High-dimensional Co-occurrence Modelling with an Application in Disease Co-morbidity

Cassandra Posthumus

Assignment presented in the partial fulfilment
of the requirement for the degree of
BComHons (Mathematical Statistics)
at the University of Stellenbosch

Supervisor: Dr. D.P. Hofmeyr

Degree of confidentiality: A

November 2021

PLAGIARISM DECLARATION

1. Plagiarism is the use of ideas, material and other intellectual property of another's work and to present it as my own.
2. I agree that plagiarism is a punishable offence because it constitutes theft.
3. Accordingly, all quotations and contributions from any source whatsoever (including the internet) have been cited fully. I understand that the reproduction of text without quotation marks (even when the source is cited) is plagiarism.
4. I also understand that direct translations are plagiarism.
5. I declare that the work contained in this assignment, except otherwise stated, is my original work and that I have not previously (in its entirety or in part) submitted it for grading in this assignment or another assignment.

20687842	
Student number	Signature
C. Posthumus	5 November 2021
Initials and surname	Date

ACKNOWLEDGEMENTS

I would like to acknowledge Insight Actuaries & Consultants for the use of their data which was instrumental to this study. Additionally I would like to acknowledge and thank my supervisor, Dr Hofmeyr for his continuous guidance and support throughout this process.

The Department of Statistics and Actuarial Science wishes to acknowledge David Rodwell for generously creating a template based off the University of Stellenbosch Business School guidelines which have been adapted for the purposes of the department.

ABSTRACT

The primary focus of this study is to use statistical means to identify groups of chronic conditions that commonly co-occur from a dataset comprising of a large number of conditions.

In determining the inter-relatedness of conditions, the co-occurrence probabilities of each condition pair needs to be estimated. The co-occurrence probabilities are estimated in a pairwise manner, because allowing for higher-order interactions results in a computationally intractable problem when a large number of conditions are simultaneously considered. To this extent, the multivariate Bernoulli distribution can be studied.

This study proposes three models which can be used to estimate pairwise co-occurrence probabilities. The first model, the incomplete maximum likelihood estimation (MLE) model, is unregularised. The following two models proposed regularise the incomplete MLE model by means of penalisation and rank reduction.

The estimated co-occurrence probabilities are used to inform a similarity matrix, which is then used as an input to the spectral clustering algorithm. The normalised mutual information score is chosen as the similarity metric because it provides a measure of dependency between the conditions and is therefore appropriate for this problem.

The models are applied to observational data, but also to simulated data. Because clustering is an unsupervised learning problem, the model performance needs to be validated through simulation studies whereby the underlying dependence structure of the data are known. This study discusses and implements the simulation of dependent binary co-occurrence data.

The study found that two of the models showed favourable results, and can be used to cluster high-dimensional co-occurrence data.

Key words:

Co-occurrence modelling; multi-morbidity; dependent binary simulation; spectral clustering

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF ABBREVIATIONS AND/OR ACRONYMS

ICD	International Statistical Classification of Diseases and Related Health Problems
MLE	Maximum Likelihood Estimation
RRMM	Rank Regularised Mixture Model
NMI	Normalised Mutual Information
CM	Clustering Metric

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Co-occurrence modelling is an instrumental tool in the context of unsupervised learning. There are many applications where a network structure is to be analysed, and the components grouped. In this study we construct a network of chronic conditions and investigate the pairwise relationships between these conditions. The aim of this is to determine the inter-relatedness between the chronic conditions so that they can be clustered using spectral clustering. In order to characterise the similarity matrix for clustering, we estimate the co-occurrence probabilities using three techniques, two of which are regularised estimation approaches.

These methods are applied to both real co-occurrence data and simulated data. The aim of including a simulation study is to validate the performance of the models on data with a known underlying distribution.

This study aims to apply clustering to a large number of chronic conditions. The majority of publicised multi-morbidity studies involve only a small subset of conditions so a goal of this study is to analyse multi-morbidity patterns across a multitude of conditions. The nature of this data is that it is inherently sparse and so the proposed models should be computationally efficient and able to perform adequately when faced with high-dimensional sparse data.

1.2 PROBLEM STATEMENT

The primary focus of this study is to cluster chronic conditions into inter-related groups. The study thus aims to answer the following question.

In the data, do disease clusters exist, and if so what chronic conditions are inter-related?

The structure of co-occurrence data is high-dimensional, sparse and binary. Manipulating this type of data to reveal clusters poses many challenges, which this study aims to overcome through the use of simplified and regularised models.

1.3 CHAPTER PLAN

In Chapter 2, we provide a justification for the use of multi-morbidity as an application area, and describe existing co-occurrence models.

In Chapter 3, we propose and discuss three approaches to estimate co-occurrence probabilities and discuss the importance of regularisation in co-occurrence modeling.

In Chapter 4, we discuss the importance of simulation studies for model validation and describe two techniques of simulating co-occurrence data that has an underlying dependence structure.

In Chapter 5, we outline the methodology used to cluster inter-related chronic conditions. In this chapter we also describe how model parameters were tuned.

In Chapter 6, we discuss the results of the analysis on the simulated and real data.

1.4 CLARIFICATION OF KEY CONCEPTS

1.4.1 ICD 10 codes

ICD stands for 'International Statistical Classification of Diseases and Related Health Problems', however it is more commonly known as 'International Classification of Diseases'. Healthcare professionals use ICD codes to classify conditions and causes of injury for standardized and consistent reporting. ICD 10 refers to a specific version of the ICD codes.

In this study we define chronic conditions by their corresponding ICD 10 code.

1.4.2 Chronic condition

The Centers for Disease Control and Prevention (CDC) defines chronic conditions as 'conditions that last 1 year or more and require ongoing medical attention or limit activities of daily living or both' (Centers for Disease Control and Prevention, 2021).

The way that chronic conditions are defined in the study must be aligned with the data provided to us. The definition we then use to describe an individual with a chronic condition is as follows,

> *"An individual is defined as having a chronic condition if they have ever claimed for medication that was paid out from a chronic benefit by their medical scheme. The*

beneficiary becomes chronic on the date which they first claimed for chronic medication, and are not assumed to recover."

This definition makes use of the simplifying assumption that a chronic condition cannot be cured.

1.4.3 Co-morbidity and multi-morbidity

A disease co-morbidity is the simultaneous existence of two chronic conditions in an individual. The term 'multi-morbidity' is an extension of the term 'co-morbidity', in that it also allows for the presence of more than two chronic conditions co-existing in an individual.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

In this chapter we use literature to justify the use of investigating co-morbidities as the application to a co-occurrence model. Additionally, we investigate existing co-occurrence models such as the multivariate Bernoulli distribution, the Ising model and a co-clustering approach proposed by Ng (2015).

2.2 JUSTIFICATION OF DISEASE CO-OCCURRENCE APPLICATION

Co-occurrence modelling has typically been applied in modelling species co-occurrence, or in the context of natural language processing, where word co-occurrence within documents is analysed. These applications would be equally acceptable in the context of this research. However, the application in co-morbidities provides certain advantages.

Firstly, multi-morbidity is a societal problem with far reaching impacts, and through this study we hope to better understand the inter-relatedness of chronic conditions. Education can be considered one of the first steps towards policy and societal change. Secondly, model validation with co-morbidity data is comparatively unchallenging. Lastly, this study is novel in that few studies have been conducted which model disease co-occurrence of a large number of conditions.

Justification through the Widespread Adverse Impacts of Multi-morbidity

Studies have shown that multi-morbidity results in poor healthcare outcomes, including a reduced life quality and higher mortality rates (Arokiasamy *et al.*, 2015). Additionally, multi-morbidity leads to obstacles in working environments. A cross-sectional study in Australia and Japan by Sum *et al.* (2020) revealed that multi-morbid individuals have lower work productivity and worse performance. This is due to the increased number of sick leave days required by multi-morbid people and the lower odds of a multi-morbid person being employed despite being in the labour force (Sum *et al.*, 2020).

As well as being a healthcare issue, multi-morbidity comes with substantial economic ramifications

due to the increased costs associated with disease co-occurrence. Multi-morbid individuals have an increased healthcare utilization which is compounded by health care costs associated with the interactions between conditions. The increased healthcare utilization of multi-morbid individuals results in significant out-of pocket expenditure.

To quantify the cost of multi-morbidity to the healthcare system we need to quantify:

1. What proportion of individuals face multi-morbidity?
2. To what extent does the healthcare system carry the costs associated with multi-morbidity? Does most of the expenditure come from out-of-pocket payments or healthcare utilisation?

In evaluating the prevalence of multi-morbidity, can review various studies. A South African study estimated the prevalence of multi-morbidity of the adult population to be 4%, however the authors acknowledge that this may be an underestimation of the true prevalence because only a few chronic diseases were included in the analysis (Alaba and Chola, 2013). Another South African study was conducted to determine the prevalence of chronic multi-morbidity among older adults in rural South Africa. This study included HIV, which is an important condition to consider in the South African context. The estimated multi-morbidity prevalence was 69.4% with 95% confidence interval (68.0%, 70.9%) (Chang *et al.*, 2019). Only 10 conditions were considered in this study. Another study estimated the prevalence of non-communicable disease multi-morbidity in South Africa as 32% (Agrawal and Agrawal, 2016). This study only analysed nine chronic conditions, and HIV was not part of the study. A study by Peltzer (2018) was conducted to estimate the prevalence of non-communicable disease (NCD) multi-morbidity among tuberculosis patients in South Africa. The prevalence of co-morbidity with one NCD was 26.9% and the prevalence of multi-morbidity with two or more NCDs was 25.3%.

The above studies show great disparity in their estimation of the prevalence of multi-morbidity in South Africa with estimates ranging from 4% to 69.4%. It is therefore difficult to surmise the proportion of South Africans which face multi-morbidity. The study which produced the highest estimate of multi-morbidity included HIV in the set of diseases considered. Given the large scale of the HIV-epidemic in Southern Africa, including HIV in the study of multi-morbidity seems intuitive. However, this study focuses on older adults in rural regions. Lower levels of income and increasing age are factors which increase the probability of an individual.

In quantifying the cost of multi-morbidity to the healthcare system, we must consider what proportion of multi-morbidity related costs are paid out-of-pocket and what proportion is covered by the healthcare system. Larkin *et al.* (2020) conducted a systematic review of qualitative research relating to the financial burden associated with multimorbidity. This section summarises findings of the review. Many studies have attributed the increased medical expenses associated with multi-morbidity to poor coordination between healthcare providers. Indirect costs associated with poor coordination will fall on the multi-morbid individual - i.e. increased travel expenditure and time off work. The review indicated that whilst most multi-morbid study participants had supplemental healthcare insurance, few participants reported that the level of coverage was sufficient, with co-payments forming a large portion of the out-of-pocket expenditure. Cases are referenced where the high cost of care lead to poorer medical outcomes - the high-cost of medicine resulted in medicine non-adherence and high consultation costs have resulted in patients not seeking further care.

According to a publication from the World Health Organisation, multi-morbidity is more common in disadvantaged groups (World Health Organization, 2016*b*). This contributes to health inequality because economically disadvantaged individuals may not be able to afford the high out-of-pocket costs associated with multi-morbidity.

The healthcare objective of this study is to improve understanding of common co-morbidities. This endeavor is worthwhile since multi-morbidity is a societal problem. As outlined previously, many individuals have co-morbidities and this results in significant costs to these individuals as well as medical schemes and other funding parties. Multi-morbid individuals experience worse mortality and quality of life. An improved understanding of disease co-occurrence could potentially mitigate these effects through education of prevention methods on part of medical aid schemes.

Justification through Model Validation

The application in multi-morbidity allows for easy model validation. There is an abundance of literature available on common co-morbidities, so the model results should mimic the literature. For example, the connection between diabetes and hypertension is well documented. A medically reviewed article states that 'most people with diabetes will eventually have high blood pressure' (WebMD, 2021). Therefore, if our model is applied to medical data, we expect the model to

indicate a strong pairwise connection between diabetes and hypertension. Similarly, our model should show no connection between antagonistic chronic conditions. Intuitively, individuals will not be simultaneously affected by high blood pressure (hypertension) and low blood pressure (hypotension). These known antagonistic diseases should be expressed accordingly in our study.

Whilst it would be beneficial if the co-occurrence model validates common co-morbidities, it would also be valuable if the model highlighted statistically significant co-morbidities which are not well-documented.

Justification through Novelty

To my knowledge, publicised multi-morbidity studies often include very few chronic conditions. In a South African context, few multi-morbidity studies have been published. Studies of multi-morbidity in a South African context include papers by Alaba and Chola (2013), Chang *et al.* (2019), Agrawal and Agrawal (2016) and Peltzer (2018). A limitation shared by these studies is that only small sets of chronic conditions were included in the analyses. Co-morbidity data was made available to us that includes a comprehensive list of chronic disorders. The high-dimensionality of the data used in this study could result in the study being a valuable addition to existing literature.

Therefore, the multi-morbidity application is justified through the severity of the problem, unchallenging model validation and the novelty of high-dimensional disease co-occurrence modelling.

2.3 EXISTING MODELS

2.3.1 Multivariate Bernoulli Distribution

The multivariate Bernoulli distribution provides a comprehensive framework through which a binary graph structure can be estimated (Dai *et al.*, 2013). Applying this framework to the context of disease co-occurrence modelling we can define the joint probability density function of a multivariate Bernoulli distribution.

Denote the D-dimensional vector of potentially correlated Bernoulli random variables by $\mathbf{Y}$ such

that,

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_D \end{bmatrix}$$

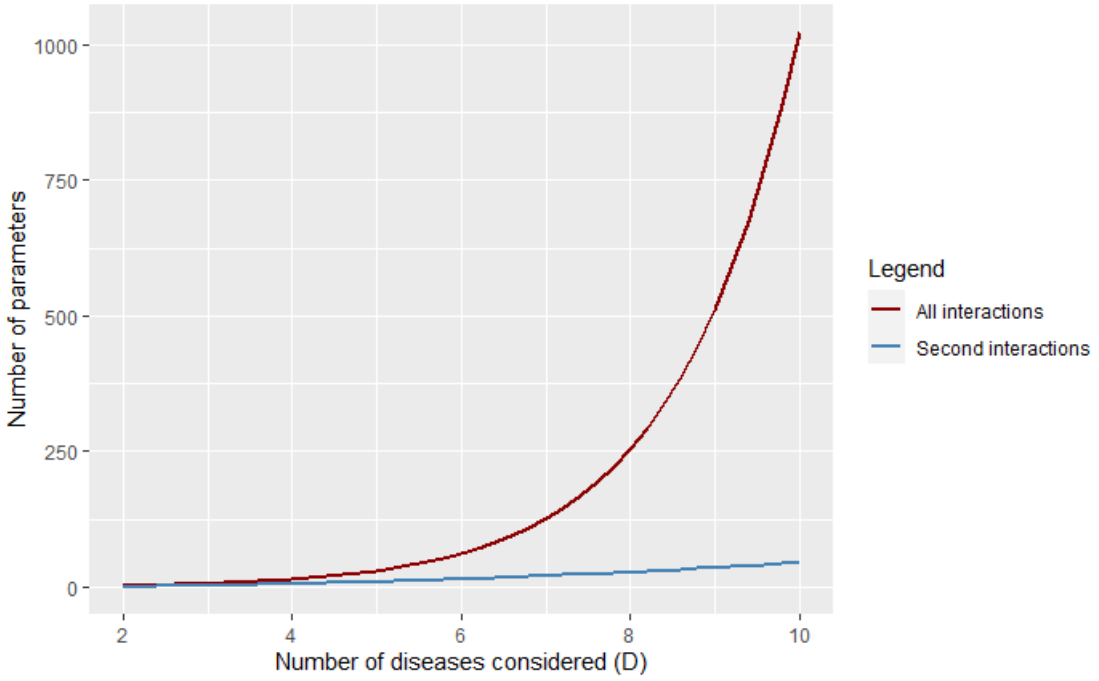
Suppose we denote a realisation of Y_i to be y_i, then the set of values which y_i can take on is $\{0,\ 1\}$. Applying this to the application of multi-morbidity modelling, we can interpret each realisation of $\mathbf{Y}$ to be an expression of an individual's chronic diseases. If an individual presents the i^{th} chronic condition, then $y_i = 1$, otherwise $y_i = 0$. The number of chronic diseases simultaneously occurring in an individual is the sum of the elements in $\mathbf{y}$, i.e. $\sum_{i=1}^{D} y_i$. The dimension of $\mathbf{Y}$ is the number of chronic diseases considered in the analysis, denoted D.

The joint probability density function of a multivariate Bernoulli distribution is,

$$\begin{aligned} &P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_D = y_d) \\ &= P(Y_1 = 0, Y_2 = 0, \ldots, Y_D = 0)^{\prod_{i=1}^{D}(1-y_i)} \times P(Y_1 = 1, Y_2 = 0, \ldots, Y_D = 0)^{y_1 \times \prod_{i=2}^{D}(1-y_i)} \\ &\quad \times P(Y_1 = 0, Y_2 = 1, \ldots, Y_D = 0)^{(1-y_1)y_2 \times \prod_{i=3}^{D}(1-y_i)} \times \ldots \times P(Y_1 = 1, Y_2 = 1, \ldots, Y_D = 1)^{\prod_{i=1}^{D} y_i} \end{aligned}$$

The multivariate Bernoulli distribution is clearly incredibly comprehensive, due to all combinations of diseases being considered. The total number of estimable parameters in the multivariate Bernoulli distribution is $2^D - 1$ (Dai *et al.*, 2013). The models we suggest only consider the pairwise interactions between diseases. The number of unique second interactions in the multivariate Bernoulli distribution is $\frac{D(D-1)}{2}$, which is equivalent to the number of pairwise co-morbidities (Dai *et al.*, 2013). Figure 2.1 illustrates how quickly the number of parameters in the multivariate distribution increases as more diseases are considered. The number of parameters corresponding to second interactions, the pairwise connection between diseases, makes up an increasingly small proportion of the total number of parameters in the multivariate Bernoulli distribution. For large values of D, the estimation of parameters in the multivariate Bernoulli distribution will be computationally infeasible. Allowing higher order interactions provides a more holistic view of disease interactions. The data utilised in this study comprises of more than 500 conditions. If all pairwise conditions are to be considered simultaneously, the multivariate

Figure 2.1: The number of parameters associated with the multivariate Bernoulli distribution



Bernoulli model will be unsuitable compared to a simplified model.

2.3.2 Ising Model

The Ising model is a mathematical model which was originally designed to describe magnetic phase transitions (Singh, 2020). This model has applications in statistics as it can be utilised in estimating graph structures with binary nodes. In fact, the multivariate Bernoulli distribution described in the previous section is an extension of the Ising model (Dai *et al.*, 2013). We can show that the Ising model is a simplified version of the multivariate Bernoulli distribution.

Consider the multivariate Bernoulli distribution defined previously. Under the multivariate

Bernoulli distribution framework described by Dai *et al.* (2013), allow

$$\begin{aligned} f^{c_i} &= \text{the probability density function of the } i^{th} \text{ condition} \\ f^{c_i,c_j} &= \text{the joint probability density function of the } i^{th} \text{ and } j^{th} \text{ conditions} \\ &\vdots \\ f^{\{c_1,c_2,\ldots,c_r\}} &= \text{the joint probability density function of the set of conditions } \{c_1, c_2, \ldots, c_r\} \\ \tau &= \text{the set of all possible superscripts of } f \end{aligned}$$

Then we can write,

$$S^{\tau} = \sum_{\tau_0 \subseteq \tau} f^{\tau_0}$$

The assumption underlying the Ising model is that $S^{\tau} = 0$ for any set τ which has a minimum of 3 elements. Additionally, the parameter coefficients of the Ising model can be represented using the S variables. Continuing with the notation used in Dai *et al.* (2013), we can write the probability mass function of $\mathbf{Y}$ under the Ising model as follows,

$$P(Y_1, Y_2, \ldots, Y_D) = \frac{1}{Z(\Theta)} \exp\left(\sum_{i=1}^{D} \theta_{ii} Y_i + \sum_{1 \leq i < i' \leq D} \theta_{ii'} Y_i Y_{i'} \right)$$

where $\theta_{ii'} : ii' \in \{1, 2, \ldots, D\}$ are a set of parameters and $Z(\Theta)$ is a partition function which ensures that the sum of the probabilities equals one (De Canditiis, 2020). The Ising model parameters and the S values under the multivariate Bernoulli distribution are linked since,

$$\theta_{ii'} = S^{ii'} \ \forall \ i, i' \in \{1, 2, \ldots, D\} \qquad\qquad Dai et\ al.\ (2013)$$

The difference between the two models is the order of the interactions considered and therefore the number of estimable parameters. Suppose we continue notating the number of diseases in a model by D. The Ising model considers only pairwise interactions and thus requires the estimation of $\frac{D(D+1)}{2}$ parameters, $\theta_{ii'} : ii' \in \{1, 2, \ldots, D\}$. In comparison, the multivariate Bernoulli model includes all possible combinations of chronic conditions leading to a much higher

number of estimable parameters, $2^D - 1$ (Dai *et al.*, 2013).

An interesting extension of the Ising model is outlined by Ravikumar *et al.* (2010). The Ising model only accounts for pairwise connections, however higher-order interactions can be introduced into the Ising model through the introduction of auxiliary variables. This will, however, increase the number of estimable parameters which may not be appropriate when a large number of conditions are considered.

2.3.3 Co-Clustering

Co-clustering is a method that simultaneously performs clustering on both the rows and the columns of a matrix. This clustering technique is beneficial when what you are trying to group is dependent on two separate variables. A common example is co-clustering on user ratings for a set of films. The user ratings are dependent on user preference and the films' characteristics; thus, user preference could form the rows of a matrix, and the columns could represent film characteristics (Reshef, 2015).

A paper by S.K. Ng used co-clustering to identify groups of commonly co-occurring illnesses. Co-morbidities are identified through the simultaneous clustering of individuals and health conditions (Ng, 2015). This approach is beneficial because disease clustering would be dependent on both the nature of the conditions being clustered as well as the demographics of the multi-morbid group of individuals. Diseases do not affect the whole population identically. For instance, males are more prone to developing colour blindness (National Eye Institute, 2019), and only females will experience postnatal depression. Therefore, clustering diseases based only on the expression of diseases may result in a loss of information because demographical information has the ability to inform the clustering solution.

This method was applied to data consisting of 24 conditions, of which only 77 pairs of conditions were found to be significant at a 1% significance level. The data being used in our research consists of 508 conditions after data cleaning which is exceedingly more conditions than what is utilised in the study by S.K. Ng. It is not apparent whether this method could be applied to high-dimensional data.

The paper is summarized by the following algorithm.

Algorithm 1 Co-Clustering Algorithm by Ng (2015)

procedure CLUSTERING HEALTH CONDITIONS(B)

Represent the data by an $n \times p$ binary matrix B with the rows being individuals (n) and the columns being the health conditions (p).

if $B_{ij} = 1$ **then**

Individual i presents condition j

else if $B_{ij} = 0$ **then**

Individual i does not present condition j

Use the similarity measure Somers' D statistic to obtain a $p \times p$ matrix with entries ranging between -1 and 1. These entries represent the direction and magnitude of the association between each pair of conditions.

Apply the Benjamini–Hochberg procedure to the $p \times p$ matrix to reduce the false discovery rate. The resulting matrix is a binary, symmetric $p \times p$ matrix which is denoted as M.

Make use of a clumping clustering method to identify groups of co-morbid health conditions in the matrix M. This step will result in overlapping clusters of illnesses.

Follow an iterative procedure to obtain unique non-overlapping groups of conditions by maximising the strength of the clusters.

procedure CLUSTERING INDIVIDUALS(B)

Convert the data matrix B to an $n \times q$ matrix where q denotes the number of non-overlapping groups of co-morbid health conditions.

Allow $\mathbf{y}_j = (y_{1j}, ..., y_{qj})', \ j = 1, ..., n$ where

$$y_{ij} = \begin{cases} 1, & \text{if no conditions in } i^{th} \text{ group} \\ 2, & \text{if one condition in } i^{th} \text{ group} \\ 3, & \text{if } > 1 \text{ condition in } i^{th} \text{ group} \end{cases}$$

Use a finite mixture model of multivariate generalized Bernoulli distributions to cluster the $\mathbf{y}_j$.

CHAPTER 3

CO-OCCURRENCE MODELLING

3.1 INTRODUCTION

In this chapter, we propose three methods that can be used to estimate co-occurrence probabilities. The first method is called the incomplete maximum likelihood estimation approach and it is an unregularised model. Two regularised versions of this model are discussed which perform regularisation by penalisation and by rank reduction.

The purpose of estimating co-occurrence probabilities is to inform the similarity matrix used in clustering conditions.

3.2 PROBLEM FORMULATION AND NOTATION

The primary focus of this study is to identify groups of chronic conditions which appear to be inter-related. The similarity metric used in clustering chronic conditions makes use of the pairwise relationships between conditions to infer dependency between the conditions. This way of grouping conditions has the potential to expose interactions between multiple conditions due to the weak transitivity of dependence among random variables.

Suppose condition X is found to be related to condition Y, and condition Y is found to be related to condition Z. Then by transitive dependency, we can infer that condition X is most likely related to condition Z. Networks of inter-related conditions can be built using this principle.

In order to inform the similarity matrix, we need to model the dependency between conditions by identifying pairs of chronic conditions that co-occur at a higher or lower rate than what would be expected under the assumption that conditions occur independently.

In order to determine whether conditions co-occur at a different rate than under the independence assumption, we must define what we hereafter refer to as the null model. The null model describes the occurrence of disease co-morbidities when it is assumed that there is no association between diseases and co-occurrences of diseases are random.

In describing chronic disease co-occurrence, the estimation of the following two matrices is vital:

1. The matrix of conditional co-occurrence probabilities, denoted as Π^C.
2. The matrix of joint co-occurrence probabilities, denoted as Π^J.

The collection of conditions considered in this study is extensive, so regularised forms of the underlying statistical model will be investigated in order to estimate these matrices. If we assume that D chronic conditions are considered in the analysis, then allow $X_i,\ i = \{1, ..., D\}$ to define the event that an individual presents chronic condition i. Then, the entries of the conditional and joint probability matrices take the form,

$$\begin{aligned}
[\Pi^C]_{ij} &= P(X_i|X_j) \\
&= \text{the probability that an individual has condition } i \text{ given that the individual} \\
&\quad \text{has condition } j \\
[\Pi^J]_{ij} &= P(X_i \cap X_j) \\
&= \text{the probability that an individual has both condition } i \text{ and condition } j
\end{aligned}$$

Under the null model it is assumed that,

$$\begin{aligned}
\Pi^J_{ij} &= P(X_i \cap X_j) \doteq P(X_j)P(X_j) \\
\Pi^C_{ij} &= \frac{P(X_i \cap X_j)}{P(X_j)} \doteq \frac{P(X_i)P(X_j)}{P(X_j)} = P(X_i)
\end{aligned}$$

Because of the large number of conditions under consideration, we estimate these matrices using only the matrix of co-occurrence counts. This matrix gives an indication of the prevalence of each condition within the study sample and describes the pairwise co-occurrence of conditions. Multivariate Bernoulli models take into account higher order interactions between conditions, as opposed to only the pairwise interactions, however they are found to be computationally intractable when a large number of conditions are included in the analysis.

We denote the matrix of co-occurrence counts as N such that $[N]_{ij}$ represents the number of individuals in the sample with chronic condition i as well as chronic condition j, where $i, j \in \{1, ..., D\}$. The diagonal of this matrix, $\{N_{11}, N_{22}..., N_{DD}\}$, gives the prevalence of each chronic condition, or equivalently the number of individuals in the sample with each condition.

For simplicity, denote $N_{ii} \equiv N_i$.

3.3 INCOMPLETE MAXIMUM LIKELIHOOD ESTIMATION

As a baseline for further analysis, we consider a simple estimation approach based on pairwise maximum likelihood. In this approach, we estimate the elements in the conditional and joint probability matrices by their standard maximum likelihood estimates. This simplistic approach treats each of the counts in N as independent of the others, which is clearly not a true reflection of the underlying mechanism generating co-occurrence counts. Modelling the dependence structure of conditions more accurately whilst adhering to a valid probability model results in an intractable problem as seen in the Multivariate Bernoulli model. This limitation is acceptable for our objectives given that our primary interests are in pairwise comparisons with the null model for clustering.

To reflect our acknowledgement that the estimation we perform does not accommodate the full underlying distribution of conditions, we refer to this model as the "incomplete maximum likelihood estimation model" and we refer to its objective functions as "incomplete log-likelihood functions".

The incomplete log-likelihood functions for the conditional and joint probability matrices are denoted $f(\Pi^C)$ and $f(\Pi^J)$ respectively. These functions form part of the objective functions of regularised co-occurrence models proposed in later sections. The reason that these log-likelihood functions are considered 'incomplete' is because they only consider the pairwise interactions between chronic conditions. This is in contrast with the likelihood function associated with the multivariate Bernoulli distribution which incorporates all combinations of diseases.

We first consider the estimation of the conditional probability matrix. For simplicity, denote $[\Pi^C]_{ij} = \pi_{ij}^C$ for $i, j \in \{1, ..., D\}$. The incomplete log-likelihood function associated with the conditional probability matrix is,

$$f(\Pi^C) = \sum_{i \neq j = 1}^{D} N_{ij} \ln(\pi_{ij}^C) + (N_j - N_{ij}) \ln(1 - \pi_{ij}^C)$$

or equivalently,

$$f(\Pi^C) = \sum_{i \neq j=1}^{D} N_{ij} \ln\left(\frac{\pi_{ij}^C}{1-\pi_{ij}^C}\right) + N_j \ln(1-\pi_{ij}^C)$$

The maximum likelihood estimates for $\pi_{ij}^C,\ i,j \in \{1,...,D\}$ can now be derived.

$$\begin{aligned}\frac{\partial f(\Pi^C)}{\partial \pi_{ij}^C} &= \frac{\partial}{\partial \pi_{ij}^C}\left(\sum_{i \neq j} N_{ij}\ln(\pi_{ij}^C) + (N_j - N_{ij})\ln(1-\pi_{ij}^C)\right) \\ &= \frac{N_{ij}}{\pi_{ij}^C} - \frac{N_j - N_{ij}}{1-\pi_{ij}^C}\end{aligned}$$

Setting $\frac{\partial f(\Pi^C)}{\partial \pi_{ij}^C} = 0$ yields,

$$\begin{aligned}\frac{N_{ij}}{\hat{\pi}_{ij}^C} - \frac{N_j - N_{ij}}{1-\hat{\pi}_{ij}^C} &= 0 \\ N_{ij} - N_{ij}\hat{\pi}_{ij}^C &= N_j\hat{\pi}_{ij}^C - N_{ij}\hat{\pi}_{ij}^C \\ \hat{\pi}_{ij}^C &= \frac{N_{ij}}{N_j}\end{aligned}$$

We can show that this estimate maximises the incomplete log-likelihood function.

$$\begin{aligned}\frac{\partial}{\partial \pi_{ij}^C}\left(\frac{N_{ij}}{\pi_{ij}^C} - \frac{N_j - N_{ij}}{1-\pi_{ij}^C}\right) &= -\frac{N_{ij}}{{\pi_{ij}^C}^2} - \frac{N_j - N_{ij}}{\left(1-\pi_{ij}^C\right)^2} \\ \implies \frac{\partial^2}{\partial {\pi_{ij}^C}^2} f(\Pi^C) &< 0\end{aligned}$$

Hence, $f(\Pi^C)$ is maximised by maximum likelihood estimate $\hat{\pi}_{ij}^C = \frac{N_{ij}}{N_j}$.

Similarly we can define the incomplete log-likelihood function associated with the joint probability

matrix, Π^J.

$$\begin{aligned} f(\Pi^J) &= \sum_{i,j=1}^{D} N_{ij} \ln(\pi_{ij}^J) + (N_{sample} - N_{ij}) \ln(1 - \pi_{ij}^J) \\ &= \sum_{i,j=1}^{D} N_{ij} \ln\left(\frac{\pi_{ij}^J}{1 - \pi_{ij}^J}\right) + N_{sample} \ln(1 - \pi_{ij}^J) \end{aligned}$$

where $\pi_{ij}^J = [\Pi^J]_{ij}$, and N_{sample} is the total number of beneficiaries in the study sample. We can derive the maximum likelihood estimates of π_{ij}^J in a similar manner to the conditional maximum likelihood estimates. We find,

$$\hat{\pi}_{ij}^J = \frac{N_{ij}}{N_{sample}}$$

It can be shown that $\hat{\pi}_{ij}^J$ maximises $f(\Pi^J)$, the incomplete log-likelihood function. The maximum values of the incomplete log-likelihood functions are the matrices of empirical proportions as shown above. The incomplete log-likelihood functions then reward solutions which align closely with the data.

Estimating the conditional probability matrix using this method results in the estimation of $D \times (D-1)$ probabilities, and estimating the joint probability matrix will result in the estimation of $\frac{D(D+1)}{2}$ parameters. The incomplete MLE model is therefore relatively simple in that only $\frac{D(D+1)}{2}$ parameters are required to describe the pairwise joint distributions of condition occurrences. The full multivariate Bernoulli model requires considerably more parameters. However, when D is large the number of parameters underlying the incomplete MLE model is potentially problematic as the model may have a high estimation variance unless the co-occurrence counts for all pairs of conditions are large enough. To reduce this variation, and also introduce dependence between the estimates in the joint and conditional matrices, we consider two regularised forms of the incomplete MLE model.

3.4 REGULARISATION METHODS

We consider two approaches to regularisation, namely:

1. Regularisation by penalisation
2. Regularisation by rank reduction

3.5 REGULARISATION BY PENALISATION

In the unregularised, incomplete MLE approach we seek to estimate parameters by solving the following optimisation function.

$$\begin{aligned}&\arg\max_{\pi_{ij}^J}\left(f\left(\pi_{ij}^J\right)\right)\\&\equiv\arg\max_{\pi_{ij}^J}\left(\sum_{i,j=1}^{D}N_{ij}\ln\left(\pi_{ij}^J\right)+\left(N_{sample}-N_{ij}\right)\ln\left(1-\pi_{ij}^J\right)\right)\end{aligned}$$

Regularisation by penalisation applies a penalty to the optimisation function in order to reduce estimation variance, and to improve the generalisation ability of the model. By constraining the optimisation problem, the model will be able to better accommodate unseen data. Suppose we formulate an appropriate penalty function $Q\left(\pi_{ij}^J\right)$, then the optimisation function becomes,

$$\arg\max_{\pi_{ij}^J}\left(f\left(\pi_{ij}^J\right)-\lambda Q\left(\pi_{ij}^J\right)\right)$$

where λ is a tuning parameter controlling how large the penalty should be. The choice of the penalty function should depend on the context of the study. In this study we regularised the incomplete MLE model by applying an L_1 penalty which penalises deviations in the co-occurrence probabilities from the null model. Under the null model, conditions are assumed to occur independently of one another.

The L_1 penalty makes use of the absolute value function which enables it to produce sparse solutions. This property is specifically advantageous in this study. The co-occurrence model should be able to detect large deviations from the null model. Equivalently, the model should be able to detect pairs of conditions that have a higher than random chance of co-occurring, or a lower than random chance of co-occurring. It would therefore be beneficial if the co-occurrence probabilities corresponding to effectively independent conditions were to shrink to zero exactly.

This method can be utilised to estimate either the joint probability matrix or the conditional probability matrix. We first wish to estimate the conditional probability matrix, Π^C. Under the independence assumption of the null model, $\pi_{ij}^C = P(X_i) = \pi_i$. Since the conditional co-

occurrence probability π_{ij}^C under the null model is only dependent on the i^{th} condition, we denote this probability by π_i^0.

To achieve regularisation, we maximise the objective function:

$$g\left(\pi_{ij}^C,\ \pi_i^0\right) = f(\Pi^C) - \lambda \sum_{i \neq j=1}^{D} |\pi_{ij}^C - \pi_i^0|$$

where λ is a tuning parameter and $f(\Pi^C)$ represents the incomplete log-likelihood function previously described.

We can also estimate the joint probability matrix, Π^J using this method. We denote the vector of marginal probabilities by $\boldsymbol{\pi}$ where,

$$\Pi^J = \begin{bmatrix} \pi_1 & \pi_{12}^J & \dots & \pi_{1D}^J \\ \pi_{21}^J & \pi_2 & \dots & \pi_{2D}^J \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{D1}^J & \pi_{D2}^J & \dots & \pi_D \end{bmatrix} = \begin{bmatrix} \pi_1 & \pi_{12}^J & \dots & \pi_{1D}^J \\ \pi_{12}^J & \pi_2 & \dots & \pi_{2D}^J \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{1D}^J & \pi_{2D}^J & \dots & \pi_D \end{bmatrix} \qquad \textit{Symmetry of probabilities}$$

$$\boldsymbol{\pi} = \operatorname{diag}\left(\Pi^J\right) = \begin{bmatrix} \pi_1 & \pi_2 & \dots & \pi_D \end{bmatrix}'$$

In the estimation of the conditional probability matrix, we jointly estimate the conditional probability matrix Π^C and the vector of marginal probabilities $\boldsymbol{\pi}$. In the estimation of the joint probability matrix, however, the vector of marginal probabilities is the diagonal of the joint probability matrix and so the only object requiring estimation is the joint probability matrix.

Under the assumption that chronic conditions X_i and X_j occur independently of one another,

$$\pi_{ij}^J \doteq \pi_i \pi_j$$

We can use this to construct the following objective function:

$$\begin{aligned} g(\pi_{ij}^J, \pi_i, \pi_j) &= f\left(\Pi^J\right) - \lambda \sum_{i \neq j=1}^{D} |\pi_{ij}^J - \pi_i \pi_j| \\ &= \sum_{i,j=1}^{D} \left(N_{ij} \ln \left(\frac{\pi_{ij}^J}{1 - \pi_{ij}^J} \right) + N_{sample} \ln \left(1 - \pi_{ij}^J\right) \right) - \lambda \sum_{i \neq j=1}^{D} |\pi_{ij}^J - \pi_i \pi_j| \end{aligned}$$

where N_{sample} denotes the number of individuals in the study sample, λ is a tuning parameter and $f\left(\Pi^J\right)$ is the incomplete log-likelihood function previously defined. This objective function penalises deviations from the independence model. The L_1 penalty is advantageous because it has the potential to obtain sparse solutions with respect to the matrix

$$\hat{\Pi}^J - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}'$$

This matrix represents the probability that the i^{th} and j^{th} chronic conditions co-occur over and above the co-occurrence probability under the assumption that the conditions occur independently of one another. The larger the ij^{th} element of the matrix is, the more likely the pairwise conditions form a co-morbidity. Similarly, we can infer that if the ij^{th} element of the matrix is largely negative, the i^{th} and j^{th} conditions are likely antagonistic. Sparsity in the above matrix is therefore advantageous because it allows us to infer where the important interactions between conditions exist.

3.5.1 Maximisation

Where possible the objective functions $g\left(\pi_{ij}^C,\ \pi_i^0\right)$ and $g\left(\pi_{ij}^J, \pi_i, \pi_j\right)$ are maximised by differentiation. Otherwise the objective functions are maximised using the properties of the L_1 penalty. Utilising the L_1 penalty as opposed to the L_2 penalty is advantageous because it can produce sparse solutions. However, the L_1 penalty makes use of the non-differentiable absolute value function and so maximisation of the objective function by gradient-based methods is not always possible.

In Table 3.1, we provide the partial derivatives of the objective functions of the penalised co-occurrence models. The calculation of the partial derivatives can be found in the Appendix.

Table 3.1: Partial derivatives of the objective functions for penalised co-occurrence models

Matrix	Partial derivative	Value
Conditional Probability Matrix, Π^C	$\frac{\partial g(\pi_{ij}^C,\ \pi_i^0)}{\partial \pi_{ij}^C}$	$\begin{cases} \frac{N_{ij}-N_j\pi_{ij}^C}{\pi_{ij}^C(1-\pi_{ij}^C)} - \lambda & \text{for } \pi_{ij}^C > \pi_i^0 \\ \frac{N_{ij}-N_j\pi_{ij}^C}{\pi_{ij}^C(1-\pi_{ij}^C)} + \lambda & \text{for } \pi_{ij}^C < \pi_i^0 \\ \text{undefined} & \text{for } \pi_{ij}^C = \pi_i^0 \end{cases}$
	$\frac{\partial g(\pi_{ij}^C,\ \pi_i^0)}{\partial \pi_i^0}$	$\begin{cases} \lambda & \text{for } \pi_{ij}^C > \pi_i^0 \\ -\lambda & \text{for } \pi_{ij}^C < \pi_i^0 \\ \text{undefined} & \text{for } \pi_{ij}^C = \pi_i^0 \end{cases}$
Joint Probability Matrix, Π^J	$\frac{\partial g(\pi_{ij}^J,\pi_i,\pi_j)}{\partial \pi_{ij}^J}$	$\begin{cases} \frac{N_{ij}-N_{sample}\pi_{ij}^J}{\pi_{ij}^J(1-\pi_{ij}^J)} - \lambda, & \text{for } \pi_{ij}^J > \pi_i\pi_j \\ \frac{N_{ij}-N_{sample}\pi_{ij}^J}{\pi_{ij}^J(1-\pi_{ij}^J)} + \lambda, & \text{for } \pi_{ij}^J < \pi_i\pi_j \\ \text{undefined}, & \text{for } \pi_{ij}^J = \pi_i\pi_j \end{cases}$
	$\frac{\partial g(\pi_{ij}^J,\pi_i,\pi_j)}{\partial \pi_i}$	$\begin{cases} \lambda\pi_j, & \text{for } \pi_{ij}^J > \pi_i\pi_j \\ -\lambda\pi_j, & \text{for } \pi_{ij}^J < \pi_i\pi_j \\ \text{undefined}, & \text{for } \pi_{ij}^J = \pi_i\pi_j \end{cases}$

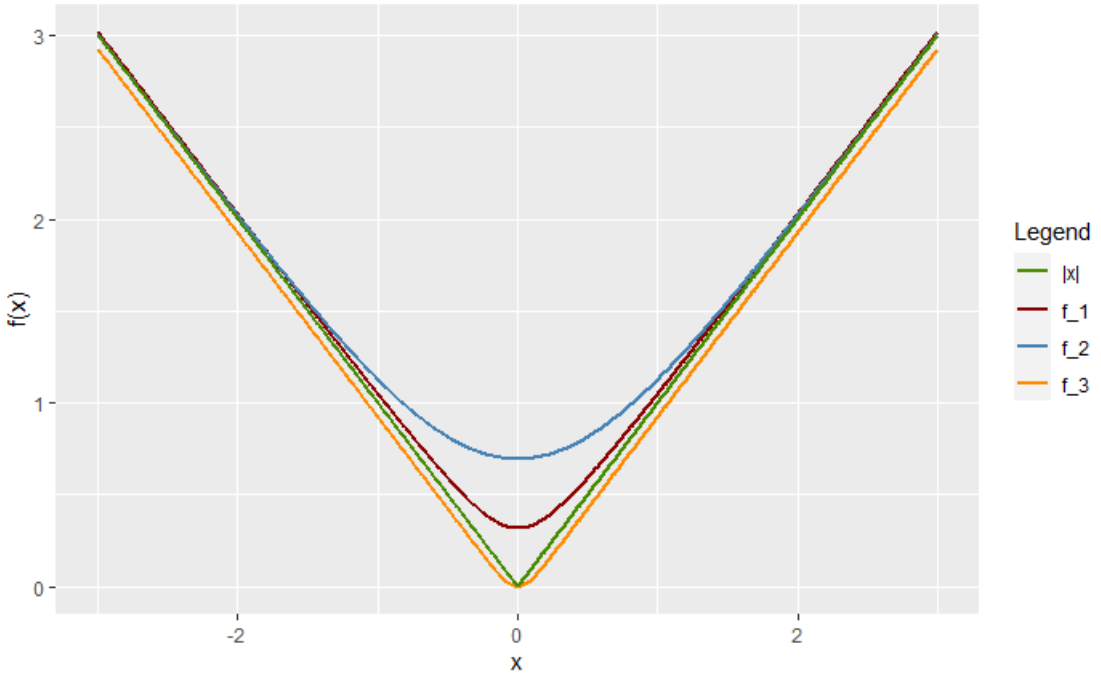
Methods of optimising an objective function containing the L_1 penalty must be discussed. One method includes performing gradient-based optimisation where the L_1 penalty is replaced with a differentiable approximation. Many approximations have been proposed to approximate the absolute value function and so a few methods will be discussed (Mathematics Stack Exchange User, 2021).

Figure 3.1 depicts approximations of $|x|$. This includes functions,

$$f_1(x) = \sqrt{x^2 + c},\ c > 0$$
$$f_2(x) = \ln\left(\exp\left(2x\right) + 1\right) - x$$
$$f_3(x) = \frac{1}{\alpha}\ln\left(\cosh\left(\alpha x\right)\right),\ \alpha > 0$$

The dark green line is the line we wish to approximate, and the other lines depict various approximations. The parameters used to create this plot are $\alpha = 9$ and $c = 0.1$. If a differentiable approximation of the absolute value function is used then gradient-based optimisation can be

Figure 3.1: Approximations of $|x|$



utilised to maximise,

$$g^A\left(\pi_{ij}^C,\ \pi_i^0\right) = f(\Pi^C) - \lambda \sum_{i \neq j=1}^{D} h^A(\pi_{ij}^C - \pi_i^0)$$

$$g^A(\pi_{ij}^J, \pi_i, \pi_j) = f\left(\Pi^J\right) - \lambda \sum_{i \neq j=1}^{D} h^A\left(\pi_{ij}^J - \pi_i \pi_j\right)$$

where $g^A\left(\pi_{ij}^C,\ \pi_i^0\right)$ and $g^A\left(\pi_{ij}^J, \pi_i, \pi_j\right)$ approximate $g\left(\pi_{ij}^C,\ \pi_i^0\right)$ and $g\left(\pi_{ij}^J, \pi_i, \pi_j\right)$ respectfully, and $h^A(x)$ is a differentiable approximation for $|x|$. Whilst utilising an approximation of the L_1 penalty is convenient because it allows for gradient-based optimisation, it loses the advantages that the L_1 penalty provides. The L_1 penalty has the potential to produce sparse solutions and this advantage is lost when an approximation is utilised.

Alternative methods can be utilised to optimise the non-differentiable concave optimisation functions, $g\left(\pi_{ij}^C,\ \pi_i^0\right)$ and $g\left(\pi_{ij}^J, \pi_i, \pi_j\right)$. One such method is the proximal gradient method. Under the proximal gradient method, the estimates can be iteratively updated even when parts of the objective function are non-differentiable (Parikh and Boyd, 2014).

3.6 REGULARISATION BY RANK REDUCTION

Using rank reduction as a regulariser is advantageous because it can be interpreted as a mixture model containing simplified mixture component distributions.

This section outlines two models within the mixture model framework - a rank regularised model and an extension of the rank regularised model which includes penalisation.

3.6.1 Rank Regularised Mixture Model

The rank regularised model jointly estimates the conditional probability matrix, Π^C, and the vector of marginal probabilities, $\boldsymbol{\pi}$. This approach achieves regularisation through a simplifying assumption. Suppose that each individual in the sample is of one of K person types. It is then assumed that given the person type of an individual, the events of having each of the conditions are independent. If T represents the person type, where $T \in \{1, ..., K\}$, then for any two conditions i and j, we have

$$P(X_i|X_j, T) \doteq P(X_i|T).$$

We can therefore write,

$$\begin{aligned} P(X_i|X_j) &= \sum_{t=1}^{K} P(X_i, T = t|X_j) \\ &= \sum_{t=1}^{K} P(X_i|T = t, X_j)P(T = t|X_j) \\ &= \sum_{t=1}^{K} P(X_i|T = t)P(T = t|X_j) \\ &= \Pi_{i:}^{X|T}\Pi_{:j}^{T|X}, \end{aligned}$$

where $\Pi_{ij}^{X|T} = P(X_i|T = j)$ and $\Pi_{ij}^{T|X} = P(T = i|X_j)$.

This is equivalent to modelling X as having a mixture distribution consisting of K components. The elements of X are independent, conditional on the component of the mixture distribution. The random variable which identifies the component which gives rise to a realisation of X is T,

the 'person type' of an individual. Making use of the mixture model framework is advantageous in that it produces a reduced rank estimator for Π^C, and it is also interpretable. The rank of the conditional probability matrix Π^C is at most equal to the number of allocated 'person types', K. Suppose we have a set of D chronic diseases and K person types. Then, we can show that the rank is at most K since,

$$\begin{aligned}
\hat{\Pi}^C &\doteq \hat{\Pi}^{X|T}\hat{\Pi}^{T|X} \\
\text{rank}\left(\hat{\Pi}^{X|T}\right) &\leq \min\left(D, K\right) \\
\text{rank}\left(\hat{\Pi}^{T|X}\right) &\leq \min\left(D, K\right) \\
\text{rank}\left(\hat{\Pi}^{C}\right) &\leq \min\left(\text{rank}\left(\hat{\Pi}^{T|X}\right), \text{rank}\left(\hat{\Pi}^{X|T}\right)\right) \leq \min\left(D, K\right)
\end{aligned}$$

This follows because the rank of an $n \times p$ matrix is less than the minimum of n and p (Taboga, 2017*b*), and the rank of a product of two matrices is bounded by the minimum rank of the two matrices (Taboga, 2017*a*). Since it can be assumed that for this application, $D > K$,

$$\text{rank}\left(\hat{\Pi}^{C}\right) \leq K < D$$

It is then shown that the mixture model approach produces a low-rank approximation for the conditional probability matrix, therefore performing regularisation through rank reduction.

Given that the entries of the $\Pi^{X|T}$ and $\Pi^{T|X}$ matrices represent probabilities, constraints imposed on these matrices are such that $0 \leq \Pi_{ij}^{X|T} \leq 1$, and $0 \leq \Pi_{ij}^{T|X} \leq 1$. Additionally, $\sum_i \Pi_{ij}^{T|X} = \sum_i P(T = i|X_j) = 1$. To adhere to the above constraints, a convenient parameterisation of the $\Pi^{X|T}$ and $\Pi^{T|X}$ matrices is,

$$\begin{aligned}
\Pi_{ij}^{X|T} &= \frac{\exp(Z_{ij})}{1 + \exp(Z_{ij})} \qquad \{Z_{ij} \in \mathbb{R}\} \\
\Pi_{ij}^{T|X} &= \frac{\exp(Q_{ij})}{\sum_{k=1}^{K} \exp(Q_{kj})} \qquad \{Q_{ij} \in \mathbb{R}\}
\end{aligned}$$

We want to maximise objective function,

$$\hat{f} = \sum_{i \neq j} N_{ij} \log\left(\widehat{P(X_i|X_j)}\right) + (N_j - N_{ij}) \log\left(1 - \widehat{P(X_i|X_j)}\right)$$

$\widehat{P(X_i|X_j)}$ can be rewritten in terms of variables $Z_{ij},\ i = \{1, 2, ..., D\},\ j = \{1, 2, ..., K\}$ and $Q_{ij},\ i = \{1, 2, ..., K\},\ j = \{1, 2, ..., D\}$.

$$\begin{aligned}
\mathbf{\Pi}^{X|T}_{i:} &= \left[\Pi^{X|T}_{i1},\ \Pi^{X|T}_{i2},\ \cdots\ , \Pi^{X|T}_{iK}\right] \\
\mathbf{\Pi}^{X|T}_{i:} &= \left[\frac{\exp(Z_{i1})}{1+\exp(Z_{i1})},\ \frac{\exp(Z_{i2})}{1+\exp(Z_{i2})},\ \cdots\ , \frac{\exp(Z_{iK})}{1+\exp(Z_{iK})}\right] \\
\mathbf{\Pi}^{T|X}_{:j} &= \begin{bmatrix} \Pi^{T|X}_{1j} \\ \Pi^{T|X}_{2j} \\ \vdots \\ \Pi^{T|X}_{Kj} \end{bmatrix} = \frac{1}{\sum_{t=1}^{K} \exp(Q_{tj})} \begin{bmatrix} \exp(Q_{1j}) \\ \exp(Q_{2j}) \\ \vdots \\ \exp(Q_{Kj}) \end{bmatrix}
\end{aligned}$$

Therefore we can write,

$$\begin{aligned}
\widehat{P(X_i|X_j)} &= \mathbf{\Pi}^{X|T}_{i:} \mathbf{\Pi}^{T|X}_{:j} \\
&= \frac{1}{\sum_{t=1}^{K} \exp(Q_{tj})} \left(\sum_{t=1}^{K} \frac{\exp(Z_{it}) \exp(Q_{tj})}{1+\exp(Z_{it})} \right) \\
&= \frac{1}{\sum_{t=1}^{K} \exp(Q_{tj})} \left(\sum_{t=1}^{K} \frac{\exp(Z_{it} + Q_{tj})}{1+\exp(Z_{it})} \right)
\end{aligned}$$

The objective then becomes to maximise,

$$\begin{aligned}
\hat{f} = \sum_{i \neq j} N_{ij}\ \ln &\left(\sum_{t=1}^{K} \frac{\exp(Z_{it} + Q_{tj})}{(1+\exp(Z_{it})) \sum_{t=1}^{K} \exp(Q_{tj})} \right) \\
&+ (N_j - N_{ij})\ \ln \left(1 - \sum_{t=1}^{K} \frac{\exp(Z_{it} + Q_{tj})}{(1+\exp(Z_{it})) \sum_{t=1}^{K} \exp(Q_{tj})} \right)
\end{aligned}$$

Partial Derivative of $\hat{f}$ with respect to Z_{it}

First we calculate:

$$
\begin{aligned}
&\frac{\partial}{\partial Z_{it}}\left(\frac{\exp(Z_{it}+Q_{tj})}{(1+\exp(Z_{it}))\sum_{t=1}^{K}\exp(Q_{tj})}\right)\\
&=\frac{\exp(Q_{tj})}{\sum_{t=1}^{K}\exp(Q_{tj})}\frac{\partial}{\partial Z_{it}}\left(\exp(Z_{it})(1+\exp(Z_{it}))^{-1}\right)\\
&=\frac{\exp(Q_{tj})}{\sum_{t=1}^{K}\exp(Q_{tj})}\left[-\exp(Z_{it})\left(1+\exp(Z_{it})\right)^{-2}\exp(Z_{it})+\left(1+\exp(Z_{it})\right)^{-1}\exp(Z_{it})\right]\\
&=\frac{\exp(Q_{tj})}{\sum_{t=1}^{K}\exp(Q_{tj})}\left[\frac{\exp(Z_{it})}{1+\exp(Z_{it})}-\frac{\exp(2Z_{it})}{(1+\exp(Z_{it}))^{2}}\right]\\
&=\frac{\exp(Z_{it}+Q_{tj})}{(\sum_{t=1}^{K}\exp(Q_{tj}))(1+\exp(Z_{it}))}\left(1-\frac{\exp(Z_{it})}{1+\exp(Z_{it})}\right)\\
&=\frac{\exp(Z_{it}+Q_{tj})}{(\sum_{t=1}^{K}\exp(Q_{tj}))(1+\exp(Z_{it}))}\left(\frac{1}{1+\exp(Z_{it})}\right)\\
&=\frac{\exp(Z_{it}+Q_{tj})}{(\sum_{t=1}^{K}\exp(Q_{tj}))(1+\exp(Z_{it}))^{2}}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\frac{\partial\hat{f}}{\partial Z_{it}}\\
&=N_{ij}\left(\sum_{t=1}^{K}\frac{\exp(Z_{it}+Q_{tj})}{(1+\exp(Z_{it}))\sum_{t=1}^{K}\exp(Q_{tj})}\right)^{-1}\frac{\exp(Z_{it}+Q_{tj})}{(\sum_{t=1}^{K}\exp(Q_{tj}))(1+\exp(Z_{it}))^{2}}\\
&\quad+(N_{j}-N_{ij})\left(1-\sum_{t=1}^{K}\frac{\exp(Z_{it}+Q_{tj})}{(1+\exp(Z_{it}))\sum_{t=1}^{K}\exp(Q_{tj})}\right)^{-1}\frac{-\exp(Z_{it}+Q_{tj})}{(\sum_{t=1}^{K}\exp(Q_{tj}))(1+\exp(Z_{it}))^{2}}\\
&=\left(\frac{\exp(Z_{it}+Q_{tj})}{(\sum_{t=1}^{K}\exp(Q_{tj}))(1+\exp(Z_{it}))^{2}}\right)\times\\
&\quad\left(\frac{N_{ij}}{\sum_{t=1}^{K}\frac{\exp(Z_{it}+Q_{tj})}{(1+\exp(Z_{it}))\sum_{t=1}^{K}\exp(Q_{tj})}}-\frac{N_{j}-N_{ij}}{1-\sum_{t=1}^{K}\frac{\exp(Z_{it}+Q_{tj})}{(1+\exp(Z_{it}))\sum_{t=1}^{K}\exp(Q_{tj})}}\right)
\end{aligned}
$$

Partial Derivative of $\hat{f}$ with respect to Q_{tj}

The additional constraint on the $\Pi^{T|X}$ matrix whereby,

$$\sum_{i=1}^{K} \Pi_{ij}^{T|X} = \sum_{i=1}^{K} \left(\frac{\exp(Q_{ij})}{\sum_{k=1}^{K} \exp(Q_{kj})} \right) = 1$$

results in the partial derivative of the objective function with respect to Q_{tj} being more complicated than the previously calculated partial derivative. The calculations are therefore shown in the first appendix.

The partial derivative of the objective function $\hat{f}$ with respect to Q_{tj} is given as,

$$\begin{aligned}
\frac{\partial \hat{f}}{\partial Q_{tj}} = & \sum_{i \neq j} N_{ij} \left(\sum_{k=1}^{K} \frac{\exp(Z_{ik} + Q_{kj})}{(1 + \exp(Z_{ik})) \sum_{k=1}^{K} \exp(Q_{kj})} \right)^{-1} \left(\frac{\exp(Z_{it} + Q_{tj})}{(1 + \exp(Z_{it})) \left(\sum_{k=1}^{K} \exp(Q_{kj}) \right)} \right. \\
& \left. \times \left(1 - \frac{\exp(Q_{tj})}{\sum_{k=1}^{K} \exp(Q_{kj})} \right) - \frac{\exp(Q_{tj})}{\left(\sum_{k=1}^{K} \exp(Q_{kj}) \right)^2} \sum_{k \neq t}^{K} \left(\frac{\exp(Z_{ik} + Q_{kj})}{1 + \exp(Z_{ik})} \right) \right) \\
& + (N_j - N_{ij}) \left(1 - \sum_{k=1}^{K} \frac{\exp(Z_{ik} + Q_{kj})}{(1 + \exp(Z_{ik})) \sum_{k=1}^{K} \exp(Q_{kj})} \right)^{-1} \left(- \frac{\exp(Z_{it} + Q_{tj})}{(1 + \exp(Z_{it})) \left(\sum_{k=1}^{K} \exp(Q_{kj}) \right)} \right. \\
& \left. \times \left(1 - \frac{\exp(Q_{tj})}{\sum_{k=1}^{K} \exp(Q_{kj})} \right) + \frac{\exp(Q_{tj})}{\left(\sum_{k=1}^{K} \exp(Q_{kj}) \right)^2} \sum_{k \neq t}^{K} \left(\frac{\exp(Z_{ik} + Q_{kj})}{1 + \exp(Z_{ik})} \right) \right)
\end{aligned}$$

3.6.2 Rank Regularised Mixture Model with Penalisation

An extension of the rank regularised mixture model includes penalisation of deviations from the independence model. The rank regularised mixture model is enhanced through penalisation because it adds further regularisation and aids in inference. The conditional matrix and marginal probabilities are still the focus of the estimation, since the formulation of the conditional probability matrix $\hat{\Pi}^{C} \doteq \hat{\Pi}^{X|T} \hat{\Pi}^{T|X}$ is already restricted, whereas it is not immediately clear whether the matrix $\Pi^{J} - \boldsymbol{\pi}\boldsymbol{\pi}'$ will have a regularised form under the mixture model framework.

Previously, the L_1 penalty has been selected for penalisation due to its ability to introduce sparsity

into a system. In the implementation of this model, gradient-based optimisation was used to determine the elements in the matrices $\hat{\Pi}^{T|X}$ and $\hat{\Pi}^{X|T}$. Given that the L_1 penalty uses the non-differentiable absolute value function, a differentiable approximation of the penalty term is utilised. Various approximations are discussed in section 3.5.1.

Suppose the function $f(x) = \frac{1}{\alpha}\ln(\cosh(\alpha x))$ is selected to approximate $f(x) = |x|$. The parameter α is constrained to positive values, $\alpha > 0$. The function of the parameter α regulates how closely $\frac{1}{\alpha}\ln(\cosh(\alpha x))$ approximates $|x|$, with larger values of α resulting in a closer approximation.

Using $\frac{1}{\alpha}\ln(\cosh(\alpha x))$ as the approximation function for the L_1 penalty, we have objective function,

$$f(\Pi^{X|T}, \Pi^{T|X}, \boldsymbol{\pi}) = f_{RRMM}\left(\Pi^{X|T}, \Pi^{T|X}\right) - \frac{\lambda}{\alpha}\sum_{i\neq j}\ln\left(\cosh\left(\alpha\left(\boldsymbol{\Pi}_{i:}^{X|T}\boldsymbol{\Pi}_{:j}^{T|X}\pi_j - \pi_i\pi_j\right)\right)\right)$$

which should be maximised with respect to $\{\Pi^{X|T}, \Pi^{T|X}, \boldsymbol{\pi}\}$. The function $f_{RRMM}\left(\Pi^{X|T}, \Pi^{T|X}\right)$ is the objective function from the rank regularised mixture model,

$$\begin{aligned} f_{RRMM}\left(\Pi^{X|T}, \Pi^{T|X}\right) &= \sum_{i\neq j} N_{ij}\log\left(\widehat{P(X_i|X_j)}\right) + (N_j - N_{ij})\log\left(1 - \widehat{P(X_i|X_j)}\right) \\ &= \sum_{i\neq j} N_{ij}\ \ln\left(\sum_{t=1}^{K}\frac{\exp(Z_{it}+Q_{tj})}{(1+\exp(Z_{it}))\sum_{t=1}^{K}\exp(Q_{tj})}\right) \\ &\quad + (N_j - N_{ij})\ \ln\left(1 - \sum_{t=1}^{K}\frac{\exp(Z_{it}+Q_{tj})}{(1+\exp(Z_{it}))\sum_{t=1}^{K}\exp(Q_{tj})}\right) \end{aligned}$$

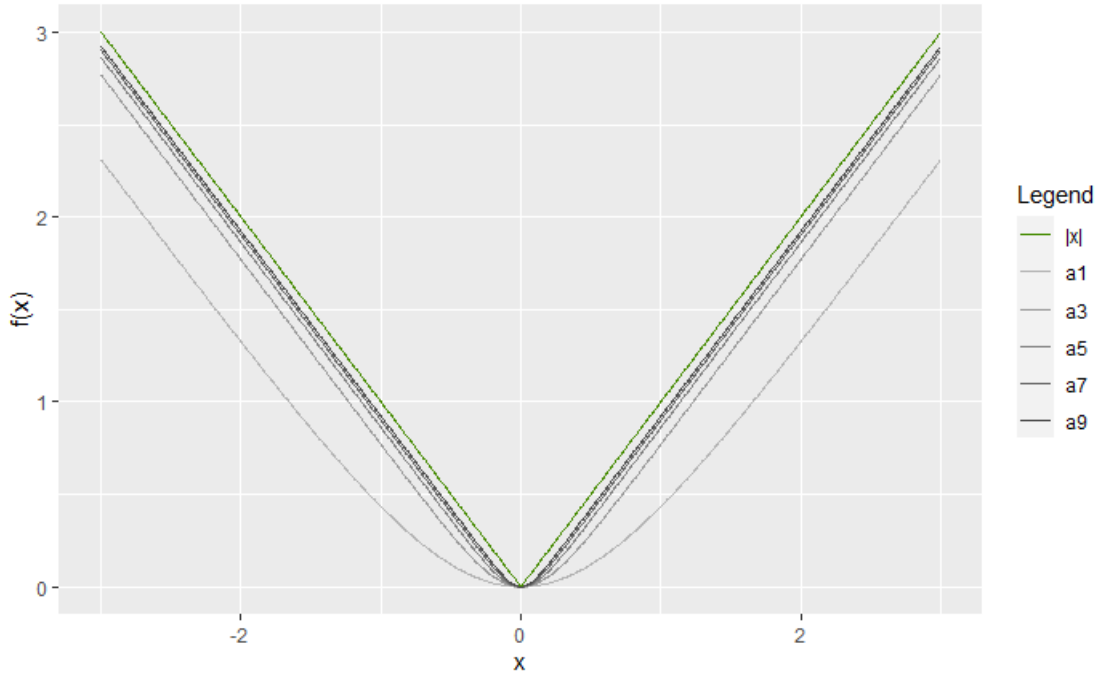
As before, $\boldsymbol{\pi}$ is the vector of marginal probabilities. Matrices $\Pi^{X|T}$ and $\Pi^{T|X}$ are as defined previously and the parameter λ acts as a tuning parameter. The penalty restricts deviations from the independence model,

$$\begin{aligned} \boldsymbol{\Pi}_{i:}^{X|T}\boldsymbol{\Pi}_{:j}^{T|X}\pi_j - \pi_i\pi_j &= P(X_i|X_j)P(X_j) - P(X_i)P(X_j) \\ &= P(X_i, X_j) - P(X_i)P(X_j) \end{aligned}$$

For the purpose of this analysis, a large value for α is required since, for the most part, it is expected that $\boldsymbol{\Pi}_{i:}^{X|T}\boldsymbol{\Pi}_{:j}^{T|X}\pi_j - \pi_i\pi_j$ will produce very small values. This is due to the fact that when selecting a value for α a trade-off exists between selecting a large enough value that the

approximation is close enough, and selecting a small enough value that the approximation does not make the gradient-based optimisation too unstable. Figure 3.2 depicts how the approximation,

Figure 3.2: Effect of α



$f(x) = \frac{1}{\alpha} \ln(\cosh(\alpha x))$, changes with parameter α. The green line depicts $|x|$, and line '$a.$' depicts approximation $f(x) = \frac{1}{.} \ln(\cosh(. \times x))$.

We see that when $\alpha = 1$, the lightest grey line, the approximation is not very close. As the value for α increases, the function better approximates the absolute value function.

CHAPTER 4

SIMULATING CO-OCCURRENCES

4.1 INTRODUCTION

In this chapter we discuss the necessity of simulation studies for model validation. A discussion ensues outlining methods of simulating co-occurrence data that has an underlying dependence structure.

4.2 MODEL VALIDATION BY SIMULATION STUDIES

In this study, it is imperative that the co-occurrence models should be tested on simulated data. In order to validate the models, real data cannot be used since the 'truth' in the data is not known. The distribution of chronic diseases in the entire population of individuals may not be accurately represented through the sample of individuals in the data set which we have. In order to verify that the co-occurrence models perform adequately, we must apply them to data whereby the underlying distribution of the data is known through its simulation.

The simulated data can specify the dependence of conditions in such a way that the true underlying cluster structure is known. The clustering accuracy of each of the models applied to simulated data can be assessed because the true clustering is known. When we apply the models to real data, we do not know the truth in the underlying clusters and hence we cannot truly validate the clustering performance of the models. We can only validate the results by investigating literature on common multi-morbidities.

Another advantage of applying the models to simulated data is that simulation provides scalability. The dimension and size of the simulated data set can be altered. As a result, the computational expense of each of the models can be compared through applying the models to increasingly large data sets. Investigating how well the various algorithms fare when the dimension of the data increases is required in the comparison of algorithms.

4.3 SIMULATION METHODS

In this section we consider techniques of simulating co-occurrence data that has an underlying dependence structure. This has not proven to be a trivial task due to the complexity of the full multivariate Bernoulli distribution. Accurately simulated data should incorporate the dependency between all combinations of conditions, including the dependencies of higher-order interactions as stipulated in the multivariate Bernoulli distribution.

The data being simulated is stored in a binary matrix wherein the columns represent the chronic conditions and the rows represent the study sample. A '1' in the $(ij)^{th}$ cell of this matrix implies that the i^{th} individual has been diagnosed with the j^{th} chronic condition, and a '0' implies the absence of the condition.

Three simulation methods are presented in this study, two of which are implemented.

4.3.1 Re-sampling from Observational Data

Given that data are available to us for use, the first simulation technique investigated is random sampling from the data. Accurately modelling the chronic status of a group of individuals is incredibly complex due to the intricate interactions that exist between conditions. To this extent, re-sampling from a real dataset frees us from making any distributional assumptions about the data. However, this comes at a trade-off. Since this approach uses observational data, we do not know the true underlying cluster structure. This shortcoming somewhat defeats the purpose of using simulated data for model validation, and hence this technique will not be considered further.

4.3.2 Disease Prototype Simulation Method

This simulation approach entails developing a set of disease prototypes. The assumption underlying this approach is that sets of diseases commonly co-occur. We then encode these common co-occurrences in the set of disease prototypes. The prototypes can be designed in such a way that disease clusters are formed, where diseases within a cluster are dependent, but diseases from different clusters co-occur independently.

Mathematical Formulation of the Disease Prototype Simulation Method

Suppose we wish to simulate the chronic status of N_{sample} individuals across a set of D conditions, which form K disease clusters. The simulated data is stored in binary matrix B where,

$$B = \begin{bmatrix} B_{1,1} & \dots & B_{1,D} \\ \vdots & \ddots & \vdots \\ B_{N_{sample},1} & \dots & B_{N_{sample},D} \end{bmatrix}$$

$$[B]_{i,j} = \begin{cases} 1 & \text{if the } i^{th} \text{ individual has the } j^{th} \text{ condition} \\ 0 & \text{if the } i^{th} \text{ individual does not have the } j^{th} \text{ condition} \end{cases}$$

We then construct a set of p disease prototypes that adhere to an underlying cluster structure, and denote these prototypes by $\{\mathbf{P}_1, \mathbf{P}_2, \dots, \mathbf{P}_p\}$. We can randomly assign each of the N_{sample} observations to a disease prototype. The values in each of the prototypes correspond to a probability of exhibiting a chronic condition. Mathematically,

$$\mathbf{P}_r = \begin{bmatrix} P(\text{an individual from prototype } \mathbf{P}_r \text{ exhibits condition 1}) \\ P(\text{an individual from prototype } \mathbf{P}_r \text{ exhibits condition 2}) \\ \vdots \\ P(\text{an individual from prototype } \mathbf{P}_r \text{ exhibits condition } D) \end{bmatrix}$$

Now, suppose that the i^{th} individual expresses diseases in accordance with disease prototype $\mathbf{P}_r$. Then, $B_{i,j}$ is a simulated Bernoulli realisation where the probability of success is the j^{th} element of $\mathbf{P}_r$, the probability of expressing the j^{th} condition under disease prototype $\mathbf{P}_r$.

4.3.3 Gaussian Latent Variable Simulation Method

A latent variable is a variable that is not directly observable, however it produces observable effects (Salkind, 2010). Latent variable models attempt to portray some hidden structure in the data. In this instance, the underlying structure is the dependency between diseases.

The Gaussian latent variable simulation method entails simulating a $D \times D$ covariance matrix Σ, which describes the dependency structure between the chronic conditions. In simulating the

data, N_{sample} multivariate normal realisations were simulated from the distribution $N_D(\mathbf{0}, \Sigma)$. The multivariate normal realisations were thresholded such that values exceeding a threshold were assigned a value of '1' and a '0' otherwise. The latent structure should be encapsulated in the simulated realisations of the multivariate normal distribution, which is in turn used to construct the binary simulated data.

The simulated covariance matrix can be structured in such a way that disease clusters are formed. Diseases which have a strong positive dependency on one another should form part of the same cluster and so the covariance between the conditions will be large and positive. The increased covariance will induce a higher than random chance of disease co-occurrence. Antagonistic diseases will have a large negative covariance to reflect the lower than random chance of disease co-occurrence. Diseases in different clusters are assumed to be independent and will therefore be assigned a zero covariance.

CHAPTER 5

CLUSTERING AND PARAMETER TUNING

5.1 INTRODUCTION

In this chapter we outline the methodology followed to cluster co-morbid conditions. In clustering conditions we used spectral clustering with normalised mutual information scores as a similarity metric. Metrics to evaluate clustering performance, and parameter tuning, are also discussed.

5.2 CLUSTERING OF CONDITIONS

The main investigation of this study is to distinguish groups of chronic conditions that are interrelated. The models discussed in Chapter 3 aid us in describing the relationships between conditions. The output from these models can be used to construct similarity matrices which are in turn used as the input to a clustering algorithm. In this study, conditions are clustered using spectral clustering.

The method of clustering conditions is outlined below:

1. The joint or conditional probability matrix is estimated using one of the models discussed in Chapter 3.
2. A similarity matrix is computed from the estimated probability matrix. In this study, normalised mutual information was used as a similarity metric. This will be discussed further.
3. The graph Laplacian of the similarity matrix is then calculated.
4. Spectral clustering is achieved by applying k-means clustering to the normalised eigenvectors of the graph Laplacian.

An elaboration of these steps follows.

5.2.1 Normalised Mutual Information

Broadly speaking, the normalised mutual information provides a measure of dependency in the sense that it computes the deviation from the independence assumption. If D denotes the number of

chronic conditions considered, then the normalised mutual information matrix is a $D \times D$ symmetric matrix. The ij^{th} element of this matrix represents the normalised mutual information between the i^{th} and j^{th} chronic conditions. Denoting the normalised mutual information matrix by NMI, then NMI_{ij} represents the normalised 'statistical distance' between the joint probability $P(X_i, X_j)$ and the product of the marginal probabilities, $P(X_i) \times P(X_j)$ (Kvalseth, 2017). The normalised mutual information score is a normalisation of the mutual information score, thus the normalised mutual information score ranges from '0' to '1'. A score of '0' implies that there is no mutual information between variables and a score of '1' implies that variables are perfectly correlated. If there is no mutual information between chronic conditions, the conditions are said to be independent. A large normalised mutual information score implies that there is a high level of dependency between conditions.

Using normalised mutual information as a similarity metric is advantageous since we want to determine whether diseases co-occur at a rate different to the co-occurrence rate under the assumption that there is no association between diseases.

5.2.2 Spectral Clustering

Suppose we wish to partition the set of D chronic conditions into K clusters. Denote the $D \times D$ symmetric similarity matrix by G. In literature this is often referred to as the 'affinity matrix'. Define $D \times D$ diagonal matrix F such that,

$$F = \text{diag}\left(\sum_{j=1}^{D} G_{1j}, \sum_{j=1}^{D} G_{2j}, \ldots, \sum_{j=1}^{D} G_{Dj}\right)$$

In literature this matrix is often denoted by D, however in our study D is reserved for the number of chronic diseases. The symmetric normalised Laplacian is defined as,

$$L_{norm} = F^{-\frac{1}{2}} G F^{-\frac{1}{2}}$$

An eigen-decomposition is then performed on L_{norm}. Denote the eigenvalues of L_{norm} by $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_K \leq \lambda_{K+1} \leq \ldots \leq \lambda_D$, and the eigenvectors corresponding to these eigenvalues by $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_K, \boldsymbol{v}_{K+1}, \ldots, \boldsymbol{v}_D\}$. The graph Laplacian can now be constructed and is denoted by

V where,

$$V = [\boldsymbol{v}_2, \boldsymbol{v}_3 \ldots, \boldsymbol{v}_K]$$

Denote the normalised eigenvectors of the graph Laplacian by V_{norm}. Both V and V_{norm} are $D \times (K-1)$ matrices. We normalise V as follows:

$$\begin{aligned} R &= [R_1, R_2, \ldots, R_D]' \\ &= \left[\sqrt{\sum_{j=1}^{k-1} ([V]_{1j})^2}, \sqrt{\sum_{j=1}^{k-1} ([V]_{2j})^2}, \ldots, \sqrt{\sum_{j=1}^{k-1} ([V]_{Dj})^2} \right]' \\ [V_{norm}]_{ij} &= \frac{[V]_{ij}}{R_i} \end{aligned}$$

K-means clustering is then performed on the matrix V_{norm}, the normalised eigenvectors of the graph Laplacian. Plotting V_{norm} gives us a Laplacian eigenmap. The cluster structure extracted from the co-occurrence model can be visually represented by a Laplacian eigenmap. Laplacian eigenmaps perform non-linear dimensionality reduction by representing a high-dimensional structure in a low dimensional space.

If one were to apply k-means clustering to the similarity matrix or the estimated probability matrix directly then the dimension of the data being clustered is $D \times D$. Spectral clustering performs dimensionality reduction since the data being clustered is the $D \times K$ Laplacian eigenmap.

Spectral clustering is unlike centroid-based clustering methods, like k-means clustering, in that it clusters observations based on connectivity as opposed to compactness (Singh *et al.*, 2010). Spectral clustering makes use of local 'neighbourhoods' in the data thus preserving local information, as opposed to incorporating the similarity structure between all points (Singh, 2010).

5.2.3 Thresholding the Similarity Matrix

It can be hypothesized that conditions with a low level of dependency on other conditions have no material impact on the investigation and should be omitted from the analysis. If the sum of the i^{th} row, and column, of the similarity matrix is very low, then the i^{th} chronic condition has very weak connections to other conditions and is effectively independent. Conditions with weak associations can be considered 'noisy features'.

We may exclude these conditions from the analysis by applying a threshold to the similarity matrix which in this study is the normalised mutual information matrix. Suppose we enforce a threshold of τ. The i^{th} condition is excluded from the analysis if $\sum_{j=1}^{D} G_{ij} \leq \tau$, or equivalently, $\sum_{j=1}^{D} G_{ji} \leq \tau$. Thresholding conditions performs dimension reduction through feature selection which may be advantageous given the high-dimensionality of the data.

A minimum threshold of 0 needs to be applied. This is because spectral clustering cannot be performed when the similarity matrix has rows, and columns, containing only zero values. If we have a disconnected similarity graph, G, then some of the diagonal elements in F will be zero, and the matrix $L_{norm} = F^{-\frac{1}{2}} G F^{-\frac{1}{2}}$ will have non-real entries.

Increasing the value of τ will perform more feature selection and the remaining features, or conditions, will have a higher level of dependency.

5.2.4 Clustering Metric and Cluster Ratio

Evaluating the clustering performance in an unsupervised learning problem is not a trivial task. In determining whether the model extracted cohesive cluster structures from the data, we may investigate the normalised total within-cluster sum of squares. Denote:

- K: The number of clusters.
- C_k: The k^{th} cluster where $k \in \{1, 2, \ldots, K\}$.
- μ_k: The centroid corresponding to the k^{th} cluster.
- $v_{norm}^{(i)}$: The i^{th} row of V_{norm}, the normalised eigenvectors of the graph Laplacian. This corresponds to the i^{th} chronic condition, $i \in \{1, 2, \ldots, D\}$.

The normalised total within-cluster sum of squares is calculated as,

$$\text{CM} = \frac{\text{Total within-cluster sum of squares}}{\text{Between-cluster sum of squares}}$$

where,

$$\text{Total within-cluster sum of squares} = \sum_{k=1}^{K} \sum_{v_{norm}^{(i)} \in C_k} \left(v_{norm}^{(i)} - \mu_k \right)^2$$

and,

$$\text{Between-cluster sum of squares} = \text{Total sum of squares} - \text{Total within-cluster sum of squares}$$

We can calculate the total sum of squares as,

$$\text{Total sum of squares} = \sum_{i=1}^{D} \left(v_{norm}^{(i)} - \frac{1}{D} \sum_{i=1}^{D} v_{norm}^{(i)} \right)^2$$

The total within-cluster sum of square measures how compact observations are in their clusters (Boehmke). The between sum of square value measures how compact all the observations are to each other. Cluster solutions which produce clusters that are compact and well-separated are preferable. We cannot use the total within-cluster sum of square as a clustering metric because it provides no indication of how separated the clusters are. In comparing models, we may use the normalised total within-cluster sum of squares as a clustering metric (CM) to evaluate how well the model extracts cohesive clusters in the data. The model minimising this clustering metric exhibits the best ability to produce compact and well-separated clusters from the data.

A limitation of this clustering metric is that it can only be used to compare solutions with the same number of clusters, k. Additionally, if the number of conditions being clustered is inconsistent under different models then the clustering metric will also be skewed. A model with fewer conditions may have an improved clustering performance because there are fewer interactions to consider between conditions and hence there will likely be less noise in the model.

This presents a challenge as the similarity matrix has a threshold of at least 0 and so the output of models with potentially differing number of conditions is to be compared. We expect models with fewer conditions to show a lower clustering metric and so this must be considered when evaluating clustering performance. As the threshold to the similarity matrix increases, we expect the model to include fewer conditions and hence the clustering metric will likely decrease.

To mitigate this limitation, we may consider viewing the ratio of the clustering metric to the number of conditions in the model. An optimal clustering solution would minimise this ratio. This is illustrated with two examples. Suppose we have two models and wish to investigate the effect that the number of conditions has on this ratio. Model A and Model B both have a clustering

metric of 0.5, but the number of conditions in Model A is 100 and the number of conditions in Model B is 300. The cluster ratio for Model A is $r_A = \frac{0.5}{100} = 0.005$ and the cluster ratio for Model B is $r_B = \frac{0.5}{300} = 0.002$. We expect the clustering metric to be lower when fewer conditions are considered, and so by Model B achieving the same clustering metric as Model A despite having more conditions, Model B is superior. As a result it produces a lower clustering ratio. Now we investigate the effect of the clustering metric on the cluster ratio. If Model A has a clustering metric of 0.1 and has 100 conditions then it will produce a cluster ratio of $r_A = 0.001$. If Model B also has 100 conditions but a clustering metric of 0.5, it will have a cluster ratio of 0.005. Both models have the same number of conditions and Model A achieved a lower clustering metric rendering it superior to Model B. Consequently it produces a lower cluster ratio.

5.3 PARAMETER TUNING

In this section we discuss how the model hyper-parameters and parameters were computed. The models implemented were the L_1-penalisation model, the rank regularised mixture model with penalisation and the incomplete MLE model.

L_1-Penalisation Model

The objective function for the regularised model with an L_1 penalty is,

$$g\left(\pi_{ij}^J, \pi_i, \pi_j\right) = \sum_{i,j=1}^{D} N_{ij} \ln\left(\pi_{ij}^J\right) + (N_{sample} - N_{ij}) \ln\left(1 - \pi_{ij}^J\right) - \lambda \sum_{i \neq j=1}^{D} |\pi_{ij}^J - \pi_i \pi_j|$$

The marginal and joint probabilities form the model parameters and are estimated through maximising the objective function. The parameter, λ, is a tuning parameter and requires optimisation. Given that clustering is an unsupervised learning problem we cannot optimise λ by minimising some error rate. The optimal value for λ was then selected as the value which resulted in the smallest clustering metric.

Values for the thresholds of the similarity matrix were specified. A set of four thresholds were proposed, $\{0, 0.10, 0.15, 0.20\}$.

Rank Regularised Mixture Model with Penalisation

The objective function for the rank regularised mixture model with penalisation is,

$$h\left(\Pi^{X|T}, \Pi^{T|X}, \pi_i\right) = \sum_{i \neq j=1}^{D} N_{ij} \ln\left(\boldsymbol{\pi}_{i\cdot}^{X|T} \boldsymbol{\pi}_{\cdot j}^{T|X}\right) + (N_j - N_{ij}) \ln\left(1 - \boldsymbol{\pi}_{i\cdot}^{X|T} \boldsymbol{\pi}_{\cdot j}^{T|X}\right) - \frac{\lambda}{\alpha} \sum_{i \neq j=1}^{D} \ln\cosh\left(\alpha\left(\boldsymbol{\pi}_{i\cdot}^{X|T} \boldsymbol{\pi}_{\cdot j}^{T|X} \pi_j - \pi_i \pi_j\right)\right)$$

The marginal probabilities and the elements of matrices $\Pi^{X|T}$ and $\Pi^{T|X}$ form the model parameters and are estimated through maximising the objective function. The parameter, λ, is a tuning parameter and is optimised to produce the smallest clustering metric. The parameter α is a hyper-parameter which controls how accurately the penalty term approximates the L_1 penalty and is not estimated from the data. This parameter was selected as 50 in the implementation of the models. The value for α was chosen to be relatively large given that most of the joint probabilities will be very small. The number of person-types is also a hyper-parameter and was selected to be 100. Under the rank reduced model, the rank of the estimated conditional probability matrix is restricted to be at most as large as the number of person-types. The simulated data has 200 conditions and the cleaned data has 508 conditions, thus 100 person-types seemed to be an appropriate choice as it still allows for some flexibility in the model.

The same set of thresholds were applied to the rank reduced models, however the mean level of similarity scores under the rank reduction models greatly exceeded the mean level of similarity scores under the penalisation model and so the thresholds had no effect.

Incomplete MLE Model

The objective function for the incomplete MLE model is,

$$f\left(\pi_{ij}^{J}\right) = \sum_{i,j=1}^{D} N_{ij} \ln\left(\pi_{ij}^{J}\right) + (N_{sample} - N_{ij}) \ln\left(1 - \pi_{ij}^{J}\right)$$

The marginal and joint probabilities form the model parameters and are estimated by $\hat{\pi}_{ij}^{J} = \frac{N_{ij}}{N_{sample}}$ and $\hat{\pi}_i = \frac{N_i}{N_{sample}}$. The same set of thresholds were applied to the incomplete MLE Model.

CHAPTER 6

RESULTS AND DISCUSSION

6.1 INTRODUCTION

In this chapter we discuss the results of the analysis when the various models are applied to both simulated and real data.

6.2 SIMULATION STUDIES RESULTS

6.2.1 Disease Prototype Simulation

Data were simulated using the disease prototype simulation approach and the models described in Chapter 3 were applied to the data. This experiment was repeated 50 times to account for random variation and provide a more consistent comparison between models.

- The set of possible λ values ranged from 0 to 1000, with the optimal value chosen through validation as described in Chapter 5.
- D = the number of diseases simulated = 200
- N_{sample} = the number of people simulated = $10,000$
- K = the number of true clusters in the data = 10
- The total number of person-types under the rank reduced model was set to 100.
- The parameter α under the rank reduced model which controls how well the penalty approximates the L_1-penalty was set to 50.
- 20 iterations were performed per model fit

Experiments were only accepted if they didn't require a threshold to be applied to the similarity matrix. This was to ensure that the number of conditions considered in each experiment was consistent. Not thresholding the similarity matrix allows us to use the clustering metric to evaluate clustering performance, as opposed to computing both the clustering metric and the cluster ratio. The cluster ratio must be computed if the number of conditions considered differs between models, and so if no thresholding is applied, no feature selection takes place and the clustering metric is

sufficient to evaluate cluster performance. This results in fewer necessary computations and speeds up computational time.

Figure 6.1: Clustering Metrics for the Models Applied to Disease Prototype Simulated Data

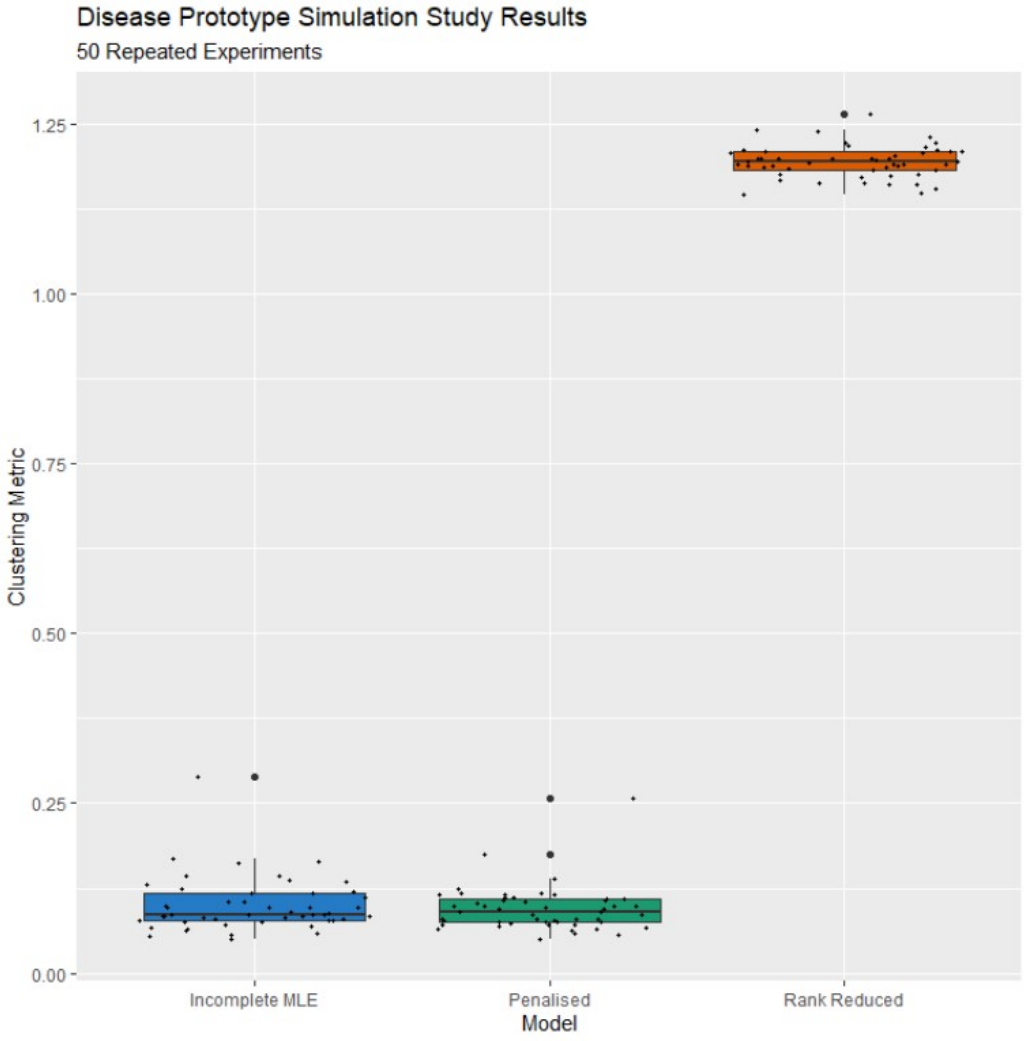


Figure 6.1 allows us to compare the variation in the clustering metric between the models applied to disease prototype simulated data. It is evident that the rank reduction model had the worst clustering performance. Figure 6.1 illustrates that the penalised model performed similarly to the incomplete MLE model, however Table 6.1 demonstrates that the L_1-penalised model slightly outperforms its unregularised counterpart. The mean and standard deviation of the clustering metric under the L_1-penalised model is lower than under the incomplete MLE model.

Table 6.1: Mean and Standard Deviation of Model Outputs Using Disease Prototype Simulated Data

	Model	Mean	Standard Deviation
Optimal λ	Penalised Model	337.9592	242.6371
	Rank Reduced Model	509.3878	320.6721
Clustering metric	Incomplete MLE Model	0.1002	0.0402
	Penalised Model	0.0945	0.0332
	Rank Reduced Model	1.1943	0.0242

We may wish to investigate how the choice of the penalty parameter, λ, affects the clustering performance of the regularised models. To this extent, we consider Figure 6.2. As seen previously, the rank reduced model is shown to perform poorly in comparison to the penalised model. Interestingly, the linear regression lines applied to the scatter-plots are almost linear for both of the regularised models, which indicates that the choice of λ does not have a material impact on the clustering performance of the model. Under the penalised model, the optimised λ values fall predominantly on the lower half of the set of possible values. There are, however, instances where a large penalty is selected as optimal under the framework of the penalised model. Under the rank reduced model, the optimal choice of λ is almost evenly distributed across the range of possible values.

We may investigate Figures 6.3, 6.4 and 6.5 which represent the Laplacian eigenmaps from the various models applied to a single set of disease prototype simulated data. The observations are coloured according to the cluster structure detected by the model. These eigenmaps confirm that the penalised model and incomplete MLE model have the potential to extract cohesive clusters from the disease prototype simulated data. The rank reduced model, however, fails in this regard.

Figure 6.2: The Optimized λ Values and Clustering Metrics for the Regularised Models Applied to Disease Prototype Simulated Data

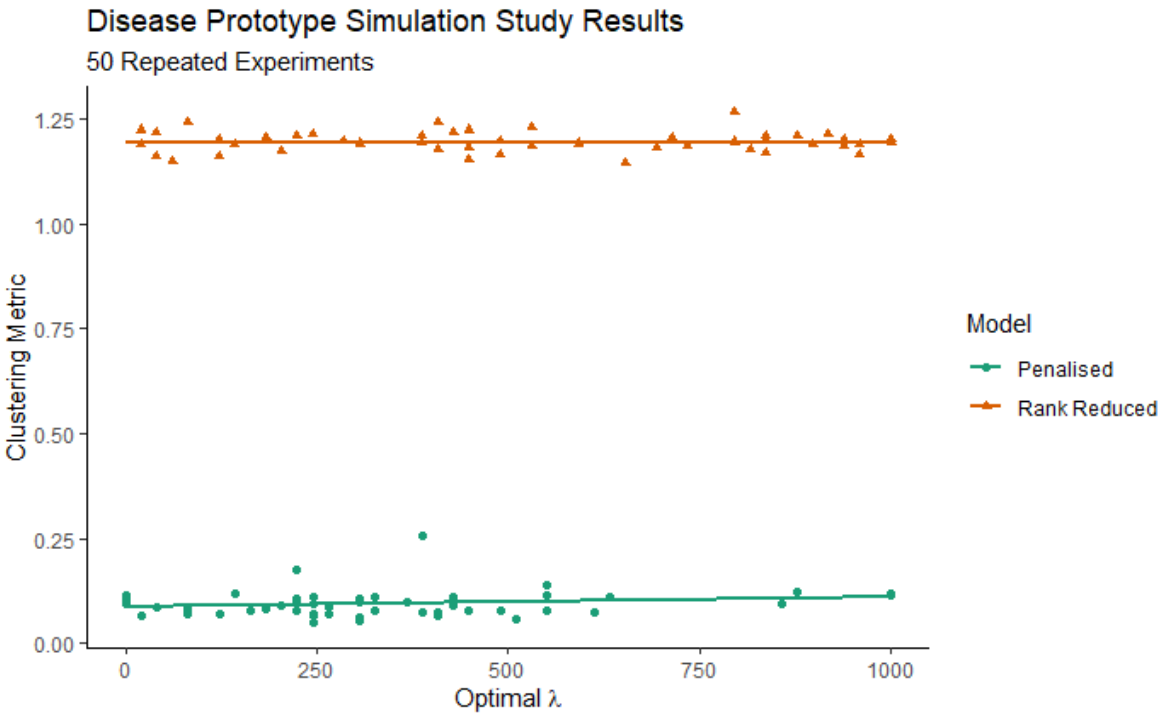


Figure 6.3: Laplacian Eigenmap of Conditions under the Penalisation Model Applied to Disease Prototype Simulated Data

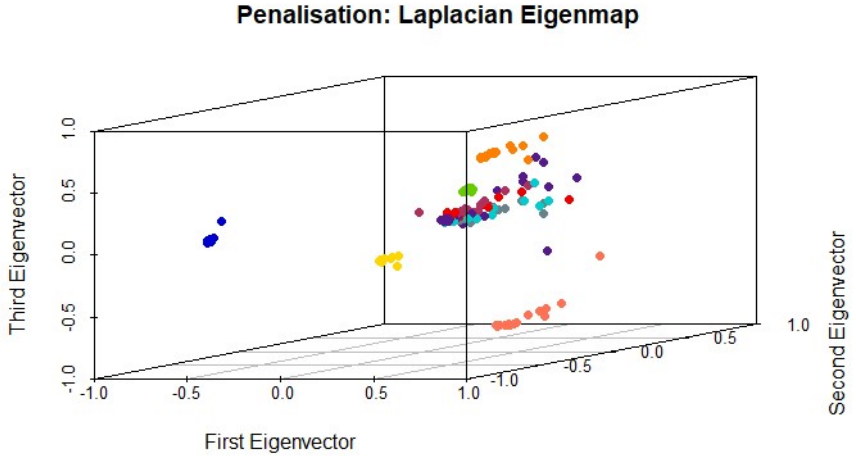


Figure 6.4: Laplacian Eigenmap of Conditions under the Rank Reduced Model Applied to Disease Prototype Simulated Data

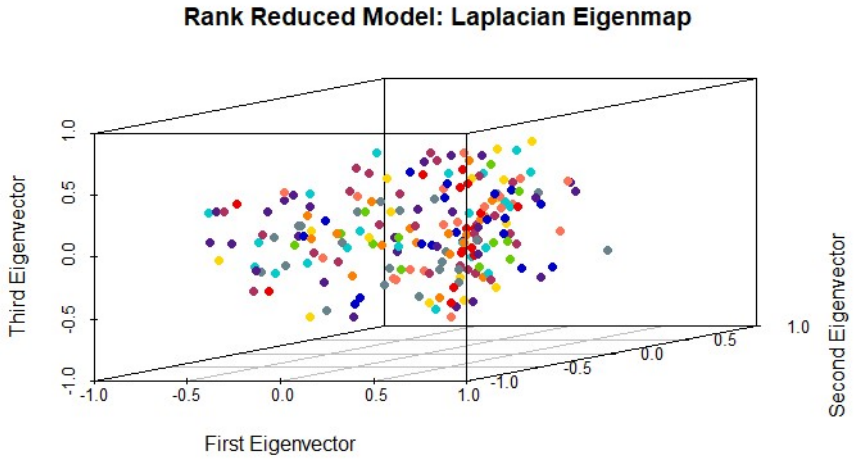
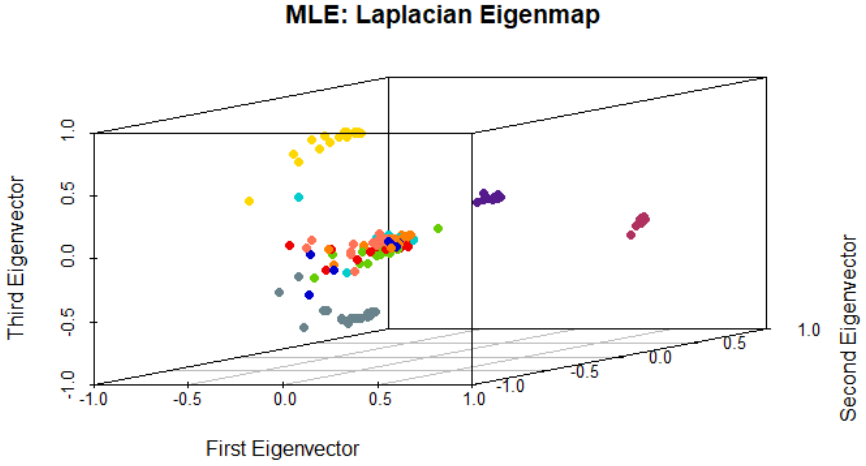


Figure 6.5: Laplacian Eigenmap of Conditions under the MLE Model Applied to Disease Prototype Simulated Data

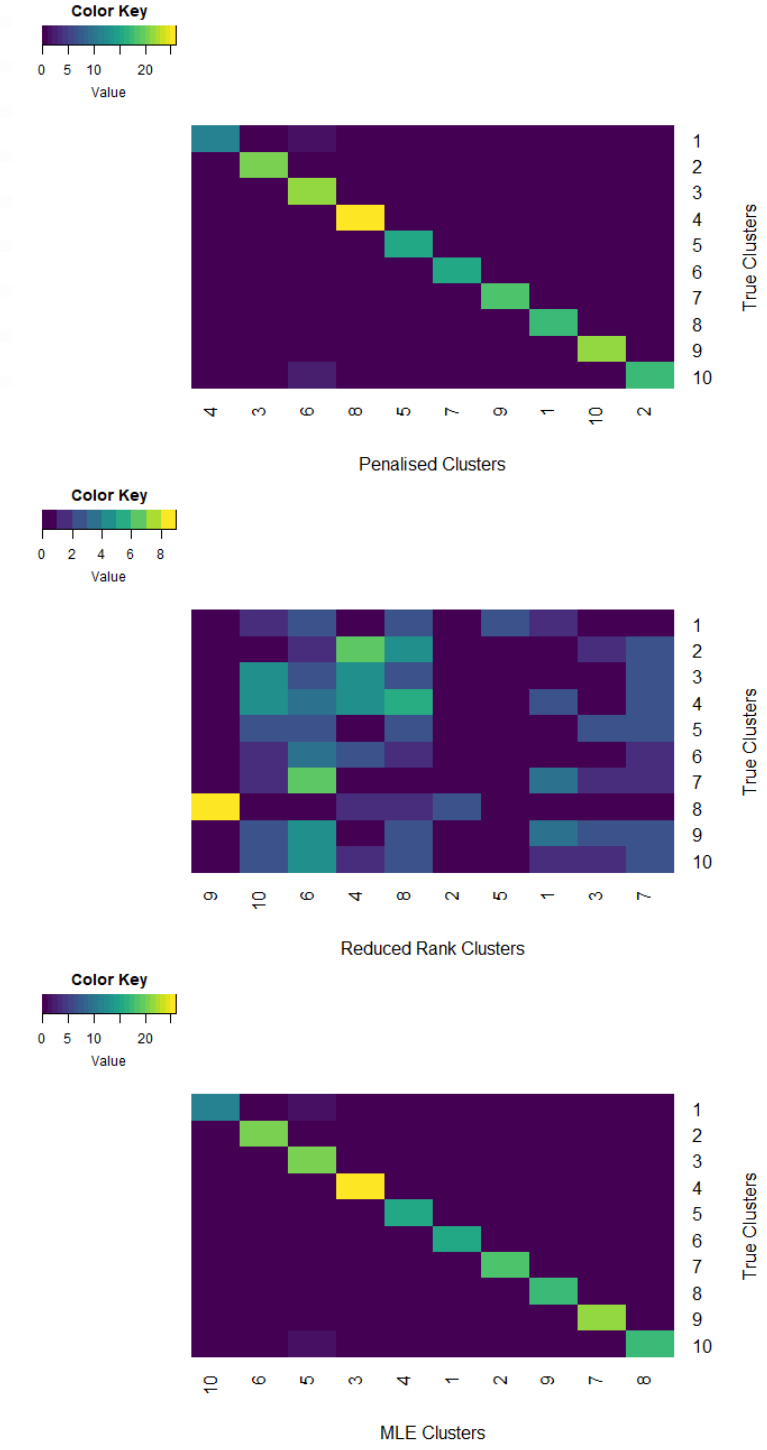


Lastly, we may wish to investigate how correctly the models clustered conditions. Because the true cluster structure is incorporated into the simulation of the data, we are able to compare the extracted clusters with the true clusters. To visually assess the clustering accuracy of the models we consider Figure 6.6.

Each of the models were applied to a single set of disease prototype simulated data, and for each model a confusion matrix was constructed which compared the cluster membership of the derived clusters and the actual clusters. The confusion matrix was reordered for increased interpretability of the results. Heatmaps of the reordered confusion matrices were constructed and are displayed in Figure 6.6. Cells with a more yellow hue represent a large overlap in the cluster membership of the derived and actual clusters. Cells with a more purple hue indicate little to no overlap in cluster membership.

The heatmaps corresponding to the penalised model and the incomplete MLE model include a strong diagonal which indicates that the clusters are well identified under these models. The heatmap for the rank reduced model displays no distinguishable pattern and thus the clusters were not correctly extracted by the reduced rank method. Because the true number of clusters is quite small, the clustering accuracy can be determined by visual means. Clustering accuracy can also be computed through numerical means such as using the adjusted Rand index.

Figure 6.6: Heatmaps of Reordered Confusion Matrices to Determine Clustering Performance under Disease Prototype Simulated Data



The results are in agreement and we can conclude that the penalised model and the incomplete MLE model perform sufficiently well to data simulated through the disease prototype simulation approach. Regularising the incomplete MLE model by means of applying the L_1-penalty marginally improves the clustering performance. The rank reduced model fares poorly in comparison.

6.2.2 Gaussian Latent Variable Simulation Method

Gaussian latent variable data were simulated and the models described in Chapter 3 were fit to the data. As before, this experiment was repeated 50 times to account for any random variation, and to provide a more consistent comparison between models. The parameters of the estimation are as follows:

- The set of possible λ values ranged from 0 to 1000, with the optimal value chosen through validation as described in Chapter 5.
- D = the number of diseases simulated = 200
- N_{sample} = the number of people simulated = $10,000$
- K = the number of true clusters in the data = 10
- The total number of person-types under the rank reduced model was set to 100.
- The parameter α under the rank reduced model which controls how well the penalty approximates the L_1-penalty was set to 50.
- 20 iterations were performed per model fit

As previously, to ensure a consistent comparison of models, experiments were only accepted if they didn't require a threshold to be applied to the similarity matrix.

In comparing the clustering metric across the three models, we can look to Figure 6.7. On average, the incomplete MLE model had the best clustering performance, followed by the penalised model and lastly by the rank reduced model. The clustering metric of the rank reduced model varied the least, and the clustering metric of the incomplete MLE model varied the most. The mean and standard deviation of the clustering metric under the three models are given in Table 6.2. These values corroborate the findings illustrated in Figure 6.7.

Figure 6.7: Clustering Metrics for the Models Applied to Gaussian Latent Variable Simulated Data

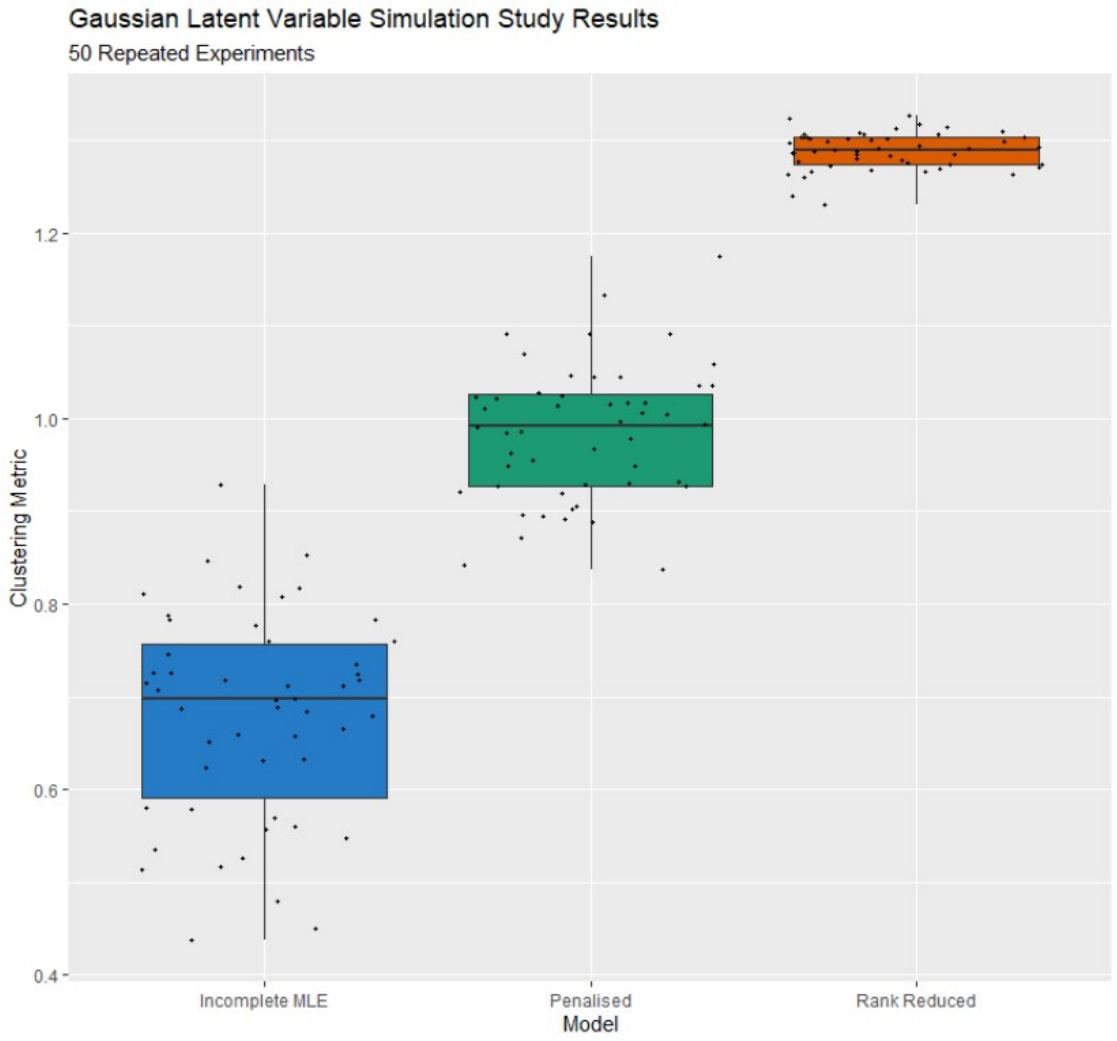


Table 6.2 also provides the mean and standard deviation of the optimal penalty parameter, λ, for each of the regularised models. On average, the penalised model selected slightly larger λ values than the rank reduced model. Figure 6.8 illustrates the relationship between the optimised λ values and the clustering metric under the regularised models. A linear regression line has been included under each model to describe how the clustering metric varies as a function of the optimised λ value. Under the penalised model, the clustering metric decreases slightly as λ increases. This seems to indicate that stricter regularisation improves clusterability of the penalised model. However, the incomplete MLE model achieves the lowest average clustering metric and performs no regularisation. Under the rank reduced model framework, increasing the penalty parameter marginally increases

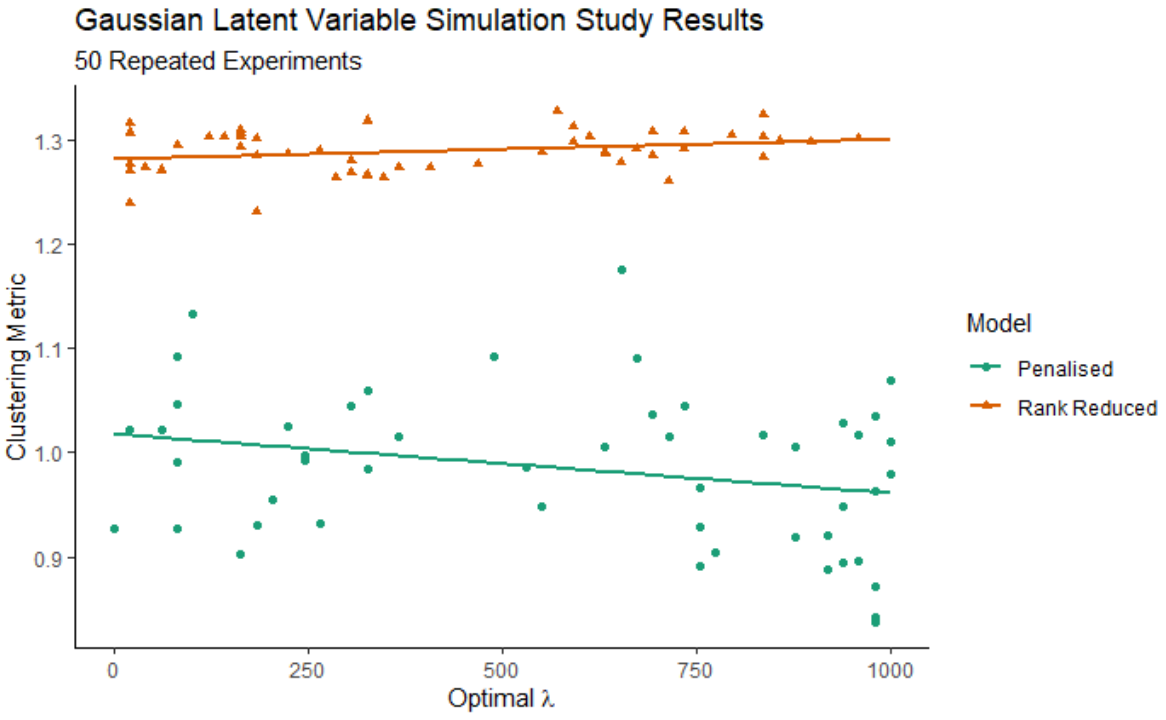
Table 6.2: Mean and Standard Deviation of Model Outputs Using Gaussian Latent Variable Simulated Data

	Model	Mean	Standard Deviation
Optimal λ	Penalised Model	582.8571	349.8581
	Rank Reduced Model	416.3265	288.4383
Clustering metric	Incomplete MLE Model	0.6795	0.1114
	Penalised Model	0.9845	0.0733
	Rank Reduced Model	1.2888	0.0201

the clustering metric.

Figure 6.8 also indicates that the penalised model produces lower clustering metrics than the rank reduced model, although the clustering metrics for the penalised model are more volatile than the clustering metrics under the rank reduced model.

Figure 6.8: The Optimized λ Values and Clustering Metrics for the Regularised Models Applied to Gaussian Latent Variable Simulated Data



As before, we may wish to investigate how the Laplacian eigenmaps from the three models fare against each other when applied to a single common set of Gaussian latent variable simulated data. For this particular dataset, the clustering metric of the incomplete MLE model was 0.7668, the

metric under the penalisation model was 0.9096 and under the rank reduced model the metric was 1.2860. The optimal λ value under the penalised model was 583.2735 and under the rank reduced model λ this value was 631.8796.

These Laplacian eigenmaps are shown in Figures 6.9, 6.10 and 6.11. The colour of the points correspond to the clusters detected by the model.

Figure 6.9: Laplacian Eigenmap of Conditions under the Penalisation Model Applied to Gaussian Latent Variable Simulated Data

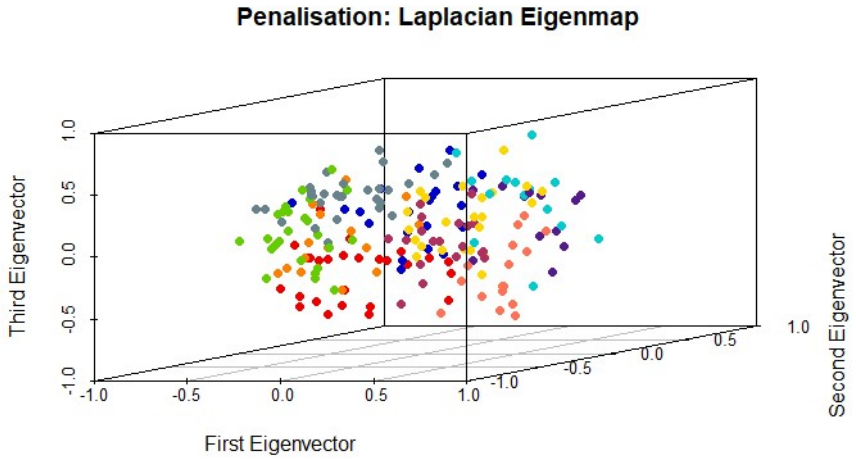


Figure 6.10: Laplacian Eigenmap of Conditions under the Rank Reduced Model Applied to Gaussian Latent Variable Simulated Data

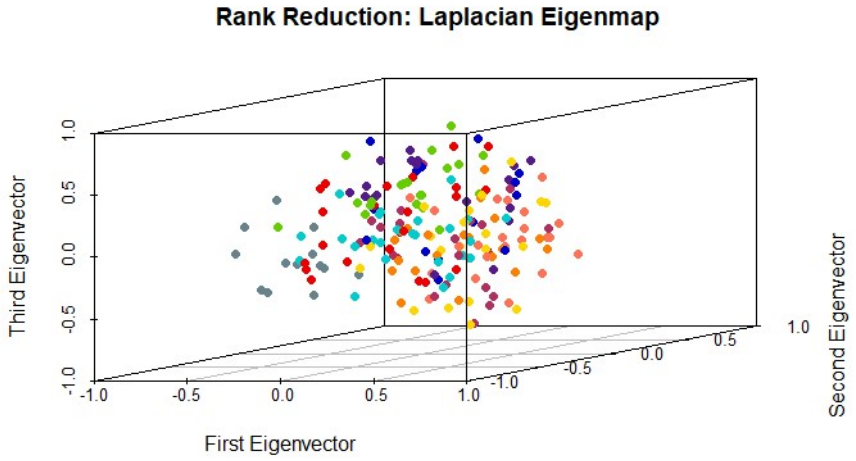
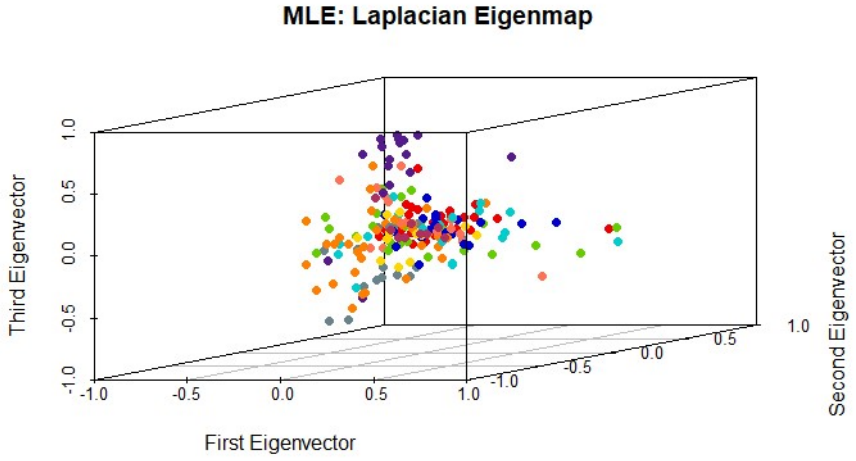


Figure 6.11: Laplacian Eigenmap of Conditions under the MLE Model Applied to Gaussian Latent Variable Simulated Data

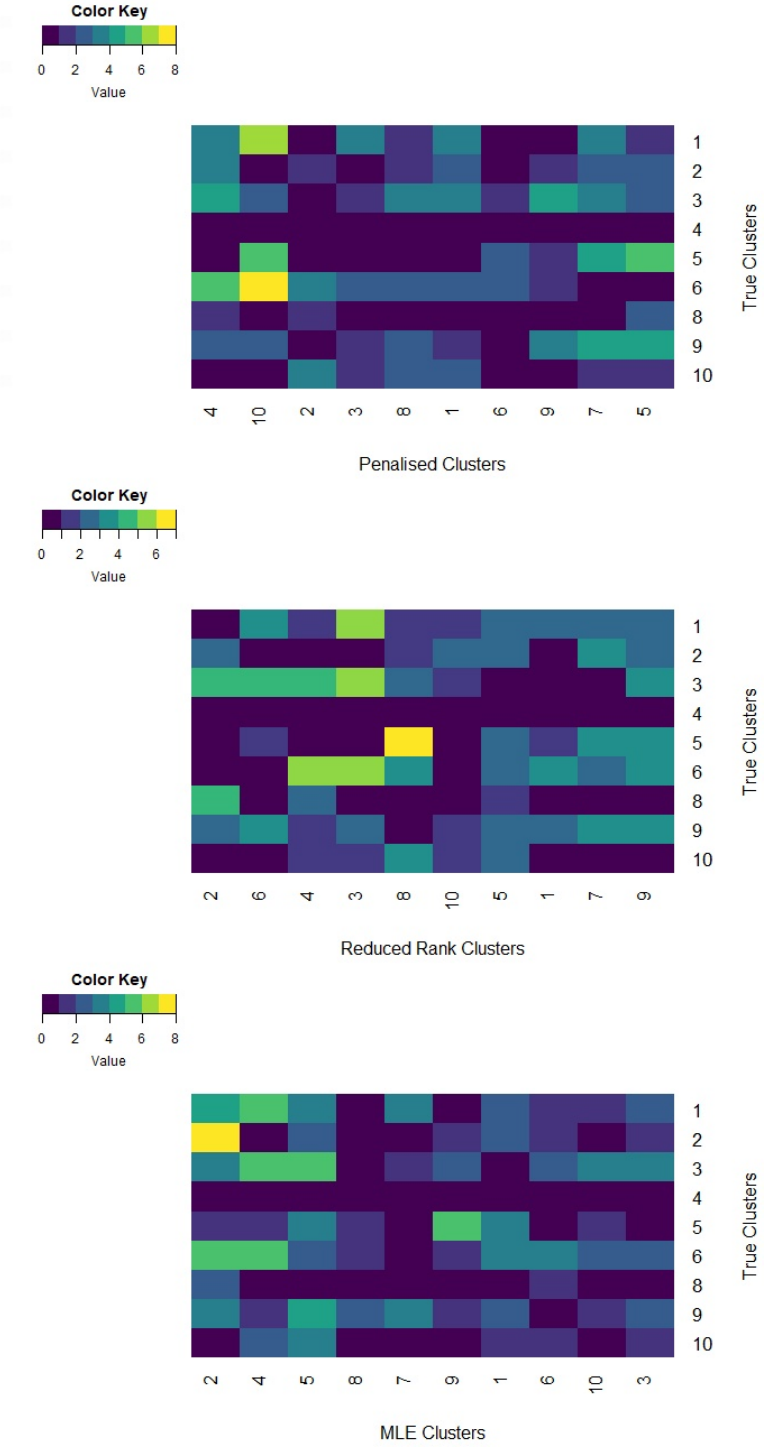


Whilst none of the models extract specifically cohesive clusters when applied to Gaussian latent variable simulated data, the penalised model seems to perform much worse when applied to this data in comparison to its performance when applied to disease prototype simulated data.

Lastly, because the data are simulated we can evaluate the clustering accuracy of the models

applied. To this extent, we refer to Figure 6.12. These heatmaps are constructed in the same way as the heatmaps in Figure 6.6. The heatmaps illustrated in Figure 6.12 show no distinguishable patterns in the cluster membership between the true clusters and the derived clusters. This may indicate that either all of the models performed poorly on this type of simulated data, or the data was simulated in such a way that the dependency between conditions was not clearly translated into the simulated data.

Figure 6.12: Heatmaps of Reordered Confusion Matrices to Determine Clustering Performance under Gaussian Latent Variable Simulated Data



6.3 DATA RESULTS

6.3.1 Data Cleaning

After data cleaning, 508 chronic conditions remained. The original data set included 1,382 conditions, therefore the data was reduced drastically through data cleaning.

Data cleaning involved removing chronic conditions which fell into an ICD 10 Code category which was unquestionably unrelated to chronic conditions. This included codes relating to pregnancy, childbirth, conditions specific to the perinatal period, chromosomal abnormalities, congenital malformations, injury, poisoning and administrative codes.

Table 6.3 illustrates the ICD 10 code categories that remained in the dataset. Further data cleaning

Table 6.3: ICD 10 Code Categories with Descriptions (World Health Organization, 2016*a*)

ICD 10 Code Category	Description
A, B	Infectious and parasitic diseases
C	Malignant neoplasms
D	Benign and in situ neoplasms, and neoplasms of unknown and uncertain behaviour. Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism.
E	Endocrine, nutritional and metabolic diseases
F	Mental and behavioural disorders
G	Diseases of the nervous system
H	Diseases of the eye and adnexa. Diseases of the ear and mastoid process.
I	Diseases of the circulatory system.
J	Diseases of the respiratory system.
K	Diseases of the digestive system.
L	Diseases of the skin and subcutaneous tissue.
M	Diseases of the musculoskeletal system and connective tissue.
N	Diseases of the genitourinary system.
R	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified.

involved removing codes that didn't follow the format of the ICD 10 codes. The format of the ICD 10 codes is a character followed by two numbers. Codes which didn't meet this formatting requirement were omitted. Additionally, conditions with a prevalence of fewer than 10 study participants were omitted.

6.3.2 Results

The model output was extracted when fitting the L_1-penalisation model, the rank reduction model with penalisation and the incomplete MLE model. The regularised models were fitted with 20 iterations per model fit. Under the rank reduced models, the number of person-types was selected to be 100 and the α parameter which controls how well the penalty approximates the L_1 penalty was chosen to be 50.

These results are summarised in Table 6.4.

Table 6.4: Results of the Models Applied to the Data

K	Model	Optimal λ				Clustering Metric				Conditions			
		Threshold				Threshold				Threshold			
		0.00	0.10	0.15	0.20	0.00	0.10	0.15	0.20	0.00	0.10	0.15	0.20
5	P	300	300	300	300	0.434	0.405	0.393	0.344	508	450	354	273
	R	150				0.077				508			
	M	×				0.503	0.436	0.356	0.356	508	434	300	182
10	P	50	50	50	300	0.480	0.454	0.395	0.348	508	450	355	273
	R	200				0.710				508			
	M	×				0.500	0.478	0.428	0.323	508	434	300	182
15	P	150	200	200	50	0.515	0.495	0.459	0.403	508	450	354	274
	R	50				1.316				508			
	M	×				0.514	0.607	0.481	0.313	508	434	300	182

The model output comprises of the optimal value for λ, the clustering metric and the corresponding number of chronic conditions considered. The number of clusters is represented by K where $K \in \{5, 10, 15\}$. Each of the models were run with thresholds $\{0, 0.1, 0.15, 0.2\}$. The minimum threshold applied was $\tau = 0$ and this was to ensure that spectral clustering could take place. In Table 6.4 the models applied are abbreviated such that the L_1-penalisation model is denoted 'P', the rank reduction model with penalisation is denoted 'R' and the incomplete MLE model is denoted 'M'.

The tuning parameter λ controls the level of penalisation applied. The values which λ could take on ranged from 0, where no penalisation was applied, to 300. Both the L_1-penalisation model and

the rank reduction model with penalisation make use of this parameter, however the incomplete MLE model does not. Therefore, the table entries corresponding to λ for the incomplete MLE model are left blank.

The mean level of similarity scores under the rank reduction models greatly exceeded the mean level of similarity scores under the other two models and hence no conditions were removed with thresholding. The results of the rank reduction model therefore remained the same when each of the thresholds were applied.

As described previously, the clustering metric is not entirely suitable for cluster performance comparison and hence we must consider the cluster ratio. The details of the clustering metric and cluster ratio are fully described in Chapter 3. The cluster ratios for each of the models are given in Table 6.5.

Table 6.5: Cluster Ratios of the Models Applied to the Data

K	Model	Cluster Ratio			
		Threshold			
		0.00	0.10	0.15	0.20
5	P	0.00085	0.00090	0.00111	0.00126
	R	0.00015			
	M	0.00099	0.00101	0.00119	0.00195
10	P	0.00095	0.00101	0.00111	0.00127
	R	0.00140			
	M	0.00098	0.00110	0.00143	0.00178
15	P	0.00101	0.00110	0.00130	0.00147
	R	0.00259			
	M	0.00101	0.00140	0.00160	0.00172

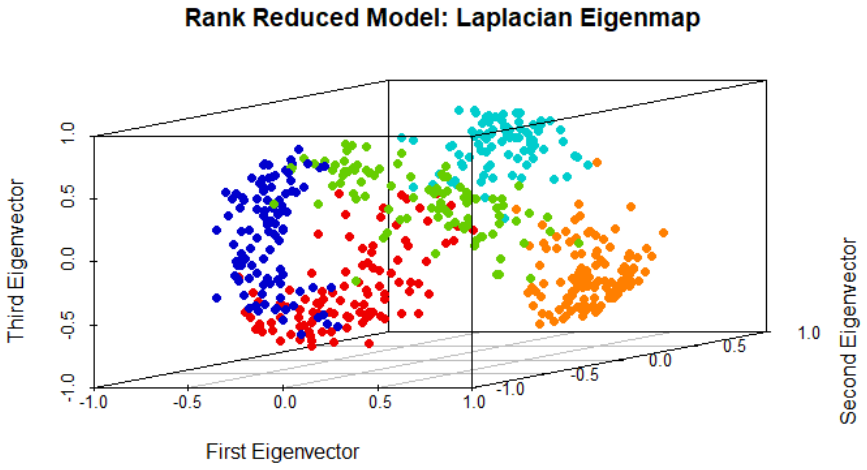
To visually compare the clustering ratios, a heatmap has been applied to the values in the table. The lowest cluster ratios correspond to cells with the most green hue and the highest cluster ratios correspond to cells with the most red hue. Models with the same K value can be compared.

Results: $K = 5$

The rank reduction model produced the lowest cluster ratio out of the three models. Regularisation improves the clusterability of the model outputs. The penalisation model outperforms its unregularised counterpart, the incomplete MLE model, for all values of the threshold, τ. Looking at the clustering metrics of the penalised model with $K = 5$, we see that the clustering metric decreases as τ increases for the penalised and incomplete MLE models. Looking at the corresponding cluster ratios, we can see that the cluster ratios increase as τ increases. This implies that the feature selection did not reduce the clustering metric enough to make it worthwhile.

The clusterability of the rank reduced model is reinforced in the model's Laplacian eigenmap. The Laplacian eigenmap of the fitted rank reduced model with 5 clusters exhibits a pattern in the data and the clusters are quite well defined. This is shown in Figure 6.13.

Figure 6.13: Laplacian Eigenmap of Conditions under the Rank Reduced Model with $k = 5$

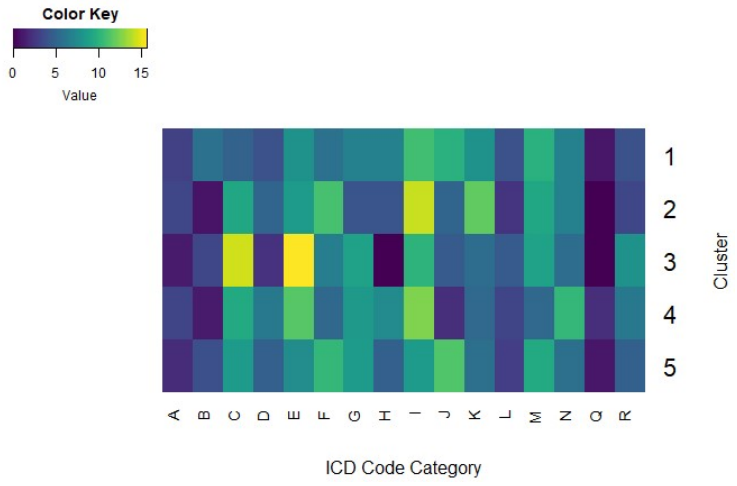


We may wish to investigate how well the clusters grouped chronic conditions together. Whilst the true underlying cluster structure is unknown, we may use the distribution of the clusters across the ICD Code categories as a proxy for how correctly the conditions are clustered. In doing so we investigate Figure 6.14.

Figure 6.14 illustrates the proportion of ICD Code categories in each of the clusters. The matrix underlying Figure 6.14 tabulates the membership of the ICD Codes across the clusters. Conditions are defined by their ICD Codes and these codes are categorised as shown in Table 6.3. The matrix underlying Figure 6.14 modifies the membership matrix by dividing the membership in each cluster by the total number of conditions classified to the cluster. The rows of the matrix underlying Figure 6.14 sum to 100%. This figure allows us to visualise the composition of each cluster.

In analysing the heat-map, we may notice that it is quite murky. There is no discernible trend in the membership composition of the clusters. This implies that the conditions from different ICD Code categories is almost equally distributed across the clusters.

Figure 6.14: Rank Reduced Model: Heatmap of the Percentage of ICD Codes Falling into Each Cluster when $k = 5$, by cluster



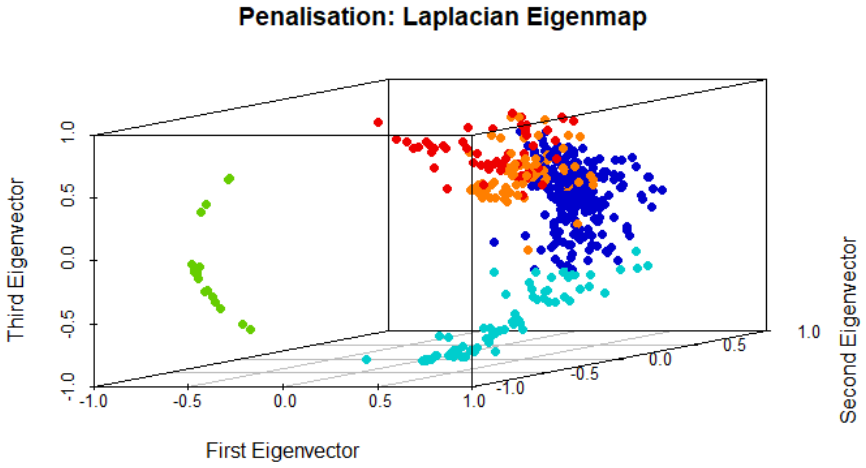
There is a large body of literature which suggests that mental-health related conditions commonly co-occur. One study by Plana-Ripoll *et al.* (2019) analysed a set of mental and behavioural conditions from ICD Codes F00-F09 and F10-F19. The study analysed the disorders in a pairwise manner and found that mental-health related co-morbidities are common. We can use the results of this study, and other studies which support it, as a proxy in determining whether the models correctly cluster conditions. This study would lead us to believe that mental health conditions in ICD Code category F should typically fall into the same cluster. Under the rank reduced model there is no evident clustering of conditions from ICD Code category F. This may suggest that although the clustering metric and cluster ratio under the rank reduced model with 5 clusters was

very low, the actual cluster performance may be poor in terms of its accuracy.

There is no evident clustering of the ICD Code categories under the rank reduction model which is not as one might anticipate. We will also then show output from the L_1-penalisation model with $\tau = 0$ which is the model with the second lowest cluster ratio when $K = 5$.

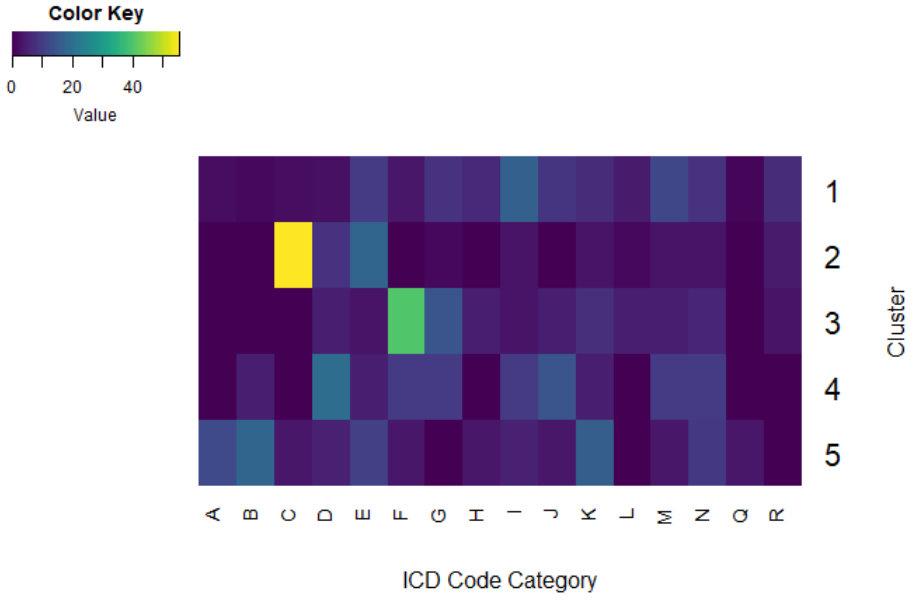
The Laplacian eigenmap of the L_1-penalisation model with $\tau = 0$ and $K = 5$ is shown in Figure 6.15. Figure 6.15 illustrates that a cluster structure is discernible under this model.

Figure 6.15: Laplacian Eigenmap of Conditions under the L_1-Penalisation Model with $\tau = 0$ and $K = 5$



In investigating the membership composition of the clusters we look to Figure 6.16. Under the L_1-penalisation model with $\tau = 0$ and $K = 5$, we notice a strong clustering of conditions from ICD Code category C, malignant neoplasms. Additionally, mental and behavioural conditions from ICD Code category F appear to cluster together. This model was able to extract a cluster of predominantly mental-health related conditions which is in support of medical literature such as the study by Plana-Ripoll *et al.* (2019). Given the nature of cancer, the clustering of malignant neoplasms is intuitive. A malignant neoplasm is a cancerous tumour. Advanced cancer has the ability to spread across an individuals organs. It is then medically intuitive that ICD Codes corresponding to cancerous tumours would form a cluster.

Figure 6.16: L_1-Penalisation Model: Heatmap of the Percentage of ICD Codes Falling into Each Cluster when $\tau = 0$ and $K = 5$, by cluster

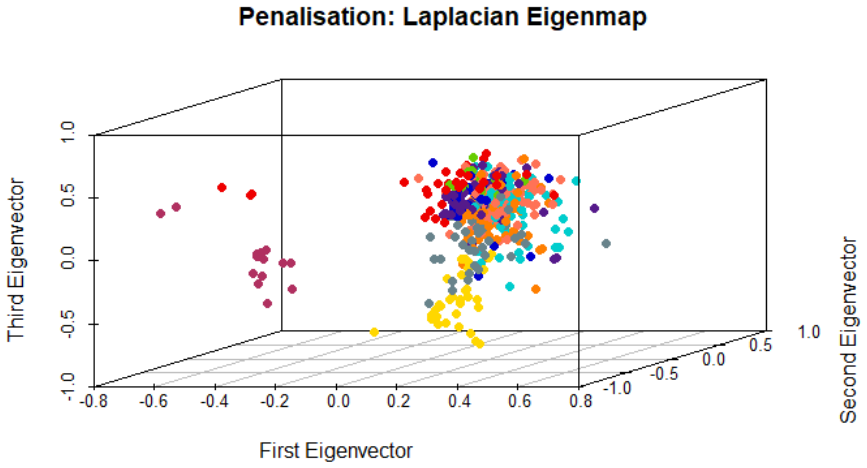


Results: $K = 10$

The L_1-penalised model with $\tau = 0$ produced the lowest cluster ratio out of the three models. Again, regularisation notably improves the clusterability of the model outputs. The penalisation model outperforms its unregularised counterpart, the incomplete MLE model, for all values of the threshold, τ. As before, as τ increases, the clustering metric of the penalised and incomplete MLE models decreases and the corresponding cluster ratios increase. This implies that the feature selection did not reduce the clustering metric enough to make it worthwhile.

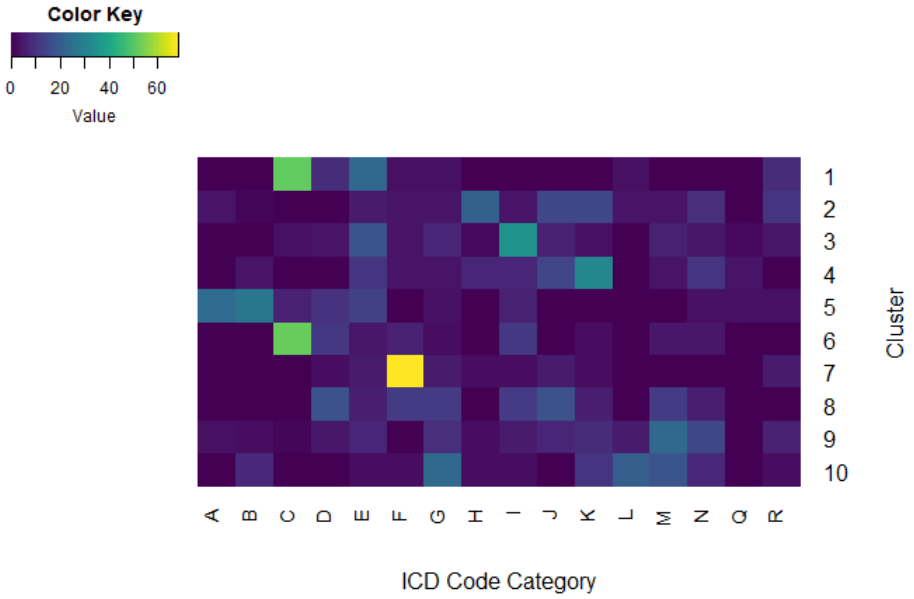
We now investigate the Laplacian eigenmap of the L_1-penalised model with $\tau = 0$ and $K = 10$. Figure 6.17 indicates that segregation of observations is present for some clusters, however there is a mass of observations whose cluster structure is indistinguishable. This may be due to the fact that the data is projected to a lower-dimension which results in a loss of information. The input to the k-means clustering algorithm is the normalised eigenvectors of the graph Laplacian which is a $D \times (K-1)$ matrix. The column dimension of this matrix increases as we introduce more clusters. The result of this is an increasing loss of information projected in the Laplacian Eigenmap. It is also easier to discern a smaller number of clusters.

Figure 6.17: Laplacian Eigenmap of Conditions under L_1-penalised model with $\tau = 0$ and $K = 10$



As before, we can investigate the membership heat-map to determine whether there is a cluster structure in ICD Code categories. Figure 6.18 can be used in this regard.

Figure 6.18: L_1-penalised model: Heatmap of the Percentage of ICD Codes Falling into Each Cluster when $k = 10$ and $\tau = 0$, by cluster



From Figure 6.18 we can deduce,

- The mental and behavioural disorders from ICD Code category F make up most of the composition of cluster 7. This supports a wide body of literature that proposes that mental-health related conditions form common co-morbidities.
- Cluster 1 predominantly consists of conditions in ICD Code categories C and E. ICD Code category C is malignant neoplasms and category E includes endocrine, nutritional and metabolic diseases.
- Cluster 5 predominantly consists of conditions in ICD Code categories A and B. These are the infectious and parasitic diseases, like bacterial and viral infections.

Results: $K = 15$

Table 6.5 reveals that when $K = 15$ the L_1-penalised model with $\tau = 0$ and the incomplete MLE model with $\tau = 0$ produce the lowest cluster ratios. Increasing the number of decimals reveals that the incomplete MLE model has the lowest cluster ratio. For $\tau > 0$, the L_1-penalised model outperforms the incomplete MLE model. The rank reduced model produces the highest cluster ratio. The Laplacian eigenmap of the incomplete MLE model with $\tau = 0$ is given in Figure 6.19 and the membership heat-map in Figure 6.20.

Figure 6.19: Laplacian Eigenmap of Conditions under Incomplete MLE model with $\tau = 0$ and $K = 15$

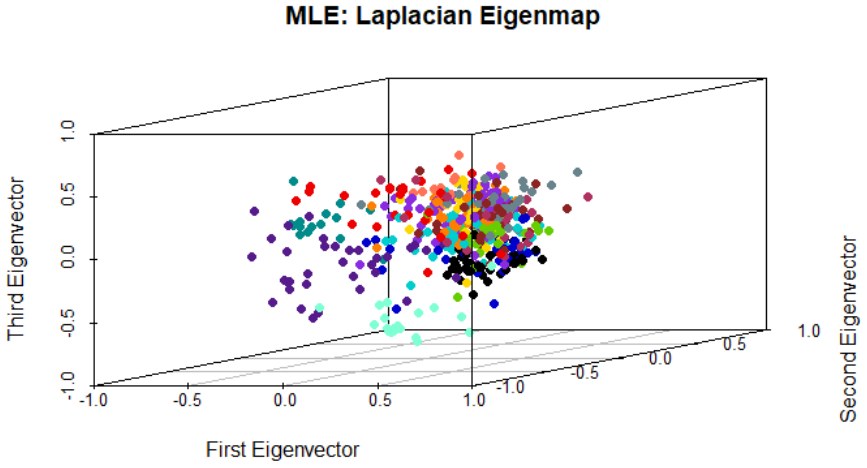
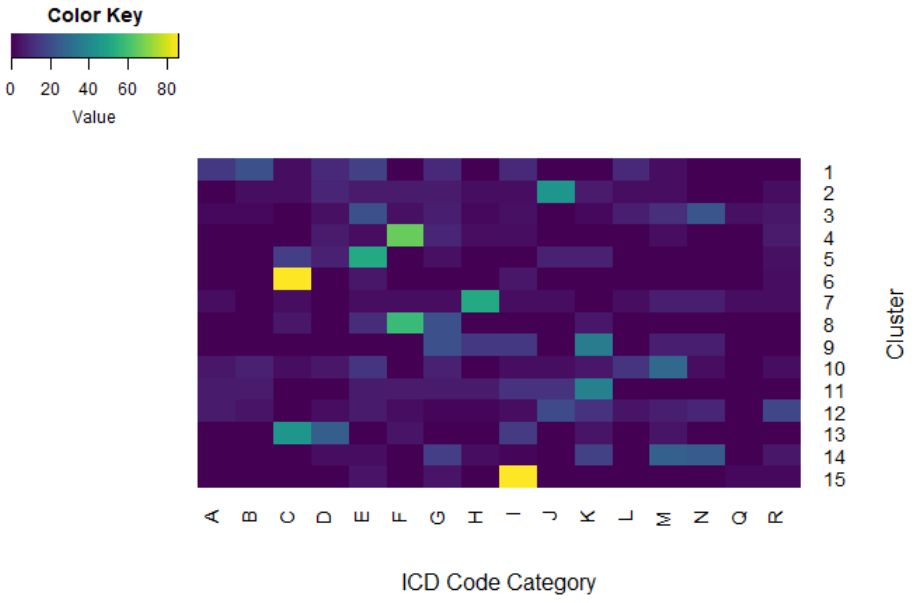


Figure 6.19 is not wholly interpretable which may be due to the loss of information from data projection. Increasing the number of clusters reduces the interpretability of the Laplacian eigenmap visual.

Figure 6.20 allows us to visualise the composition of each cluster.

- Cluster 6 is composed predominantly of conditions from ICD Code category C, malignant neoplasms.
- Cluster 15 is primary composed of conditions from ICD Code category I, diseases of the circulatory system.
- Cluster 4 is mostly made up of mental and behavioural disorders, ICD Code category F, as is cluster 8. One would expect these conditions to form clusters given the extensive body of literature regarding mental-health related co-morbidities.
- Conditions from ICD Code categories C and D appear to cluster together in cluster 13. ICD Code category C corresponds to malignant neoplasms and ICD Code category D corresponds to all other neoplasms as well as immune disorders.

Figure 6.20: Incomplete MLE model: Heatmap of the Percentage of ICD Codes Falling into Each Cluster when $k = 15$ and $\tau = 0$, by cluster



6.4 DISCUSSION

When the models were applied to disease prototype simulated data, both the incomplete MLE model and the penalised model performed well, with the penalised model slightly outperforming the incomplete MLE model. The Gaussian latent variable simulation study showed that the incomplete MLE model produced the most cohesive cluster structure from the data, however none of the models were able to accurately cluster the data to match the underlying dependency structure. The failure of the models to determine the clustering structure may be due to poor model performance or it may be due to how the data was simulated. It is possible that the simulation did not encapsulate a strong dependency between conditions. Under both simulation studies, the rank reduction model performed the worst.

When the models were applied to observational data, the penalised model appeared to perform better than the unregularised incomplete MLE model. The rank reduction model appears to perform the best when we consider 5 clusters because it produces a very low cluster ratio. However, when we investigate the membership composition of the rank reduced model, the results are not intuitive. We expect that mental-health related conditions would form a cluster and that types of cancers would also cluster together. These clusters were not evident in the rank reduced model with $K = 5$, however they were present in the penalised model with $K = 5$. When $K = 10$, the penalised model produced the best cluster ratio and formed a cluster of mental-health related conditions as expected. When $K = 15$, the incomplete MLE model performed the best and a cluster structure was observable across the ICD Code categories. These results should ideally be verified by a medical practitioner.

Overall, we may conclude that the model performance is dependent on the underlying data, however the penalised model and incomplete MLE models appear to perform better than the rank reduced model.

6.5 LIMITATIONS AND IMPROVEMENTS

Model fitting

The rank reduced model parameters α and the number of person-types were fixed at 50 and 100 respectfully. Exploring the effect that these parameters have on the cluster performance of the models would be beneficial. In the simulation study, 200 conditions and 10 clusters were simulated. It would be advantageous to investigate how each of the models fare when the dimensionality of the problem changes, and when the number of underlying clusters change.

The number of iterations per model fit for the regularised models was selected to be 20. Each model fit requires the estimation of a large number of parameters and so increasing the number of iterations per model fit would result in a substantial computational burden. If these models were to be run on a supercomputer, the number of iterations could be increased which may result in better model performance.

Different Clustering Techniques

A potential extension of this study is to consider multiple similarity-based clustering methods, or alternatively probabilistic graphical models. We used spectral clustering that makes use of a similarity matrix, however other similarity-based clustering methods exist. Because this study involves constructing a network, the problem can also be addressed from the viewpoint of graph theory. In graph theory, a graph is constructed from an adjacency matrix which is analogous to the similarity matrix used in similarity-based clustering methods. Chronic conditions can then be clustered using a probabilistic graphical model.

Numerical Clustering Performance Metric

We used visual means to determine the clustering accuracy of the models applied to simulated data. Ideally, a numerical method would also be used to assess how accurately the models detected the underlying dependency structure. The numerical and visual methods could be used in conjunction to assess clustering accuracy.

Medical Expert Input

The results of this study could be assessed by a medical professional to determine the validity of the extracted clusters from the observational data.

REFERENCES

Agrawal, S. and Agrawal, P.K. (2016). Association between body mass index and prevalence of multimorbidity in low-and middle-income countries: a cross-sectional study. *International journal of medicine and public health*, vol. 6, no. 2, p. 73.

Alaba, O. and Chola, L. (2013). The social determinants of multimorbidity in south africa. *International journal for equity in health*, vol. 12, no. 1, pp. 1–10.

Arokiasamy, P., Uttamacharya, U., Jain, K., Biritwum, R.B., Yawson, A.E., Wu, F., Guo, Y., Maximova, T., Espinoza, B.M., Rodríguez, A.S. *et al.* (2015). The impact of multimorbidity on adult physical and mental health in low-and middle-income countries: what does the study on global ageing and adult health (sage) reveal? *BMC medicine*, vol. 13, no. 1, pp. 1–16.

Boehmke, B. (). *UC Business Analytics R Programming Guide: K-means Cluster Analysis*.
Available at: `https://uc-r.github.io/kmeans_clustering`

Centers for Disease Control and Prevention (2021 April). *About Chronic Diseases*.
Available at: `https://www.cdc.gov/chronicdisease/about/index.htm`

Chang, A.Y., Gómez-Olivé, F.X., Payne, C., Rohr, J.K., Manne-Goehler, J., Wade, A.N., Wagner, R.G., Montana, L., Tollman, S. and Salomon, J.A. (2019). Chronic multimorbidity among older adults in rural south africa. *BMJ Global Health*, vol. 4, no. 4, p. e001386.

Dai, B., Ding, S. and Wahba, G. (2013). Multivariate bernoulli distribution. *Bernoulli*, vol. 19, no. 4, pp. 1465–1483.

De Canditiis, D. (2020). A global approach for learning sparse ising models. *Mathematics and Computers in Simulation*, vol. 176, pp. 160–170.

Kvalseth, T.O. (2017). On normalized mutual information: Measure derivations and properties. *Entropy*, vol. 19, no. 11. ISSN 1099-4300.
Available at: `https://www.mdpi.com/1099-4300/19/11/631`

Larkin, J., Foley, L., Smith, S.M., Harrington, P. and Clyne, B. (2020 Nov). The experience of financial burden for people with multimorbidity: A systematic review of qualitative research.

Health Expectations, vol. 24, no. 2, p. 282–295.
Available at: https://onlinelibrary.wiley.com/doi/epdf/10.1111/hex.13166

Mathematics Stack Exchange User (2021). Differentiable approximation of the absolute value function. Mathematics Stack Exchange. URL:https://math.stackexchange.com/q/1172472 (version: 2015-03-02), https://math.stackexchange.com/q/1172472.
Available at: https://math.stackexchange.com/q/1172472

National Eye Institute (2019 June). *Causes of Color Blindness*.
Available at: https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/color-blindness/causes-color-blindness

Ng, S. (2015). A two-way clustering framework to identify disparities in multimorbidity patterns of mental and physical health conditions among australians. *Statistics in medicine*, vol. 34, no. 26, pp. 3444–3460.

Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Foundations and Trends in optimization*, vol. 1, no. 3, pp. 127–239.

Peltzer, K. (2018). Tuberculosis non-communicable disease comorbidity and multimorbidity in public primary care patients in south africa. *African Journal of Primary Health Care and Family Medicine*, vol. 10, no. 1, pp. 1–6.

Plana-Ripoll, O., Pedersen, C.B., Holtz, Y., Benros, M.E., Dalsgaard, S., De Jonge, P., Fan, C.C., Degenhardt, L., Ganna, A., Greve, A.N. *et al.* (2019). Exploring comorbidity within mental disorders among a danish national population. *JAMA psychiatry*, vol. 76, no. 3, pp. 259–270.

Ravikumar, P., Wainwright, M.J. and Lafferty, J.D. (2010). High-dimensional ising model selection using l1-regularized logistic regression. *The Annals of Statistics*, vol. 38, no. 3, pp. 1287–1319.

Reshef, R. (2015). How do you build a "people who bought this also bought that"-style recommendation engine. *URL https://datasciencemadesimpler. wordpress. com/tag/alternating-least-squares*.

Salkind, N. (2010). Latent Variable. *Encyclopedia of Research Design*.
Available at: https://dx.doi.org/10.4135/9781412961288.n213

Singh, A. (2010). *Nonlinear Methods.*
Available at: https://www.cs.cmu.edu/~aarti/Class/10701/slides/Lecture21_1.pdf

Singh, A., Xing, E., Hein, M. and Luxburg, U. (2010). *Spectral Clustering.*
Available at: http://www.cs.cmu.edu/~aarti/Class/10701/slides/Lecture21_2.pdf

Singh, S.P. (2020). The ising model: Brief introduction and its application. In: *Solid State Physics-Metastable, Spintronics Materials and Mechanics of Deformable Bodies-Recent Progress.* IntechOpen.

Sum, G., Ishida, M., Koh, G.C.-H., Singh, A., Oldenburg, B. and Lee, J.T. (2020 Apr). Implications of multimorbidity on healthcare utilisation and work productivity by socioeconomic groups: Cross-sectional analyses of australia and japan. *Plos One*, vol. 15, no. 4.
Available at: https://pubmed.ncbi.nlm.nih.gov/32343739/

Taboga, M. (2017*a*). *Matrix product and rank.*
Available at: https://www.statlect.com/matrix-algebra/matrix-product-and-rank

Taboga, M. (2017*b*). *Rank of a matrix.*
Available at: https://www.statlect.com/matrix-algebra/rank-of-a-matrix

WebMD (2021). *Diabetes and High Blood Pressure.*
Available at: https://www.webmd.com/diabetes/high-blood-pressure

World Health Organization (2016*a*). *International Statistical Classification of Diseases and Related Health Problems 10th Revision.*
Available at: https://icd.who.int/browse10/2016/en

World Health Organization (2016*b*). *Multimorbidity.* Technical Series on Safer Primary Care. World Health Organization.

APPENDIX A

CALCULATIONS

A.1 PARTIAL DERIVATIVES ASSOCIATED WITH THE PENALISED CO-OCCURRENCE MODELS

A.1.1 Partial Derivative of $g\left(\pi_{ij}^C,\ \pi_i^0\right)$ with Respect to π_{ij}^C

$$\frac{\partial g\left(\pi_{ij}^C,\ \pi_i^0\right)}{\partial \pi_{ij}^C} = \frac{\partial}{\partial \pi_{ij}^C}\left[\sum_{i\neq j=1}^{D}\left(N_{ij}\ln\left(\frac{\pi_{ij}^C}{1-\pi_{ij}^C}\right) + N_j\ln\left(1-\pi_{ij}^C\right) - \lambda|\pi_{ij}^C - \pi_i^0|\right)\right]$$

where

$$\frac{\partial}{\partial \pi_{ij}^C}\ln\left(\frac{\pi_{ij}^C}{1-\pi_{ij}^C}\right) = \left(\frac{1-\pi_{ij}^C}{\pi_{ij}^C}\right)\left(\frac{(1-\pi_{ij}^C) - \pi_{ij}^C \times -1}{\left(1-\pi_{ij}^C\right)^2}\right) = \frac{1}{\pi_{ij}^C(1-\pi_{ij}^C)}$$

and

$$\frac{\partial}{\partial \pi_{ij}^C}\ln(1-\pi_{ij}^C) = -1 \times \frac{1}{1-\pi_{ij}^C}$$

Using the fact that $\frac{\partial}{\partial x}|u(x)| = \frac{\partial}{\partial x}\sqrt{u^2(x)} = \frac{u(x)}{|u(x)|}u'(x)$,

$$\frac{\partial}{\partial \pi_{ij}^C}|\pi_{ij}^C - \pi_i^0| = \frac{\pi_{ij}^C - \pi_i^0}{|\pi_{ij}^C - \pi_i^0|}$$

The partial derivative can then be written as,

$$\begin{aligned}
\frac{\partial g\left(\pi_{ij}^C,\ \pi_i^0\right)}{\partial \pi_{ij}^C} &= \frac{N_{ij}}{\pi_{ij}^C(1-\pi_{ij}^C)} - \frac{N_j}{1-\pi_{ij}^C} - \lambda\frac{\pi_{ij}^C - \pi_i^0}{|\pi_{ij}^C - \pi_i^0|} \\
&= \frac{N_{ij} - N_j\pi_{ij}^C}{\pi_{ij}^C(1-\pi_{ij}^C)} - \lambda\frac{\pi_{ij}^C - \pi_i^0}{|\pi_{ij}^C - \pi_i^0|} \\
&= \begin{cases} \frac{N_{ij}-N_j\pi_{ij}^C}{\pi_{ij}^C(1-\pi_{ij})} - \lambda & \text{for } \pi_{ij}^C > \pi_i^0 \\ \frac{N_{ij}-N_j\pi_{ij}^C}{\pi_{ij}^C(1-\pi_{ij}^C)} + \lambda & \text{for } \pi_{ij}^C < \pi_i^0 \\ \text{undefined} & \text{for } \pi_{ij}^C = \pi_i^0 \end{cases}
\end{aligned}$$

A.1.2 Partial Derivative of $g\left(\pi_{ij}^C,\ \pi_i^0\right)$ with Respect to π_i^0

$$\begin{aligned}
\frac{\partial g\left(\pi_{ij}^C,\ \pi_i^0\right)}{\partial \pi_i^0} &= \frac{\partial}{\partial \pi_i^0}\left[\sum_{i\neq j=1}^{D}\left(N_{ij}\ln\left(\frac{\pi_{ij}^C}{1-\pi_{ij}^C}\right)+N_j\ln\left(1-\pi_{ij}^C\right)-\lambda|\pi_{ij}^C-\pi_i^0|\right)\right] \\
&= \frac{\partial}{\partial \pi_i^0}\left(-\lambda|\pi_{ij}^C-\pi_i^0|\right) \\
&= -\lambda\left(\frac{\pi_{ij}^C-\pi_i^0}{|\pi_{ij}^C-\pi_i^0|}\times -1\right) \\
&= \lambda\frac{\pi_{ij}^C-\pi_i^0}{|\pi_{ij}^C-\pi_i^0|} \\
&= \begin{cases} \lambda & \text{for } \pi_{ij}^C > \pi_i^0 \\ -\lambda & \text{for } \pi_{ij}^C < \pi_i^0 \\ \text{undefined} & \text{for } \pi_{ij}^C = \pi_i^0 \end{cases}
\end{aligned}$$

A.1.3 Partial Derivative of $g(\pi_{ij}^J, \pi_i, \pi_j)$ with Respect to π_{ij}^J

$$\frac{\partial g(\pi_{ij}^J,\pi_i,\pi_j)}{\partial \pi_{ij}^J} = \frac{\partial}{\partial \pi_{ij}^J}\left(\left[\sum_{i,j=1}^{D} N_{ij}\ln\left(\frac{\pi_{ij}^J}{1-\pi_{ij}^J}\right)+N_{sample}\ln\left(1-\pi_{ij}^J\right)\right]-\lambda\sum_{i\neq j=1}^{D}\left|\pi_{ij}^J-\pi_i\pi_j\right|\right)$$

This can be calculated using a similar approach to the corresponding derivative in the estimation of the conditional probability model. It can then be calculated that,

$$\begin{aligned}
\frac{\partial g(\pi_{ij}^J,\pi_i,\pi_j)}{\partial \pi_{ij}^J} &= \frac{N_{ij}}{\pi_{ij}^J[1-\pi_{ij}^J]}-\frac{N_{sample}}{1-\pi_{ij}^J}-\lambda\frac{\pi_{ij}^J-\pi_i\pi_j}{\left|\pi_{ij}^J-\pi_i\pi_j\right|} \\
&= \frac{N_{ij}-N_{sample}\pi_{ij}^J}{\pi_{ij}^J\left(1-\pi_{ij}^J\right)}-\lambda\frac{\pi_{ij}^J-\pi_i\pi_j}{\left|\pi_{ij}^J-\pi_i\pi_j\right|} \\
&= \begin{cases} \frac{N_{ij}-N_{sample}\pi_{ij}^J}{\pi_{ij}^J\left(1-\pi_{ij}^J\right)}-\lambda, & \text{for } \pi_{ij}^J > \pi_i\pi_j \\ \frac{N_{ij}-N_{sample}\pi_{ij}^J}{\pi_{ij}^J\left(1-\pi_{ij}^J\right)}+\lambda, & \text{for } \pi_{ij}^J < \pi_i\pi_j \\ \text{undefined}, & \text{for } \pi_{ij}^J = \pi_i\pi_j \end{cases}
\end{aligned}$$

A.1.4 Partial Derivative of $g(\pi_{ij}^J, \pi_i, \pi_j)$ with Respect to π_i

Deriving the penalty term with respect to π_i yields,

$$\begin{aligned}\frac{\partial}{\partial \pi_i}\left|\pi_{ij}^J - \pi_i\pi_j\right| &= \frac{\pi_{ij}^J - \pi_i\pi_j}{\left|\pi_{ij}^J - \pi_i\pi_j\right|}\frac{\partial}{\partial \pi_i}\left(\pi_{ij}^J - \pi_i\pi_j\right) \\ &= \frac{\pi_{ij}^J - \pi_i\pi_j}{\left|\pi_{ij}^J - \pi_i\pi_j\right|}\left(-\pi_j\right)\end{aligned}$$

The partial derivative of the objective function with respect to π_i is then written as,

$$\begin{aligned}\frac{\partial g(\pi_{ij}^J, \pi_i, \pi_j)}{\partial \pi_i} &= \frac{\partial}{\partial \pi_i}\left(\left[\sum_{i,j=1}^{D} N_{ij}\ln\left(\frac{\pi_{ij}^J}{1-\pi_{ij}^J}\right) + N_{sample}\ln\left(1-\pi_{ij}^J\right)\right] - \lambda\sum_{i\neq j=1}^{D}\left|\pi_{ij}^J - \pi_i\pi_j\right|\right) \\ &= \frac{\partial}{\partial \pi_i}\left(-\lambda\left|\pi_{ij}^J - \pi_i\pi_j\right|\right) \\ &= -\lambda\left(\frac{\pi_{ij}^J - \pi_i\pi_j}{\left|\pi_{ij}^J - \pi_i\pi_j\right|}\left(-\pi_j\right)\right) \\ &= \lambda\pi_j\frac{\pi_{ij}^J - \pi_i\pi_j}{\left|\pi_{ij}^J - \pi_i\pi_j\right|} \\ &= \begin{cases}\lambda\pi_j, & \text{for } \pi_{ij}^J > \pi_i\pi_j \\ -\lambda\pi_j, & \text{for } \pi_{ij}^J < \pi_i\pi_j \\ \text{undefined}, & \text{for } \pi_{ij}^J = \pi_i\pi_j\end{cases}\end{aligned}$$

A.2 PARTIAL DERIVATIVES ASSOCIATED WITH THE RANK-REDUCED CO-OCCURRENCE MODELS

In the section, 'Regularization by Rank Reduction' the partial derivative of the objective function, $\hat{f}$, with respect to Q_{tj} is shown. Here we show the calculations associated with this partial derivative. The additional constraint on the $\Pi^{T|X}$ matrix whereby,

$$\sum_{i=1}^{K}\Pi_{ij}^{T|X} = \sum_{i=1}^{K}\left(\frac{\exp(Q_{ij})}{\sum_{k=1}^{K}\exp(Q_{kj})}\right) = 1$$

results in the partial derivative of the objective function with respect to Q_{tj} being more complicated than the previously calculated partial derivative.

We have, for $t \neq s$

$$\begin{aligned}\frac{\partial}{\partial Q_{tj}}\left(\frac{\exp(Q_{sj})}{\sum_{k=1}^{K}\exp(Q_{kj})}\right) &= -\frac{\exp(Q_{sj})\exp(Q_{tj})}{\left(\sum_{k=1}^{K}\exp(Q_{kj})\right)^2} \\ &= -\frac{\exp(Q_{sj})}{\sum_{k=1}^{K}\exp(Q_{kj})}\frac{\exp(Q_{tj})}{\sum_{k=1}^{K}\exp(Q_{kj})} \\ &= -\Pi_{sj}^{T|X}\Pi_{tj}^{T|X}\end{aligned}$$

In order to understand the effect of the case where $t \neq s$, a simple illustration where $K = 3$ will be explored. Suppose $K = 3$, then the objective function becomes,

$$\begin{aligned}\hat{f}_3 = \sum_{i\neq j} N_{ij}\ln\Bigg(&\frac{\exp(Z_{i1}+Q_{1j})}{(1+\exp(Z_{i1}))\,(\exp(Q_{1j})+\exp(Q_{2j})+\exp(Q_{3j}))} \\ &+ \frac{\exp(Z_{i2}+Q_{2j})}{(1+\exp(Z_{i2}))\,(\exp(Q_{1j})+\exp(Q_{2j})+\exp(Q_{3j}))} \\ &+ \frac{\exp(Z_{i3}+Q_{3j})}{(1+\exp(Z_{i3}))\,(\exp(Q_{1j})+\exp(Q_{2j})+\exp(Q_{3j}))}\Bigg) \\ &+ (N_j - N_{ij})\ln\Bigg(1 - \frac{\exp(Z_{i1}+Q_{1j})}{(1+\exp(Z_{i1}))\,(\exp(Q_{1j})+\exp(Q_{2j})+\exp(Q_{3j}))} \\ &- \frac{\exp(Z_{i2}+Q_{2j})}{(1+\exp(Z_{i2}))\,(\exp(Q_{1j})+\exp(Q_{2j})+\exp(Q_{3j}))} \\ &- \frac{\exp(Z_{i3}+Q_{3j})}{(1+\exp(Z_{i3}))\,(\exp(Q_{1j})+\exp(Q_{2j})+\exp(Q_{3j}))}\Bigg)\end{aligned}$$

For illustration, we find the maximum of the objective function with respect to Q_{2j}. We denote:

$$\begin{aligned}A_1 &= \frac{\exp(Z_{i1}+Q_{1j})}{(1+\exp(Z_{i1}))\,(\exp(Q_{1j})+\exp(Q_{2j})+\exp(Q_{3j}))} \\ A_2 &= \frac{\exp(Z_{i2}+Q_{2j})}{(1+\exp(Z_{i2}))\,(\exp(Q_{1j})+\exp(Q_{2j})+\exp(Q_{3j}))} \\ A_3 &= \frac{\exp(Z_{i3}+Q_{3j})}{(1+\exp(Z_{i3}))\,(\exp(Q_{1j})+\exp(Q_{2j})+\exp(Q_{3j}))}\end{aligned}$$

Taking the partial derivative with respect to Q_{2j} we have that,

$$\begin{aligned}
\frac{\partial A_1}{\partial Q_{2j}} &= \frac{\exp(Z_{i1}+Q_{1j})}{(1+\exp(Z_{i1}))}\frac{\partial}{\partial Q_{2j}}\left(\frac{1}{\exp(Q_{1j})+\exp(Q_{2j})+\exp(Q_{3j})}\right) \\
&= -\frac{\exp(Z_{i1}+Q_{1j})}{(1+\exp(Z_{i1}))}\frac{\exp(Q_{2j})}{(\exp(Q_{1j})+\exp(Q_{2j})+\exp(Q_{3j}))^2} \\
\text{Similarly, } \frac{\partial A_3}{\partial Q_{2j}} &= -\frac{\exp(Z_{i3}+Q_{3j})}{(1+\exp(Z_{i3}))}\frac{\exp(Q_{2j})}{(\exp(Q_{1j})+\exp(Q_{2j})+\exp(Q_{3j}))^2}
\end{aligned}$$

Therefore we have for $t \neq s$,

$$\frac{\partial}{\partial Q_{tj}}\frac{\exp(Z_{is}+Q_{sj})}{(1+\exp(Z_{is}))\left(\sum_{k=1}^{K}\exp(Q_{kj})\right)} = -\frac{\exp(Z_{is}+Q_{sj})}{(1+\exp(Z_{is}))}\frac{\exp(Q_{tj})}{\left(\sum_{k=1}^{K}\exp(Q_{kj})\right)^2}$$

When $t = s$,

$$\begin{aligned}
\frac{\partial A_2}{\partial Q_{2j}} &= \frac{\exp(Z_{i2})}{1+\exp(Z_{i2})}\frac{\partial}{\partial Q_{2j}}\left(\frac{\exp(Q_{2j})}{\exp(Q_{1j})+\exp(Q_{2j})+\exp(Q_{3j})}\right) \\
&= \frac{\exp(Z_{i2})}{1+\exp(Z_{i2})}\left(\frac{\exp(Q_{2j})}{\exp(Q_{1j})+\exp(Q_{2j})+\exp(Q_{3j})} - \frac{\exp(2Q_{2j})}{(\exp(Q_{1j})+\exp(Q_{2j})+\exp(Q_{3j}))^2}\right) \\
&= \frac{\exp(Z_{i2}+Q_{2j})}{(1+\exp(Z_{i2}))\,(\exp(Q_{1j})+\exp(Q_{2j})+\exp(Q_{3j}))}\left(1-\frac{\exp(Q_{2j})}{\exp(Q_{1j})+\exp(Q_{2j})+\exp(Q_{3j})}\right)
\end{aligned}$$

Therefore, more generally, when $t = s$

$$\frac{\partial}{\partial Q_{tj}}\frac{\exp(Z_{is}+Q_{sj})}{(1+\exp(Z_{is}))\left(\sum_{k=1}^{K}\exp(Q_{kj})\right)} = \frac{\exp(Z_{it}+Q_{tj})}{(1+\exp(Z_{it}))\left(\sum_{k=1}^{K}\exp(Q_{kj})\right)}\left(1-\frac{\exp(Q_{tj})}{\sum_{k=1}^{K}\exp(Q_{kj})}\right)$$

We now maximise the objective function for our simplified example,

$$\frac{\partial \hat{f}_3}{\partial Q_{2j}} = \sum_{i \neq j} \frac{N_{ij}}{A_1 + A_2 + A_3} \left(\frac{\exp(Z_{i2} + Q_{2j})}{(1 + \exp(Z_{i2})) \left(\sum_{k=1}^{K} \exp(Q_{kj}) \right)} \left(1 - \frac{\exp(Q_{2j})}{\sum_{k=1}^{K} \exp(Q_{kj})} \right) \right.$$
$$\left. - \frac{\exp(Q_{2j})}{\left(\sum_{k=1}^{K} \exp(Q_{kj}) \right)^2} \left(\frac{\exp(Z_{i1} + Q_{1j})}{1 + \exp(Z_{i1})} + \frac{\exp(Z_{i3} + Q_{3j})}{1 + \exp(Z_{i3})} \right) \right)$$
$$+ \frac{N_j - N_{ij}}{1 - (A_1 + A_2 + A_3)} \left(- \frac{\exp(Z_{i2} + Q_{2j})}{(1 + \exp(Z_{i2})) \left(\sum_{k=1}^{K} \exp(Q_{kj}) \right)} \left(1 - \frac{\exp(Q_{2j})}{\sum_{k=1}^{K} \exp(Q_{kj})} \right) \right.$$
$$\left. + \frac{\exp(Q_{2j})}{\left(\sum_{k=1}^{K} \exp(Q_{kj}) \right)^2} \left(\frac{\exp(Z_{i1} + Q_{1j})}{1 + \exp(Z_{i1})} + \frac{\exp(Z_{i3} + Q_{3j})}{1 + \exp(Z_{i3})} \right) \right)$$

Therefore more generally stated,

$$\frac{\partial \hat{f}}{\partial Q_{tj}} = \sum_{i \neq j} N_{ij} \left(\sum_{k=1}^{K} \frac{\exp(Z_{ik} + Q_{kj})}{(1 + \exp(Z_{ik})) \sum_{k=1}^{K} \exp(Q_{kj})} \right)^{-1} \left(\frac{\exp(Z_{it} + Q_{tj})}{(1 + \exp(Z_{it})) \left(\sum_{k=1}^{K} \exp(Q_{kj}) \right)} \right.$$
$$\left. \times \left(1 - \frac{\exp(Q_{tj})}{\sum_{k=1}^{K} \exp(Q_{kj})} \right) - \frac{\exp(Q_{tj})}{\left(\sum_{k=1}^{K} \exp(Q_{kj}) \right)^2} \sum_{k \neq t}^{K} \left(\frac{\exp(Z_{ik} + Q_{kj})}{1 + \exp(Z_{ik})} \right) \right)$$
$$+ (N_j - N_{ij}) \left(1 - \sum_{k=1}^{K} \frac{\exp(Z_{ik} + Q_{kj})}{(1 + \exp(Z_{ik})) \sum_{k=1}^{K} \exp(Q_{kj})} \right)^{-1} \left(- \frac{\exp(Z_{it} + Q_{tj})}{(1 + \exp(Z_{it})) \left(\sum_{k=1}^{K} \exp(Q_{kj}) \right)} \right.$$
$$\left. \times \left(1 - \frac{\exp(Q_{tj})}{\sum_{k=1}^{K} \exp(Q_{kj})} \right) + \frac{\exp(Q_{tj})}{\left(\sum_{k=1}^{K} \exp(Q_{kj}) \right)^2} \sum_{k \neq t}^{K} \left(\frac{\exp(Z_{ik} + Q_{kj})}{1 + \exp(Z_{ik})} \right) \right)$$