of "label-switching" (Redner and Walker, 1984) which is common when taking a Bayesian approach to parameter estimation of mixture models (Stephens, 2000), and it is often dealt with by imposing identifiability with the help of constraints. In other words, during Bayesian inference, if the priors given do not distinguish between the components of the mixture, then posterior distribution will turn out to be symmetric which causes problem in identifying the components. One of the common approaches in addressing label-switching in mixture modelling includes ordering the parameters.

4.10.4 Initialisation

K-means to estimate an initialisation for EM

To determine a point of initialisation for the Expectation-Maximisation (EM) algorithm, we perform K-means clustering on the centred log-ratio (CLR) transformed data. The following steps are executed:

1. Transform the trajectories to matrices of counts of observed transitions.

2. Perform CLR on each row of the matrix such that for every $i \in 1, \ldots, n$:

$$clr(x_i) = \log x_i - \frac{1}{n} \sum_{j=1}^{n} \log x_{ij}$$

3. Joining all the rows, we get vectors in $\mathbb{R}^{n \times n}$.

4. Apply K-means algorithm for clustering which returns cluster membership probabilities.

5. Fit Dirichlet distribution to the proportions of counts for every ith row of matrix for a subject in a particular component to get the ith vector of Dirichlet shape parameters for the component.

We could have stopped here and utilised the membership probabilities attained here directly to estimate the model parameters. However, when we did so, MCMC did not show convergence for some of the parameters of the Dirichlet-multinomial model. Therefore, we implemented EM to re-estimate the cluster membership probabilities and generate the point of initialisation for MCMC.

Now we re-estimate the Dirichlet shape parameters using EM algorithm as follows:

EM to estimate cluster membership probabilities

1. Recall Eq. 4.9:

$$\begin{aligned} &P(\boldsymbol{y} \mid \boldsymbol{\alpha}) \\ &= \frac{N!}{\prod_{i=1}^{n} y_i} \frac{\mathcal{B}(\boldsymbol{\alpha} + \boldsymbol{y})}{\mathcal{B}(\boldsymbol{\alpha})} \\ &= \mathrm{DirMult}(\boldsymbol{y} \mid \boldsymbol{\alpha}) \end{aligned} \tag{4.15}$$

   , and Eq. 4.10:

$$\mathrm{DirMult}(\boldsymbol{y}_s \mid \boldsymbol{\omega}, \boldsymbol{\alpha}) = \sum_{k=1}^{K} \omega_k \mathrm{DirMult}(\boldsymbol{y}_s \mid \boldsymbol{\alpha}_k) \tag{4.16}$$

2. Take log of Eq. 4.15, we get:

$$
\begin{aligned}
&\log L \\
&\quad = \log P(\boldsymbol{y} \mid \boldsymbol{\alpha}) \\
&\quad = \log N! - \log \prod_{i=1}^{n} y_i + \log \mathcal{B}(\boldsymbol{\alpha} + \boldsymbol{y}) - \log \mathcal{B}(\boldsymbol{\alpha}) \\
&\quad = \log \Gamma(\sum_{i=1}^{n} y_i + 1) - \sum_{i=1}^{n} \log \Gamma(y_i + 1) + \log \mathcal{B}(\boldsymbol{\alpha} + \boldsymbol{y}) - \log \mathcal{B}(\boldsymbol{\alpha})
\end{aligned}
$$

3. Computing the gradient by differentiating log likelihood with respect to the shape parameters:
$$\frac{\partial \log L}{\partial \boldsymbol{\alpha}} = \psi(\mathcal{B}(\boldsymbol{\alpha} + \boldsymbol{y})) - \psi(\mathcal{B}(\boldsymbol{\alpha})) \tag{4.17}$$
where $\psi(.)$ is digamma function which is derivative of Gamma function.

4. The log likelihood is maximised using numerical methods. Class-memberships are re-estimated till convergence of the algorithm. These are further used to estimate mixture weights and Dirichlet shape parameters.

Chapter 5

Dimensionality reduction on self-reported longitudinal data

Rajenki Das[1], Mark Muldoon[1], Thomas House[1]

1 – Department of Mathematics, University of Manchester, Manchester, UK

Abstract

Here we perform unsupervised learning techniques to cluster and reduce the dimension of data containing ten self-reported variables in a longitudinal data. We observe more or less consistent grouping of these variables across methods.

5.1 Introduction

Health has been often associated with physical health. Absence of an identifiable ailment often makes us conclude that a person is healthy. However, WHO defines health as "complete physical, mental and social well-being and not merely the absence of disease or infirmity" (World Health Organization et al., 1948) which looks at health at a holistic level. Good health can be characterised by adequate physical and mental health, and also physiological well-being which can bring a sense of purpose in life, enhance relationships with others and realise one's potential (Ryff and Singer, 1998). A healthy person may experience "well-being" which can improve functioning of biological system thereby preventing a person from succumbing to a disease, and in the case of an illness, it can promote rapid recovery hence potentially forming a cycle of positive health (Ryff et al., 2004). "Wellbeing" is a very subjective concept which is associated with happiness and life satisfaction (Diener, 2009), and gives importance to parameters affecting health other than the physiology of it. The WHO-5 (Topp et al., 2015) is a widely used questionnaire that helps in measuring the well-being as well as acts as a screening tool for depression thereby reflecting on the mental health.

Narrowing down to mental health, mood disorders are quite common and very prominent in those having mental health issues. Sleep is essential in maintaining mental and physical health, and can help regulate poor moods and emotions. It has been shown that anxiety and depression, which are indicative of mental health, are related with sleep quality (João et al., 2018). Its relationship with depression and other mental health disorders is complex (Fang et al., 2019), but sleep remains an indicator of mental health. Another parameter to consider while looking at mental health, and especially could be useful in building a survey is waking up tired which is associated with sleep can reflect on the mental health (Palmer, 2020; Appels and Schouten, 1991). Similarly, fatigue is another indicator which can tell how healthy one is. Fatigue is a complex concept though, as it there can be mental or/and physical fatigue (Lee and Giuliani, 2019; Rosenthal et al., 2008), nevertheless it remains an interesting variable to consider. Mental health particularly is very complex which can be affected by numerous factors and disentangling those remains a challenge.

Other than the symptoms talked till now, lifestyle based indicators like exercise and time spent outside can be measured against other parameters to see how health varies.

When talking about health, an important aspect of physical health is the musculoskeletal system of the human. Experiencing pain can affect our day to day lives and can impact negatively by interfering with our routine. For example, morning stiffness- a general increase in musculoskeletal symptoms in the morning, is a common trait in people with rheumatoid arthiritis, which is the most common type of poly-arthirtis or other chronic pain conditions. In fact, pain and mood are linked as we can find in Das et al. (2023) and their associations can help us understand health. Chronic pain has affected approximately 20 % of the population in the USA and Europe (Breivik et al., 2006) and prevails to be a global problem.

Studying health, especially mental health, is very complex problem. Usually there are several factors that affect the health and considering all these can be a difficult problem. However, interactive media is proving to have substantial impact in the field of healthcare as initially noted by Frank (2000). Digital health especially is vital in capturing granularity of individual behaviour within and amongst individuals.With the help of electronic health devices, it becomes easier to record information on a wide range of aspects pertaining to different aspects of human being. We can get data on age, sex, economic conditions, physical ailments etc. at a place and also perform longitudinal studies to track health data and other relevant symptoms. But the wide availability of data can be an overload both on user and developer ends. In such scenarios, statistical tools like feature extraction and dimensionality reduction can come into aid and help in selecting data which captures most information.

Here, we present an overall view of Cloudy with a Chance of Pain data which consists of categorical ordinal data information on some self-reported severities of 10 variables: mood, pain severity, impact of pain (on daily activity), physical activity, time spent outside, fatigue, sleep quality, morning stiffness, waking up tired.

Dealing with complex datasets with several features often requires the need to decrease the size of the dataset. Dimensionality reduction is the method of transforming high-dimensional data by reducing the number of random variables of the problem in consideration (Roweis and Saul, 2000). It can help in identifying principal variables and getting rid of redundant variables. Hence, it is vital in classification and visualisation, amongst many other applications. It is often combined with data processing steps to get the data in a more understandable format. Since the given dataset is multivariate, eventually, there is a need to adopt specific dimensionality reduction techniques.

We apply few dimensionality reduction methods to understand the relationship amongst the self-reported variables. Methods used are:

1. Principal Component Analysis (PCA)
2. Hierarchical Clustering
3. Independent Component Analysis (ICA)
4. Logistic Principal Component Analysis (LPCA)

Code for these methods will be made available at: `https://github.com/rajenkidas/`

5.2 Data

We use data from the Cloudy with a Chance of Pain study (Reade et al., 2017; Dixon et al., 2019), which was conducted to investigate the relationship between weather and pain, but in doing so created an extremely rich dataset suitable to answer a diversity of research questions. Data were collected from January 2016 to April 2017 from participants resident in the UK who were aged 17 or above and had experienced chronic pain for at least 3 months preceding the survey (Druce et al., 2017).

The cohort had 10,584 survey participants, each of whom was asked to rate their symptoms and other variables on a mobile application in five ordinal categories (e.g. pain scores ranged from 1 for no pain to 5 for very severe pain). Data were recorded for pain interference, sleep quality, time spent outside, tiredness, activity, mood, well-being, pain severity, fatigue severity and stiffness on a daily basis.

For the applications of PCA, ICA and hierarchical clustering, we perform dimensionality reduction on the self-reported symptoms, while for LPCA, we have converted the data to binary on the basis of the change of scores of each of the variables. We have loosely regrouped into two categories: 1) if the severity has increased from the previous day's, 2) if the severity has decreased. More have been discussed in the corresponding sections.

5.3 Method

We have s participants with with n observations in total that contain ordinal measurements on a set of d features which we call symptoms here. We have $d = 10$ symptoms viz: Fatigue, Mood, Morning stiffness, Pain impact, Pain severity, Well-being, Exercise, Sleep quality, Time spent outside, Waking up tired. We perform unsupervised learning techniques on these data for a better understanding of the features.

5.3.1 Principal Component Analysis

Principal components are representative variables that explain most of the variability in a given dataset. Principal Component Analysis (PCA) is a linear dimensionality reduction method used to find the principal components representing the data in which the they get embedded to a linear subspace of lower dimension. The low dimension representation describes data by maximising variance.

As we have n observations and d features, then each of the new dimension or principal component is a linear combination of the d features. The first PC of the features has the largest variance and can be given by: $Z_1 = \sum_{i=1}^{d} \phi_{i1} X_i$ where $\sum_{i=1}^{d} \phi_{i1}^2 = 1$. Here the coefficients ϕs are called loadings of the first principal component. These tell us how much a feature contributes towards the specific principal component, and the sign of a loading indicate if a feature is positive or negatively correlated with the component. Hence the loading vector for a principal component j is $\phi_j = (\phi_{1j}, \phi_{2j} \ldots, \phi_{dj})^T$ with the constraint that the sum of squares of loadings is 1, to prevent the variances from being arbitrarily large. The total variance amongst all the principal components is same as the total variance amongst the features, so the there is no loss in information. PCA simply rotates the data and gives new set of orthogonal vectors. New data Y is a result of the transformation $Y = X\phi$. We assume that the data is centred at zero, *i.e.* the mean of each of the features is zero. Keeping that in mind, finding PCA components becomes a problem of maximising variance or minimising the mean squared residuals.

5.3.2 Hierarchical clustering

Hierarchical clustering is a unsupervised learning technique that groups data to create a tree-like/ hierarchical structure with branches separating out the features. It builds a binary tree *i.e.* every node has at most two branches. There are two major types of hierarchical clustering: 1) Agglomerative clustering- it starts with the assumption that every feature is its own cluster and the method progresses by combining the features to form clusters till all features are connected. 2) Divisive clustering- it begins by considering all the features belong to one cluster and carries on by breaking down into clusters till every feature is allocated a group. The clusters are often represented in the form of tree like structures called dendrograms.

Let $X = x_1, \ldots, x_n$ be the dataset with cardinality n. Let there be K clusters represented by $C_{1 \leq k \leq K}$, then $\bigcup\limits_{1 \leq k \leq K} C_k = X$ The generic hierarchical clustering algorithm is as follows (Nielsen, 2016):

- Initialise by putting each data point $x_i \in X$ into its own cluster C *i.e* $x_i \in C_i$.
- Compute the distances between two data points and select the pair with the least distance and merge the points' clusters.
- Calculate the new pairwise inter-cluster distances for the remaining clusters.

In the case of this paper, we have a dataset d features for n observations. We perform agglomerative hierarchical clustering by taking Manhattan distance with complete linkage. Manhattan distance between two points $x(x_1, x_2)$ and $y(y_1, y_2)$ in 2-dimensional space is given by $dist(x, y) = |x_1 - y_1| + |x_2 - y_2|$. Now the distance between the clusters, graphically represented by the height of the link between two clusters in a dendrogram, is determined by the linkage specified. The complete linkage between cluster A and B is defined by $dist(A, B) = \max\limits_{x \in A, y \in B} dist(x, y)$.

5.3.3 Independent Component Analysis

Independent Component Analysis (ICA), used in image and signal processing, helps to differentiate independent sources from a mixed signal. ICA of a random vector includes finding the linear transformation which minimises the statistical dependence between its components (Comon, 1994). It is a generative model that describes how the observed data are produced by mixing the components (Hyvärinen and Oja, 2000). Let $x = (x_1, \ldots, x_n)$ be the vector of observations and $s = (s_1 \ldots s_d)$ be the vector of latent variables called independent components. An unknown constant matrix $\mathbf{A}$ is called the mixing matrix. Note that every bold capital letter is a matrix and bold small letter is for a column vector. Then, every observation x_i is a linear combination of d independent components giving us x_i= $a_{i1}s_1 + a_{i2}s_2 + \cdots + a_{id}s_d \, \forall \, 1 \leq i \leq n$. Assuming no noise in the model, the ICA can be expressed as (Hyvarinen, 1999):

$$\mathbf{x} = \mathbf{A}\mathbf{s} = \sum_{i=1}^{d} \mathbf{a_i} s_i \tag{5.1}$$

where $\mathbf{x}$ has the basis vector $\mathbf{a_i} = \{a_{1i}, \ldots, a_{ni}\}^{\mathrm{T}}$. Equation (5.1) represents ICA model where the independent components are non-Gaussian. The goal of the problem can be regarded as maximising the *non-Gaussianity* of the independent components.

The non-Gaussanity can be maximised in several ways. In this paper, we have used FastICA algorithm which measures non-Gaussanity by approximating negentropy J defined for entropy H as (Hyvärinen and Oja, 2000):

$$J(\mathbf{y}) = H(\mathbf{y}_{\mathrm{Gauss}}) - H(\mathbf{y})$$

where $\mathbf{y}_{\mathrm{Gauss}}$ is a Gaussian random variable with covariance matrix equal to that of $\mathbf{y}$.

5.3.4 Logistic Principal Component Analysis

Logistic Principal Component Analysis (LPCA) is an extension of PCA by making it more suitable for binary data. LPCA is based on Bernoulli model as described by Landgraf and Lee (2020). Let $\mathbf{X}$ be data matrix of size $n \times d$ such that each of its element x_{ij} is binary which is assumed to be withdrawn from Bernoulli distribution *i.e.* x_{ij} $\sim$ Bernoulli(p$_{ij}$). The natural parameter θ_{ij} = logit(p_{ij}) for the Bernoulli distribution describes the saturated model ($p_{ij} = x_{ij}$). To perform the equivalent PCA to the binary data, we instead minimise the Bernoulli deviance - this is done by taking the natural parameters of the saturated model and projecting them on a d-dimensional space. It can be said that the classical PCA is extended to logsitic PCA analogous to linear regression being extended to logistic linear regression (Song et al., 2020). We used the package *logisticPCA* (Landgraf and Lee, 2015) to perform LPCA on the binary data representing changes in the self-reported symptoms.

5.4 Results

Looking at the scree plot in Figure 5.7, the optimal number of principal components for PCA is taken to be 3.

5.4.1 Principal Component Analysis

In Figure 5.1, we look at the barplot of loadings for the first three principal components. The colours are based on the signs of the loadings. In PC1, we see how five of the symptoms have positive loadings, while the others have have negative values on the component. Projection onto PC1 gives an understanding of the general wellbeing based on the symptoms, where positive is good while negative has the opposite meaning. In PC2, Exercise and Time spent outside show high loadings which could imply that they vary the most irrespective of other symptoms' behaviour. It hints at the independent behaviour of the factors exercise and time spent outside. In PC3, Sleep quality has the highest loadings, and Exercise, Time spent outside and Waking up tired are the only symptoms with positive loadings which could imply that these three factors are related.

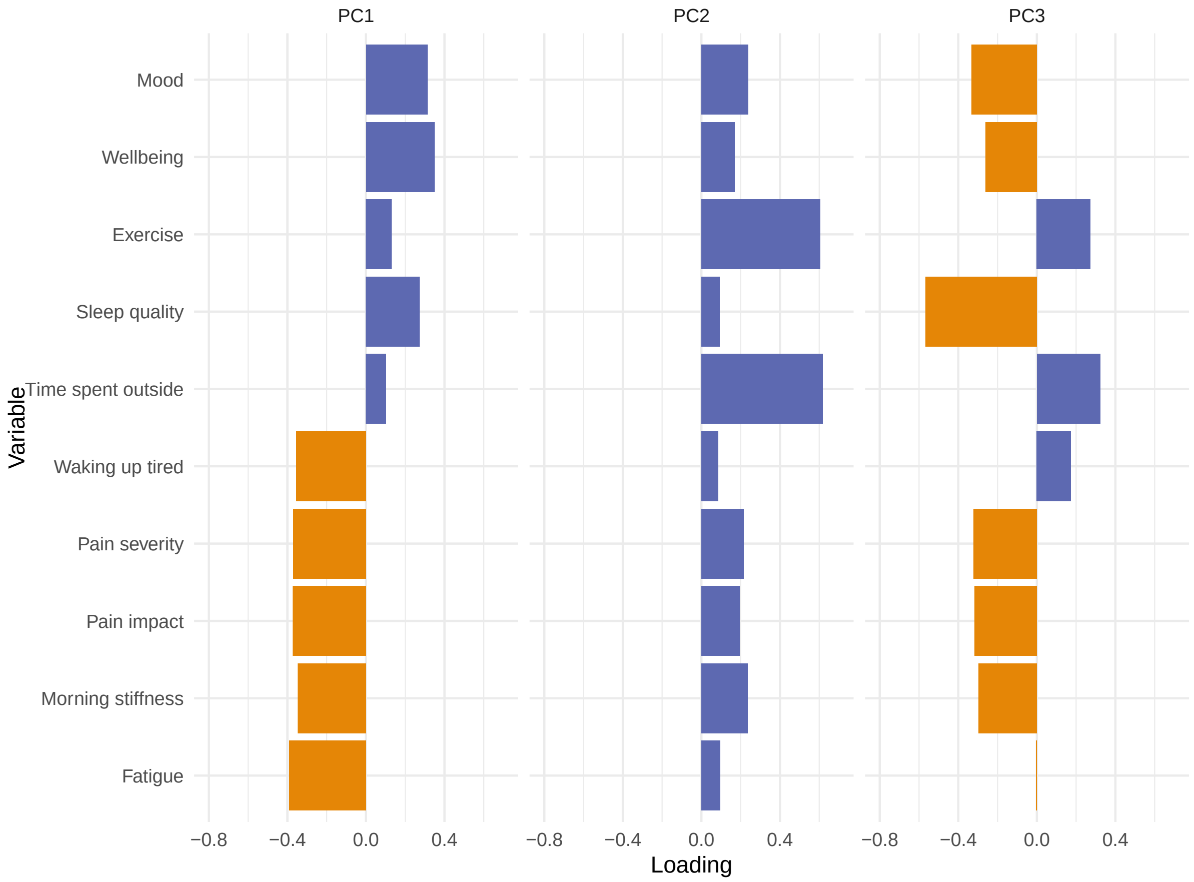


Figure 5.1: Loadings per principal component

In Figure 5.2 showing the biplot for the first two principal components, the loading plot is overlaid on the kernel density representing the distribution of scores (participants) in the background. The direction of the loadings shows how the symptoms are correlated amongst each other. Two symptoms in the opposite directions are negatively correlated while the ones perpendicular can be said to be unrelated to the other symptoms. Additional way of looking at it is by considering the acute angle between any two arrows. Lesser the angle, more related the symptoms are, and the orientation helps in understanding the direction of the correlation. We also see a higher density of scores around PC1 = −2 and PC2 = 0. But there are no clear clusters of scores.

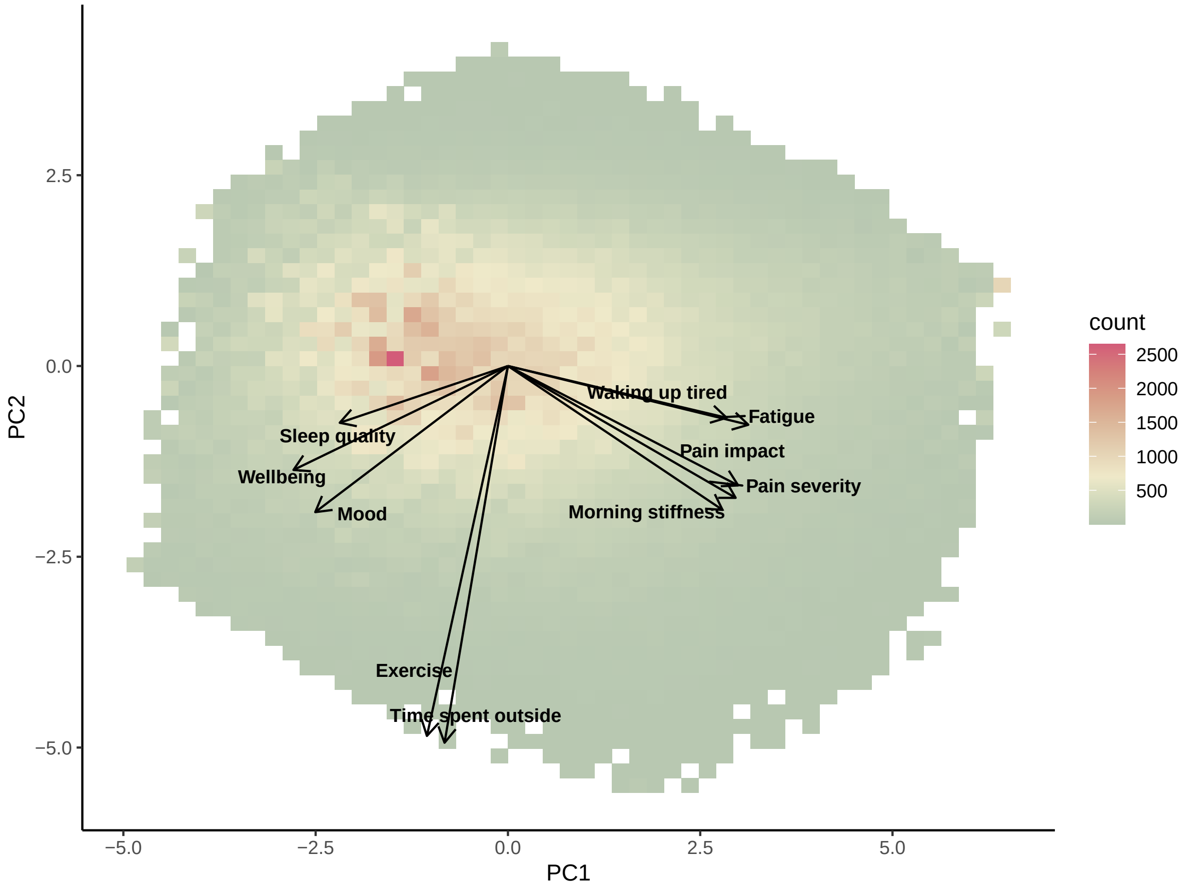


Figure 5.2: Bi-plot

5.4.2 Hierarchical clustering

The heatmap for the distance matrix and the dendrogram representing hierarchical clustering on the basis of Manhattan distance is given in Figure 5.3. The heatmap is sliced by considering three clusters. So we get the groups of symptoms as: 1) sleep, mood and wellbeing, 2) exercise and time spent out, and 3) fatigue, waking up tired, morning stiffness, pain impact and its severity.

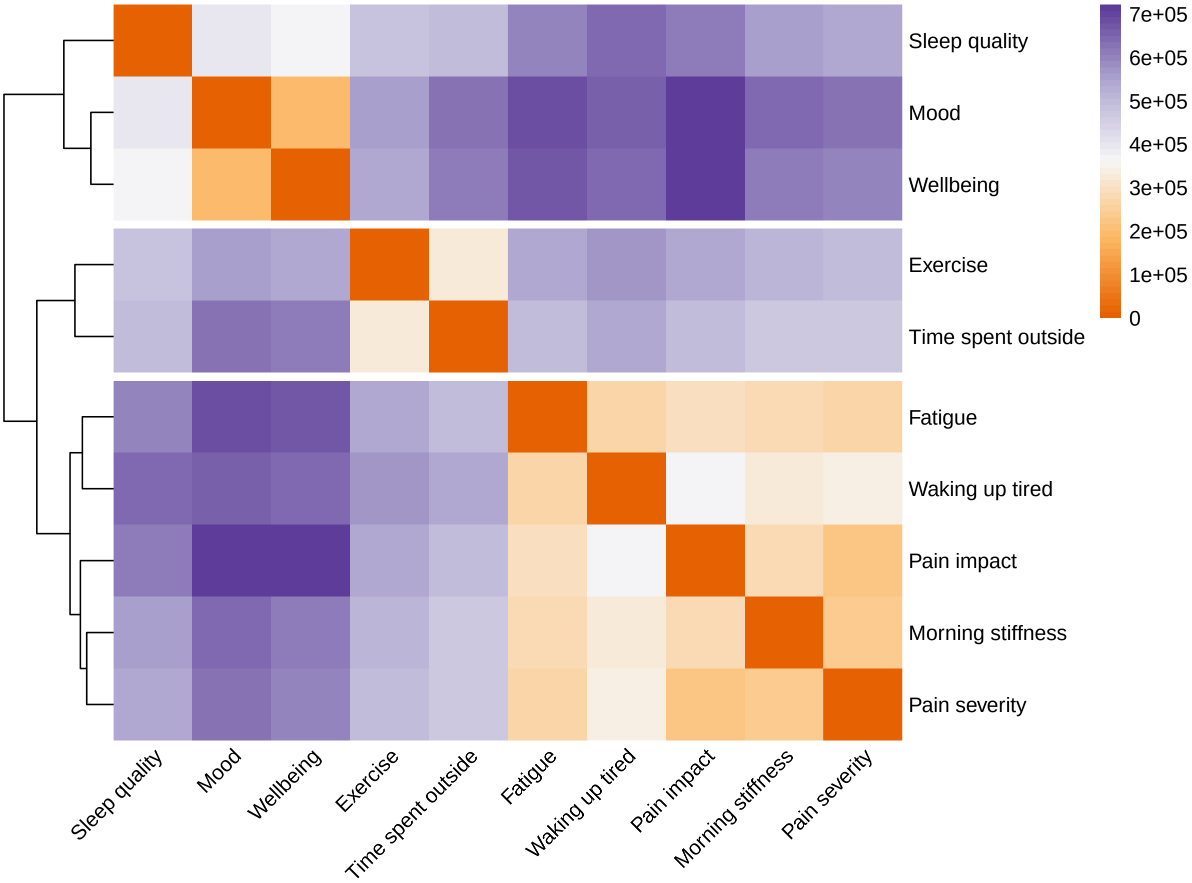


Figure 5.3: Distance matrix- Manhattan with Complete linkage dendrogram

5.4.3 Independent Component Analysis

We look at the squares of the loadings in Figure 5.4. Here the direction of the loadings are ignored. We only focus on the magnitude of the loadings. In PC1, we see Mood, Pain severity and Fatigue capture the maximum variance of the data, while Sleep quality and Well-being have least loadings indicating low information about the data. in PC2, most of the variables have low loadings other than Exercise, Pain severity and then Waking up tired and Pain impact. In PC3, all loadings are less than 0.25 except Well-being and Sleep quality which had low loadings for the previous components. So it can be said that PC3 is most represented by well-being and sleep quality.

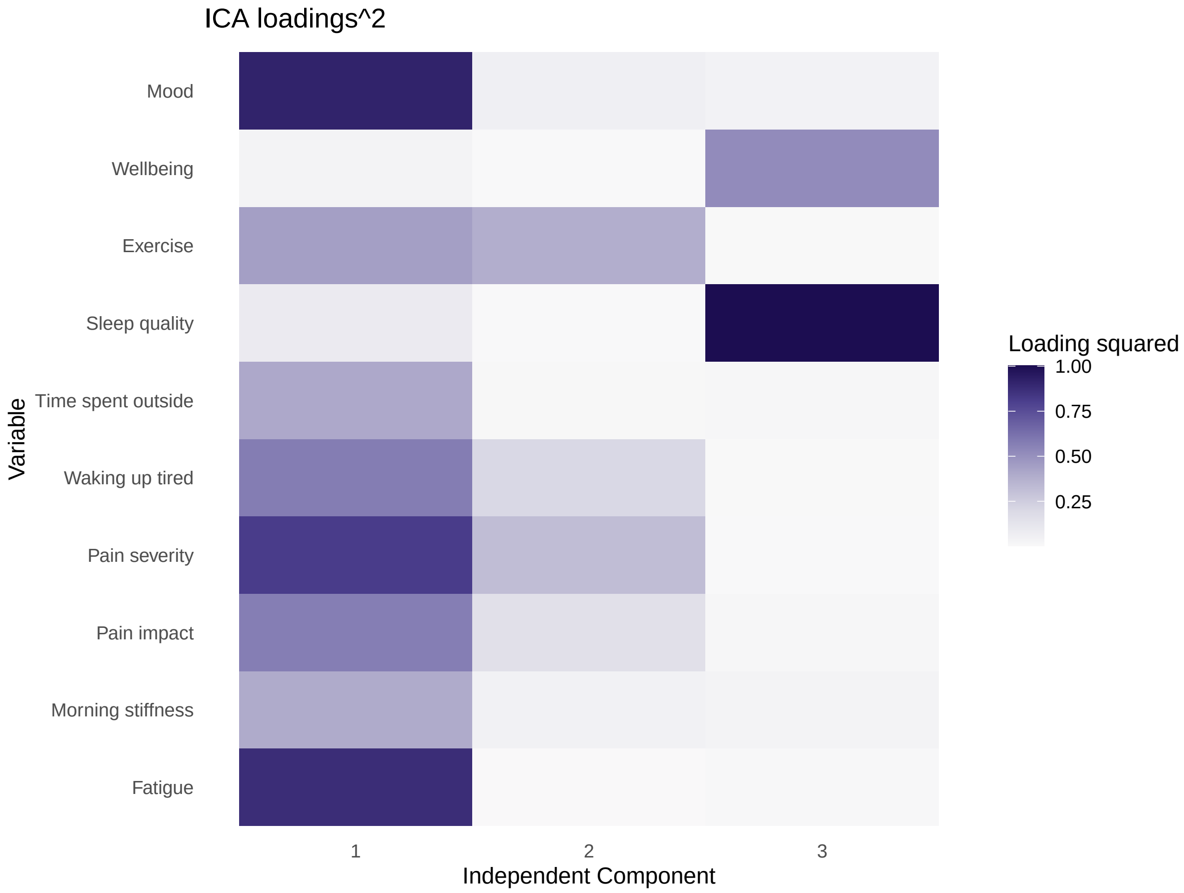


Figure 5.4: ICA commonalities

5.4.4 Logistic Principal Component Analysis

LPCA was performed on the changes of symptoms between two reports. So only positive and negative changes of the scores of the variables were retained for the application of this analysis. In Figure 5.5, we find the barplot for the loadings for the three principal components of the LPCA. The results are quite similar to what we have found in Figure 5.1.

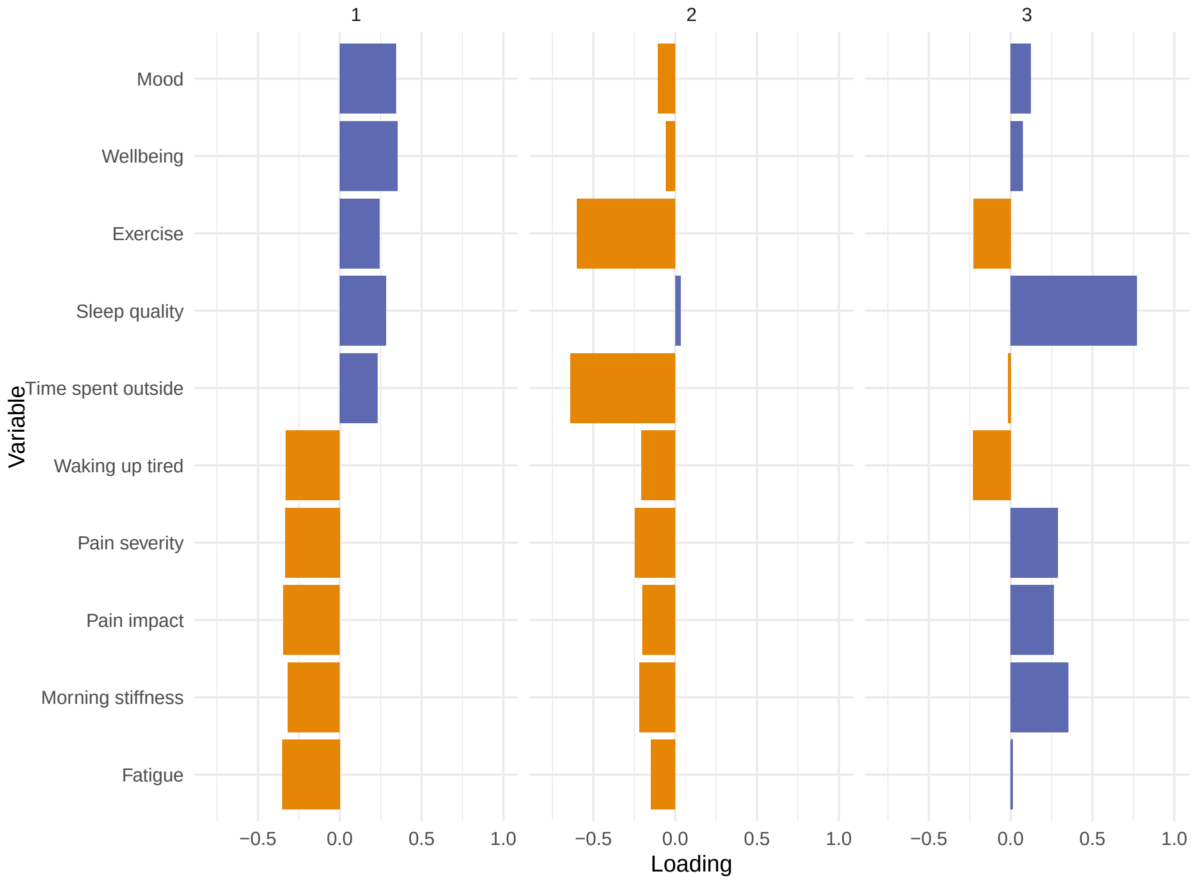


Figure 5.5: LPCA loadings based on the changes of symptoms

In Figure 5.6, the biplot containing the loadings of the principal components on top of scores of the samples is shown. This figure is also very similar to what we have observed before in Figure 5.2. The scanty scores in the background are due to the consideration of retaining only those data points which have positive and negative score changes, hence it reduced the size of the data here.

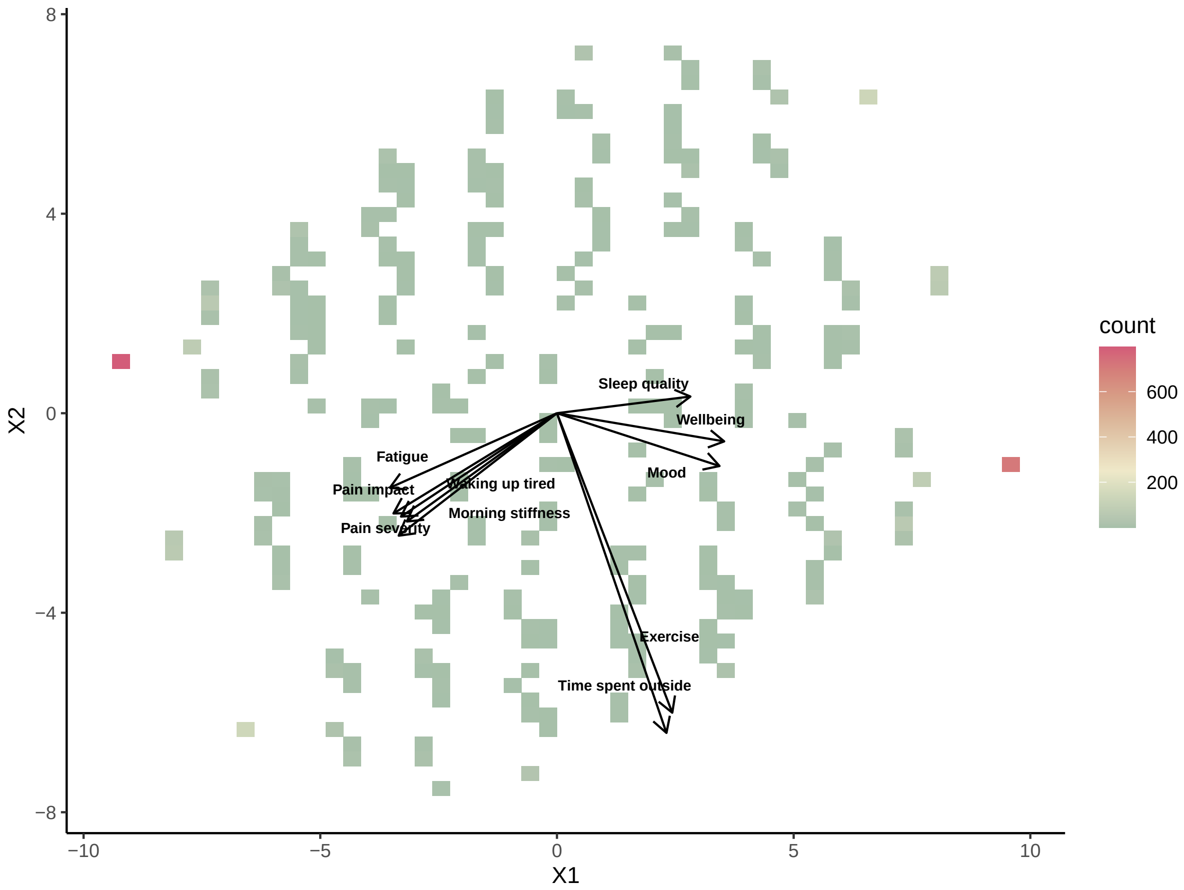


Figure 5.6: Biplot of LPCA loadings and scores based on the changes of symptoms

5.5 Discussion

Conducting Principal Component Analysis (PCA) on symptoms and Logistica Principal Component Analysis (LPCA) on the changes on the scores of symptoms, we found similar results. We found three groups of the features which seem to be inter-related:
1) sleep quality, well being, mood 2) fatigue, time spent outside 3) pain severity, pain impact, exercise, waking up tired, morning stiffness.

In this grouping, we find that the symptoms of groups 1 and 2 appear to be negatively correlated while those of group 2 are almost orthogonal to the vectors of the features belonging to the other groups. This tells us that the features 'fatigue' and 'time spent outside' are independent of the remaining features but closely correlated within themselves. This grouping was further re-established by hierarchical clustering.

In the case of Independent Component Analysis (ICA), 'wellbeing' and 'sleep quality' differentiate the first and third components where in the first one, they capture the least of the variance while in the third, they express the maximum variances in the data.

Few of the limitations of this study are: 1) as we are given longitudinal data, the results may have some biases according to those participants of the studies who recorded their values the most. 2) Existence of other inherent biases based on the population of this data may hinder in generalisation these results at a bigger scale for the overall public.

In this study, we attempted to present longitudinal data visually. We clustered the symptoms with dimensionality reduction related techniques. It gives an overview of the structure of the data by telling us the relationship amongst the variables, and can be extended to including more features and parameters for deeper understanding of health and lifestyle. It needs to be remembered that dimensionality reduction includes information loss - so a feature might get less priority in general, but for a specific individual or subject, that certain feature may have the most importance. This paper highlights an alternative way of dealing with digital health data, which can be standardised along with the traditional practices. More similar datasets can be considered in the study for comparison. More methods can also be included.

Also, we do emphasise on the findings of this report. Biplots of the first two components of PCA and LPCA, and the dendrograms of hierarchical clustering show consistency in the grouping of the symptoms. In case of lesser feasibility of building a study and availing of resources, developing similar studies asking for lesser information can be enough in understanding mental health or related matters. Suppose, we are given a clinical population of those suffering with insomnia or having other sleep problems, we can correlate it with mood and build treatments and additional research problems accordingly. This is simply an alternative use that we are suggesting, not a solution to observational data or other clinical real world based problems as there are biases involved and generalisation often becomes tough.

Supplementary Material

5.6 Materials and Methods

This is a secondary data analysis of data collected by Cloud from the residents in the United Kingdom from January 2016 to April 2017. In this study, we extracted data about ten self-reported categorical variables. Details are in the table below. Dataset was reduced to exclude missing values in the categorical variables. No other exclusions were made based on age or other characteristics.

	Fatigue	Mood	Morning stiffness	Pain impact	Pain severity
Min.	1.0	1.0	1.0	1.0	1.0
1st Qu.	2.0	3.0	2.0	2.0	2.0
Median	3.0	4.0	3.0	2.0	3.0
Mean	2.6	3.6	2.7	2.5	2.7
3rd Qu.	3.0	4.0	3.0	3.0	3.0
Max.	5.0	5.0	5.0	5.0	5.0
NA's	345136	344784	349760	346624	341850

	Patient wellbeing	Exercise	Sleep quality	Time spent outside	Waking up tired
Min.	1.0	1.0	1.0	1.0	1.0
1st Qu.	3.0	2.0	2.0	2.0	2.0
Median	4.0	3.0	3	2.0	3.0
Mean	3.5	2.5	3	2.2	2.8
3rd Qu.	4.0	3.0	4	3.0	4.0
Max.	5.0	5.0	5.0	5.0	5.0
NA's	345584	348061	351230	348076	350899

Table 5.2: Summary statistics of the self-reported symptoms

Variable	Description	Scale
Mood	How was your mood today?	1 = Depressed, 2 = Feeling low, 3 = Not very happy, 4 = Quite happy, 5 = Very happy
Well-being	How well did you feel today?	1 = Very unwell, 2 = Quite unwell, 3 = Unwell, 4 = Well, 5 = Very well
Pain severity	How severe was your pain today?	1 = No pain, 2 = Mild Pain, 3 = Moderate pain, 4 = Severe pain, 5 = Very severe pain
Fatigue	How severe was your fatigue today?	1 = No fatigue, 2 = Mild fatigue, 3 = Moderate fatigue, 4 = Severe fatigue, 5 = Very severe fatigue
Morning stiffness	How stiff did you feel on waking this morning?	1 = No stiffness, 2 = A little Stiff, 3 = Moderately stiff, 4 = Severe stiff, 5 = Very severe stiff
Pain impact	Has your pain interfered with your activities today?	1 = Not at all, 2 = A little bit, 3 = Somewhat, 4 = Quite a bit, 5 = Very much
Sleep quality	How was your sleep quality last night?	1 = Very poor, 2 = Poor, 3 = Fair, 4 = Good, 5 = Very good
Time spent outside	How much time have you spent outside today?	1 = None of the day, 2 = Some of the day, 3 = Half of the day, 4 = Most of the day, 5 = All of the day
Feeling tired	How did you feel when you woke this morning?	1 = Not at all tired, 2 = A little bit tired, 3 = Moderately tired, 4 = Quite a bit tired, 5 = Extremely tired
Exercise	How long have you exercised today?	1 = No exercise, 2 = Less than 30 minutes of light activity, 3 = 30+ minute light activity, 4 = Less than 30 minute strenuous activity, 5 = 30+ minute strenuous activity

Table 5.3: Description of self-reported symptoms

5.7 Model selection

5.7.1 PCA

Figure 5.7 gives the screeplot to choose an optimal number of principal components. We chose 3.

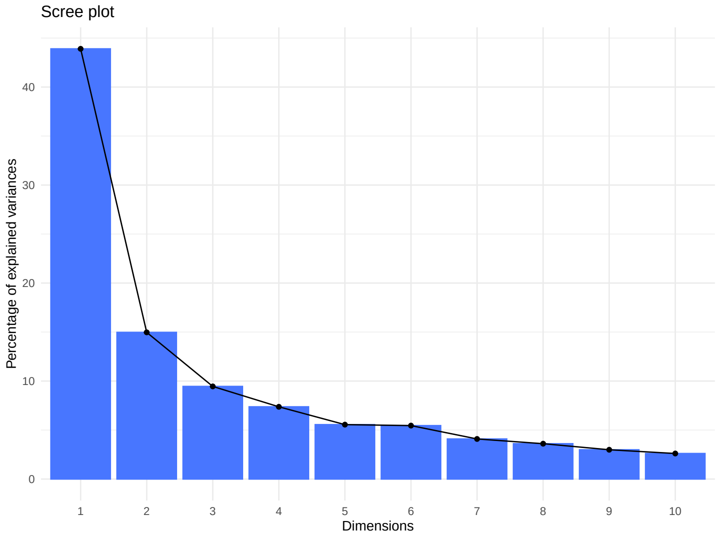


Figure 5.7: Scree Plot- Variance

5.7.2 ICA

R package by Nordhausen et al. (2022) tests for the number of Gaussian components using Fourth Order Blind Identification. We additionally look at it to select the number of independent components which are supposed to be non-Gaussian. The bars of Figure 5.8 are in the order of non-Gaussianity.

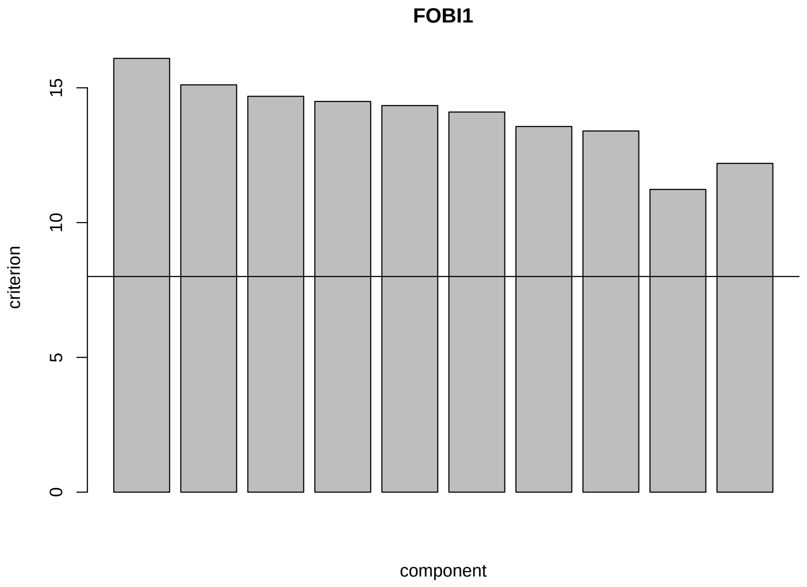


Figure 5.8: ICA screeplot

5.7.3 LPCA

Figure 5.9 shows different k values for different dimensions and m is for approximations to the saturated model. We are more interested in k as that helps in deciding the number of principal components. We took $k = 3$ and $m = 3$.

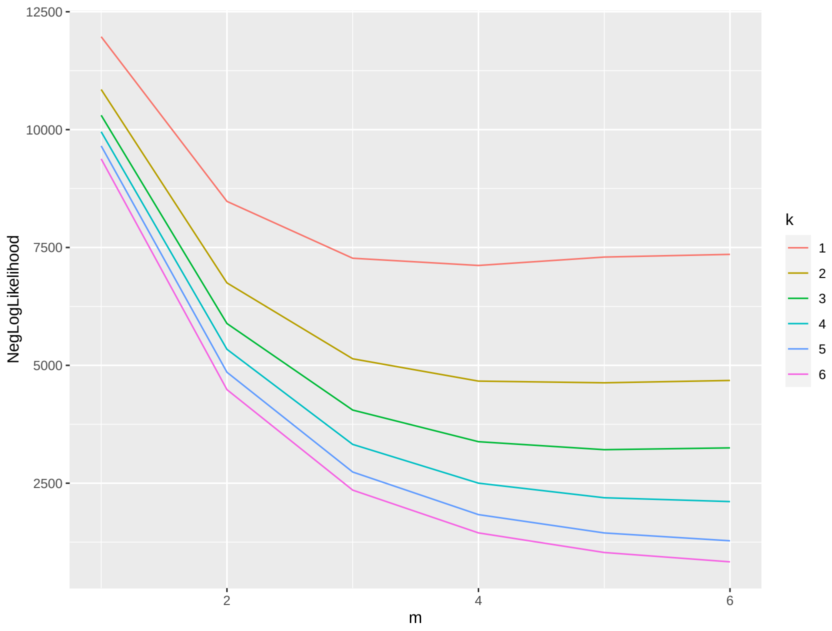


Figure 5.9: LPCA model selection

5.8 Extras

In this section, we have provided alternative plots to few of the figures in main text for more clarity. In Figure 5.10, we can see the loading vectors along with the contribution of each of those.

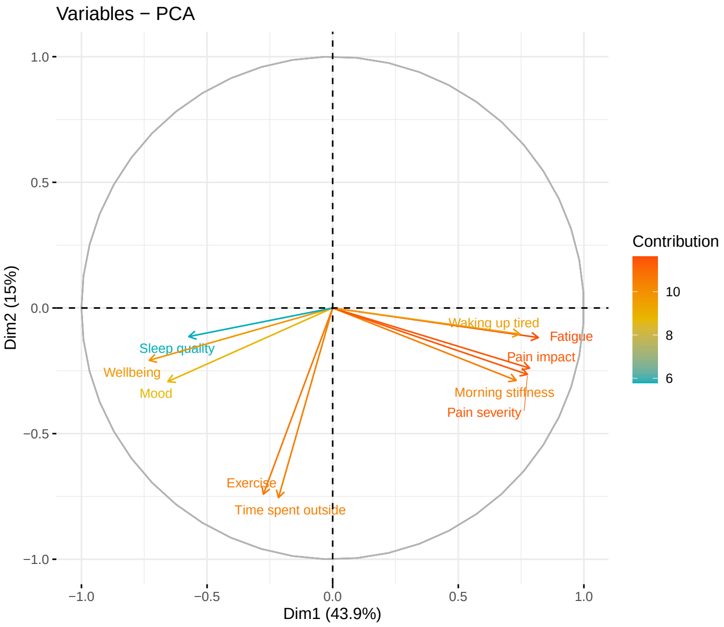


Figure 5.10: Contributions to principal components

Figure 5.11 shows the barplot of loadings for 10, same as total number of self-reported symptoms, principal components. The colour of each symptom is consistent throughout the components.

Figures 5.12a. 5.12b and 5.12c, we find the heatmaps of the loadings for first three principal components for each of PCA, ICA and LPCA. Please note that LPCA was performed on the changes of symptoms, while the other two were on symptoms as it is.

Figure 5.13 is the correlation plot for all the ten self-reported symptoms.

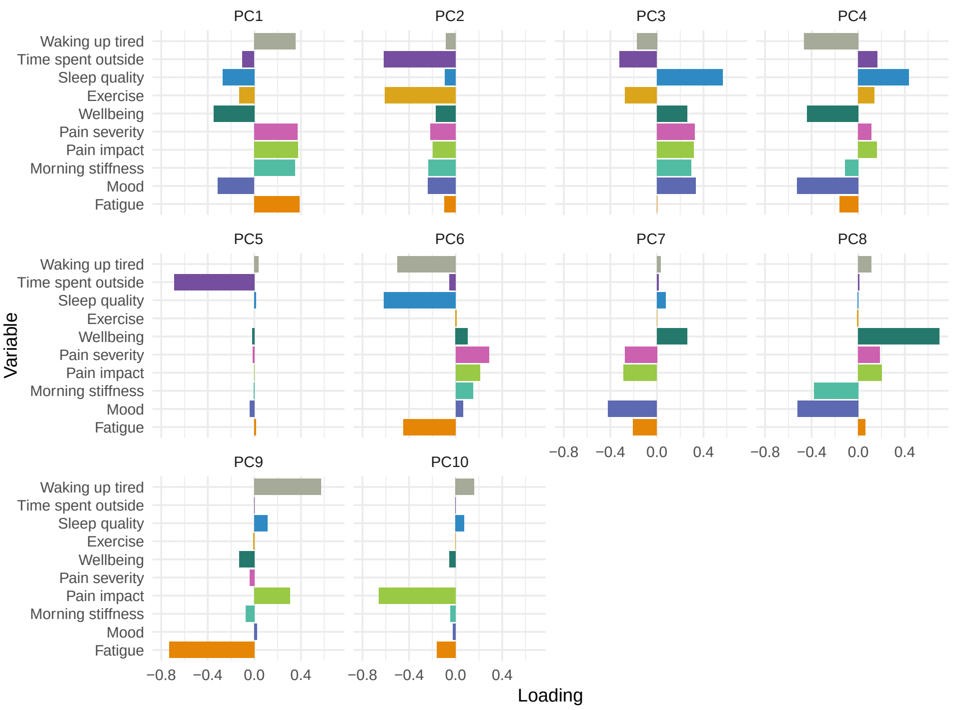
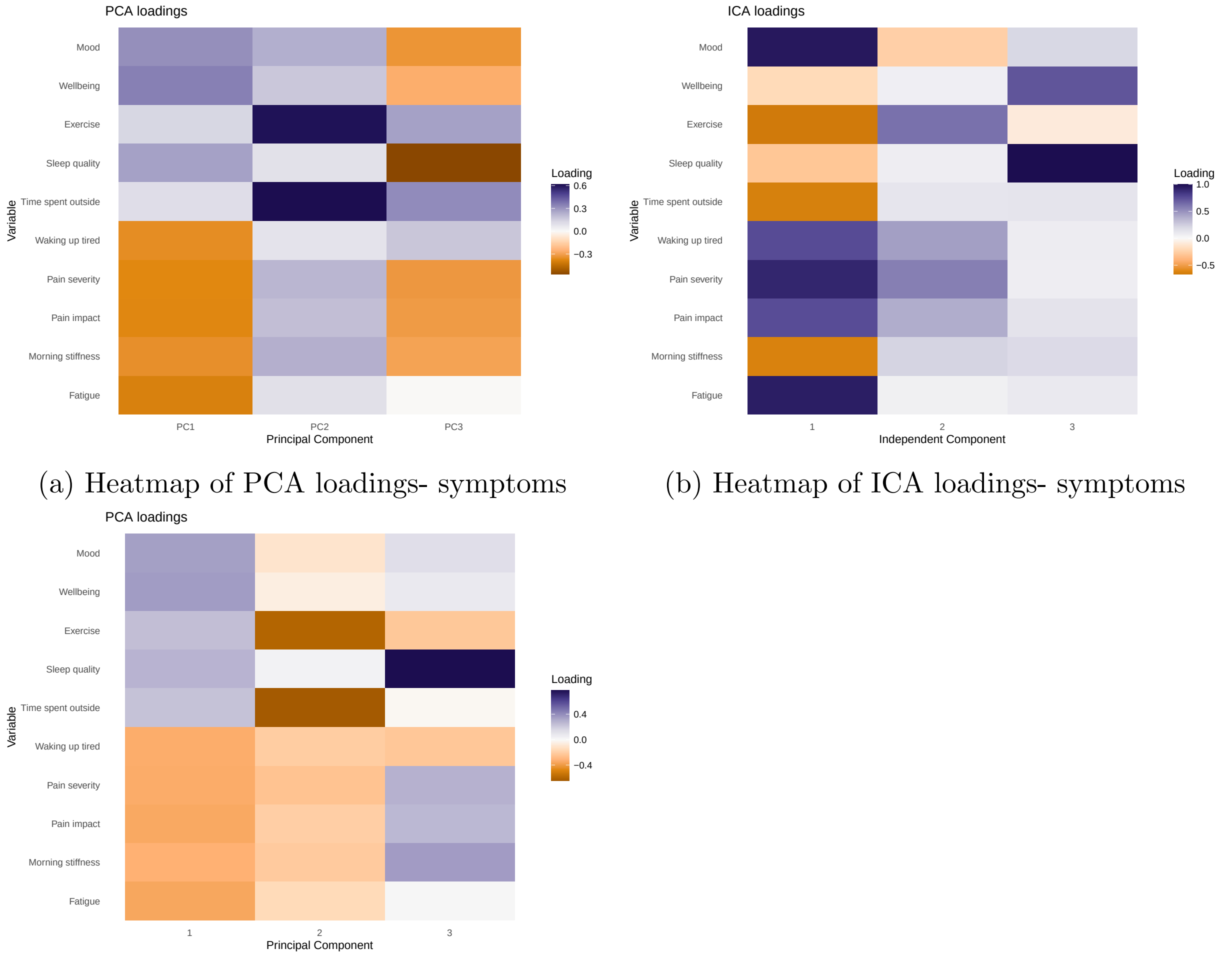


Figure 5.11: Loadings per principal component



(a) Heatmap of PCA loadings- symptoms

(b) Heatmap of ICA loadings- symptoms

(c) Heatmap of LPCA loadings- changes of symptoms

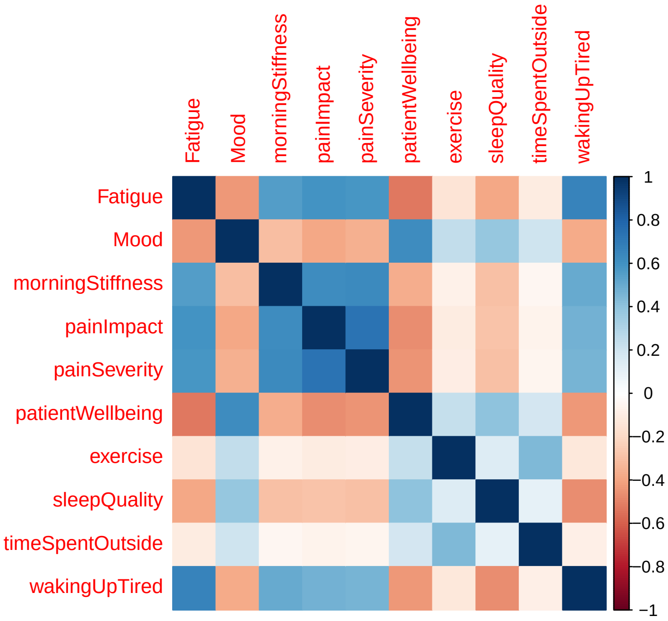


Figure 5.13: Correlation Plot

Chapter 6

Further work

A broad outcome of the thesis includes analysing longitudinal digital health data and interpreting results related to mental health, and developing and applying suitable methods. In this chapter, we discuss the possible ways of extending our work for further research. We do so by giving some suggestions and sharing additional results which can be utilised for the purpose of more related studies.

6.1 Missing data interpretations

Throughout the thesis, we performed analyses where the cleaned dataset was obtained by simply removing the data points (self-reported variables) containing NA (Not Available) values. But the missing data may have provided additional insights which have been not captured. So our work can be further extended by including the missing information and treating them appropriately, and then re-analyse and possibly compare the results. Taking an inter-disciplinary approach in handling the missing data, here, the self-reported variables, can provide meaningful interpretations.

6.2 More clusters

We have applied an Expectation-Maximisation (EM) algorithm to perform clustering on the joint longitudinal trajectories of self-reported data of mood and pain as elaborated in Chapter 3.

In this section, we specifically talk about extending the study by including clustering of trajectories containing other self-reported symptoms paired with pain. We re-run the clustering for other symptoms and show the results.

We present the cluster-heatmaps by first showing the frequency plots— Figures 6.1 to

6.9 give the frequencies of self-reported symptoms paired with pain. This helps in understanding how the states are distributed over the period of time of the mobile health study.

We can note similarities amongst some of the plots by looking at them. Few are almost centered at the middle, while few are left aligned and others are to the right. These can be analysed in relation to all the studies we have performed till now. At this point, we have just taken a glance of the frequency plots and not studied them individually. We hope the results can serve beyond the extension of our analyses, and can be utilised as observational data output for other related work.

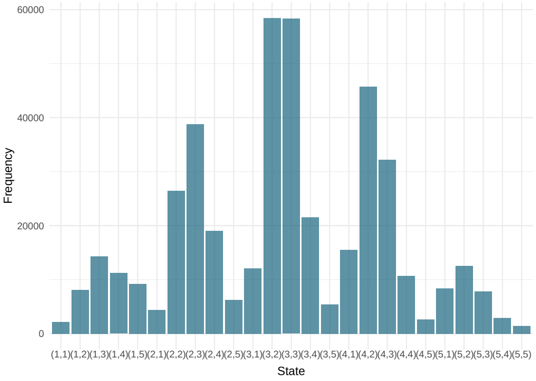


Figure 6.1: (Sleep quality, Pain)

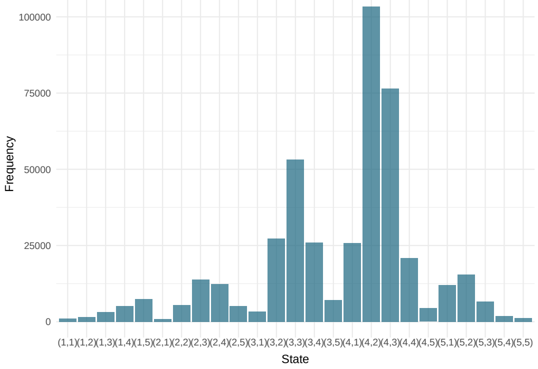


Figure 6.2: (Wellbeing, Pain)

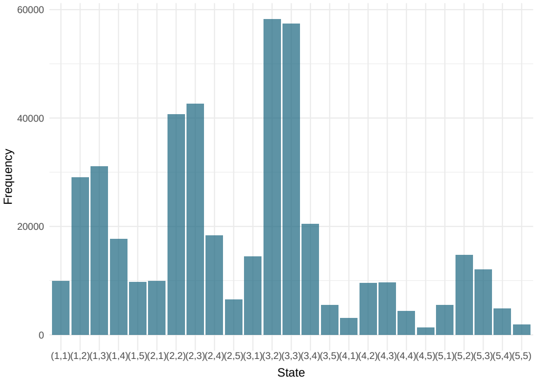


Figure 6.3: (Exercise, Pain)

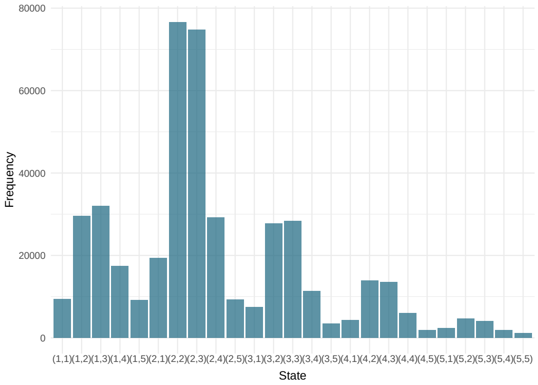


Figure 6.4: (Time spent outside, Pain)

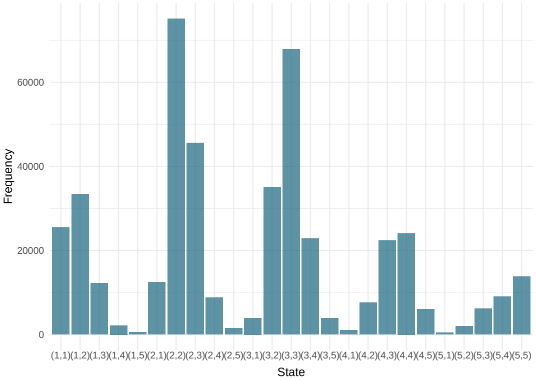


Figure 6.5: (Fatigue, Pain)

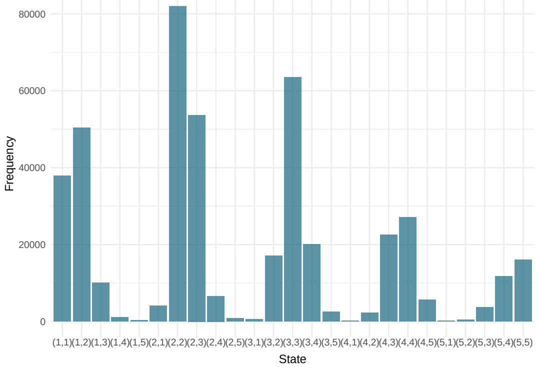


Figure 6.6: (Pain impact, Pain)

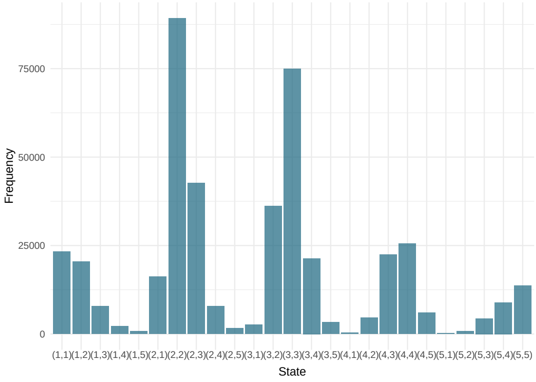


Figure 6.7: (Morning stiffness, Pain)

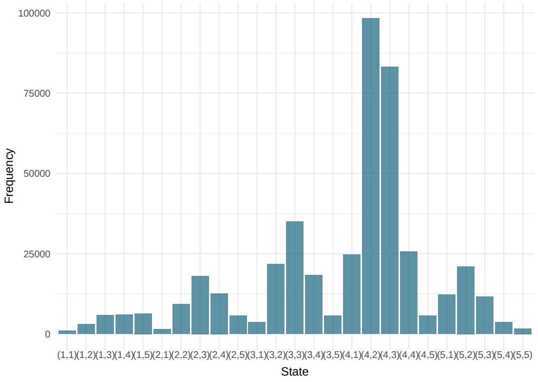


Figure 6.8: (Mood, Pain)

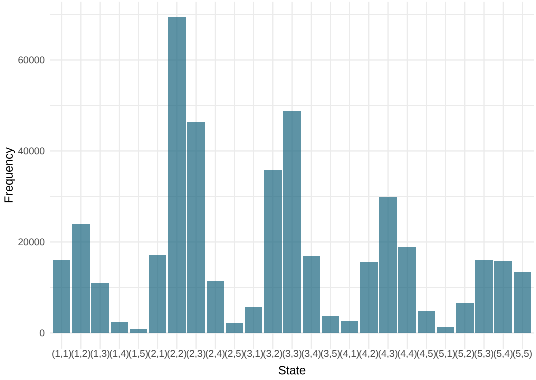


Figure 6.9: (Waking up tired, Pain)

Clusters

Figure 6.10 to Figure 6.17 present heatmaps of four clusters based on trajectories of other self-reported symptoms, which are sleep quality, wellbeing, exercise, time spent outside, fatigue, pain impact, morning stiffness and waking up tired, paired with pain severity. On this new data, we perform the same EM algorithm based clustering as described in Chapter 3 by selecting the optimal number of clusters to be 4. However, we do not regroup the individual states to binary values which resulted total $2 \times 2 = 4$ states of Markov chain earlier, because the dichotomisation of the states can't be uniformly done in similar ways as before as the meanings of the categories of other symptoms differ and require more consideration during bifurcation. So we retain all the $5 \times 5 = 25$ transition states of the Markov chain during re-running of the EM based clustering.

While looking at the following figures of heatmaps of the transition probability matrix per cluster, please note that the mid point of the range of colours in legend has been set at $1/25 = 0.04$. The granularity of the heatmaps of the clusters has not been examined at this point, but could be done as part of another analysis. The results can provide insights to the relationships between the two variables in a pair and can be useful to other related studies involving these.

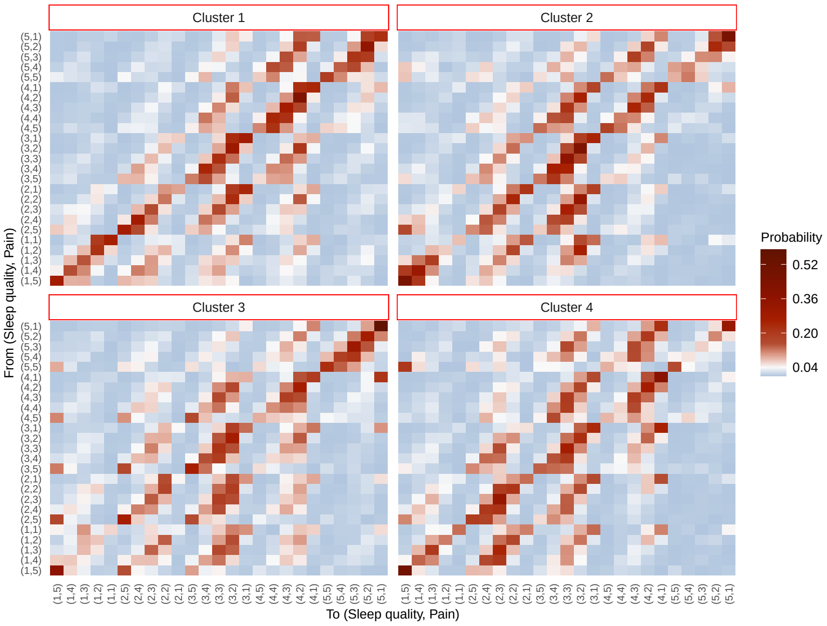


Figure 6.10: Clusters containing (Sleep quality, Pain) states

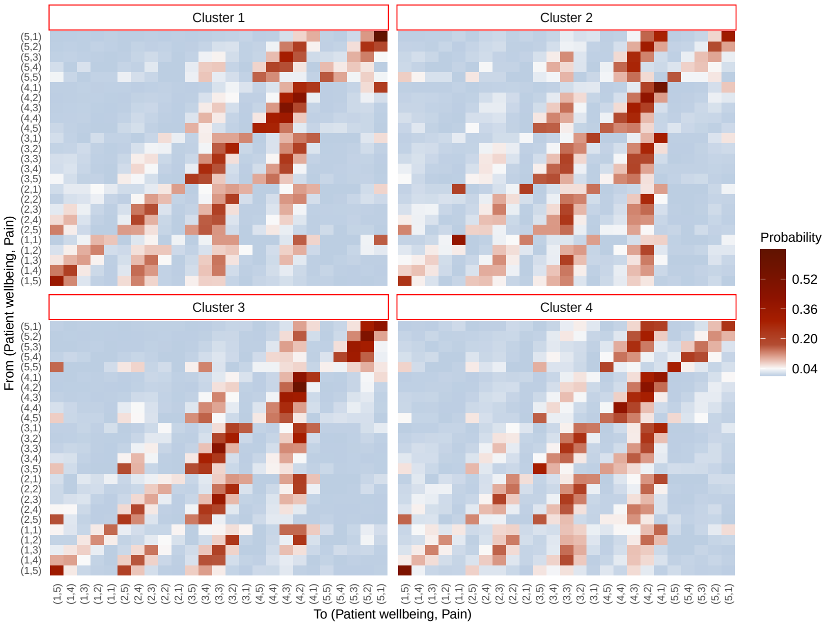


Figure 6.11: Clusters containing (Wellbeing, Pain) states

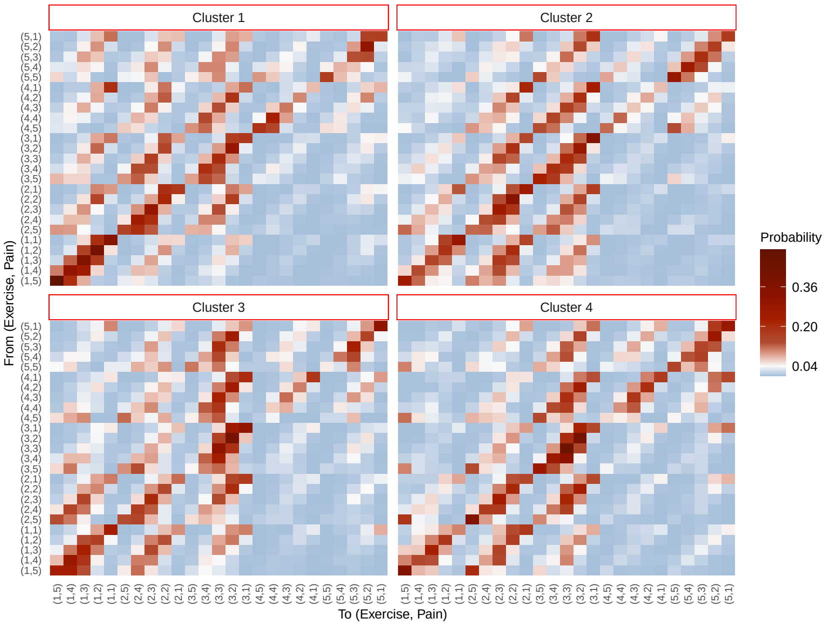


Figure 6.12: Clusters containing (Exercise, Pain) states

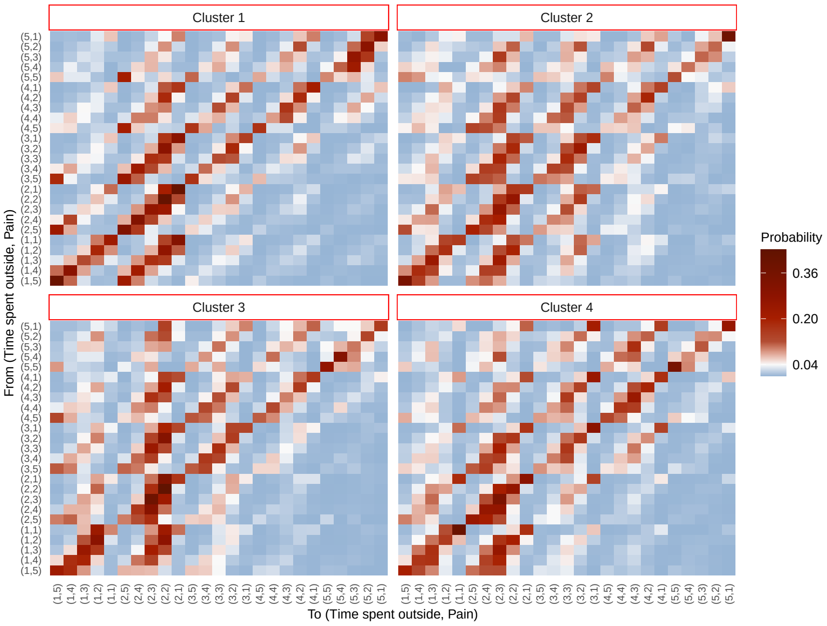


Figure 6.13: Clusters containing (Time spent outside, Pain) states

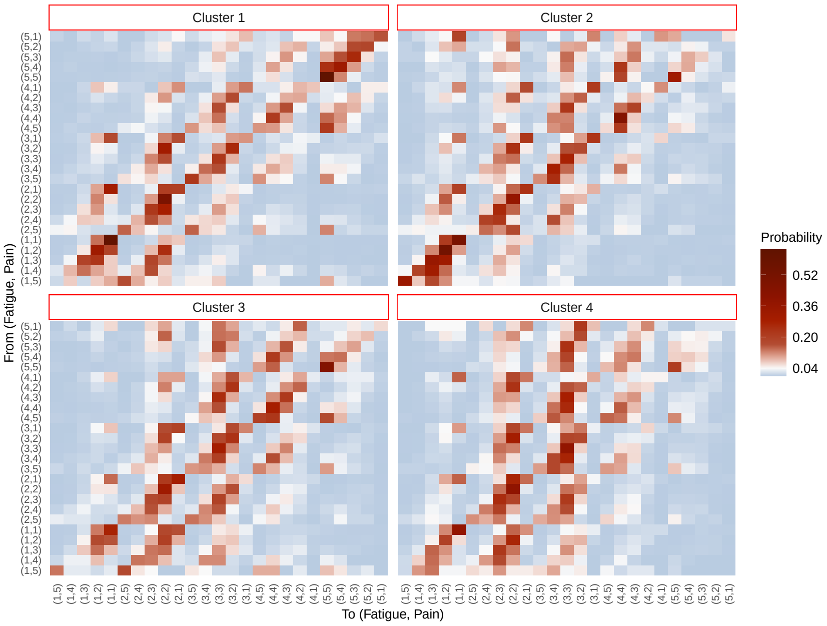


Figure 6.14: Clusters containing (Fatigue, Pain) states

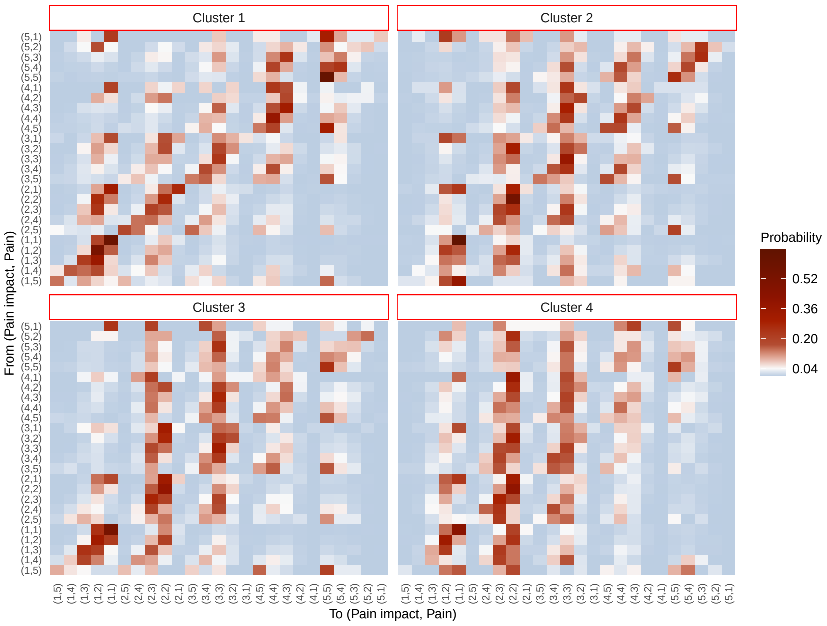


Figure 6.15: Clusters containing (Pain impact, Pain) states

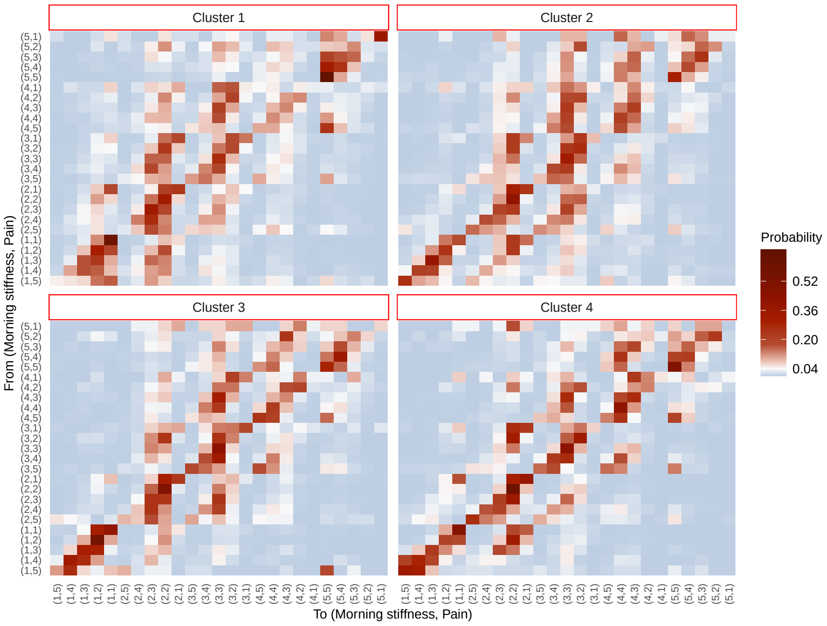


Figure 6.16: Clusters containing (Morning stiffness, Pain) states

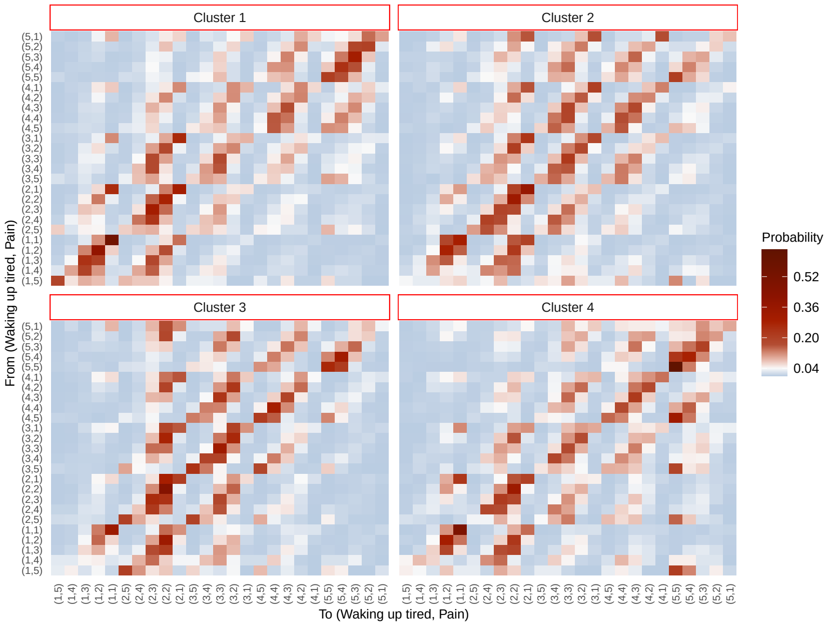


Figure 6.17: Clusters containing (Waking up tired, Pain) states

6.3 Extension to n-tuple

We have performed our EM clustering on pairs (2-tuples) of symptoms. We propose that it can be further extended to an $n-$ tuple system where $n \geq 3$ where more symptoms can be included in the model. For e.g.: (mood, pain, sleep quality) can be considered as a state.

6.4 Variations of Dirichlet-multinomial method

Motivated by the longitudinal data in hand, we proposed a Bayesian inference model for multinomial longitudinal data by assuming the data to be generated from a mixture of Dirichlet distributions parameterised by matrices of shape parameters as discussed in Chapter 4. We did not do a sensitivity analysis based on the parameters.

The method can be further extended by varying the following:

1. Constraints on parameters:
   We imposed an ordering constraint on the sum of the Dirichlet parameters to address label-switching in mixture models. We could experiment with more parameters or other label-switching methods inspired by Papastamoulis (2015) but for the Hamiltonian Monte Carlo sampling, to increase the efficiency of the simulations.

2. Initialisation for MCMC chains:
   We found an initial point for the MCMC chains with the help of EM algorithm again to help with label-switching in the MCMC sampling. However, the method can be retested with more points of initialisation by considering different parameters.

3. Number of dimensions:
   In this thesis, we applied our Dirichlet-multinomial method to 4 states *i.e.* the dimension of the simplex was $4 - 1 = 3$. This mixture model can be extended by taking more parameters to increase the dimension of the simplex, and analysed and improved further. Situations pertaining to higher state space arise often in many spheres of life and research where higher-dimensional data are presented.

4. Geometry of Dirichlet-multinomial distribution:
   Dirichlet distributions are represented by their shape parameters belonging to a simplex. As the dimensions increase, the corresponding simplices get tougher to visualise. In continuation to what has been suggested in the previous point, increasing the number of states while implementing an MCMC method may incur issues in the geometry of the simplex due to the shape constraint in its definition. To deal with this issue, the geometry of the simplex can be transformed such that the shape constraint gets softened. Betancourt (2012) shows transformation of variables to

simplify the simplex.

We would have proceeded with including higher dimensional data in Chapter 4 where we would have liked to explore the geometry of the simplex and then the efficiency of the implemented MCMC method. However, at this point, this remains as a proposed extension to the project.

6.5 More dimensionality reduction, clustering and other techniques

Based on this chapter, we could extend the analysis to consider more dimensionality reduction and clustering techniques like t-SNE (Van der Maaten and Hinton, 2008), UMAP (McInnes et al., 2018) and CLASSIX (Chen and Güttel, 2022) so we can include more data and stratify according to several other variables like a weather parameter or site of pain or baseline information like age, sex to check if the stratification influences the results and gives new clusters. In fact, where and as applicable time-series methods, some of which are listed in the review by Aghabozorgi et al. (2015) could be translated to longitudinal data implementation. Other methods to look at would include efforts to build a dynamical system e.g. Cramer et al. (2016) and Demic and Cheng (2014).

Another addition is comparing the results obtained from this thesis with similar studies or other available data– populations of different demographics can provide new insights on the behaviour of the self-reported symptoms amongst other groups at a larger scale.

Last but not the least, performing regression and comparing the results with our methods can give an overview of the methods with the corresponding results. This can help in updating traditional epidemiological or similar longitudinal data analyses' methods and can potentially propose a new set of initial check-up routines for handling such data.

Chapter 7

Conclusion

At the start of this work, I was presented with longitudinal data recording self-reported ratings of mood, pain and eight other variables. These data were of a scale and complexity such that careful consideration of methodology allowed for greater insights to be delivered. Together with collaborators, I studied the data using (and comparing) expectation-maximisation, Bayesian, and machine learning approaches. This thesis has two outcomes: (1) application wise – we find digital phenotypes based on the self-reported data of mood and pain trajectories; (2) methods wise – we develop and implement methods driven by different statistical ideologies on the longitudinal data.

The three main bodies of work in this thesis show the multiple ways of handling large, heterogeneous and complex longitudinal data. In Chapter 3, we investigated the relationship between mood and pain trajectories by performing a residual analysis first based on the observed transition probability matrix of the data. Due to high residuals, we moved on to model-based clustering of the mood and pain transitions. Since it would be difficult to perform clustering of longitudinal data transitions with a distance-measure (Aghabozorgi et al., 2015; Liao, 2005) we implemented an expectation-maximisation algorithm based clustering as elaborated in the chapter which helped in discovering the underlying phenotypes. The algorithm developed can be applied to any scenario with similar form of data containing trajectories of values over a period of time. The results in this chapter highlighted the need for personalised treatments in healthcare and emphasised that an umbrella solution to a problem may not be correct. The biases and limitations were also pointed out.

In Chapter 4, we performed Bayesian inference of the same mood, pain transitions. We developed a Dirichlet-multinomial mixture distribution to address the specific problem where the row of every individual transition matrix was sampled from a Dirichlet distribution belonging to a component. Its application on mobile health data showed similar clustering to what had been achieved before in Chapter 3, but with the help of Bayesian inference, predictions can be made which can potentially forecast mood or similar health parameters.

Finally in Chapter 5, we applied several dimensionality reduction and clustering techniques to our dataset by selecting all self-reported variables instead of focusing on only mood and pain. The visualisations obtained in these techniques showed clear clusters of the symptoms and highlighted the benefits of the method in quick assessment therefore, can be implemented as part of primary and exploratory analyses.

Overall all these three bodies of work show the applications of mathematical, statistical and computational tools in understanding health and mental health specifically. The results give us insights on the behaviour of mood, pain and other parameters experienced by a cohort of people who were already experiencing chronic pain. Application wise, the main takeaway is highlighting the complexity of mental health and chronic pain associations which need to be dealt with care thereby emphasising on making treatments personalised to a patient's history and needs. A limitation of the project is the possibility of biases (e.g. sampling bias) inherent to smartphone studies. However, if we can carry out similar studies on cohorts of different demographics and make comparisons, we can eventually learn more and examine the biases accordingly. If consistency in results is observed in other iterations of this study with varying cohorts, it would be a breakthrough to find the specific clusters of human behaviour. Also, if we are able to identify such groups based on the transitions of self-reported symptoms, it would lead to further research and designing of studies pertaining to the particular groups.

In terms of methodology presented in the thesis, we implement novel approaches on the longitudinal data and find the results accordingly. The methods can be further extended by including more variables and parameters, as well as adding more layers of constraints and assumptions. While doing so, it is important to remember that making methods too complex may present problems of over-fitting and interpret-ability, and be computationally intensive. Nevertheless, the methods introduced and discussed in this thesis are not exclusive to the Cloudy with a Chance of Pain survey and can be applied to other similarly structured data belonging to other fields of study as well.

References

Roger AH Adan, Eline M van der Beek, Jan K Buitelaar, John F Cryan, Johannes Hebebrand, Suzanne Higgs, Harriet Schellekens, and Suzanne L Dickson. Nutritional psychiatry: Towards improving mental health by what you eat. *European Neuropsychopharmacology*, 29(12):1321–1332, 2019. 15

Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering–a decade review. *Information Systems*, 53:16–38, 2015. 141, 142

Luis Agüera, Inmaculada Failde, Jorge A Cervilla, Paula Díaz-Fernández, and Juan Antonio Mico. Medically unexplained pain complaints are associated with underlying unrecognized mood disorders in primary care. *BMC Family Practice*, 11(1):1–8, 2010. 49

Laura Helena Andrade, J Alonso, Z Mneimneh, JE Wells, A Al-Hamzawi, G Borges, E Bromet, Ronny Bruffaerts, G De Girolamo, R De Graaf, et al. Barriers to mental health treatment: results from the who world mental health surveys. *Psychological Medicine*, 44(6):1303–1317, 2014. 14

Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to mcmc for machine learning. *Machine Learning*, 50(1):5–43, 2003. 42

AD Appels and Erik Schouten. Waking up exhausted as risk indicator of myocardial infarction. *The American Journal of Cardiology*, 68(4):395–398, 1991. 113

Gordon JG Asmundson and Joel Katz. Understanding the co-occurrence of anxiety disorders and chronic pain: state-of-the-art. *Depression and Anxiety*, 26(10):888–901, 2009. 49

Michael Betancourt. Cruising the simplex: Hamiltonian monte carlo and the dirichlet distribution. In *AIP Conference Proceedings 31st*, volume 1443, pages 157–164. American Institute of Physics, 2012. 140

Michael Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017. 45

Prashant Bharadwaj, Mallesh M Pai, and Agne Suziedelyte. Mental health stigma. *Economics Letters*, 159:57–60, 2017. 14

Dinesh Bhugra, Alex Till, and Norman Sartorius. What is mental health?, 2013. 15

Jeff A Bilmes et al. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510):126, 1998. 24

Christopher M Bishop and Nasser M Nasrabadi. *Pattern Recognition and Machine Learning*, volume 4. Springer, 2006. 45

Y. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete Multivariate Analysis: Theory and Practice.* Massachusetts Institute of Technology Press, Cambridge, 1975. 51, 52, 67

Sean Borman. The expectation maximization algorithm: A short tutorial. 2009. URL `http://www.seanborman.com/publications/EM_algorithm.pdf`. 32, 68

Nizar Bouguila. Clustering of count data using generalized dirichlet multinomial distributions. *IEEE Transactions on Knowledge and Data Engineering*, 20(4):462–474, 2008. 85, 86

Nicole Wallbridge Bourmistrova, Tomas Solomon, Philip Braude, Rebecca Strawbridge, and Ben Carter. Long-term effects of covid-19 on mental health: A systematic review. *Journal of Affective Disorders*, 299:118–125, 2022. 15

Harald Breivik, Beverly Collett, Vittorio Ventafridda, Rob Cohen, and Derek Gallacher. Survey of chronic pain in europe: prevalence, impact on daily life, and treatment. *European Journal of Pain*, 10(4):287–333, 2006. 48, 113

Igor Cadez, David Heckerman, Christopher Meek, Padhraic Smyth, and Steven White. Model-based clustering and visualization of navigation patterns on a web site. *Data mining and knowledge discovery*, 7:399–424, 2003. 19

Louise Camm-Crosbie, Louise Bradley, Rebecca Shaw, Simon Baron-Cohen, and Sarah Cassidy. 'people like me don't get support': Autistic adults' experiences of support and treatment for mental health difficulties, self-injury and suicidality. *Autism*, 23(6): 1431–1441, 2019. 14

Xinye Chen and Stefan Güttel. Fast and explainable clustering based on sorting. *arXiv preprint arXiv:2202.01456*, 2022. 141

Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335, 1995. 43

Pamela Y Collins, Vikram Patel, Sarah S Joestl, Dana March, Thomas R Insel, Abdallah S Daar, Isabel A Bordin, E Jane Costello, Maureen Durkin, Christopher Fairburn, et al. Grand challenges in global mental health. *Nature*, 475(7354):27–30, 2011. 48

Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36 (3):287–314, 1994. 116

Angélique OJ Cramer, Claudia D Van Borkulo, Erik J Giltay, Han LJ Van Der Maas, Kenneth S Kendler, Marten Scheffer, and Denny Borsboom. Major depression as a complex dynamic system. *PloS one*, 11(12):e0167490, 2016. 141

Rajenki Das, Mark Muldoon, Mark Lunt, John McBeth, Belay Birlie Yimer, and Thomas House. Modelling and classifying joint trajectories of self-reported mood and pain in a large cohort study. *PLoS Digital Health*, 2(3):e0000204, 2023. 94, 113

Rolando De la Cruz-Mesía, Fernando A Quintana, and Guillermo Marshall. Model-based clustering for longitudinal data. *Computational Statistics & Data Analysis*, 52(3): 1441–1457, 2008. 18

Selver Demic and Sen Cheng. Modeling the dynamics of disease states in depression. *PLoS One*, 9(10):e110358, 2014. 141

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. 18, 53

Persi Diaconis and David Freedman. Finite exchangeable sequences. *The Annals of Probability*, pages 745–764, 1980. 109

Ed Diener. Subjective well-being. *The Science of Well-being*, pages 11–58, 2009. 112

Peter Diggle, Peter J Diggle, Patrick Heagerty, Kung-Yee Liang, Scott Zeger, et al. *Analysis of Longitudinal Data*. Oxford University Press, 2002. 17

William G Dixon, Anna L Beukenhorst, Belay B Yimer, Louise Cook, Antonio Gasparrini, Tal El-Hay, Bruce Hellman, Ben James, Ana M Vicedo-Cabrera, Malcolm Maclure, et al. How the weather affects the pain of citizen scientists using a smartphone app. *NPJ Digital Medicine*, 2(1):1–9, 2019. 16, 17, 50, 86, 114

Chang CY Dorea, Catia R Goncalves, and PA Resende. Simulation results for markov model seletion: Aic, bic and edc. In *Proceedings of World Congress on Engineering and Computer Science*, volume 2, pages 899–901, 2014. 69

Katie L Druce, John McBeth, Sabine N van der Veer, David A Selby, Bertie Vidgen, Konstantinos Georgatzis, Bruce Hellman, Rashmi Lakshminarayana, Afiqul Chowdhury, David M Schultz, et al. Recruitment and ongoing engagement in a uk smartphone study examining the association between weather and pain: cohort study. *JMIR mHealth and uHealth*, 5(11):e168, 2017. 7, 18, 50, 86, 114

Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987. 44

Sean R Eddy. What is a hidden markov model? *Nature Biotechnology*, 22(10):1315–1316, 2004. 18

Hong Fang, Sheng Tu, Jifang Sheng, and Anwen Shao. Depression in sleep disturbance: a review on a bidirectional relationship, mechanisms and treatment. *Journal of Cellular and Molecular Medicine*, 23(4):2324–2332, 2019. 113

Maddalena Fiordelli, Nicola Diviani, Peter J Schulz, et al. Mapping mhealth research: a decade of evolution. *Journal of Medical Internet Research*, 15(5):e2430, 2013. 15

David A Fishbain, Robert Cutler, Hubert L Rosomoff, and Renee Steele Rosomoff. Chronic pain-associated depression: antecedent or consequence of chronic pain? a review. *The Clinical Journal of Pain*, 13(2):116–137, 1997. 48

Wilbert Evans Fordyce. *Behavioral Methods for Chronic Pain and Illness*, volume 1. Mosby St. Louis, 1976. 48

Seth R Frank. Digital health care—the convergence of health care and the internet. *The Journal of Ambulatory Care Management*, 23(2):8–17, 2000. 113

Eiko I Fried. Studying mental disorders as systems, not syndromes. 2021. 15

Sylvia Frühwirth-Schnatter and Christoph Pamminger. Model-based clustering of categorical time series. *Bayesian Analysis*, 2010. 19, 87

Silvana Galderisi, Andreas Heinz, Marianne Kastrup, Julian Beezhold, and Norman Sartorius. Toward a new definition of mental health. *World Psychiatry*, 14(2):231, 2015. 48

Tanya P Garcia and Karen Marder. Statistical approaches to longitudinal data analysis in neurodegenerative diseases: Huntington's disease as a model. *Current Neurology and Neuroscience Reports*, 17(2):14, 2017. 17

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. CRC Press, 3 edition, 2014. 45

Charles J Geyer. Markov chain monte carlo lecture notes. *Course Notes, Spring Quarter*, 80, 1998. 41

Benjamin Gibson, Jekaterina Schneider, Deborah Talamonti, and Mark Forshaw. The impact of inequality on mental health outcomes during the covid-19 pandemic: A systematic review. *Canadian Psychology/Psychologie Canadienne*, 62(1):101, 2021. 15

Ronald Gijsen, Nancy Hoeymans, François G Schellevis, Dirk Ruwaard, William A Satariano, and Geertrudis AM van den Bos. Causes and consequences of comorbidity: a review. *Journal of Clinical Epidemiology*, 54(7):661–674, 2001. 48

Stephen Ginn and Jamie Horder. "one in four" with a mental health problem: the anatomy of a statistic. *BMJ*, 344:e1302, 2012. 49

Katherine J Gold, Louise B Andrew, Edward B Goldman, and Thomas L Schwenk. "i would never want to have a mental health diagnosis on my record": a survey of female physicians on mental health diagnosis, treatment, and reporting. *General Hospital Psychiatry*, 43:51–57, 2016. 14

Scott D Grimshaw and William P Alexander. Markov chain models for delinquency: Transition matrix estimation and forecasting. *Applied Stochastic Models in Business and Industry*, 27(3):267–279, 2011. 19

Jonathan Haidt and Nick Allen. Scrutinizing the effects of digital technology on mental health, 2020. 16

Jeffrey P. Harrison and Angela Lee. The role of e-health in the changing health care environment. *Nursing Economics*, 24(6):283–8, 279; quiz 289, Nov 2006. URL `https://manchester.idm.oclc.org/login?url=https://search.proquest.com/docview/236937602?accountid=12253`. Copyright - Copyright Anthony J. Jannetti, Inc. Nov/Dec 2006; Document feature - ; Tables; Last updated - 2014-03-29. 49

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics. Springer, New York, 2 edition, 2009. 22

W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970. 43

Daniel W. Heck, Antony Overstall, Quentin F. Gronau, and Eric-Jan Wagenmakers. Quantifying Uncertainty in Transdimensional Markov Chain Monte Carlo Using Discrete Markov Models. *Statistics & Computing*, 29:631–643, 2019. doi: 10.1007/s11222-018-9828-0. 93

Moritz Herle, Nadia Micali, Mohamed Abdulkadir, Ruth Loos, Rachel Bryant-Waugh, Christopher Hübel, Cynthia M Bulik, and Bianca L De Stavola. Identifying typical trajectories in longitudinal data: modelling strategies and interpretations. *European Journal of Epidemiology*, 35(3):205–222, 2020. 17

Ian Holmes, Keith Harris, and Christopher Quince. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS One*, 7(2):e30126, 2012. 19

Po-Cheng Hou, Fang-Ju Lin, Shin-Yi Lin, Tzung-Jeng Hwang, and Chi-Chuan Wang. Risk of intracranial hemorrhage with concomitant use of antidepressants and nonsteroidal anti-inflammatory drugs: a nested case-control study. *Annals of Pharmacotherapy*, 55(8):941–948, 2021. 49

Aapo Hyvarinen. Fast ica for noisy data using gaussian moments. In *1999 IEEE International Symposium on Circuits and Systems (ISCAS)*, volume 5, pages 57–61. IEEE, 1999. 117

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000. 116, 117

Michael Irwin. Psychoneuroimmunology of depression: clinical implications. *Brain, Behavior, and Immunity*, 16(1):1–16, 2002. 48

Marie Jahoda. *Current Concepts of Positive Mental Health.* Basic Books, 1958. 48

Ru-Rong Ji, Alexander Chamessian, and Yu-Qiu Zhang. Pain regulation by non-neuronal cells and inflammation. *Science*, 354(6312):572–577, 2016. 48

Karine Alexandra Del Rio João, Saul Neves de Jesus, Cláudia Carmo, and Patrícia Pinto. The impact of sleep quality on the mental health of a non-clinical population. *Sleep Medicine*, 46:69–73, 2018. 113

Brett DM Jones, Zafiris J Daskalakis, Andre F Carvalho, Rebecca Strawbridge, Allan H Young, Benoit H Mulsant, and M Ishrat Husain. Inflammation as a treatment target in mood disorders. *BJPsych Open*, 6(4), 2020. 48

Michelle Kendall, Luke Milsom, Lucie Abeler-Dörner, Chris Wymant, Luca Ferretti, Mark Briers, Chris Holmes, David Bonsall, Johannes Abeler, and Christophe Fraser. Epidemiological changes on the Isle of Wight after the launch of the NHS Test and Trace programme: a preliminary analysis. *The Lancet Digital Health*, 2(12):e658–e666, 2020. 49

Lola Kola, Brandon A Kohrt, Charlotte Hanlon, John A Naslund, Siham Sikander, Madhumitha Balaji, Corina Benjet, Eliza Yee Lai Cheung, Julian Eaton, Pattie Gonsalves, et al. Covid-19 mental health impact and responses in low-income and middle-income countries: reimagining global mental health. *The Lancet Psychiatry*, 8(6):535–550, 2021. 15

Arnošt Komárek and Lenka Komárková. Clustering for multivariate continuous and discrete longitudinal data. *The Annals of Applied Statistics*, 7(1):177–200, 2013. 17, 63

S Yunkap Kwankam. What e-health can offer. *Bulletin of the World Health Organization*, 82(10):800–801, 2004. 49

Andrew J. Landgraf and Yoonkyung Lee. Dimensionality reduction for binary data through the projection of natural parameters. Technical Report 890, Department of Statistics, The Ohio State University, 2015. URL `http://arxiv.org/abs/1510.06112`. 117

Andrew J Landgraf and Yoonkyung Lee. Dimensionality reduction for binary data through the projection of natural parameters. *Journal of Multivariate Analysis*, 180: 104668, 2020. 117

Chieh-Hsin Lee and Fabrizio Giuliani. The role of inflammation in depression and fatigue. *Frontiers in Immunology*, 10:1696, 2019. 113

Fabio Leonardi. The definition of health: towards new perspectives. *International Journal of Health Services*, 48(4):735–748, 2018. 14

Ximing Li, Jiaojiao Zhang, and Jihong Ouyang. Dirichlet multinomial mixture with variational manifold regularization: Topic modeling over short texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7884–7891, 2019. 19

T Warren Liao. Clustering of time series data—a survey. *Pattern Recognition*, 38(11): 1857–1874, 2005. 18, 142

Jiayu Lin. On the dirichlet distribution. *Department of Mathematics and Statistics, Queens University*, 2016. 40

Joseph Liouville. Note sur la théorie de la variation des constantes arbitraires. *Journal de Mathématiques Pures et Appliquées*, pages 342–349. 45

Jan Lötvall, Cezmi A Akdis, Leonard B Bacharier, Leif Bjermer, Thomas B Casale, Adnan Custovic, Robert F Lemanske Jr, Andrew J Wardlaw, Sally E Wenzel, and Paul A Greenberger. Asthma endotypes: a new approach to classification of disease entities within the asthma syndrome. *Journal of Allergy and Clinical Immunology*, 127 (2):355–360, 2011. 20

May N Lwin, Lina Serhal, Christopher Holroyd, and Christopher J Edwards. Rheumatoid arthritis: the impact of mental health on disease: a narrative review. *Rheumatology and Therapy*, 7(3):457–471, 2020. 48

Zhihua Ma and Guanghui Chen. Bayesian methods for dealing with missing data problems. *Journal of the Korean Statistical Society*, 47(3):297–313, 2018. 33

Kathleen MacDonald, Nina Fainman-Adelman, Kelly K Anderson, and Srividya N Iyer. Pathways to mental health services for young people: a systematic review. *Social Psychiatry and Psychiatric Epidemiology*, 53(10):1005–1038, 2018. 14

David JC MacKay and Linda C Bauman Peto. A hierarchical dirichlet language model. *Natural Language Engineering*, 1(3):289–308, 1995. 86

Alexina Mason, Nicky Best, Ian Plewis, and Sylvia Richardson. Insights into the use of bayesian models for informative missing data. 2010. 33

Simon C Mathews, Michael J McShea, Casey L Hanley, Alan Ravitz, Alain B Labrique, and Adam B Cohen. Digital health: a path to validation. *NPJ Digital Medicine*, 2(1): 1–9, 2019. 16

Iain B McInnes, Christopher D Buckley, and John D Isaacs. Cytokines in rheumatoid arthritis—shaping the immunological landscape. *Nature Reviews Rheumatology*, 12(1): 63, 2016. 21

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 141

Paul D McNicholas and T Brendan Murphy. Model-based clustering of longitudinal data. *Canadian Journal of Statistics*, 38(1):153–168, 2010. 18

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953. 41

Thomas Minka. Estimating a dirichlet distribution, 2000. 86, 107

Carmen Moreno, Til Wykes, Silvana Galderisi, Merete Nordentoft, Nicolas Crossley, Nev Jones, Mary Cannon, Christoph U Correll, Louise Byrne, Sarah Carr, et al. How mental health care should change as a consequence of the covid-19 pandemic. *The Lancet Psychiatry*, 7(9):813–824, 2020. 14

Nina Morris. Health, well-being and open space. *Edinburgh: Edinburgh College of Art and Heriot-Watt University*, 2003. 15

John A Naslund, Kelly A Aschbrenner, Ricardo Araya, Lisa A Marsch, Jürgen Unützer, Vikram Patel, and Stephen J Bartels. Digital technology for treating and preventing mental disorders in low-income and middle-income countries: a narrative review of the literature. *The Lancet Psychiatry*, 4(6):486–500, 2017. 49

Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011. 44

Charles B Nemeroff and Michael J Owens. Treatment of mood disorders. *Nature Neuroscience*, 5(11):1068–1070, 2002. 49

Frank Nielsen. Hierarchical clustering. In *Introduction to HPC with MPI for Data Science*, pages 195–211. Springer, 2016. 116

Klaus Nordhausen, Hannu Oja, David E. Tyler, and Joni Virta. *ICtest: Estimating and Testing the Number of Interesting Components in Linear Dimension Reduction*, 2022. URL `https://CRAN.R-project.org/package=ICtest`. R package version 0.3-5. 127

Miranda Olff. Mobile mental health: a challenging research agenda. *European Journal of Psychotraumatology*, 6(1):27882, 2015. 15

Geneva: World Health Organization. World mental health report: Transforming mental health for all, 2022. 14

World Health Organization and Others. Depression and other common mental disorders: global health estimates. Technical report, World Health Organization, 2017. 48

Cara A Palmer. Tired teens: Sleep disturbances and heightened vulnerability for mental health difficulties. *Journal of Adolescent Health*, 66(5):520–521, 2020. 113

Panagiotis Papastamoulis. label. switching: An r package for dealing with the label switching problem in mcmc outputs. *arXiv preprint arXiv:1503.02271*, 2015. 24, 140

Kushang V Patel, Dagmar Amtmann, Mark P Jensen, Shannon M Smith, Christin Veasley, and Dennis C Turk. Clinical outcome assessment in clinical trials of chronic pain treatments. *Pain Reports*, 6(1), 2021. 49

Russell K Portenoy. Current pharmacotherapy of chronic pain. *Journal of Pain and Symptom Management*, 19(1):16–20, 2000. 49

Matthew Price, Erica K Yuen, Elizabeth M Goetter, James D Herbert, Evan M Forman, Ron Acierno, and Kenneth J Ruggiero. mhealth: a mechanism to deliver more accessible, more effective mental health care. *Clinical Psychology & Psychotherapy*, 21(5): 427–436, 2014. 15

Martin Prince, Vikram Patel, Shekhar Saxena, Mario Maj, Joanna Maselko, Michael R Phillips, and Atif Rahman. No health without mental health. *The Lancet*, 370(9590): 859–877, 2007. 14, 48

Cécile Proust-Lima, Viviane Philipps, and Benoit Liquet. Estimation of extended mixed models using latent classes and latent processes: the r package lcmm. *arXiv preprint arXiv:1503.00890*, 2015. 17, 63

Samuel Reade, Karen Spencer, Jamie C Sergeant, Matthew Sperrin, David M Schultz, John Ainsworth, Rashmi Lakshminarayana, Bruce Hellman, Ben James, John McBeth, et al. Cloudy with a chance of pain: engagement and subsequent attrition of daily data entry in a smartphone pilot study tracking weather, disease severity, and physical activity in patients with rheumatoid arthritis. *JMIR mHealth and uHealth*, 5(3):e6496, 2017. 50, 86, 114

Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM Review*, 26(2):195–239, 1984. 109

Joan M Romano and Judith A Turner. Chronic pain and depression: does the evidence support a relationship? *Psychological Bulletin*, 97(1):18, 1985. 48

Joshua D Rosenblat, Danielle S Cha, Rodrigo B Mansur, and Roger S McIntyre. Inflamed moods: a review of the interactions between inflammation and mood disorders. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 53:23–34, 2014. 48

Thomas C Rosenthal, Barbara A Majeroni, Richard Pretorious, and Khalid Malik. Fatigue: an overview. *American Family Physician*, 78(10):1173–1179, 2008. 113

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000. 114

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. 33

Michael Rutter. How the environment affects mental health. *The British Journal of Psychiatry*, 186(1):4–6, 2005. 15

Carol D Ryff and Burton Singer. The contours of positive human health. *Psychological Inquiry*, 9(1):1–28, 1998. 112

Carol D Ryff, Burton H Singer, and Gayle Dienberg Love. Positive health: connecting well–being with biology. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359(1449):1383–1394, 2004. 112

David A Scott, Bart Valley, and Brooke A Simecka. Mental health concerns in the digital age. *International Journal of Mental Health and Addiction*, 15(3):604–613, 2017. 16

Jamie Sergeant, David Schultz, Caroline Sanders, William Dixon, Samuel Reade, and Karen Spencer. Feasibility study of smartphone data collection for cloudy with a chance of pain: Sustained engagement for daily self-reporting of disease severity in rheumatoid arthritis over two months. 2015. 16

Jiyao Sheng, Shui Liu, Yicun Wang, Ranji Cui, and Xuewen Zhang. The link between depression and chronic pain: neural mechanisms in the brain. *Neural Plasticity*, 2017. 48

Ju-Young Shin, Mi-Ju Park, Shin Haeng Lee, So-Hyun Choi, Mi-Hee Kim, Nam-Kyong

Choi, Joongyub Lee, and Byung-Joo Park. Risk of intracranial haemorrhage in antidepressant users with concurrent use of non-steroidal anti-inflammatory drugs: nationwide propensity score matched study. *BMJ*, 351, 2015. 49

Kimmen Sjölander, Kevin Karplus, Michael Brown, Richard Hughey, Anders Krogh, I Saira Mian, and David Haussler. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Bioinformatics*, 12(4):327–345, 1996. 86

Yipeng Song, Johan A Westerhuis, and Age K Smilde. Logistic principal component analysis via non-convex singular value thresholding. *Chemometrics and Intelligent Laboratory Systems*, 204:104089, 2020. 117

Matthew Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000. 109

Richard A Sternbach. *Pain Patients: Traits and Treatment*. Academic Press New York, 1974. 48

David Stirzaker. *Elementary Probability*. Cambridge University Press, Mathematical Institute and St. John's College, University of Oxford, 2003. 54

Nicole KY Tang, Paul M Salkovskis, Amy Hodges, Kelly J Wright, Magdi Hanna, and Joan Hester. Effects of mood on pain responses and pain tolerance: an experimental study in chronic back pain patients. *Pain*, 138(2):392–401, 2008. 48

Jerry Tew, Shula Ramon, Mike Slade, Victoria Bird, Jane Melton, and Clair Le Boutillier. Social factors and recovery from mental health difficulties: a review of the evidence. *The British Journal of Social Work*, 42(3):443–460, 2012. 15

Christian Winther Topp, Søren Dinesen Østergaard, Susan Søndergaard, and Per Bech. The who-5 well-being index: a systematic review of the literature. *Psychotherapy and Psychosomatics*, 84(3):167–176, 2015. 112

Andreas K Triantafyllidis and Athanasios Tsanas. Applications of machine learning in real-life digital health interventions: review of the literature. *Journal of Medical Internet Research*, 21(4):e12286, 2019. 16

Ming T Tsuang. Genes, environment, and mental health wellness. *American Journal of Psychiatry*, 157(4):489–491, 2000. 15

Stephen Tu. The dirichlet-multinomial and dirichlet-categorical models for bayesian inference. *Computer Science Division, UC Berkeley*, 2, 2014. 19

Eldon R Tunks, Joan Crook, and Robin Weir. Epidemiology of chronic pain with psychological comorbidity: prevalence, risk, course, and prognosis. *The Canadian Journal of Psychiatry*, 53(4):224–234, 2008. 49

Rita J van den Berg-Emons, Fabiënne C Schasfoort, Leonard A de Vos, Johannes B Bussmann, and Henk J Stam. Impact of chronic pain on everyday physical activity. *European Journal of Pain*, 11(5):587–593, 2007. 48

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008. 141

Astrid M. Vicente, Wolfgang Ballensiefen, and Jan-Ingvar Jönsson. How personalised medicine will transform healthcare by 2030: the icpermed vision. *Journal of Translational Medicine*, 18:180, 2020. 48

Daniel Vigo, Scott Patten, Kathleen Pajer, Michael Krausz, Steven Taylor, Brian Rush, Giuseppe Raviola, Shekhar Saxena, Graham Thornicroft, and Lakshmi N Yatham. Mental health of communities during the covid-19 pandemic, 2020. 15

Michele Vitacca, Marco Mazzu, and Simonetta Scalvini. Socio-technical and organizational challenges to wider e-health implementation. *Chronic Respiratory Disease*, 6(2): 91–97, 2009. 15

Lars Von Knorring, C Perris, M Eisemann, U Eriksson, and H Perris. Pain as a symptom in depressive disorders: I. relationship to diagnostic subgroup and depressive symptomatology. *Pain*, 1983. 48

Harvey A Whiteford, Louisa Degenhardt, Jürgen Rehm, Amanda J Baxter, Alize J Ferrari, Holly E Erskine, Fiona J Charlson, Rosana E Norman, Abraham D Flaxman, Nicole Johns, et al. Global burden of disease attributable to mental and substance use disorders: findings from the global burden of disease study 2010. *The Lancet*, 382 (9904):1575–1586, 2013. 48

World Health Organization et al. Preamble to the constitution of the world health organization as adopted by the international health conference, new york, 19-22 june, 1946; signed on 22 july 1946 by the representatives of 61 states (official records of the world health organization, no. 2, p. 100) and entered into force on 7 april 1948. *http://www.who.int/governance/eb/who_constitution_en.pdf*, 1948. 14, 112

CE Wright, PC Strike, L Brydon, and A Steptoe. Acute inflammation and negative mood: mediation by cytokine activation. *Brain, Behavior, and Immunity*, 19(4):345–350, 2005. 48

Peifeng Yin, Qi He, Xingjie Liu, and Wang-Chien Lee. It takes two to tango: Exploring social tie development with both online and offline interactions. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 9(3):174–187, 2016. 69

Scott Ziegler. Current samplers and hamiltonian monte carlo: Notes for users of cosmosis. 2019. 45

Panagiotis Zis, Argyro Daskalaki, Ilia Bountouni, Panagiota Sykioti, Giustino Varrassi, and Antonella Paladini. Depression and chronic pain in the elderly: links and management challenges. *Clinical Interventions in Aging*, 12:709, 2017. 48