A pattern-clustering method for longitudinal data - heroin users receiving methadone

CHIEN-JU LIN

STATISTICAL SCIENCE

University College London

A thesis for the degree of Doctor of Philosophy

October 2014

Declaration

I confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Name: CHIEN-JU LIN

Signature:

Date:

Acknowledgements

I would like to express my sincere gratitude to everyone who supported me throughout my PhD study.

I would like to express the deepest appreciation to my PhD supervisor Dr. Christian Hennig for the continuous support of my research, for his patience, enthusiasm, and vast knowledge.

I would also like to thank my second supervisor, Dr. Ioannis Kosmidis, and thesis committee, Prof. Tom Fearn and Prof. Sabine Landau, for their encouragement and insightful comments.

I thank my fellows at the Department of Statistical Science. In particular, I am grateful to Oya Kalaycioglu and Khadijeh Taiyari for the stimulating discussions and for all the fun we had. Special thanks to Tsu-Jui Cheng, Saul Holding, Maurice Lyver, Yu-Chia Shih, and Ying-Li Wu for their encouragement and support.

Last but not the least, I owe my deepest gratitude to my family: my parents, my lovely sister and brother, my adorable nephews for supporting me spiritually throughout my life.

Abstract

Methadone is used as a substitute of heroin and there may be certain groups of users according to methadone dosage. In this work we analyze data for 314 participants of a methadone study over 180 days. The data, which is called category-ordered data throughout this study, consists of seven categories in which six categories have an ordinal scale for representing dosages and one category for missing dosages. We develop a clustering method involving the so-called p-dissimilarity, modification of Prediction Strength (PS), a null model test, and two ordering algorithms. (1) The p-dissimilarity is used to measure dissimilarity between the 180-day time series of the participants. It accommodates categorical and ordinal scales by using a parameter p as a switch between data being treated as categorical and ordinal. It measures dissimilarity between observed dosages and missing dosages by using a parameter β. Also, it could be applied in a wider field of applications, such as survey studies in which questions use choices on the Likert scales and a $don't know$-category. (2) The PS determines the number of clusters by measuring the stability of clusters, and the Average Silhouette Width (ASW) measures coherence. We propose rules to modify PS so that it can be fully applied to hierarchal clustering methods. Next, instead of preselecting a clustering method, we let the data to decide which clustering method to use based on cluster stability and cluster coherence. The partition around medoids (PAM) method is then selected. (3) We propose the null model test to determine the number of clusters (k). Many methods for the determination of number of clusters give values for $k \geq 2$ based on cluster compactness and separation, and suggest to use the k with the highest value. Viewing this question from a different perspective, for a fixed k and a selected clustering method, the null model test uses a null model and parametric bootstrap to explore the distribution of a statistic under the

null assumption. A hypothesis test for each k can then be performed. For our data, we construct a Markov null model without structure of clusters, in which the distributions of the categories are the same as those of the real data. We apply the null model test to investigate whether the clusters found according to PAM and ASW/PS can be explained by random variation. (4) We use heatplots to evaluate the quality of clustering. A heatplot is a graph that represents data by colour. It consists of horizontal lines representing the data for objects. However, the interpretability of a heatplot strongly depends on the location of the objects along the vertical-axis. We propose two algorithms to locate objects on a heatplot. The first algorithm using multidimensional scaling (MDS) is for general use. The second algorithm using projection vector is for the PAM method. Each of them locates objects in a heatplot. The heatplot can then be used for information visualisation. It displays clustering structures, relationships between objects and clusters in terms of their dissimilarities, locations of medoids, and the density of clusters. Despite the fact that no significant clustering structure is observed, the sequences of categories for clusters are clinically useful. The sequences of categories indicate detoxification. Our data shows participants with low heroin addictions attempted to reduce/quit the use of methadone at the third month. As for participants with high addictions, few attempted to reduce the use of methadone at the fifth month and most required more time to finish the detoxification process. Also, we find the heroin onset age might have an influence on the patterns of detoxification.

Notations

The following notations, abbreviations and defined terms are used throughout this thesis. They are also introduced in their first occurrence in each chapter.

Symbol	Definition and Explanation
k	denotes the number of clusters
x_{it}	denotes data for the t^{th} variable for object i.
$\mathbf{x}_i$	represents data for object i.
$d(.,.)$	denotes dissimilarity between variables
$D(.,.)$	denotes dissimilarity between objects \ clusters
C_i	denotes a category i, representing a set of dosages
$\delta_{ii'(t)}$	is equal to 1 when both objects i and i' for their t^{th} values are non-missing, and equal to 0 otherwise.
p	is a tuning constant, $0 < p < 1$, for measuring dissimilarities between objects.
$\alpha_{ii'}(t)$	refers to dissimilarity between the t^{th} values for objects i and i'. It is set to the absolute value of the difference between the t^{th} values
$\beta(t)$	refers to dissimilarity between the t^{th} values in which one or both of them are missing values.

Symbol	Explanation
MMT	Methadone Maintenance Therapy
ASW(k)	Average Silhouette Width index for k clusters
PS	Prediction Srengh
MDS	Multidimensional scaling
Dosage_{314}	dosage data for the 314 participants for 180 days
CO_{314}	category-ordered data for the 314 participants

Term	Definition
stable methadone dosage	means that categorized dosage for a participant consists of long sequences of categories.
initial date	means the first date on which a participant joined the MMT.
category-ordered data	means the data that consists of categories, referring to sets of dosage and a set of missing dosage.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Drug abuse creates problems in society and the economy. The statistical news released by the Ministry of Justice of Taiwan in 2010 showed that among all arrests for drug abuse violations, 74.2% of them were arrested on charges of Schedule I drugs, defined as drugs with a high potential for abuse and highly addictive. Schedule I drugs include heroin, opium, morphine, etc. Moreover, there was an increasing trend in the number of arrests for abusing Schedule I drugs over the years. Winick [1962] found addicts mature out of addiction as a reflection of their life cycle or they mature out of addiction as a function of the length of their addiction. Termorshuizen et al. [2005] showed that the concept of "maturing out" to a drug-free state did not apply to the majority of drug users. Also, Termorshuizen et al. [2005] examined harmfulness to drug users by the mortality rates and reported at least 27% of drug users died within 20 years of starting regular drug use.

Among the abuse Schedule I drugs, heroin is the most expensive and highly addictive. Heroin-dependent individuals who aim at overcoming their addiction are offered a methadone maintenance therapy (MMT) for many years. The main purpose of the MMT is not to help them to achieve abstinence but to minimize the harm associated with the use of heroin (Ball and Ross [1991]; Ward et al. [1999]). Research showed that MMT had a positive effect on drug users and on society (Gossop et al. [2000]; Marsch [1998]; Masson et al. [2004]; McLellan et al. [1985]; Powers and Anglin [1993]; Strain

et al. [1993a,b]). The effect of methadone lasts 24 hours and consequently it has to be taken on a daily basis. To date there is no clear principle for the determination of the methadone dosage. Physicians prescribe dosages based on their own intuition.

Some researchers studied methadone dosages. Maxwell and Shinderman [2002] reported that higher methadone dosages (above 100 mg/day) were more effective in treating heroin addicts, while Maremmani et al. [2003] observed that many heroin addicts had positive outcomes with lower dosages. Both high and low dosages are considered as good prescriptions. There is no principle for determining proper methadone dosages. Some researchers studied the association between methadone dosages and groups. Langendam et al. [1998] observed that the mean methadone dosage was higher for ethnic west Europeans, older drug users, HIV-positive drug users, longer duration of methadone use and so on. Murray et al. [2008] surveyed 54 participants from a methadone maintenance clinic and found that methadone dosage might be uniquely related to the personality disorders. On the other hand, Gossop et al. [2000] performed a one year follow-up study on 478 participants. The Euclidean distance K-Means clustering method was used to group their study participants by the frequency of their illicit drug use over time. Four groups were identified and two groups showed substantial reductions in their illicit drug use and criminality. They concluded that the methadone dosage might be related to a certain group in which MMT was appropriate. Their works were limited in using self-reported data that was not reliable but their finding might be clinically useful. It was possible that high/low dosage was better in treating some groups of people. However, due to the unreliable data, they failed to address how to find the certain group.

Ideally, drug users are expected to reduce the use of heroin by addicting to methadone and then to quit use of methadone. The dosages should consequently have a pattern in which they go up at the beginning of the treatment and later go down. This would indicate detoxification. Physicians think participants with such a dosage pattern and a high attendance rate, most will have a positive outcome. Therefore, we are interested in the participants' behaviour, that is, patterns of daily methadone dosage.

An MMT project was launched by a hospital in Taiwan. The project provided an opportunity to acquire more information about this therapy and offered an opportunity to obtain more insight into MMT by assessing patterns of daily methadone dosage administered to participants. Two types of methadone dosage were recorded by an MMT database system, one being the dosage of their weekly prescriptions prescribed by physicians and the other the daily dosage they had taken recorded by pharmacists. Participants would occasionally have multiple prescriptions but only one record of dosage taken in a single day. Besides, there were occasions on which participants abused drugs and took methadone at the same time, so participants were allowed to take dosage that was lower than the prescribed dosage to avoid overdosing. Of those who continued to abuse heroin while receiving the MMT, there were some fluctuations in their dosage taken records as a result of their demand for daily methadone differing. By and large, following a weekly prescription, a participant took methadone daily for a period of seven days. The prescription records were constant over every 7-day period, whereas the dosage taken records varied over time. Moreover, many participants dropped-off from MMT and returned days later, which resulted in lots of missing data in their dosage taken records. These missing records were not missing at random. Although the numerical daily methadone dosage contains variation and non-random missing values, these records were considered to be a more reliable data. More details of the MMT data can be found in Chapter 2.

The aim of our work is to develop a method to divide participants into groups and then find the differences between the groups. Unfortunately, without the data of whether participants achieve abstinence or not, we do not understand the relationship between treatments and final outcomes. However, by clustering, we can study about the association between dosage patterns and demographic factors, the degrees of addictions, retention of MMT. Also, the dosage patterns provides the possibility of developing a guideline for prescribing a proper methadone dosage.

1.2 Motivation

The initial prescription dosage for the participants who had no experience of methadone was 20 mg. The physicians adjusted the dosage of methadone every seven days after dis-

cussing with participants their preferred prescription dosage settings. 10 mg was often used as a unit of adjustment of prescription dosage. While receiving MMT, participants reported the frequency of their drugs use, from which physicians could measure the effectiveness of the MMT. However, this kind of self-reported data was not reliable and not validated. The physicians of the MMT project observed that abused drugs would reduce the demand of daily methadone. As a result, there were fluctuations or missing values in the dosage taken records. These fluctuations reflected the patterns of drug abuse. There were reasons for which participants abused drugs. One of which was they did not believe in the treatment. Ball and Ross [1991] said participants who remained for more than six months had a marked drop in their drug abuse. However, they found, on average, 11 % (200/1800) of people who commenced MMT after inquiry. Besides, only 38 % of them stayed in the therapy after a year. The physicians of the MMT project in Taiwan suspected that early drop out might be caused by participants having no confidence in methadone.

As aforementioned, drug users are expected to have a dosage pattern which goes up at the beginning of the treatment, followed by a period of stable dosage, and then goes down. The physicians believe that of those with such an up-stable-down pattern and a high attendance rate during the treatment period, most will have a positive outcome. On the other hand, those whose daily methadone dosage fluctuates can be interpreted as lacking motivation.

Our idea starts from identifying patterns. We define a methadone curve by joining all daily dosages with a line. Methadone curves are a remarkable tool to show detoxication. By identifying the patterns of methadone curves, we can correct participants' behaviours of taking dosage to the right track. We mean to convince participants that the dosage is right for them and lead them to have an up-stable-down curve. Since how long participants have been abusing heroin, the degrees of their addictions, the drug abuse history and some unknown factors might all have an influence on the patterns of detoxication. There might be more than one concave curve. Therefore, we attempt to develop a clustering technique that is capable of dividing the MMT participants into subgroups for finding dosage patterns of clusters. These patterns can then

be used as a guideline for determining proper methadone dosages to increase participants' trust in MMT and to reduce the rate of quitting the treatment at an early stage.

The central problems of clustering the participants in our study are the fluctuations of dosages and missing dosages. First of all, some participants who abused heroin while receiving the MMT did not need the full dosages indicated on their prescriptions to accommodate their addictions. In fact, they took a combination of drug and methadone in order for their addictions to be satisfied, so it was not guaranteed that the methadone dosages they took indeed represented detoxication. Secondly, missing dosages were not missing at random. They were recorded as zeroes but the addictions should not be zeros. These zeros appeared as sequences. In some cases, a long sequence of zeros point to more severe problems of the participant, or a tendency to leave the study, or illicit drug use. We take account of these issues and propose to categorize dosages for alleviating the fluctuations of observed dosages and for keeping the sequences of missing dosages. The ranges of observed dosages for categories are based on the recommendations of physicians. Participants whose actual dosage is in the range of 20 mg, that is, dosage between 1 and 20 mg, between 21 and 40 mg, between 41 and 60 mg, between 61 and 80 mg, between 81 and 100 mg, can be considered as the same. We define a new data format. The new data consist of seven categories in which six categories have an ordinal scale for representing dosages and one category for missing dosages. Throughout the study we use the term "category-ordered data" to refer to this new data. The methadone curves will then be represented by sequences of categories. The aim of this study is thus to find clusters in which participants have similar long sequences of categories.

Two issues arising in applied cluster analysis are the selection of the clustering method and the determination of the number of clusters. Among clustering methods, we focus on dissimilarity-based clustering methods because the features of our data make the model based clustering methods hard to applied straightforward (see Section 3.3 for details). We review the Single Linkage, the Complete Linkage, the Average Linkage, the K-Means and the partitioning around medoids (PAM) (Gordon [1999]; Hartigan and Wong [1979]; Kaufman and Rousseeuw [1990]). The Single Linkage defines the dissimilarity of two clusters as the shortest distance between two objects,

while the Complete Linkage defines the dissimilarity as the furthest distance between two objects. The Average Linkage, instead, utilizes the average of all distances of objects of two clusters. As for the K-Means and the PAM clustering method, the former partitions objects into k clusters in which each object is assigned to the cluster with the nearest mean vector, while the latter partitions objects into k clusters in which each object is assigned to the cluster with the closest medoid. Note that k is a positive integer and has to be decided first. The clustering methods group together objects that are considered as similar. The criterion of considering two objects or sets as similar is defined by dissimilarity functions. The dissimilarity functions play the role of connecting the researcher's goals, features of the data and scientific knowledge (Gordon [1990]; Hennig and Hausdorf [2006]). There is a considerable amount of literature on dissimilarity functions. Many attempts are made with respect to study purposes. For example, in the gene research of Luca and Zuccolotto [2011] and the financial research of Douzal-Chouakria et al. [2009], they proposed new dissimilarity functions which adapt the features of their data. To the best of our knowledge, a dissimilarity function for data in which variables have both categorical and ordinal characters has not yet been established. Therefore, we propose a so-called p-dissimilarity. The p-dissimilarity is used to measure dissimilarity between the 180-day time series of the participants. It accommodates categorical and ordinal scales by using a parameter p as a switch between data being treated as categorical and ordinal. It uses a parameter β to tune the dissimilarity involving missing values compared to the distances between non-missing values. Also, it could be applied in wider fields of application, such as survey studies in which questions use choices on the Likert scales and a $don't know$-category.

As for the determination of the number of clusters, it is impossible to prove which index is the best mathematically. Researchers try to use simulation studies to understand the performance of index. Milligan and Cooper [1985] examined 30 indexes and showed that the Calinski and Harabasz index (Calinski and Harabasz [1974]) had the best performance. Arbelaitz et al. [2013] carried out a similar study, which included many indexes that did not exist in 1985. They found that six indexes had better performance. Our cluster analysis is performed based on the p-dissimilarity, so indexes that can cooperate with it will be considered. The Average Silhouette Width (ASW) (Kaufman and Rousseeuw [1990]; Rousseeuw [1987]), which is one of the six indexes, is thus

used in our study. This index measures coherence of clusters. Besides, a more recent index called Prediction Strength (PS) (Tibshirani et al. [2001]) is also used. This index measures stability of each cluster in terms of similarity between clustering results. We focus on these two index, one measuring coherence and the other measuring stability. The PS index can cooperate with the p-dissimilarity, albeit with a modification. The concept of the PS was to view an analysis of clustering as an analysis of classification. At the beginning of the algorithm, a dataset is partitioned into a training set and a test set. Then, for objects in the test set, the algorithm compares their "predicted class" and their "true class". If a hierarchical clustering method is used, the true class is built based on the hierarchical clustering method. However, the predicted class is built on the basis of the K-Means method. This brings an issue of measuring stability of clustering results obtained by hierarchical clustering methods. Therefore, we propose new rules for modifying the PS when the hierarchical clustering methods and the PAM method are used. Also, the ASW and modified PS are used for selecting clustering methods. We let the data to select the clustering method based on cluster stability and coherence.

Moreover, the information of k is rarely previously known, so indexes for determination of the number of clusters produce values for every $k > 1$ on the basis of cluster compactness and separation, and yet only one k will be used. For some indexes, the k which scores the highest value is used, while for other indexes, the first k with a value above a threshold is used and, for other indexes, the k for which there is a gap between its value and that of $(k+1)$ is used. An area of rationale behind decisions of which k to use is not widely understood. We attempt to view the question of determining the number of clusters from a different position. Instead of comparing values for $k = 2, \ldots, K$, for a fixed k, we compare its value to the distribution of the test statistic under the null assumption. The null assumption is that there is no cluster. We propose a null model test to test if the dataset is homogeneous. The null model test involves a null model and parametric bootstrap. The null model fits all non-clustering aspects of the real dataset, such as relationships between variables, time dependency, marginal distributions and etc. The parametric bootstrap is used to draw reference datasets from the null model. These reference datasets are used to construct the distribution of the test statistic. The distribution of the test statistic is used to explore whether the found number of clusters can be explained by random variation. Also, we define a single test of the homogeneity

hypothesis against a clustering alternative by aggregating the test results for different k.

In addition, assessing the quality of clustering results is also of interest. Some research has been conducted on information visualisation via heatplots of row data matrices and of proximity matrices (Chen [2002]; Hahsler and Hornik [2011]; Hahsler, Hornik, and Buchta [2008]; Wu, Tien, and Chen [2010]). A heatplot is a graph that represents data by colours. It consists of horizontal lines, each representing the data for a study object. Its interpretability strongly depends on the location of the objects along the vertical-axis. Although the aforementioned approaches are interesting, those studies tend to focus on preservation of clustering structure. We attempt to use the heatplots to visualise more information on clustering structures, relationships between objects, relationships between clusters, and relationships between an object and a cluster with respect to the dissimilarities, locations of medoids, and the density of clusters. Moreover, we can assess the quality of clustering result by looking at the changes in colours in the heatplot. Also, one can then make statements about whether there really is some clustering which is visible by looking at the border regions of the clusters on the heatplot. Therefore, we propose two algorithms, one using multidimensional scaling (MDS) (Cox and Cox [1990]; Coxon and Davies [1982]) and the other using projection vectors. Both can be used to generate orders of the objects. By orders of the objects, we mean the locations of objects on the vertical-axis on the heatplot. The first algorithm is for general use. The second algorithm is for the PAM method. As the PAM method works on the basis on medoids, we use projections to quantify information in one dimension, so that information about medoids, such as locations of the medoids, the distance between medoids, and the relationships of the dissimilarities among the participants can be viewed on the heatplot. Also, the colour gradient around the medoids indicates the density the clusters. By which, we can see whether a cluster has its objects being scattered or not. Two algorithms are proposed. Both of them can be used for information visualisation with heatplots.

1.3 Outline

This study is divided into 9 chapters. Figure 1.1 shows the association between chapters. Chapter 2 gives a brief overview of Methadone Maintenance Therapy Data and literature reviews of research on the MMT. The data of the daily methadone dosages for 314 participants for a period of 180 days is selected. We propose a so-called category-ordered data in Section 2.4.1. The category-ordered data for 314 participants for 180 days is denoted by CO_{314}, and will be used throughout this study.

In Chapter 3 we review clustering methods and dissimilarity functions.

In Chapter 4 we propose the p-dissimilarity. The p-dissimilarity is used to measure dissimilarity for data whose variables have characters of categorical and ordinal scales. Also, it can be used for incomplete data. The p-dissimilarity is based on the assumption that it is the neighbouring categories which contribute the most to distinguish the target category. The purpose of the assumption is to find clusters whose participants share similar dosage patterns in terms of sequences of categories. The p-dissimilarity involves two parameters. p is a switch between data being treated as categorical and ordinal and β is for measuring dissimilarity when missing values occur.

In Chapter 5 we review indexes for the determination of the number of clusters, namely the Calinski and Harabasz (CH) (Calinski and Harabasz [1974]), the Average Silhouette Width (ASW) (Kaufman and Rousseeuw [1990]; Rousseeuw [1987]) and the Prediction Strength (PS) (Tibshirani et al. [2001]). The ASW and the PS are used in this study. We propose rules to modify the Prediction Strength in Section 5.3.1. Also, in order to avoid confusion, we call the equations for computing the dissimilarity between objects "functions" and those for determining the number of clusters "index".

Chapter 6 begins by selecting the value for β, p and the clustering method. We apply the p-dissimilarity to CO_{314} and compare the Single Linkage, the Complete Linkage, the Average Linkage and the PAM method by their values of the ASW and the modified PS. The values for the PAM method are higher and therefore the PAM method

is selected. In Section 6.2 we propose a null model test involving a null model and parametric bootstrap (Efron and Tibshirani [1993]). The purpose is to investigate whether the clusters found according to the value of the index can be explained by random variation. The process of the null model test is as follows. A null model is constructed to represent a real data. But the null model has no structure of clusters, it is unknown whether there exist clusters in the real data. Then, the null model and boostrap are used to explore the distribution of a statistic such as values of the ASW. Next, the value of the ASW for the real data is compared with the distribution of the values for the null model. In Section 6.3 we show an application of the null model test to CO_{314}. We use relative frequencies to describe the movements from categories to categories over time. Except for the category for missing dosages, we find that the dosage in categories following a valid prescription is stable and the movements between categories in accordance with a weekly prescription. We then construct a Markov null model without structure of clusters, in which the distributions of the categories are the same as those of CO_{314}. Five clusters are selected.

In Chapter 7 we propose two ordering algorithms for the heatplot in order to evaluate the quality of the clustering. The algorithm of MDS is for general use and the algorithm of projection vectors is for the PAM method. We first use MDS to decide the location of clusters on a heatplot in order to preserve clusters. We then apply either the MDS method or projection vector method to order participants within a cluster. Also, the ordering algorithm of projections with the heatplot is used for a visual significance test.

In Chapter 8 we present a sensitivity analysis of clusters and list demographic data of the five clusters. Some conclusions are drawn in Chapter 9. Despite the fact that no significant clustering structure is observed, the sequences of categories for clusters are clinically useful to prescribe a proper dosage to increase the efficiency of methadone maintenance therapy.

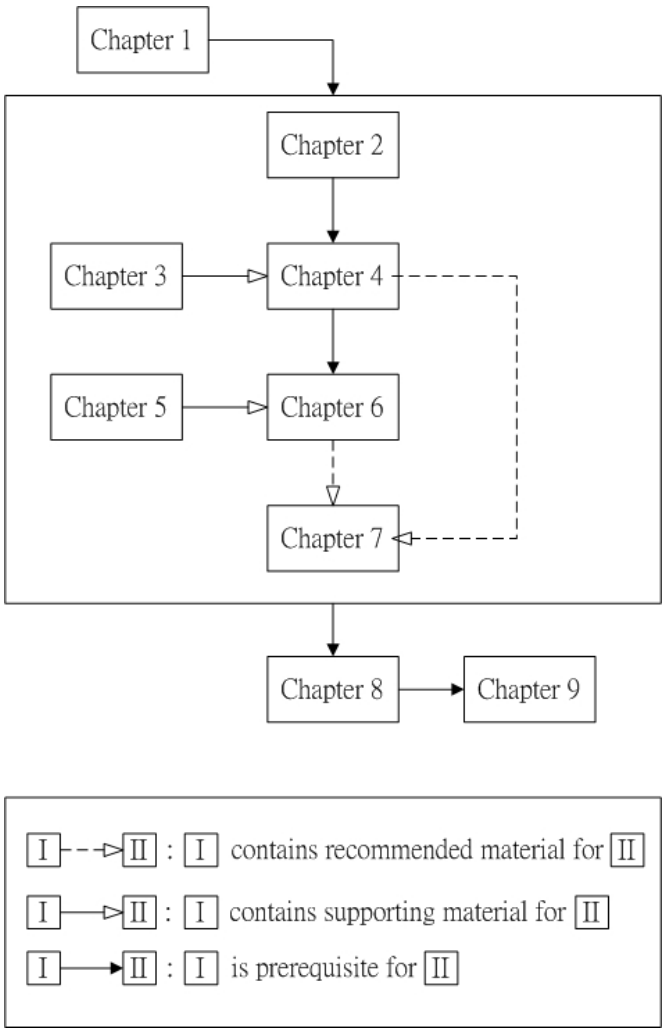


Figure 1.1: Schematic representation of the organization of the contents of the thesis -

Chapter 2

Methadone Maintenance Therapy (MMT)

In this chapter a quick overview of the Methadone Maintenance Therapy (MMT) is given, followed by some research that has been carried out on it. Section 2.2 introduces the MMT database. The initial sample is composed of demographic details and the dataset of records of dosage taken. The study period is set to 180 days. For modelling daily methadone taken by participants for 180 days, a subset of 314 participants is selected from the initial sample (see Section 2.3). In Section 2.4.1 a new data format is created by transforming dosages into categories. The data consist of seven categories in which six categories have an ordinal scale for representing dosages and one category for missing dosages. We call it category-ordered data. The records of dosage taken for the 314 participants for 180 days is denoted by Dosage_{314}. The category-ordered data for the 314 participants for 180 days is denoted by CO_{314}. A heatmap plot is used to have an overview of the datasets. In this study, we perform a clustering analysis on CO_{314}.

2.1 Literature review of methadone

Methadone was developed in 1934 to relieve pain. It was cheaper and less addictive. Later on, it was used as a heroin substitute in a treatment called Methadone Maintenance Therapy (MMT). In Taiwan, MMT was introduced in 2005. The main purpose

of MMT is to minimize the harm associated with heroin use (Ward et al. [1999]). The idea of MMT is to let drug users reduce the use of heroin by addicting to methadone and then quit the use of methadone. The effect of methadone lasts 24 hours and consequently it has to be taken on a daily basis. Ball and Ross [1991] reported, on average, a clinic received 1800 inquiries a year, but fewer than 200 people commenced MMT. In addition, only 38 percent of the 200 stayed in the therapy after a year. Also, participants who remained for more than six months had a marked drop in their drug abuse.

In some studies, the effectiveness of methadone maintenance was measured by mortality rates (Termorshuizen et al. [2005]), number of times illicit drugs (Marsch [1998]; Strain et al. [1993a,b]), frequency of criminal activity (Gossop et al. [2000]; Marsch [1998]; McLellan et al. [1985]; Powers and Anglin [1993]), cost (Masson et al. [2004]) etc. Research showed that methadone did indeed have a positive effect on drug users and on society.

More research has been done on daily methadone dosage taken by participants. Strain et al. [1993a] studied treatment retentions and illicit drugs use. They compared the groups of low to moderate doses of methadone and found that low dose of methadone (≤ 20 mg) may improve retention but were inadequate for suppressing illicit drug use. Langendam et al. [1998] regarded dosage greater than 60 mg as high. They observed that participants requested to stay at lower dosage because of fear of double addiction and of using drugs other than methadone. Bellin et al. [1999] studied associations between criminal activity and methadone dosage. They found drug user on high dose (≥ 60 mg) were less likely to return to jail than those on low dose. Maxwell and Shinderman [2002] compared high dose participants (≥ 100 mg/day, mean 211 mg/day) with control participants (< 100 mg/day, mean 65 mg/day). Their result showed that high dose was more effective in treating heroin addicts. While Maremmani et al. [2003] observed that many heroin addicts had positive outcomes with lower dosages. The basic issue of summarizing their studies is that they have different definitions for high dosage. To date there is no clear principle for the determination of the methadone dosage. However, if there were one, response to methadone could be significantly improved (Maremmani et al. [2003]; Maxwell and Shinderman [2002]).

There are more research on associations between types of methadone programmes and participants' characteristics. Murray et al. [2008] considered methadone dosage and personality disorders. The American Psychiatric Association divided personality disorders into three groups. Cluster B was one of them. It included histrionic, narcissistic, antisocial and borderline personality disorders. Murray et al. [2008] surveyed 54 participants from a methadone maintenance clinic and found that methadone dosage might be uniquely related to the personality pathology. They suggested that methadone dosage might be a response to misery and physicians might need to communicate with heroin addicts with Cluster B pathology for methadone dosage to some extent. Peles et al. [2007] reported the major risk factors for depression were female gender and high dose ($>$ 120 mg). Pud et al. [2012] took account that participants in MMT frequently experienced pain, depression and sleep disorders. They attempted characterize clusters of MMT participants and studied the association between these clusters and quality of life measures. Participants were grouped into three clusters, one of which had highest severity levels of pain, depression and sleep disorders. This cluster scored lowest on all quality of life measures. Also, they reported pain was the most important symptom differentiating MMT patients. Gossop et al. [2000] studied patterns of improvement after receiving MMT for a year. They performed a one year follow-up study on 478 participants. They found that daily methadone dosage of participants who continued to use the drugs during the treatment had a large variation. They used the Euclidean distance K-Means clustering method to group the participants according to their frequency of illicit drug use, including opiates, stimulants and benzodiazepines. Four groups were identified. Two groups showed substantial reductions in their illicit drug use and criminality. They concluded that methadone dosage might be related to a certain group and taking methadone might be of benefit to some groups. This suggested that it might be possible to develop a principle to prescribe the best dosage for certain subgroups whether dosage be high or low, which will help participants have positive outcome.

2.2 MMT database

An MMT project was launched by a hospital in central Taiwan. Due to concerns about confidentiality, the name of the hospital is not given here. The project provided an

opportunity to acquire more information about this therapy and offered the possibility of developing a principle to prescribe dosage by assessing patterns of daily methadone dosage administered to participants. As part of the MMT project, an MMT database was developed to manage the records of its participants. The MMT database system was a system for storing the demographic details, medical history and methadone dosage records of participants. Firstly, the demographic details, which included age, gender, education, etc., were recorded when participants visited the hospital for the first time. Secondly, the medical history, which included the frequency of heroin use, urine drug tests, an HIV test, etc., was recorded when participants re-visited the hospital. However, not all of the participants underwent the urine drug tests and the HIV test. Thirdly, the methadone dosage records, which included records of prescription and records of dosage taken by participants, were recorded in two steps. (1) A participant visited a doctor and received a 7-day prescription. Subsequently, the system generated seven records, one record for each day of the prescription, with respect to the dosages shown on it. At the same time, the system generated seven zeroes for records of dosage taken. (2) In the following seven days, zeros would be changed to the actual dosage taken by participants every single day they visited the hospital.

Three datasets were recorded; however, they were not synchronized. There were cases where nurses accidently forgot to file participants' data when they visited the hospital for the first time. As a result, these participants had no demographic details. There were also cases where participants registered with doctors but failed to see their doctors. Consequently, they had no methadone dosage records. These kinds of mismatch happened quite often when merging several datasets. Another issue of this system was the process of recording daily methadone dosage. Seven zeroes for records of dosage taken were generated in advance. If participants did not go to the clinic to take methadone, their records remained zeroes. Later, we will discuss these zeros from an aspect of addictions should not be treated as zeroes.

2.2.1 Data for prescription and dosage taken

In this section we detail the difference between prescription and dosage taken in terms of restriction and daily record.

Table 2.1: Illustration of the data recording process of a participant over a period of 8 days.- The table illustrates how prescriptions and daily methadone dosage taken of a participant are recorded by the MMT database in response to events over a period of 8 days. The two dosage records are recorded in two steps. (1) Seven records of prescription are generated once a prescription is received. At the same time, seven zeroes for records of dosage taken corresponding to the valid dates of prescription are generated. (2) In the following seven days, the zero records will be changed to the actual dosage taken by the participant. In the last column, RP represents the set of the records of the prescriptions dosage, while RD represents the set of records of dosage taken by this participant.

Day	Event	Response	Database
1	Receive a 7-day prescription of 50 mg Take 40 mg methadone	Generate 7 prescription records Generate 7 zero records Change the 1^{st} zero to 40 mg	RP={50, 50, 50, 50, 50, 50, 50} RD={40, 0, 0, 0, 0, 0, 0}
2	Take 45 mg methadone	Change the 2^{nd} zero to 45 mg	RP={50, 50, 50, 50, 50, 50, 50} RD={40, 45, 0, 0, 0, 0, 0}
3	Fail to take methadone	Remained 0 mg	RP={50, 50, 50, 50, 50, 50, 50} RD={40, 45, 0, 0, 0, 0, 0}
4	Take 50 mg methadone	Change the 4^{th} zero to 50 mg	RP={50, 50, 50, 50, 50, 50, 50} RD={40, 45, 0, 50, 0, 0, 0}
5	Receive a 7-day prescription of 60 mg Take 55 mg methadone	Generate 7 prescription records Generate 7 zero records Change one of the 5^{th} zeros to 55 mg	RP={50, 50, 50, 50, (50,60), (50,60), (50,60), 60} RD={40, 45, 0, 50, (55, 0), (0, 0), (0, 0), 0}
6	Take 55 mg methadone	Change one of the 6^{th} zeros to 55 mg	RP={50, 50, 50, 50, (50,60), (50,60), (50,60), 60} RD={40, 45, 0, 50, (55, 0), (55, 0), (0, 0), 0}
7	Take 60 mg methadone	Change one of the 7^{th} zeros to 60 mg	RP={50, 50, 50, 50, (50,60), (50,60), (50,60), 60} RD={40, 45, 0, 50, (55, 0), (55, 0), (60, 0), 0}
8	Take 45 mg methadone	Change the 8^{th} zero to 45 mg	RP={50, 50, 50, 50, (50,60), (50,60), (50,60), 60} RD={40, 45, 0, 50, (55, 0), (55, 0), (60, 0),45}

There was no restriction on getting prescriptions, but participants were allowed to take methadone once per day. The prescription dosage was the maximum dosage that a participant could take in a day. The initial prescription dosage for the participants who had no experience of methadone was 20 mg. Then, doctors adjusted methadone dosage every seven days according to their subjective judgement and participants' preference of prescription dosage settings. In the MMT project, 10 mg was used as a unit of adjustment of prescription dosage in practice. However, addictions to heroin varied from participant to participant. Some participants might find that their unexpired prescriptions were not high enough to compensate the need for heroin, so they went to their doctors for new prescriptions with higher dosages. As a result, some participants had more than one prescription at the same time. To avoid participants overdosing, they were limited to use at most one prescription a day.

Only one of the multiple prescriptions was used on a day. Unfortunately, the system failed to indicate which one was used. Of these participants who had multiple prescriptions, they had more than one record of prescription but at most one nonzero record of dosage taken a day. In addition, values for those nonzero records varied from day to day. The variation of the values was a result for allowing participants to take a dose lower than what was indicated on their prescriptions. The reasons of this were as follows. Heroin users took MMT because of lack of money for drugs, court orders, determination of quitting drug etc. Some participants abused heroin while receiving the MMT, so they did not need a full prescribed dosage to accommodate their additions. In contrast, some tried to reduce methadone dosage to defeat their addictions. Therefore, each participant had at most one nonzero record of dosage taken a day and those nonzero records varied over time.

Table 2.1 illustrates the recording process of prescription and that of dosage taken of a participant over a period of 8 days since they first joined MMT. The first column indicates the eight days. The second column shows the explanation of events of getting prescriptions and taking dosage, while the third column shows the response of the recording system to the events. The last column shows the records of prescription, denoted by RP, and the records of dosage taken, denoted by RD. At the beginning, the participant in Table 2.1 visits their doctor and receives a prescription of 50 mg. Then,

this participant decides to take 40 mg methadone. Firstly, the system generates seven records of 50 for the prescription records and seven zeroes corresponding to the valid dates of their prescription. Once the participant takes 40 mg methadone, the first zero in DD is changed to 40. On day 2, 45 mg methadone is taken, so the second zero in DD is changed to 45. On day 3, no methadone is taken, so no change is made. On day 4, 50 mg methadone is taken, so the fourth zero is changed to 50. This participant has only one prescription from day 1 to day 4, but has multiple prescriptions from day 5 to day 7. On day 5, the participant visits their doctor for a new prescription before their current prescription expired. The new prescription is 60 mg. Seven records of prescription with a dose of 60 mg and seven zeros of records of dosage taken are generated. As a result, the RP for day 5 is (50, 60). Later on, 55 mg methadone is taken, so the RD for day 5 is (55, 0). On days 6 and 7, 50 mg and 60 mg methadone are taken, respectively. Consequently, the records of dosage taken on day 6 and 7 are (50, 0) and (60, 0), respectively. Note that on days 5 to 7, the database dose not record which prescription is used. Based on limited information, our knowledge of record of dosage taken is that one of the zeros is changed. On day 8, the first prescription expires and 45 mg methadone is taken. The eighth zero is changed to 45.

Participants should have at most one nonzero record. Therefore, in later analysis, of those who had multiple records of dosage taken, the nonzero record would be considered first. For example, the records of dosage taken on day 5 were (0, 55) and it was 55 mg that would be used in the analysis.

2.3 Number of participants

Three datasets were collected from 1st January 2007 to 31st Dec 2008. However, they were not synchronized. The dataset of medical history showed that among those who took a blood test when they re-visited the hospital, 14 % took an HIV test. Also, 4914 urine drug tests were performed, 8 % of drug tests for morphine came positive and 1 % of drug tests for amphetamine came positive. Note that participants underwent more than one urine drug test. Some participants dropped and later returned to the treatment. Their demographic details were re-collected as participants who enter the MMT for their first time in practice. However, in our study, we appended their methadone records to

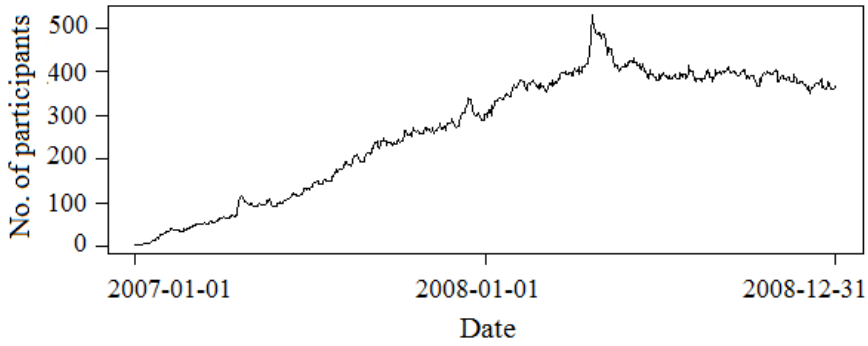


Figure 2.1: The number of participants over 732 days. - The x-axis is the date and the y-axis is the number of participants. This figure shows the number of participants that were found to have a record of dosage taken over the period between 1st January 2007 and 31st Dec 2008.

the existing dataset of dosages. With this dataset, Figure 2.1 shows the numbers of participants over time. As can be seen, at the beginning of the MMT project, there are only few people. Later on, the numbers of participants goes up as the hospital advertised MMT. Physicians considered participants who stayed in MMT more than six months as candidates being able to achieve abstinence. We limited participants to those who commenced MMT from 1st January 2007 to 30th June 2008 in order to ensure that participants should be able to stay in MMT for six months. A total of 1302 participants was selected. Twenty-one of them had only one nonzero record when they stayed in the MMT. They were eliminated in consideration of their non-contribution to form patterns of dosage taken. Taking into account their demographic details, a total of 1257 participants was obtained in which the two datasets could be matched. The initial sample was composed of demographic details and dataset of records of dosage taken.

2.3.1 Records of prescription and dosage taken for the initial sample

The records of prescriptions and records of dosage taken were stored separately. 1252 out of 1257 participants were found in the prescription dataset. A crucial issue of the records of prescription was that codings of prescription dosages were inconsistent. By coding, we meant the values that were used to indicate dosages. For instance, in the prescription dataset, 2 was used to indicate 10 mg, but 2 also referred to 20 mg.

Table 2.2: Frequency of prescription dosage of the 1252 participants.- A total of 21465 prescriptions is selected. The first column refers to the coding in the records of prescription dosages. Unfortunately, we found that the codings referring to dosage are inconsistent. For example, a value 2 refers to 10 mg and it also refers to 20 mg. By and large, most physicians prescribe in intervals of 10 mg and some in intervals of 20 mg.

dosage	frequency	dosage	frequency	dosage	frequency
1	133	30	27976	75	3198
2	49	32.5	21	80	15634
5	841	34	21	85	1807
8	7	35	8879	90	9277
10	4981	37.5	7	95	831
11	14	40	38714	100	9086
13	14	42.5	14	105	387
15	7376	45	6021	110	5282
17	7	50	32045	115	530
17.5	7	55	4185	120	3641
18	14	57	7	125	78
20	27515	60	27114	130	1270
23	8	62.5	7	140	532
25	10411	65	2955	150	219
28	7	70	16467		

This kind of inconsistency happened when creating database without a complete coding book in which a value that is unique to one dosage is listed. Regardless of the inconsistent codings, we attempted to have a better view of prescriptions by showing the frequency of prescribed dosage. This is shown in Table 2.2. Five dosages have their relative frequencies greater than 10%. They are 20 mg (12.20%), 30 mg (10.75%), 40 mg (14.93%), 50 mg (12.41%) and 60 mg (10.14%). Most physicians prescribe in intervals of 10 mg, some in intervals of 20 mg and some 5 mg.

All participants joined MMT on different dates and durations of their staying in MMT were different. Once participants joined MMT, days on which they took

Table 2.3: Number of participants of various attendance rates.-The attendance rates is the proportion of days of receiving methadone therapy to 180 days. Their corresponding numbers of participants are listed in the table.

Attendance rate (%)	Maximum number of missing records	Number of participants
≥ 90	18	169
≥ 80	36	242
≥ 70	54	314
≥ 60	72	359
≥ 50	90	412

methadone became important because the durations and days were the keys to form dosage patterns over time. So we attempted to create a picture to display the durations and days with/without taking methadone. Therefore, we defined a term "initial date", denoted day 1, as referring to the first date on which the participants joined MMT. Figure 2.2 shows the information on durations and days for the 1257 participants. The x-axis represents the day and the y-axis represents the participant. The colour indicates whether participants took methadone. Black refers to nonzero records and white refers to missing dosages. The participants are ordered by the numbers of their nonzero records. Note that if the participants stay in the study, they should be able to provide at least 180 records of dosage taken. A reference line that indicates the 180^{th} day is drawn. As can be seen, some participants have a chunk of missing records followed by nonzero records. This is because their drops out of and returns to MMT. A total of 1257 participants commenced MMT within dates ranging from 1^{st} January 2007 to 30^{th} June 2008, but only some of them provide nearly complete records for 180 days.

2.3.2 Selection of the meaningful sample

Ball and Ross [1991] reported, on average, only 38 percent of the participants stayed in the therapy after a year. A clustering result for the all 1257 participants will definitely give at least one group in which participants have most of their records appearing as missing dosages. Such a group has no contribution to the study at all. Therefore, instead of using 1257 participants, we should perform the analysis on a subset.

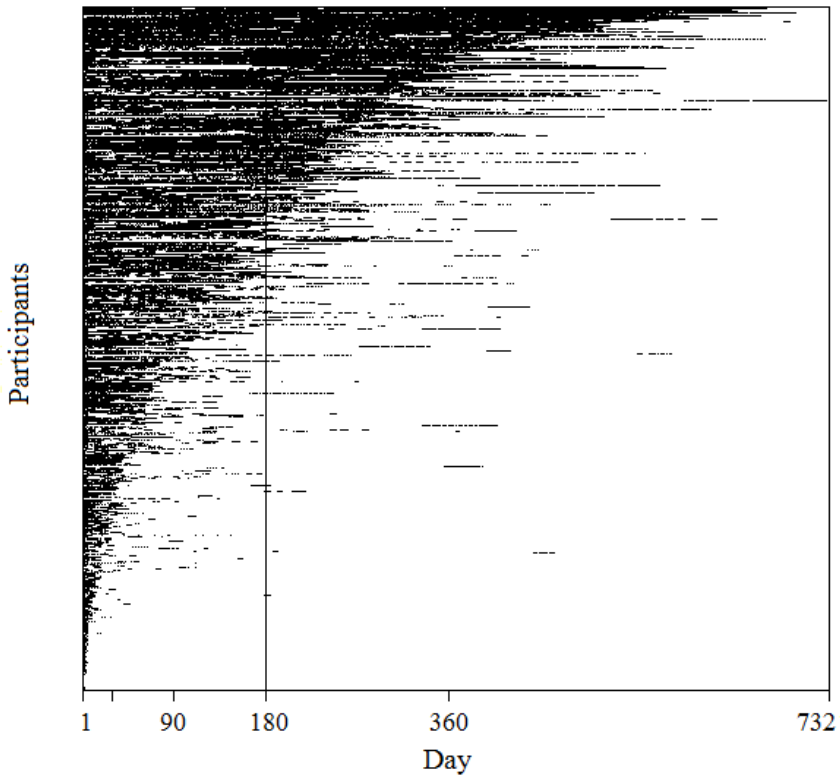


Figure 2.2: Records of dosage taken for the 1257 participants for 732 days. - The x-axis is the day and the y-axis is the participant. The colour indicates where the missing value occurred, appearing in white. The participants are ordered by the numbers of their nonzero records. A vertical line at x=180 is drawn for the reason that the length of studying period was set at 180 days. Note that if the participants did not leave MMT, they should be able to provide at least 180 records of dosage taken.

This subset needs to be considered as a more purposeful sample for modelling daily methadone taken by participants.

The selection of such subset was done based on participants' attendance to the clinic where they took methadone dosage. First of all, as six months was often used in MMT studies (Ball and Ross [1991]; Masson et al. [2004]), the length of the studying period was set to 180 days. Table 2.3 shows the number of missing records out of the 180 days and the number of participants for the attendance rates from 90% to 50%. At an attendance rate of 90 % or more, there are 169 participants who have at most 18 missing records. On the other hand, at a rate of 50 %, there are 412 participants who provide at least 90 records out of 180. The size of the former subset is too small and the proportion of missing records of the latter subset is slightly too high. Both of the two subsets are not good enough. As attendance rate goes down from 90 % to 80 %, an additional 73 participants are recruited. With another decrement in the attendance rate to 70 %, the number of participants goes up to 314. The expected maximum number of missing records is 54. For two participants whose 54 missing records are aligned on different days, their dissimilarity would then depend on 180-54-54=72 records of observed dosages. However, at 60 %, the expected maximum number of missing records is 72. At the same situation, dissimilarity between two participants would then depend on at least 36 records of observed dosages. We assumed that 30% missing records would not affect the cluster analysis too much. Therefore, a total of 314 participants was used. The dataset of the records of dosage taken of the 314 participants over 180 days was denoted by Dosage_{314}.

Of these 314 participants, 262 (83 %) were males and 52 (17 %) were females. Mean age at admission was 37 $\pm$ 7 years (range 23 to 60) and mean age of onset heroin was 25 $\pm$ 6 years (range 13 to 50). One hundred and fifty-six (50 %) participants had attended to high schools or universities. Seventy-seven participants were married or lived with a partner and 234 (75 %) participants were single or divorced. One hundred and ninety-nine (63 %) participants were occupied. Figure 2.3 shows the max, min, mean and mean $\pm$ SD daily dosage records for Dosage_{314} from day 1 to day 180. Mean dosage is 51 mg. As for these unselected 943 participants, mean age of commencing MMT was 36 $\pm$ 8 years (range 19 to 96) and mean age of onset heroin was 25 $\pm$ 7

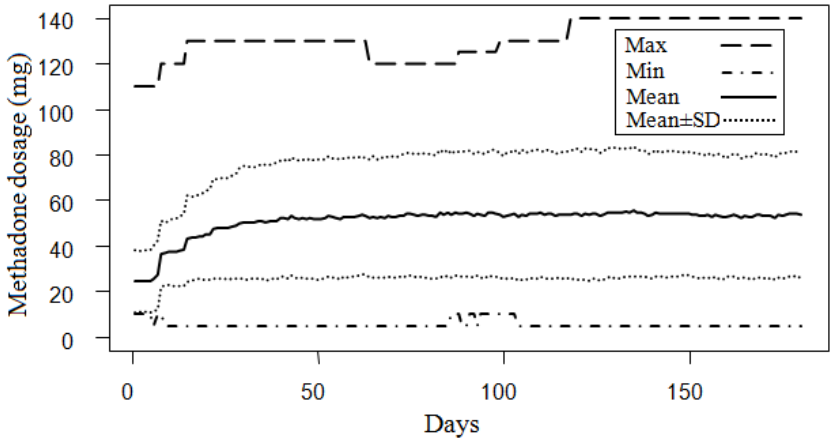


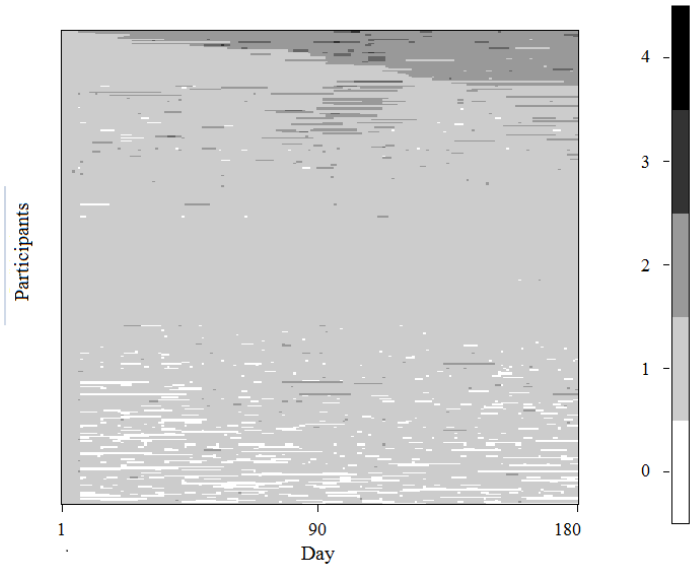
Figure 2.3: The dosage taken records for 314 participants over 180 days.

years (range 11 to 57). Also, 760 (81 %) were males, 428 (45 %) received high school or higher education, 739 (78 %) were single or divorced, 591 (63 %) were occupied.

2.4 A new data format : category-ordered data

In our study, we observed that physicians thought categorically about prescriptions. From their point of view, prescribing a higher dosage meant that participants had their levels of methadone dosage moved from one to another. Such a movement should be, therefore, captured by categories. Moreover, prescription came with physicians' assessments, a zero dosage should not be treated as zero, because participants' addictions were not zero. Valid prescriptions meant that participants needed methadone to accommodate their addictions.

In the following sections, we introduce a so called category-ordered data. The category-ordered data for the 314 participants is denoted by CO_{314}, the most important dataset that is used throughout this study. Then, we introduce imputation methods for the zero records of the category-ordered data. The imputed datasets are used to see the influence of missing values, which is evaluated by comparing the clustering result of CO_{314} and that of imputed datasets.

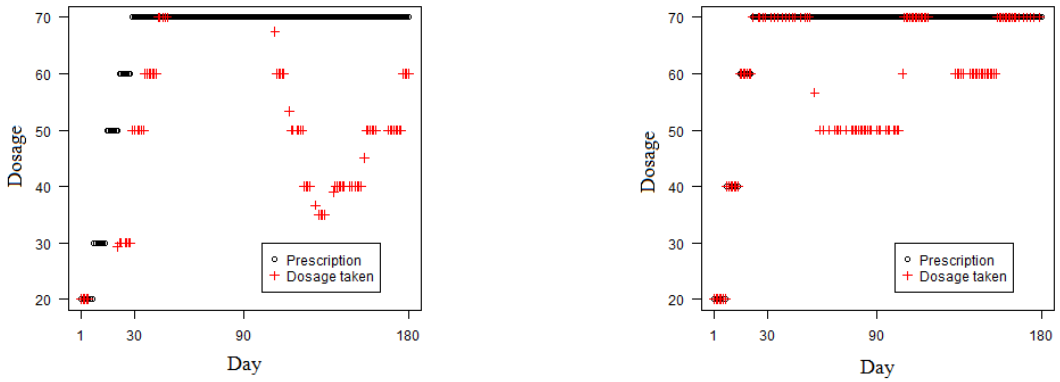


(a)

Figure 2.4: Number of prescriptions for the 313 participants over 180 days. - There are 313 out of 314 participants whose records could be found in the prescription dataset. This figure shows number of prescriptions over 180 days. The x-axis is the day and the y-axis is the participant. The colour represents the number of prescriptions. The max number of prescriptions on a single day for one participant is 4. The participants are ordered by the numbers of their nonzero prescription records.

2.4.1 Category-ordered data: CO_{314}

Of these 314 participants in Dosage_{314}, records of 313 participants are found in the prescription dataset. Figure 2.4 shows the number of prescriptions of the 313 participants over 180 days. The participants are ordered by their total of prescriptions in 180 days. As seen, participants occasionally have multiple prescriptions. In order to display the possible association between prescribed dosage and dosage taken, one participant is randomly selected from those who have only one prescription on every single day. Figure 2.5(a) shows the records of prescriptions and dosage taken of that participant from day 1 to day 180. The record of prescription dosage is indicated by circles, while record of dosage taken is indicated by crosses. The prescription dosage starts from an initial level of 20 mg/day, then there is an upward trend; moreover, it is a constant from day 29 to day 180 with a dosage of 70 mg. However, the records



(a) (b)

Figure 2.5: Records of prescriptions and dosage taken for two selected participants from day 1 to day 180. - (a) shows a participant whose records of dosage taken fluctuates during the period of receiving prescriptions of 70 mg. (b) shows the records of another participant who also receives prescriptions of 70 mg; however, the records of dosage taken of this participant are more stable in comparison to those of the participant in (a).

of dosage taken fluctuate. This might be explained by the participant abusing drugs while receiving MMT. Next, in order to compare the records of dosage taken for two participants, a participant with a long sequence of prescriptions of 70 mg is selected. Figure 2.5(b) shows records of this participant. Following the same prescriptions of 70 mg, the records of dosage taken of these two participants are different.

There are two problems with Dosage_{314}. Firstly, using methadone dosages to quantify additions, some degrees of dosage fluctuation are not meaningful. Failure to account for fluctuations which are caused by abusing illicit drugs might result in identifying false detoxification patterns. Therefore, given the same prescription with various dosage taken, participants should be considered as similar. However, there are participants with multiple prescriptions and there is no indication of which prescriptions were used. If no drugs are abused, methadone taken by participants should show long sequences of stability. Secondly, zero dosages need to be taken into account. Zero dosages mean participants did not show up for receiving methadone but not participants had no addictions. Zero dosages mean participants' dosage taken records are missing. It is

reasonable to impute these missing dosages. However, from a medical point of view, a continuous 14 records of zero indicates that participants have left the study. Of those participants who are considered as having left the study temporarily, it is reasonable to impute their records by using the observed dosage. In contrast, of these who are considered as having left the study for good, the records should remain as they are. So, some missing dosages remain after imputation. But again, these records should not be treated as zeros. Also, the dissimilarity between a missing dosage and an observed dosage is not defined in most dissimilarity functions.

We attempted to construct a new data format by categorizing daily methadone dosage. The new data was used as a solution for the aforementioned issues. The purpose of categorizing dosage was to alleviate the impact caused by fluctuation, to consider participants with same prescriptions as similar, and to keep missing dosages. The missing dosages often occurred as a sequence. The pattern of missing dosages was of interest. In the new data, a participant who regularly took dosage in an interval was considered as having stable dosage. In contrast, changing to another interval regardless how far the movement meant that their dosage was moved to another dosage level. The advantages were as follows. The impaction of the fluctuations were minimized. It was not guaranteed that of these participants who used the same prescriptions, all were then regarded as similar; but, at least, most were considered as similar. Moreover, the missing dosages were categorized. By categorizing, we could define dissimilarity of missing dosages for distinguishing the missing dosages and observed dosages. Then, a dissimilarity-based clustering could be performed.

In order to categorize dosages, we needed cut points of dosages. Here are how the dosage levels were used in the research about methadone. In the research of Mattick et al. [2009], doses between 20 and 35 mg were classified as low dose, between 50 and 80 mg as medium dose and 120 mg or more as high dose. Johnson et al. [2000] used high dose (60 to 100 mg) and low dose (20 mg). In the research of D'Aunno and Pollack [2002], patients were classified into three groups by methadone dosage, less than 40, 60, and 80 mg. Although doses were often increased in 10 mg increments in MMT, the doses of 20, 60, 100 mg were most likely to be used to classify dosage levels in the

literature.

In our study, cut points for categorizing dosages were defined by the physician. He suggested that dosage in the range of 20 mg could be considered virtually the same. This meant that the qualitative difference between two dosages in the same interval could be treated as irrelevant. Therefore, observed dosages were transformed into several ordered sets, being intervals with the width of 20 mg, and the missing dosages were categorized, being represented by a category. The new data consisted of seven categories in which six categories had an ordinal scale for representing dosages and one category for missing dosages. The six categories represented for dosages smaller than or equal to 20 mg, 21-40 mg, 41-60 mg, 61-80 mg, 81-100 mg and greater than 100 mg. The new data was called "category-ordered data" throughout this study. The new dataset, denoted by CO_{314}, was transformed from $Dosage_{314}$.

To explore the uncategorized dataset $Dosage_{314}$ and the categorized dataset CO_{314}, a heatplot is used. It is a technique to represent data by color and each horizontal line in a heatplot represents data of each participant. However, locations of participants determine the efficiency of the heatplot in terms of displaying data with respect to a purpose. Research on ordering participants for increasing the efficiency of heatplots is carried out in Chapter 7 in which we use heatplots for viewing data and evaluating clustering results. Figure 2.6 shows the heatplot of $Dosage_{314}$ and that of CO_{314}. Each horizontal line represents records of a participant from day 1 to day 180. In the graph on the left, the 314 participants are ordered by the average of their dosages. In the graph on the right, the order of participants mirrors that on the left. Figure 2.6(b) shows the colour spectrum of dosage and that of category. The values of dosage, ranging from 1 to 140, appear in a sequence of green, black and red. The values of category, ranging from 1 to 6, appear in black, red, green, blue, cyan and purple. Note that the colour white represents for missing values and category 7. What can be observed is that most of dosage records in the first week are in category 1, as the initial prescription dosage for participants, most of which have no previous experience of the MMT, is 20 mg. Subsequently, the colours of dosage start to change, reflecting the fact that the doctors started adjusting the dosage. In the figure on the right, about one-third of the participants shows dosage below 40 mg, one-third shows dosage from 41 to 60 mg and

one-third shows dosage greater than 61 mg. Among those with dosage higher than 61 mg, only few of them takes dosage more than 100 mg. Besides, it can be seen that, of these movements from categories to categories, most of them move to the next nearest categories.

2.4.2 Imputation of the category-ordered data

In this section we attempt to impute the missing dosages. Three datasets are generated but it is CO_{314} being used throughout this study. We perform imputation because the missing values were all treated the same, being categorized to one category, but we are curious about the influence of the missing values on clustering results. In order to see how much difference it makes, we attempt to construct datasets with imputation. Then, they can be used to see the influence of having treated the missing values the same in CO_{314}. The effect is evaluated by comparing the clustering result of CO_{314} and that of imputed dataset. Results for the comparison are shown in Section 8.2.

The length of the sequences of category 7, which represented missing values, determined whether participants temporarily left the study or not. A sequence with length less than 14 means that the participant temporarily left the study. Because they temporarily left, we attempted to construct a dataset with imputation only on these sequences of category 7 with length less than 14. Denote the value of category on day i by c_i, where $c_i \in \Theta = \{1, 2, \ldots, 7\}$. The imputation method works as follows.

1. identifying the days having category 7.
2. labeling the each long sequence of the category 7.
3. distinguishing sequences to which imputation might apply. Sequences should have length less than 14.
4. identifying the closest known values of category of each of the sequences. The imputation will only be applied to the sequences in which its closest known values of category are the same.

In Step 2, let $\psi = \{s_1, s_2, \ldots, s_a\}$, $a < 180$, be a collection of these sequences where $s_i \cap s_j = \phi$, for all $i \neq j$. Let $n_1, n_2, \ldots, n_a$ be the length of these sequences. By length,

is meant the number of category 7 in the sequence.

In Step 3, for those sequences with length greater than or equal to 14, because participants are considered as having left the study in days on which the sequence occurred, there is no way to assign the category to which they belong. Therefore, these records remain category 7. Denote the selected sequences for which the lengthes smaller than 14 by $s_j = \{c_i, \ldots, c_{i+n_j-1}\}$.

In Step 4, for each sequence found in Step 3, the two closest known values are the one before and the one after, that is, c_{i-1} and c_{i+n_j}. For the sequences of which c_{i-1} is equal to c_{i+n_j}, the records are replaced by the value c_{i-1}. In contrast, for the remaining sequences, because there is no clear decision about whether the category 7 should be replaced by a value close to c_{i-1} or a value close to c_{i+n_j}, we decide to let the records remain category 7.

Here is an example of the aforementioned imputation method. Figure 2.7 shows records of a participant from day 1 to day 8. The y-axis is the values of category. As can be seen, there are several records of category 7. Also, there exist long sequences of the category 7. After applying the imputation method, Figure 2.8 shows records of this participant. ImpCO_{314} was created from CO_{314} by using the imputation method for categories. The heatplot for ImpCO_{314} is shown in Figure 2.9(a). Although 14 days were used in practice, we were interested in the situation of using a more strict criterion. ImpCO^7_{314} was then created from CO_{314} in which participants continuously lacking 7 days records were considered as having left the study. So sequences of category 7 with length greater than 7 were not imputed. The heatplot for ImpCO^7_{314} is shown in Figure 2.9(b).

Moreover, an imputation method for dosage was used in order for applying clustering analysis on Dosage_{314}. The imputed dataset could be then used to evaluate the difference between clustering result for CO_{314}. The idea was to impute dataset of Dosage_{314} in which missing dosages for each participant were replaced by a linear interpolation. The imputed dosages were within the range of their closest known

dosage record (as shown in Figure 2.10). All missing dosages in Dosage_{314} were imputed. Denote the imputed dataset by ImpDosage_{314}. Figure 2.11 shows the heatplot for ImpDosage_{314}.

From Figure 2.8, we observe that some sequences with length shorter than 14 days were not imputed. Given that the participant had temporarily left the study, it was reasonable to consider the missing records being similar to the two closest known records. One solution to improve the imputation method was to impute the category 7 with the value by transforming the dosage in ImpDosage_{314} into categories. But we decided to use the ImpCO_{314}, ImpCO^{7}_{314} and ImpDosage_{314}.

Daily dosages in mg for 314 participants who received MMT between 01 January 2007 and 31 December 2008 were collected. These participants were selected from a larger study using the criterion that they had not left the study before the completion of 180 days and that they had at least 70% nonzero records of taking methadone. Dosages in mg were converted for better interpretability to seven categories in which six categories have an ordinal scale for representing dosages and one category for missing dosages. The resulting dataset was called "category-ordered data", denoted by CO_{314}. Also, few schemes for imputation were defined based on the non-missing values surrounding the missing days, and on length of periods of missingness. In the next chapter we review dissimilarity function with respect to types of data, including categorical scale, ordinal scale, etc., and dissimilarity-based clustering methods. In Chapter 8 we use these imputed datasets to study about the influence of the missing values on clustering results.

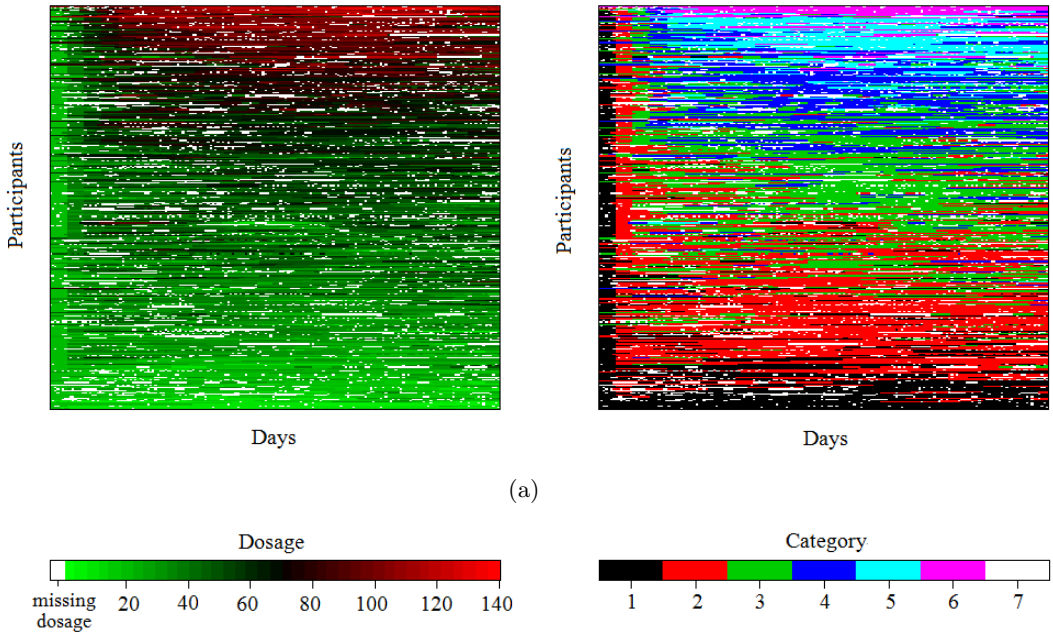


Figure 2.6: Heatplot of Dosage_{314} and heatplot of CO_{314} - (a) shows the heatplot of Dosage_{314} and the heatplot of CO_{314}. Each horizontal line represents records of a participant from day 1 to day 180. The 314 participants are ordered by the average of their dosages. (b) shows the colour spectrum of dosage, ranging from 1 to 140 mg, and that of category, ranging from 1 to 6. Note that the colour white represents for missing values in both two colour spectrums.

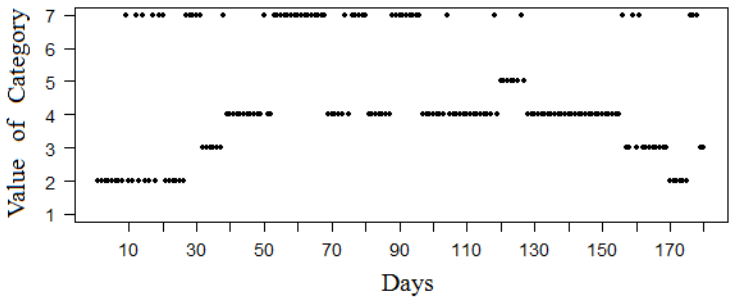


Figure 2.7: Illustration of imputation: original data. - The values of category of a participant from day 1 to day 180.

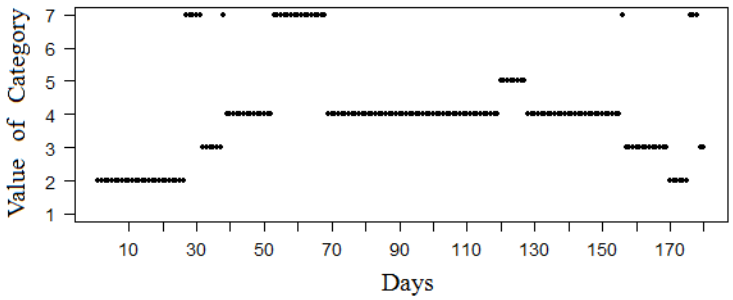
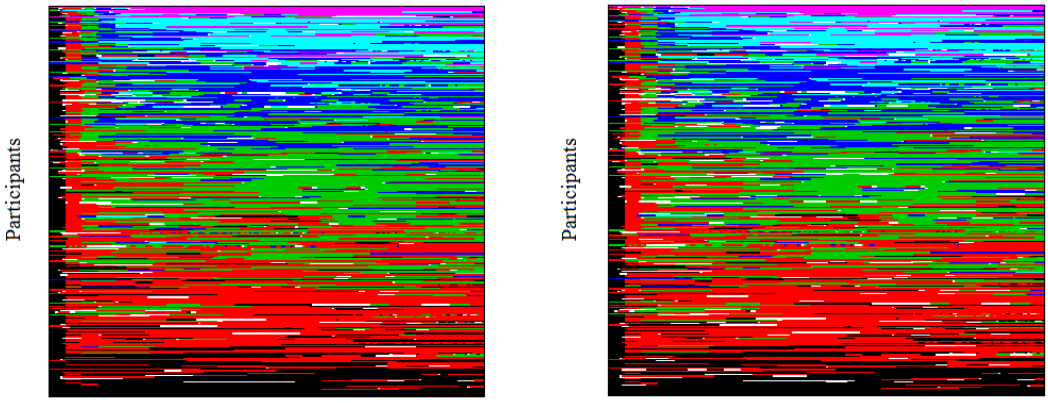


Figure 2.8: Illustration of imputation: imputed data in which among the records of category 7, some of them are imputed. - The imputation method is applied to the data of the participant whose original record is shown in Figure 2.7. This figure shows the record with imputation of this participant.



(a) (b)

Figure 2.9: Heatplot of $\mathbf{ImpCO}_{314}$ and heatplot of $\mathbf{ImpCO}^{7}_{314}$ - (a) shows the heatplot of ImpCO_{314} in which sequences of category 7 with length greater than 14 were not imputed. (b) shows the heatplot of ImpCO^{7}_{314} in which sequences of category 7 with length greater than 7 were not imputed.

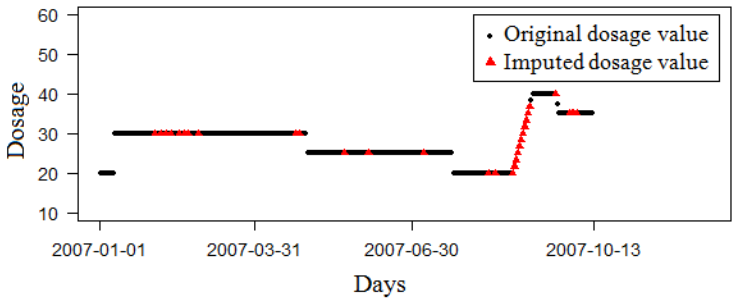


Figure 2.10: Illustration of imputation for missing records - The missing records were replaced by a linear interpolation.

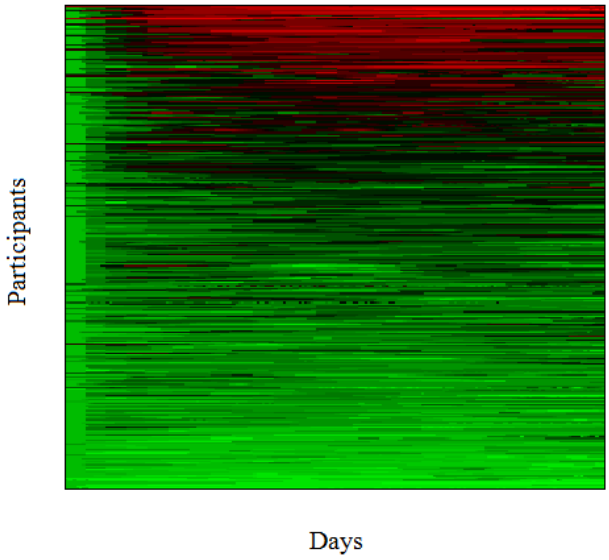


Figure 2.11: Heatplot of ImpDosage$_{314}$ - ImpDosage$_{314}$ was created from Dosage$_{314}$ by the linear interpolation.

Chapter 3

Dissimilarity functions and clustering methods

In this chapter we review dissimilarity functions and clustering methods. A model based clustering is also reviewed. The dissimilarity functions are used for nominal scale and ordinal scale. The clustering methods include the Single Linkage, Complete Linkage, Average Linkage, K-Means and partition around mediods (PAM).

3.1 Dissimilarity functions

The purpose of this study is to cluster the MMT participants using the category-ordered data. Performing a cluster analysis on a dataset with no cluster information other than the observed values is called unsupervised classification (clustering). In contrast, an analysis of a dataset which consists of groups to which objects of study belong is called supervised classification. Note that our study is a case of clustering.

Clustering methods aim to group together objects so that objects within a cluster are considered to be similar. The degree of similarity and dissimilarity between objects is measured by dissimilarity functions. Following are two of the most widely used dissimilarity functions, namely the Euclidean distance and the Manhattan distance.

Notation

Denote the observed value of the t^{th} variable of the object i by x_{it}, $t = 1, \ldots, T$. There-

fore, the data for each object over T variables can be represented by a T-dimensional vector $\mathbf{x}_i = [x_{i1}, \ldots, x_{iT}]$. Denote the dissimilarity between variables by $d(\cdot, \cdot)$ and the dissimilarity between objects by $D(\cdot, \cdot)$.

The Euclidean distance between two objects i and i' is defined by

$$D(i, i') = d(\mathbf{x}_i, \mathbf{x}_{i'}) = \sqrt{\Sigma_{j=1}^{T}(x_{ij} - x_{i'j})^2}. \tag{3.1}$$

The Manhattan distance between two objects i and i' is defined by

$$D(i, i') = d(\mathbf{x}_i, \mathbf{x}_{i'}) = \Sigma_{j=1}^{T}|x_{ij} - x_{i'j}|. \tag{3.2}$$

There are more dissimilarity functions with respect to different types of data. Regarding the type of data, Stevens [1946] defined several ones called scale types of measurements according to the arithmetic operations and meaning of measurements. The category-ordered data has features of categorical and ordinal scales. Therefore, we first review dissimilarity functions that can be applied to these two scales. Then, we review dissimilarity functions for data with a time series structure.

Nominal scale

A categorical variable, which consists of θ, $\theta \in \mathbb{N}^+$, categories, has its measurement type classified as nominal scale. The categorical variable has meanings for each of the categories. These categories can be represented by values of 0, 1, 2, ..., θ. However, these values can not be used for arithmetic operations. They are rather symbols than real numbers. Two different values can be regarded either as equal or as different. An example of this is dosage that is categorized into observed dosage and missing dosage. We can denote these two sets by any two different values, such as, by 1 and 2; by 4 and 7; by 6 and 3. The value 1 in the former is equal to the values 4 and 6 in the latter. The values/symbols are used to preserve the information of dosage but they do not carry any numerical meanings.

For this type of scale, the Simple Matching Coefficient is the simplest way of computing the dissimilarity between two objects. The concept of the Simple Matching

Coefficient is to aggregate all matched pairs of objects i and i'. The term "matched pair" means that the t^{th} variable of the object i and that of the object i' are the same. The dissimilarity between the two objects i and i' is defined by (Sokal and Michener 1958)

$$D(i, i') = 1 - \frac{\{\text{number of the matched pairs}\}}{\{\text{total number of variables}\}}. \tag{3.3}$$

For binary variables, which take values of "+" and "-". The Jaccard's coefficient between two objects i and i' is defined by

$$D(i, i') = 1 - \frac{\{\text{number of the matched pairs with "+"}\}}{\{\text{total number of variables - number of the matched pairs with "-"}\}}. \tag{3.4}$$

The difference between these two dissimilarity functions is that the Simple Matching Coefficient takes all matched pairs, while the Jaccard's coefficient takes a part of the matched pairs. In both coefficients, the larger the value is, the higher the degree of dissimilarity between objects.

Ordinal scale

An ordinal variable, which consists of θ, $\theta \in \mathbb{N}^+$, categories, has its measurement classified as ordinal scale by Stevens [1946]. The ordinal variable has meanings for each of the categories. Also, the values of an ordinal variable are more than just symbols. The values carry information about order by which they are comparable. In other words, two different values can be considered to be either equal or one is smaller than the other. An example of this is dosage levels: low dosage, mild dosage and high dosage. These three sets can be denoted by three ordered numbers, such as, 1, 2 and 3; 4, 6 and 10; 7, 8 and 9. The values indicate the dosage levels.

For this type of scale, a method of computing the dissimilarity between objects shown by Gordon [1999, p.20] and Gower [1971] is as follows. Firstly, the θ values of the ordinal variable are re-coded by $(\theta - 1)$ binary variables (Sneath and Sokal, 1973). Afterwards, the Simple Matching Coefficient (Eq 3.3) is applied to the binary data for

computing the dissimilarity.

Here is an illustration of how to compute the dissimilarity between objects by transforming the ordinal variable to binary variables. Following the example of the three dosage levels of low dosage, mild dosage and high dosage, indicated by 1, 2 and 3, the codings of the dosage levels by two binary variables, L1 and L2, are as below:

$$\begin{array}{cc} & \text{Binary variable} \\ & \begin{array}{ccc} & \text{L1} & \text{L2} \end{array} \\ \text{Ordinal variable} \left\{ \begin{array}{ccc} 1 & - & - \\ 2 & + & - \\ 3 & + & + \end{array} \right. \end{array}$$

When using the values 1,2,3 for dosage levels, the Simple Matching Coefficient shows that any two dosage levels are completely different, that is, $d(x, y) = 1, x = 1, 2, 3; y = 1, 2, 3$, and $x \neq y$. On the other hand, when taking into account of ordinality of the values, the Simple Matching Coefficient shows that the low dosage and mild dosage are slightly similar, that is, d(low dosage, mild dosage)$=1 - \frac{1}{2} = 0.5$. The method of recording keeps the ordinality and shows that one value is closer to another in comparison to the others. However, the numerical difference in the ordinal variable can not be interpreted. For example, The dissimilarity between low dosage and high dosage can not be interpreted as double of that between low dosage and mild dosage.

Another method to compute the dissimilarity for ordinal variables works as follows. Let F be a random variable and f be the observed values. Note that the values carry information about order. Step 1 is to assign ranks to f, denoted by r. Denote the highest rank value by M. Step 2 is to transform the observed values by mapping the ranks onto $[0, 1]$. It can be written as (Kaufman and Rousseeuw [1990])

$$x_f = \frac{r_f - 1}{M_f - 1}, \tag{3.5}$$

Next, the dissimilarity between two objects is computed by applying the usual formulas, such as, the Euclidean distance and the Manhattan distance, to x_f.

The ranks refer to relative positions, which are determined by quantities. In our case,

the width for categories $(i-1)$, i, and $(i+1)$ are the same. However, using the above method, the Manhattan distances from category i to categories $(i-1)$ and $(i+1)$ are different.

Pattern recognition methods

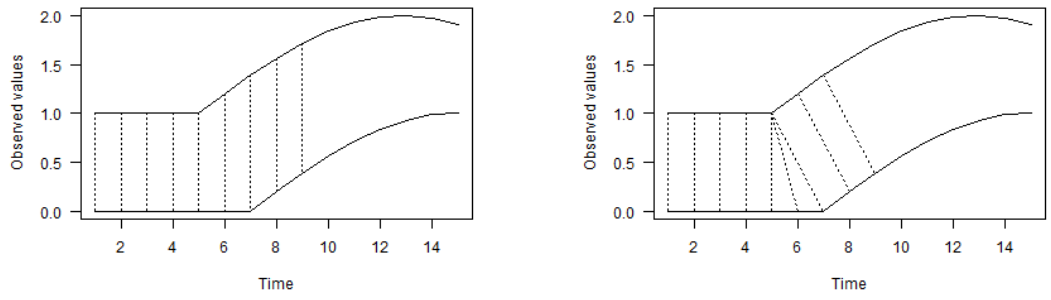
Levenshtein distance (LD)

A string variable is a sequence of characters. **Levenshtein distance** (devised by V. Levenshtein in 1965)is used to measure the difference between two sequences. The concept of the Levenshtein distance is to compute how many steps of insertions, deletions, and substitutions are needed to transform a string into the target string. For example, "extinct" and "instinct". The Levenshtein distance between these two strings is 3, since two substitution (change "e" to "i" and change "x" to "n") and one insertion (add "s" to the third place) are required to transform the string "extinct" into the target string "instinct".

Dynamic Time Warping (DTW)

DTW (Berndt and Clifford [1994]; Ratanamahatana and Keogh [2004]) is a pattern matching technique. It computes dissimilarity according to time series patterns. It seems relevant to the purpose of this study and might be useful to our research. Below is its concept and equation. DTW employs the Euclidean distance to compute the dissimilarity between two sequences. But it allows that the t^{th} point in one sequence not to align with the t^{th} point in the other sequence in order to match shapes of two sequences along the time axis.

Figure 3.1 illustrates the concept of the DTW. Let vectors $\mathbf{x} = \{x_i : i = 1, \ldots, t_x\}$ and $\mathbf{y} = \{y_j : j = 1, \ldots, t_y\}$ be the records of two objects, which are represented by two solid lines. As can be seen, the shape of the two solid lines are the same. The Euclidean distance between these two objects is the square root of the sum of the square difference between the paired t^{th} points as shown in Figure 3.1(a). On the other hand, the DTW between two objects is the square root of the sum of the square distances between the non-aligned points as shown in Figure 3.1(b). For instance, in Figure 3.1, the Euclidean distance is $\sqrt{\sum_{t=1}^{T}(x_t - y_t)^2}$, and the DTW distance is $\sqrt{(x_1-y_1)^2 + (x_2-y_2) + \cdots + (x_5-y_5) + (x_5-y_6) + (x_5-y_7) + \ldots}$. The



(a) (b)

Figure 3.1: Illustration the Euclidean distance and the DTW - The two solid lines in (a) and (b) represent the record of two objects over time. (a) illustrates the Euclidean distance between the two objects that is the sum of the distances between the first points, the the second points and so on. On the other hand, (b) illustrates the DTW between the two objects that is computed by summing up the distances between either aligned or non-aligned points of the two objects in the time axis.

fifth point of the vector $\mathbf{x}$, x_5, corresponds to multiple points of $\mathbf{y}$, y_5, y_6 and y_7. The DTW takes into account the matter of shapes and computes the dissimilarity based on the matched portions with respect to the relative time.

The dissimilarity between two objects by the DTW is defined by

$$DTW(\mathbf{x}, \mathbf{y}) = \min_{\mathbf{w}} \sqrt{\sum_{t=1}^{T} (x_{w_t} - y_{w_t})^2}. \tag{3.6}$$

The $\mathbf{w} = \{w_i : i = 1, \ldots, T\}$, $\max(t_x, t_y) \leq T < (t_x + t_y + 1)$ is called "path" which is used to indicate the points that are used to compute the dissimilarity between objects in order to minimize the total cumulative distance. As an example, the "path" is represented by the dotted lines in Figure 3.1. Here is the explanation of how to find the path of DTW. Imaging a 2-dimensional space with x-axis, ranging from 1 to t_x, and y-axis, ranging from 1 to t_y, the DTW computes the distance that would be traveled from $w_1 = (1, 1)$ to $w_T = (t_x, t_y)$. Note the 2-dimensional space, the next travel point from point (i, j) can only increase by 0 or 1 on each step along the grid-like path. Therefore, the cumulative distance at the first step $w_1 = (1, 1)$ is $d(x_1, y_1) = (x_1 - y_1)^2$. Next,

three distances are compared, namely, $d(x_1, y_2)$, $d(x_2, y_1)$, and $d(x_2, y_2)$. If $d(x_1, y_2)$ is the smallest, the second step will be $w_2 = (1, 2)$ whereby the cumulative distance up to the second step will be $(x_1 - y_1)^2 + (x_1 - y_2)^2$. Moreover, it can be seen that the first element of the vector $\mathbf{x}$, x_1, corresponds to multiple elements of $\mathbf{y}$, y_1 and y_2. The algorithm goes on until a warping path $\mathbf{w} = \{w_i : i = 1, \ldots, T\}$ is constructed. Furthermore, the Euclidean distance between two sequences with the same length can be regarded as a special case of the DTW in which $\mathbf{w}$ shows the path along the diagonal line.

The DTW allows similar shapes to be matched and is widely used in science; however, it does not obey the triangle inequality (Ratanamahatana and Keogh [2004]). The triangle inequality of a distance function states that a triangle is constructed by three objects (h, i, j) where the length of the three sides of the triangle are the dissimilarities among objects (h, i, j); in addition, the sum of any two sides of the triangle must be greater than the remaining side. In other words, $D(i, h) + D(i, j) \geq D(h, j)$ hold for all triples (h, i, j) where $D(\cdot, \cdot)$ is the dissimilarity between objects.

A dissimilarity function is said to be metric if it satisfies the triangle inequality. Gower and Legendre [1986] said that "metric methods often have a geometric rationale that implies that a metric and possibly a Euclidean coefficient should be chosen, thus disfavoring non-metric coefficients". In their paper, they investigated the metricity for several of well-known similarity coefficients for binary variables and dissimilarities for quantitative variables. The result showed that some of them were not metric. A dissimilarity function should satisfy three requirements: non-negativity in which the dissimilarity between objects is always greater or equal to zero; identity in which the dissimilarity between a object and itself is zero; symmetry in which the dissimilarity from object i to i' is equal to that from object i' to i. Note that the triangle inequality is not required for dissimilarity functions Luca and Zuccolotto [2011]. In this thesis, we call dissimilarity functions that do not satisfy triangle inequality "dissimilarity functions" and those that satisfy the triangle inequality "distance functions".

There are many dissimilarity functions for time series. In financial analysis, Luca and Zuccolotto [2011] proposed a dissimilarity function base on tail dependence coef-

ficients. The purpose of their study was to group time series data with an association between extremely low values. Douzal-Chouakria and Nagabhushan [2007] proposed a dissimilarity function based on values and behaviour over time. Also, an application on gene analysis can be found in the study of Douzal-Chouakria, Diallob, and Giroudb [2009]. In this paper, they suggested using the partitioning around medoids (PAM) clustering method with the proposed dissimilarity function for identifying expressed genes of a specific cell's function.

3.2 Hierarchical clustering and partitioning methods

In this Section we focus on dissimilarity-based clustering methods. The dissimilarity-based clustering methods employ the dissimilarity functions to measure dissimilarity between two objects, two clusters or one object and one cluster. Subsequently, the clustering methods group objects into clusters in which objects within clusters are similar. In the following sections we review the hierarchical and partitioning clustering methods and show how the clustering methods work on the basis of the dissimilarity functions.

One of the approaches in the hierarchical clustering methods is the "bottom-up" approach that works as follows. At the beginning, each object is treated as one cluster. Then, the dissimilarities between any two clusters are calculated. Later, clusters with the smallest dissimilarity are linked to each other as one cluster, which contains two objects at this moment. Afterwards, the method repeats the calculation of dissimilarity between clusters with a selected linkage method, described in the following section. Again, two clusters with the smallest dissimilarity are merged. The process of agglomeration goes on until all clusters are merged together as one cluster. Each time of merging two clusters indicates one step up the hierarchy. Note that two merged clusters can not be separated at the next step of hierarchy. As a result, the process of agglomeration of each step can be displayed by a dendrogram from which clusters are obtained. Figure 3.2 shows a dendrogram of the clustering process. The vertical-axis represents the dissimilarity. On horizontal-axis, the bottom end of each node indicates objects. At the beginning of clustering, each object is considered as one cluster.

Next, the dissimilarity between objects 1 and 2 is the shortest one among those between any two clusters, so objects 1 and 2 are merged as one cluster, denoted by G_1. The process stops when all objects are merged together. In this study the hierarchical clustering methods, namely the Single Linkage, the Complete Linkage and the Average Linkage method, are used and described in Section 3.2.1. On the other hand, in the partition methods, the number of clusters, denoted by k, has to be decided first, and only then will the algorithms partition objects into specified number of clusters in which each object will be assigned to the cluster with the nearest centroid. Two of the partition clustering methods, namely the K-Means clustering method and the partition around medoids (PAM) clustering method are introduced in Section 3.2.2. The K-Means method aims at partitioning objects into k clusters in which each object is assigned to the cluster with the nearest mean vector. The elements of a mean vector are the average of each variable over the objects of a group. The PAM method aims at partitioning objects into k clusters in which each object is assigned to the cluster with the closest medoid. Selecting one object from each of the k clusters, the selected objects are called the k medoids. They are the representative objects of the clusters.

Notation

Let x_{it}, $t = 1, \ldots, T$, be the value of the t^{th} variable for the object i. Hence, the data for T variables for each object can be represented by a T-dimensional vector $\mathbf{x}_i = [x_{i1}, \ldots, x_{iT}]$. Let X be the set of data for n objects. Let n be the total of objects and k the number of clusters. Assume that the n objects are clustered into k clusters, $(k \leq n)$, let $\psi = \{G_1, G_2, \ldots, G_k\}$ be a collection of these k clusters where $G_i \cap G_j = \phi$, for all $i \neq j$ and $X = \{G_1 \cup G_2 \cup \ldots \cup G_k\}$; $n_1, n_2, \ldots, n_k$ be the number of objects in these k clusters. Denote the dissimilarity between values of variables by $d(\cdot, \cdot)$ and the dissimilarity between two objects, two clusters or one object and one cluster by $D(\cdot, \cdot)$. $D(i, j)$ represents the dissimilarity between objects i and j. Also, $D(G_i, j)$ represents the dissimilarity between a cluster G_i and an object j.

3.2.1 Linkage methods

Single Linkage

The Single Linkage is known as using the nearest neighbour rule, which defines the

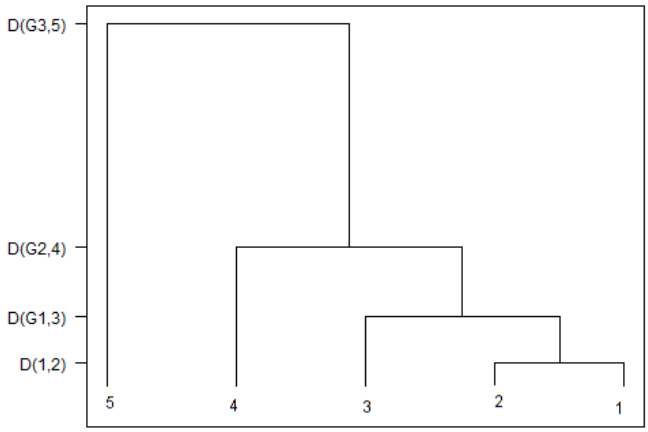


Figure 3.2: An illustration of a dendrogram - The y-axis is the dissimilarity between objects or clusters. On x-axis, the bottom end of each node indicates each object i, where $i = 1, \ldots, 5$. G_j, $j = 1, 2, 3$ represents cluster j.

dissimilarity between clusters G_i and G_j as the shortest distance between a pair of objects. The term pair means one object in G_i and one in G_i. Of these n_i times n_j dissimilarities, the smallest dissimilarity is defined as the dissimilarity between the two clusters. Therefore, the dissimilarity between clusters G_i and G_j is defined by (Gordon [1999])

$$D(G_i, G_j) = \min_{i' \in G_i, j' \in G_j} D(i', j') = d(\mathbf{x}_{i'}, \mathbf{x}_{j'}). \tag{3.7}$$

The advantage of using this method is that the found clusters are often separated in respect of dissimilarity. However, because this method uses a close pair of objects regardless of one of the pairs of objects might be far from each other, it produces clusters in which objects might be far apart and sometimes clusters that contain few objects if these few objects are isolated.

Complete Linkage

The Complete Linkage is known as using the furthest neighbour rule, which defines the dissimilarity between clusters G_i and G_j as the largest distance between an object in G_i and an object in G_j. Therefore, the dissimilarity between clusters G_i and G_j is defined by (Gordon [1999])

$$D(G_i, G_j) = \max_{i' \in G_i, j' \in G_j} D(i', j') = d(\mathbf{x}_{i'}, \mathbf{x}_{j'}). \tag{3.8}$$

In the Complete Linkage method, the dissimilarities of all pairs of objects of G_i and and G_j are computed. Of those dissimilarities, the largest one is used. By this method, clusters will not be merged together if there exists one pair of their objects that are far away from each other. The advantage is that the found clusters are compact and have similar diameters. However, the found clusters are not necessarily well separated.

Average Linkage

The Average Linkage defines the dissimilarity between clusters G_i and G_j based on the average of all distance between all pairs of objects in G_i and G_j. Therefore, the dissimilarity between clusters G_i and G_j is defined by (Gordon [1999])

$$D(G_i, G_j) = \frac{1}{n_i \times n_j} \sum_{i' \in G_i} \sum_{j' \in G_j} d(\mathbf{x}_{i'}, \mathbf{x}_{j'}). \tag{3.9}$$

In the Average Linkage method, the dissimilarities of all pairs of objects of G_i and and G_j are computed and the average of those dissimilarities is used. This method is regarded a compromise between the Single Linkage and the Complete Linkage methods. The dissimilarity between clusters is not determined by two objects but it depends on all objects in the clusters.

3.2.2 Partition clustering methods

K-Means method

The K-Means method aims at partitioning n objects into k clusters in which each object is assigned to the cluster with the nearest mean. In other words, it aims at minimizing an objective function L_{KM} which is defined by (Hartigan and Wong [1979])

$$L_{KM} = \min_{\psi=\{G_1,G_2,\ldots,G_k\}} \min_{\{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k\}} \sum_{i=1}^{k} \sum_{j \in G_i} d_E(\mathbf{x}_j, \boldsymbol{\mu}_i). \tag{3.10}$$

The $d_E(\cdot, \cdot)$ in the equation above denotes the square Euclidean distance. One way of writing an algorithm to perform the K-Means clustering method is as follows: Step 1, objects are randomly split into k initial sets. Step 2, each of the mean vectors of the k sets, $\boldsymbol{\mu}_i = \frac{1}{n_i} \sum_{j \in G_i} \mathbf{x}_j; i = 1, \ldots, k$, is calculated and treated as the centre of each of the k clusters. Step 3, each object is assigned to the cluster with the smallest Euclidean distance from the centre. Step 4, each of the mean vectors of the k clusters

is recalculated. Then, the method repeats Step 3 to Step 4 until a local minimum of within-cluster sum of squares is reached.

In the Euclidean space the mean vector of each cluster is calculated by taking the aggregate all values of each variable and divided by the number of objects, so the K-Means method can be written as a function. Thus, strictly speaking, it is regarded as neither a dissimilarity-based nor a model-based clustering method. However, when no row data but only dissimilarity matrix is provided, mean vectors can not be extracted from a dissimilarity matrix. Therefore, the K-Means is not for clustering dissimilarities.

Partition around medoids (PAM) method

The PAM method aims at partitioning n objects into k clusters in which each object is assigned to the cluster with the closest medoid. These k medoids are regarded as representative objects among the objects of the dataset. In other words, it aims at minimizing an objective function L_{PAM} which is defined by (Kaufman and Rousseeuw [1990])

$$L_{PAM} = \min_{\psi=\{G_1,G_2,\ldots,G_k\}} \min_{\{\mathbf{x}_{(1)},\ldots,\mathbf{x}_{(k)}\}} \sum_{i=1}^{k} \sum_{j\in G_i} d(\mathbf{x}_j, \mathbf{x}_{(i)}). \tag{3.11}$$

where $\mathbf{x}_{(i)}$ is the record for the medoid of G_i. The PAM clustering method consists of phases I and II. Phase I in which initial k medoids are found is called BUILD and phase II in which final k medoids are found is called SWAP. The phase I works as follows. First, the object i for which the sum of the dissimilarities to all other objects is the smallest is selected. Second, consider the nonselected objects, the object j for which the sum of the dissimilarity of objects to their closest representative objects (i, j) is the smallest is selected. The process is continued until k objects are selected. In phase I, k objects, $\{\mathbf{x}_{(i)} : i = 1, \ldots, k\}$, are selected as the initial k medoids for each of the k clusters. The phase II works as follows. First, each object is assigned to the cluster with the closest medoid. Next, in each of the k clusters, the method swaps object who is considered as the medoid with one of the remaining objects, so that the new medoid makes a minimum of sum of distances. The process repeats until the k medoids stay without change. As the medoid vector of each cluster is the record of the selected object, this clustering method works both on raw data and on a dissimilarity

matrix. In addition, the PAM method includes a selection of the initial vector, so it gives a consistent clustering result.

To sum up, the Single Linkage method tends to chain two clusters if they have one pair of points with the shortest distance. As a consequence, objects within any single cluster maybe far from each other. The Complete Linkage method ensures that the widths of dissimilarities between objects in a cluster are similar, but the clusters are not necessarily separated. As a compromise solution, the Average Linkage method uses dissimilarities of all pairs of participants. The K-Means method uses means and the PAM clustering method works with mediods. The mediods can be obtained from a dissimilarity matrix. A new dissimilarity function is proposed in this thesis, and therefore the linkage methods and the PAM are considered to be used.

3.3 Model-based clustering

The clustering methods described in the previous section, except the K-Means, are dissimilarity-based clustering methods. Apart from it, another clustering approach often being used is called model-based clustering method. It assumes that there is a distribution for each cluster; therefore, a dataset can be represented by a finite mixture of these distributions. Model-based clustering is flexible in choosing the individual distribution which is used to model each cluster. Mixture models allow the observed features of data to be continuous or discrete (Duda, Hart, and Stork [2001]; Fraley and Raftery [2002, 2010]; Liao [2006]; McLachlan and Peel [2000]). Let X be the data of all n observations $\{\mathbf{x}_i : i = 1, \ldots, n\}$. The likelihood for a mixture model with k components is

$$L_{MIX}(\theta_1, \ldots, \theta_k; \tau_1, \ldots, \tau_k|\mathrm{X}) = \prod_{i=1}^{n} \sum_{g=1}^{k} \tau_g f_g(\mathbf{x}_i|\theta_g) \tag{3.12}$$

where f_g and θ_g are the density and parameters of the g^{th} component in the mixture model and τ_g is the probability that an observation belongs to the g^{th} component. Let the vector $\phi = (\boldsymbol{\tau}, \boldsymbol{\theta})$ be the all unknown parameters of Eq 3.12. It can be fitted by using the expectation-maximization (EM) algorithm (McLachlan and Basford [1988]).

The EM algorithm for the mixture model alternates between two steps. At the "E" step, the conditional expectation of the log-likelihood is computed based on the observed data, averaging over the estimated distribution of the missing component memberships. At the "M" step, the parameters that maximize the expected log-likelihood from the E step are determined. Once an estimate of ϕ is obtained, estimates of the posterior probabilities of component memberships can be formed for each object. Each object is classified into the group to which it has the highest estimated posterior probability.

Model-based clustering has several advantages. With the underlying distributions, the questions of determining the number of clusters and selecting the clustering methods becomes a question of model selection Li [2006]. Also, it measures the uncertainty for component memberships of objects by τ_g. Fraley and Raftery [2002] showed model-based clustering has more advantages in medical data, gene expression data, spatial data, etc. Model-based clustering is flexible and can accommodate with non-Gaussian data. But it can be limited on high-dimensional data as if the dimension of the data is relatively large to the number of objects, the covariance might be singular, which causes the EM algorithm to break down.

The main reason for which the model-based clustering is not considered for the category-ordered data is the missing records. The missing values are missing not at random. In some cases, particularly if there were longer absences, missing values point to more severe problems of the participant, or a tendency to leave the study, or illicit drug use. These make the implemented model-based clustering methods hard to use as straight forward. Therefore, we focus on the dissimilarity-based clustering methods. For whom may be interested in applying the model-based clustering method, here are some ideas for making the method to accommodate the category-ordered data. First of all, a sensible scheme for imputation is required. As the aforementioned, the EM algorithm is used to take care of missing component memberships in mixture models. Since the EM algorithm can also be used to estimate missing values, it would be interesting to include a missing imputation process by finding the distribution of the complicated patterns of missing data using an EM algorithm in the mixture models. Another thought is to use the EM algorithm and the multidimensional scaling (MDS)

(Cox and Cox [1990]) (see 7.2 for details of MDS). MDS represents a dissimilarity matrix of high dimensional data to a lower dimensional space. It would also be interesting to include the EM algorithm to estimate missing values in the dissimilarity matrix in the MDS algorithm and then to apply model-based clustering to the MDS solution. These approaches are another research topics. In this study, we focus on the dissimilarity-based clustering methods and propose a so-called p-dissimilarity in the next chapter.

Chapter 4

New dissimilarity function : the p-dissimilarity

In this chapter we propose a new dissimilarity function, the p-dissimilarity. It can be used to measure dissimilarity for category-ordered data. It accommodates ordinal and categorical scales by using two parameters p and β. For convenience, we use the term "values" to refer to $\{1, 2, \ldots, 7\}$, indicating the seven categories, and we use "categories" to refer to the six categories of dosages and one category of missing dosages. Section 4.1 explains the motivation and Section 4.2 shows the assumption and requirements for dissimilarity between categories. Section 4.3 gives the definition of the p-dissimilarity and Section 4.4 discusses the advantages.

4.1 Motivation of dissimilarity design

The dissimilarity functions are a fundamental aspect of the dissimilarity-based clustering methods. Many dissimilarity functions are made with respect to study purposes (see Section 3.1). However, the existing ones are not suitable for category-ordered data. The problem of measuring the dissimilarity for the category-ordered data is complicated. The problem can be outlined in terms of ordinality and dissimilarity for missing values.

The Simple matching coefficient ignores the ordinality of the values of the categories. The approach of using binary variables takes into account the ordinality. By

this approach, the dissimilarities between value 1 and values 2 to 6 are $\frac{1}{5}$, $\frac{2}{5}$, $\frac{3}{5}$, $\frac{4}{5}$ and 1. It shows that value 2 is more similar to value 1 than to the larger values. This approach assumes that the dosage intervals of the categories are equal. However, the dosage width of category 6 is not 20 mg. Another approach of assigning ranks (Eq 3.5) also considerers ordinality. The trouble with this approach is that, for values 1 to 5 that refer to categories for which the dosage width is 20 mg, the dissimilarities between value 1 and values 2 to 5 are different. Moreover, both of these approaches define dissimilarity for observed data but not for unobserved data. This means that the dissimilarities between value 7 and values 1 to 6 are not yet defined.

The aforementioned dissimilarity functions deal with ordinal variables in a quantitative way. They compute dissimilarity by quantifying categories. However, it is not about which distance value to use for quantifying category, it is about which distance value to use as an interpretative dissimilarity (Hennig and Hausdorf [2006]). As mentioned in Section 2.4, there are two important characteristics of category-ordered data. Firstly, physicians thinks categorically in prescriptions. Secondly, the missing dosage is treated as one category. The missing value refers to unobserved dosage. It should thus not be treated as closer to any of the categories. These need to be taken into account. Our purpose is to find clusters in which participants within a cluster have high agreement with regards their behaviours while receiving methadone. The behaviour is represented by sequences of categories. The sequences of categories take in a pattern. Our data shows that participants sometimes had their dosage changed from one category to another, but what matters is the length of staying in the same category. The changing of one category to another regardless of the distance of the move means the dosage is not stable. We may want patterns that are insensitive to the sudden changes which appear as short sequences of categories. For instance, suppose data for three participants (A,B,C) for 7 days are [1, 1, 1, 1, 1, 2, 2], [1, 1, 3, 1, 1, 2, 2] and [1, 2, 1, 2, 1, 3, 2]. A cluster including participants A and B can be presented as the cluster with dosage pattern of [1, 1, (1,3), 1, 1, 2, 2]. It is more useful than a cluster produced by participants A and C.

To find clusters whose participants have similar dosages in a longer time, a dissimilarity function should give a larger dissimilarity for participants whose values on

the same day are in different categories and a small dissimilarity for participants most of whose values on the same day are in the same category. We define a "neighbouring category" as referring to category which is considered closest to a target category with respect to dosages in the intervals. We use a term " distant-neighbouring categories" to refer to the categories that are not neighbouring categories. The dissimilarity function should then focus more on distinguishing categories and less on showing how different the neighbouring categories and the distant-neighbouring categories are in relation to the target category. The dissimilarity between neighbouring category should contribute to the dissimilarity between participants the most. Also, the dissimilarity function should have the dissimilarity between two categories goes larger as the categories go further apart. A concave dissimilarity function that dominate by neighbouring category and keeps ordinality is thus ideal. However, such a dissimilarity function that enables us to distinguish categories of observed dosage from missing values, that takes into account ordinality, that can be used for data where each variable has mixed information on nominal and ordinal, and that enables a proper interpretation, does not currently exist. Therefore, we propose the p-dissimilarity function. It gives a solution for the aforementioned problems and it can be applied to the category-ordered data.

4.2 Dissimilarity between categories

The category-ordered data has a time series form, so we consider three approaches: time warping, autoregressive model and aggregating dissimilarity by days. First of all, the time warping focuses on patterns of the dosage taken by two participants, that is, the shapes formed by their daily dosage records. However, this ignores the retention in MMT. Retention in MMT is meaningful; a dosage pattern of a participant over 3 months cannot be compressed to one month or expanded to six months. So, we rather keep the length of treatment as it is. Secondly, time series clustering works based on the auto-correlations of participants' daily dosage. Assume one participant has dosage in categories 5, 4, 3, each of which has length for one month, and another participant has dosage in categories 3, 2, 1, each of which has length for one month. They will be grouped as one cluster because they have the same auto correlation matrix structure. This method is not ideal in our case. The categories are essentially different and we want to cluster on the absolute level. Also, we want to keep the length for categories.

To achieve these, we define the dissimilarity between two time series by aggregating daily dissimilarity between categories.

We start from the dissimilarly between categories. Assuming for the moment that there is no category for missing dosages. We assume the following:

Assumption
It is the neighbouring category which contributes most to distinguish the target category.

This assumption allows that the distance between a target category and its neighbouring category provides the most information for separating the target category from the others. The more distant the comparison category, the lower the rate at which the dissimilarity increases. Hence, the dissimilarity function is not linear.The dissimilarity function is concave from the target category to its distant-neighbouring categories.

Notation
The dosage, a numeric variable, is partitioned into θ categories denoted by $C_i, i \in \Theta = \{1, \ldots, \theta\}$. For convenience, value i refers to the i^{th} category, the subscript of C_i. Also, C_i refers to a category, which is a set of dosages. Let $D(\cdot, \cdot)$ denote the dissimilarity between participants and $d(\cdot, \cdot)$ denote the dissimilarity between categories.

Following are requirements for dissimilarity between categories.
Requirement 1
The dissimilarities of all paired neighbouring categories should be equal.

$$d(C_i, C_{i+1}) = d(C_{i'}, C_{i'+1}), \text{ for } i, i' \in \Theta. \tag{4.1}$$

Requirement 2
For $i, i', j \in \Theta$, if $i < i' < j$, the dissimilarity between categories C_i and $C_{i'}$ is smaller than that between categories C_i and C_j.

$$d(C_i, C_{i'}) < d(C_i, C_j). \tag{4.2}$$

Requirement 2 is set up in order to show ordinality. Requirement 2 means that given a target category and another two categories, one being closer to it than the other, the dissimilarity between the target category and the closer category should be smaller than the dissimilarity between the target category and the more distant category. In other words, the closer category has a smaller dissimilarity and the dissimilarity increases as the differences between the values of categories grow larger.

Requirement 3

For $i, i', j \in \Theta$ and $i \leq i' \leq j$, the dissimilarity between categories C_i and C_j is less than or equal to the sum of the dissimilarity between categories C_i and $C_{i'}$ and the dissimilarity between categories $C_{i'}$ and C_j.

$$d(C_i, C_j) \leq d(C_i, C_{i'}) + d(C_{i'}, C_j). \tag{4.3}$$

Requirement 3 means the dissimilarity between categories does not necessarily go up linearly.

General requirements

1. $d(C_i, C_{i'}) > 0$, if $i \neq i'$, for $i, i' \in \Theta$,
2. $d(C_i, C_{i'}) = d(C_{i'}, C_i)$,
3. $d(C_i, C_i) = 0$.

This is a set of standard requirements for dissimilarity functions. A dissimilarity function should satisfy three requirements: non-negativity in which the dissimilarity between categories is always greater or equal to zero; symmetry in which the dissimilarity from categories i to i' is equal to that from categories i' to i; identity in which the dissimilarity between a category and itself is zero (Gordon [1999]; Kaufman and Rousseeuw [1990]).

4.3 Design of the p-dissimilarity

In this section we propose the p-dissimilarity that is designed on the basis of these requirements. We give the definition of the p-dissimilarity in two steps. We start from a situation in which there are no missing values in any observation (4.3.1). Then, we move to a situation in which there are missing values (4.3.2).

4.3.1 The p-dissimilarity without missing values

Let $x_{it} \in \Theta = \{1, \ldots, \theta\}$ be the category-ordered data for the participant i on the t^{th} day since they joined the MMT. Assuming for the moment that there are no missing dosages, the p-dissimilarity between participants i and i' is defined by

$$D(i, i') = \Sigma_{t=1}^{T} d(x_{it}, x_{i't}) = \Sigma_{t=1}^{T} (1 - p^{\alpha_{ii'}(t)}) \tag{4.4}$$

where $0 < p < 1$ and $\alpha_{ii'}(t) = |x_{it} - x_{i't}|$. The meaning of $(1 - p^{\alpha_{ii'}(t)})$ is the contribution of the t^{th} dosage taken record to the dissimilarity $D(i, i')$.

The role of $\alpha_{ii'}$ is aimed at indicating ordinality of values of categories. It should guarantee that the difference between categories becomes large as the values of the categories go further away from each other. Therefore, $\alpha_{ii'}(t)$ is defined as the absolute value of the difference between the ordinal values of the categories of participants i and i' on day t. The values of categories are $\{1, \ldots, \theta\}$. The minimum difference between the values is 0, and the maximum is $(\theta - 1)$; hence, $\alpha_{ii'}$ ranges from 0 to $(\theta - 1)$.

The tuning constant p is for measuring dissimilarity between categories. The dissimilarity between two neighbouring categories is $(1 - p)$. The advantage of the parameter p is that the p can be used as a switch between data being treated as categorical and ordinal. A small p indicates that the target category is considered to be very different from its neighbouring categories. Using a small p, data are treated as rather categorical. A large p indicates that the target category is considered to be not so different from its neighbouring categories. Using a large p, data are treated as rather ordinal. In order to determine the value of p, some subjective judgement has to be used about how sensitive the p-dissimilarity needs to be for separating categories, and how important it

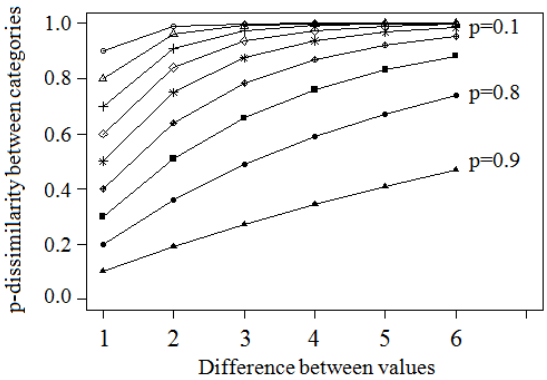


Figure 4.1: An illustration of p-dissimilarities - The p-dissimilarities is monotonic in absolute difference between values and concave.

is for data to be considered as more ordinal or as more categorical in practice. To highlight the advantage, Figure 4.1 shows the p-dissimilarities between values of categories. The dissimilarity is non-linear and the dissimilarity increases monotonically from the neighbouring category to the distant-neighbouring categories. The smaller the p, the larger the dissimilarity between categories is. Moreover, the dissimilarity between C_i and C_{i+1} is $(1-p)$. The dissimilarity between C_i and its distant-neighbouring categories can be written by an equation involving $(1-p)$ (Eq 4.5).

The following are the properties of the p-dissimilarity for a single day with respect to the aforementioned requirements.

Proposition 1. The differences between any two neighbouring categories are equal.

Proof

Suppose data of a single variable of two participants falls in categories $(C_i, C_{i'}), i, i' \in \Theta$.

$$d(C_i, C_{i+1}) = 1 - p^{|i-(i+1)|} = 1 - p^{|i'-(i'+1)|} = d(C_{i'}, C_{i'+1}).$$

Proposition 2. The dissimilarity between a target category and its neighbouring category is smaller than that between it and its distant-neighbouring category.

Proof

For $g, h, i \in \Theta, g < h < i$, and $0 < p < 1$

$$\begin{aligned} d(C_h, C_i) &= 1 - p^{|i-h|} \\ &< 1 - p^{|i-g|} \quad = d(C_g, C_i). \end{aligned}$$

Proposition 3. For $i, i' \in \Theta$ and $i \neq i'$, the dissimilarity between categories C_i and $C_{i'}$ is

$$d(C_i, C_{i'}) = (1-p) \sum_{l=0}^{|i'-i|-1} p^l. \tag{4.5}$$

Proof by induction

For $i, i' \in \Theta$ and $i \neq i'$, let $|i' - i| = n_0$,

When $n_0 = 1$, $d(C_i, C_{i'}) = 1 - p = (1-p) \sum_{l=0}^{1-1} p^l$.

assume $n_0 = n$, $d(C_i, C_{i'}) = (1 - p^n) = (1-p) \sum_{l=0}^{n-1} p^l$.

When $n_0 = n + 1$,

$$\begin{aligned} d(C_i, C_{i'}) &= 1 - p^{n+1} \\ &= (1-p) + (p - p^{n+1}) \\ &= (1-p) + p(1 - p^n) \\ &= (1-p) + p(1-p) \textstyle\sum_{l=0}^{n-1} p^l \\ &= (1-p)(1 + p \textstyle\sum_{l=0}^{n-1} p^l) \\ &= (1-p)(1 + \textstyle\sum_{l=1}^{n} p^l) \\ &= (1-p) \textstyle\sum_{l=0}^{n} p^l. \end{aligned}$$

Therefore, $d(C_i, C_{i'}) = (1-p) \sum_{l=0}^{|i'-i|-1} p^l$, $\forall i, i' \in \Theta$ and $i \neq i'$.

By proposition 3, $d(C_i, C_{i+1}) = (1-p)$, $d(C_i, C_{i+2}) = (1-p)(1+p)$, and $d(C_i, C_{i+3}) = (1-p)(1+p+p^2)$. Also, the numerical difference between $d(C_i, C_{i+1})$ and $d(C_i, C_{i+2})$ is $(1-p)p$, while the difference between $d(C_i, C_{i+2})$ and $d(C_i, C_{i+3})$ is $(1-p)p^2$ and so on. A quantity of increment of dissimilarity when the values of the categories go larger is in proportion to $(1-p)$. Also, the term $(1-p)$ is the contribution of the neighbouring category to the dissimilarity in distinguishing a category.

Proposition 4. For all triples of participants (g, h, i), $D(g, h) + D(g, i) \geq D(h, i)$. This is also known as the triangle inequality.

Proof

It is sufficient to establish the metricity for a single variable of the p-dissimilarity defined in Eq 4.4. Suppose a variable takes values C_g, C_h, C_i for three participants. Without loss of generality, let $g \leq h \leq i$. Then the dissimilarities of the three participants are

$$\begin{aligned} D(g,h) &= d(C_g, C_h) &= 1 - p^{h-g}, \\ D(g,i) &= d(C_g, C_i) &= 1 - p^{i-g}, \\ D(h,i) &= d(C_h, C_i) &= 1 - p^{i-h}. \end{aligned}$$

It is trivial that $D(g,i)$ is the longest side of the triangle and the metric property is valid for all permutations if $D(g,h) + D(h,i) \geq D(g,i)$.

$$\begin{aligned} D(g,h) + D(h,i) - D(g,i) &= (1 - p^{h-g}) + (1 - p^{i-h}) - (1 - p^{i-g}) \\ &= (1 - p^{h-g}) - p^{i-h} + p^{i-g} \\ &= (1 - p^{h-g}) - p^{i-h}(1 - p^{h-g}) \\ &= (1 - p^{h-g})(1 - p^{i-h}) \geq 0 \end{aligned}$$

hence, the p-dissimilarity defined in Eq 4.4 is metric.

4.3.2 The p-dissimilarity with missing values

In this section we take into account missing values and give a full definition of the p-dissimilarity. Although the dosage records are real zeros, the addiction should not be treated as zero. It is normal that participants on specific prescriptions miss some days. This should not spoil the dissimilarity computation. The concept of handling the missing values is to view them as if they were observed. If missing dosages were observed, they would be re-coded as one of the θ categories. Consequently, the dissimilarity between missing dosages and the θ categories would be within the range of $[1-p^0, 1-p^{\theta-1}]$, depending on the categories to which the missing dosages belong. However, missing dosages are, in fact, not observed. They should not be considered more or less similar to any of categories. The dissimilarity between a category of observed dosages and the category of the missing dosages is then set from $(1-p)$. A parameter β is used. For day t when a missing value occurs, the contribution of the t^{th} records to $D(i,i')$ is $(1 - p^{\beta_{ii'}(t)})$. The selection of $\beta(t)$ depends on the degree of difference between the category of missing dosages and the categories of observed dosages as reckoned by researchers. Using $\beta = 1$ means that the category of the missing dosages is considered to be a neighbouring category of all the categories of observed dosages. Using β greater than 1, the category of the missing dosages is considered to be a distant-neighbouring

category. Moreover, we define all $\beta_{ii'}(t)$ equal to β, that is, only one value in the range [1, (θ-1)] will be used in the analysis. A discussion about choices of β can be found in Section 6.1.1.

There are two situations when computing the dissimilarity of two participants i and i' on day t. One situation is that both participants i and i' have their t^{th} dosage taken records non-missing. Another one is that there is a missing dosage. An indicator $\delta_{ii'(t)}$ is then used to distinguish these situations. $\delta_{ii'(t)}$ is equal to 1 when both observations i and i' for their t^{th} dosage taken records were non-missing, and it is equal to 0 otherwise. The p-dissimilarity between participants i and i' with a category for missing values is defined by

$$D(i,i') = \Sigma_{t=1}^{T}[\delta_{ii'(t)}(1-p^{\alpha_{ii'}(t)}) + (1-\delta_{ii'(t)})(1-p^{\beta})], \tag{4.6}$$

where $\delta_{ii'(t)}$ is equal to 1 when both participants i and i' for their t^{th} dosage taken records are non-missing and equal to 0 otherwise, $0 < p < 1$, $\alpha_{ii'}(t) = |x_{it} - x_{i't}|$, and $1 < \beta < (\theta - 1)$. The meaning of p and α can be found in Section 4.3.1. The following are the proofs of the three general requirements for the p-dissimilarity to be a dissimilarity function.

Proposition 1. For any two participants (i, i'), the following are true for the p-dissimilarity defined in Eq 4.6:

1. $D(i,i') \geq 0$,
2. $D(i,i') = D(i',i)$,

Proof

Let $\mathbf{x}_i = [x_{i1}, \ldots, x_{iT}]$ and $\mathbf{x}_{i'} = [x_{i'1}, \ldots, x_{i'T}]$ be the category-ordered data of participants i and i'. The dissimilarity between the two participants $D(i,i')$ is equal to $\sum_{t=1}^{T} d(x_{it}, x_{i't})$.

(1) The dissimilarity of the values x_{it} and $x_{i't}$, $t = 1, \ldots, T$ is

$$d(x_{it}, x_{i't}) = \begin{cases} 1-p^{|x_{it}-x_{i't}|}, & \text{if both are non-missing,} \\ 1-p^{\beta}, & \text{otherwise.} \end{cases}$$

Table 4.1: The p-dissimilarity matrix of the seven categories with $\beta = 2$. The first six categories represent the ordered set of dosages indicated in the first column. The p-dissimilarity is the one minus absolute value of the difference of the values of the categories. The p-dissimilarities between category 7 and categories 1 to 7 are all $(1 - p^{\beta})$.

Dosage values	Category	1	2	3	4	5	6	7
$\leq$ 20 mg	1	0	1-p	1-p^2	1-p^3	1-p^4	1-p^5	1-p^2
21-40 mg	2	1-p	0	1-p	1-p^2	1-p^3	1-p^4	1-p^2
41-60 mg	3	1-p^2	1-p	0	1-p	1-p^2	1-p^3	1-p^2
61-80 mg	4	1-p^3	1-p^2	1-p	0	1-p	1-p^2	1-p^2
81-100 mg	5	1-p^4	1-p^3	1-p^2	1-p	0	1-p	1-p^2
$>$ 100 mg	6	1-p^5	1-p^4	1-p^3	1-p^2	1-p	0	1-p^2
missing dosages	7	1-p^2	1-p^2	1-p^2	1-p^2	1-p^2	1-p^2	1-p^2

Since $0 < p < 1$, the dissimilarity of any two values x_{it} and $x_{i't}$ is greater than or equal to 0. This implies that

$$D(i, i') = \Sigma_{j=1}^{T} d(x_{it}, x_{i't}) \geq 0.$$

(2)
Case I: both x_{it} and $x_{i't}$ are non-missing

$$d(x_{it}, x_{i't}) = 1 - p^{|x_{it} - x_{i't}|} = 1 - p^{|x_{i't} - x_{it}|} = d(x_{i't}, x_{it}).$$

Case II: one of x_{it} and $x_{i't}$ is missing or both of them are missing

$$d(x_{it}, x_{i't}) = 1 - p^{\beta} = d(x_{it}, x_{i't})$$

This implies that

$$D(i, i') = \Sigma_{j=1}^{T} d(x_{it}, x_{i't}) = \Sigma_{j=1}^{T} d(x_{i't}, x_{it}) = D(i', i).$$

(3) If the object is identified to be the same participant, we define the p-dissimilarity to be 0.

4.4 Advantages and disadvantages of the p-dissimilarity

The discussion is carried out by comparing the p-dissimilarities with the Euclidean distance. A quick review of the p-dissimilarities with $\beta = 2$ among seven categories on a

single day is given in Table 4.1. The first column shows the dosage intervals and the second column shows the corresponding categories. Each row shows the p-dissimilarities between the category shown in the second column and categories 1 to 7.

Why do we not simply apply the Euclidean distance to the category-ordered data as has been done with the p-dissimilarity? The Euclidean distance focuses on difference between the values of categories, which fails it in matching physicians' perspective. Physicians focus more on sequence of constancy and less on sudden changes in categories. Following the example in Section 4.1, data for three participants (A,B,C) for 7 days are [1, 1, 1, 1, 1, 2, 2], [1, 1, 3, 1, 1, 2, 2] and [1, 2, 1, 2, 1, 3, 2]. Denote the Euclidean distance between participants by $D_E(\cdot,\cdot)$. The Euclidean distances between participants are $D_E(A,B) = 2$, $D_E(B,C) = 2.65$ and $D_E(A,C) = 1.73$. Participants A and C who have most of their values on the same day in different categories are clustered. On the other hand, denote the p-dissimilarity between participants by $D_p(\cdot,\cdot)$, the p-dissimilarities of the three participants with $p = 0.6$ are $D_p(A,B) = 0.64$, $D_p(B,C) = 1.84$ and $D_p(A,C) = 1.2$. A and B are then grouped. Most of their values on the same day are in the same category. This cluster can be presented as the one cluster with dosage pattern of [1, 1, (1,3), 1, 1, 2, 2]. Such a pattern is captured by the p-dissimilarity function.

Next, to use Euclidean distance for data containing missing values, an imputation method is required. Although imputations can be used to deal with missing values, they depend on data structure and the proportion of missing values over a study period. In the MMT data, there are many missing dosages, and these missing dosages are missing not at random. An additional problem is the medical consideration behind the missing dosage. It is assumed that participants who continuously lacked 14 days' dosage records have practically left the study. If no further non-zero dosage record can be found, the participant is considered as having left the study. Since they have left the study, it is reasonable not to impute their dosage and keep zero dosage as missing. In order to apply the Euclidean distance to the category-ordered data which contains one category for missing dosages, one needs to define the distance between the category of the missing dosage and the categories of observed dosage.

In contrast, the concept behind the p-dissimilarity in respect of the missing dosages is to make use of the information on the observed dosages. We define the dissimilarity between the category of the missing dosage and all categories to be $(1-p^{\beta})$, so there is no need for imputations. Also, the p-dissimilarity allows us to observe the duration of having missing dosages which is represented by the sequence of category. The disadvantage of using $(1-p^{\beta})$ is that it destroys the metric property of the p-dissimilarity. Note that the triangle inequality is not a requirement for a dissimilarity function. Suppose category-ordered data for three participants take the values (x_t, y_t, z_t). We assume the value of x_t is missing and $|y_t - z_t| = 3$. Let $p = 0.9$ and $\beta = 1$, then

$$d(x_t, y_t)+d(y_t, z_t)-d(x_t, z_t) = (1-p^{\beta})+(1-p^{\beta})-(1-p^{|y_t-z_t|}) = 1-2p^{\beta}+p^3 = -0.07 < 0.$$

For a dataset without missing dosages, the p-dissimilarity is metric, while for a dataset with missing dosages, the p-dissimilarity is not metric. The p-dissimilarity is a dissimilarity function but not a distance function.

To sum up, the p-dissimilarity assigns a quantitative value to distances between neighbouring categories and categories further apart in a concave monotonic way, that is, further categories are further away, according to the dissimilarity. Also, the increase of the distance becomes smaller moving further away from a category and its neighbours. This implies information that can be seen as stronger than ordinal. The quantitative distance between ordinal scales is governed by the meaning of the categories with a tuning constant p. The category of missing values is treated in a specific way, as having the same dissimilarity from all other categories, that is, p^{β}. In addition, the principle behind the p-dissimilarity can be used in a wider field of applications where researchers have a quantitative idea about the interpretative distance between categories, such as studies that use questionnaires with choices on Likert scales and a $don't know$-category.

In the next chapter we will move on to the question of determination of the number of clusters.

Chapter 5

Determination of the number of clusters

In this chapter we review indexes for the determination of the number of clusters, namely the Calinski and Harabasz (CH) (Calinski and Harabasz [1974]) and the Average Silhouette Width (ASW) (Kaufman and Rousseeuw [1990]). Each of which is found to be the best indexes by simulation study (Arbelaitz et al. [2013]; Milligan and Cooper [1985]). Also, the Prediction Strength (PS) (Tibshirani et al. [2001]) is reviewed. In Section 5.3.1 we discuse more about the PS. A crucial issue of using the PS is that the PS uses classification to the closest mean, which does not work for dissimilarity data and is connected to K-Means method. However, the linkage methods work in a substantially different way. Therefore, we propose new rules for modifying Prediction Strength, so that it can be fully applied when hierarchical clustering methods and the PAM method are used. Also, we call the PS without new rules "original PS", and we call the PS with new rules "modified PS". The limitation of using CH is that the CH uses the Euclidean distance. To allow us to use the p-dissimilarity, we decide to use the ASW and the modified PS for this study.

5.1 Indexes for finding of number of clusters

Many studies have been published on indexes for determining the number of clusters (Calinski and Harabasz [1974]; Hartigan [1975]; Kaufman and Rousseeuw [1990];

Krzanowski and Lai [1988]; Tibshirani et al. [2001]). A study on the performance of the indexes was carried out by Milligan and Cooper [1985]. They examined 30 indexes on hierarchical clustering on artificial data sets on their performances. They found that the best index was Calinski and Harabasz (Calinski and Harabasz [1974]).

Notation

Suppose a dataset contains T variables and n objects. Assume that the n objects are clustered into k clusters, $(k \leq n)$. Denote the k clusters by $G_i, i = 1, \ldots, k$. Denote the number of objects in cluster G_i by n_i. Denote the data for an object r by $\mathbf{x}_r = [x_{r1}, \ldots, x_{rT}]$. Denote the dissimilarity between variables by $d(\cdot, \cdot)$ and the squared distance between variables by $d_E(\cdot, \cdot)$. Denote the dissimilarity between two objects, two clusters or one object and one cluster by $D(\cdot, \cdot)$.

Calinski and Harabasz (CH)

The index referred to as CH(k) was proposed by Calinski and Harabasz [1974]. The concept was to maximize the ratio of between-cluster sum of squares $B(k)$ and within-cluster sum of squares $W(k)$ over the k clusters, which has the same form as the F test statistic in ANOVA. The CH(k) is defined by

$$\mathrm{CH}(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)}, \tag{5.1}$$

where,

$$B(k) = \sum_{i=1}^{k} n_i \; d_E(\overline{G_i}, \boldsymbol{\mu}),$$

$$W(k) = \sum_{i=1}^{k} \sum_{i' \in G_i} d_E(\overline{G_i}, \mathbf{x}_{i'}).$$

The grand centre $\boldsymbol{\mu}$ is $\frac{1}{n}\sum_{r=1}^{n} \mathbf{x}_r$, while the cluster centres $\overline{G_i}$ are $\frac{1}{n_i}\sum_{i' \in G_i} \mathbf{x}_{i'}; i = 1, \ldots, k$. The $B(k)$ is the sum of distances of all cluster centres from the grand centre, while $W(k)$ is the sum of distances of all objects from their cluster centres. The values of CH are computed for $k > 1$. Note that the CH(k) is not defined for $k = 1$. The k which has the maximum CH value is suggested to be used.

More recently, Arbelaitz et al. [2013] carried out a similar study, which included many indexes that did not exist in 1985. The Average Silhouette Width (Kaufman and Rousseeuw [1990]; Rousseeuw [1987]) was in the group of the best indexes.

5.2 Average Silhouette Width

The index was first proposed by Rousseeuw in 1987. The idea was to display how similar an object r is to the cluster where it belongs and how similar the object r is to the remaining clusters. The Silhouette Width for an object r in cluster $G_i, i = 1, \ldots, k$ is defined by

$$s(r,k) = \frac{b(r,k) - a(r,k)}{\max(b(r,k), a(r,k))}, \tag{5.2}$$

where

$$a(r,k) = \frac{1}{n_i - 1} \sum_{r' \in G_i} d(\mathbf{x}_r, \mathbf{x}_{r'}),$$

$$b(r,k) = \min_{r \notin G_j} D(r, G_j) = \min_{r \notin G_j} \frac{1}{n_j} \sum_{r' \in G_j} d(\mathbf{x}_r, \mathbf{x}_{r'}).$$

$a(r,k)$ is the dissimilarity between object r and cluster G_i where object r belongs. It is defined as the average dissimilarity between object r and the other $(n_i - 1)$ objects in the same cluster. $b(r,k)$ is the dissimilarity between object r and one of the remaining clusters which has a smallest dissimilarity from object r. In other words, dissimilarities between r and G_j, where $j = \{1, \ldots, k\} \setminus \{i\}$, are computed, and then the minimum of the dissimilarities is $b(r,k)$.

The Average Silhouette Width (ASW) for k clusters is defined by

$$k_{ASW} = \frac{1}{n} \sum_{r=1}^{n} s(r,k).$$

It is the average of all Silhouette Width for n objects. Note that k_{ASW} is not defined for $k = 1$ because of $a(r,1) = b(r,1)$. For $k > 1$, the k which has the maximum Average Silhouette Width is suggested to be used.

5.3 Prediction Strength

This index was proposed by Tibshirani and Walther [2005]. The concept was to view an analysis of clustering as an analysis of classification. The difference between these two analyses is that the "true" class to which objects belong is known in the classification but that is unknown in the clustering. The purpose of classification is to identify the cluster where a new object belongs. It works as follows. Firstly, a dataset is split into two subsets, one being training set and the other being test set. Secondly, the training set is used to construct a classification model. The model is displayed as classification rules, decision trees, or equations, which will be used to predict the cluster for a new object. Thirdly, the classification model is applied to objects in the test set, and then the predicted cluster for each of the objects is called "predicted" class of the object. Because the "true" class of the objects in the test set is known, the "predicted" and "true" classes can then be compared. A proportion of objects whose true class and predicted class are the same is reported. Also, it is used as an index for the performance of the classification model. A basic problem about viewing clustering as classification was the unknown "true" class, and Tibshirani and Walther [2005] proposed a solution to build "true" class for the test set by which the comparison between the true and predicted classes could be done.

Figure 5.1 shows the flowchart of the calculation of the Prediction Strength. First of all, in order to use this index, one has to decide a clustering method ($Method$), and only then can the Prediction Strength produce values for $k = 1, 2, \ldots$. The options are hierarchical clustering methods, the K-Means and the PAM method. With a selected $Method$, the calculation of Prediction Strength for k clusters, denoted by $ps(k)$, involves three processes, denoted by process I, process II and process III.

First of all, the dataset is split into a training set and a test set, denoted by $\mathrm{X_{tr}}$ and $\mathrm{X_{te}}$. Assume that there are n objects in $\mathrm{X_{te}}$.

The outcome of the process I is the predicted class for the objects in $\mathrm{X_{te}}$. It works as follows. The clustering $Method$ with k clusters is applied to $\mathrm{X_{tr}}$. Denote the clustering result by $C(\mathrm{X_{tr}}, k)$. Given training clusters that are obtained from $C(\mathrm{X_{tr}}, k)$, a term

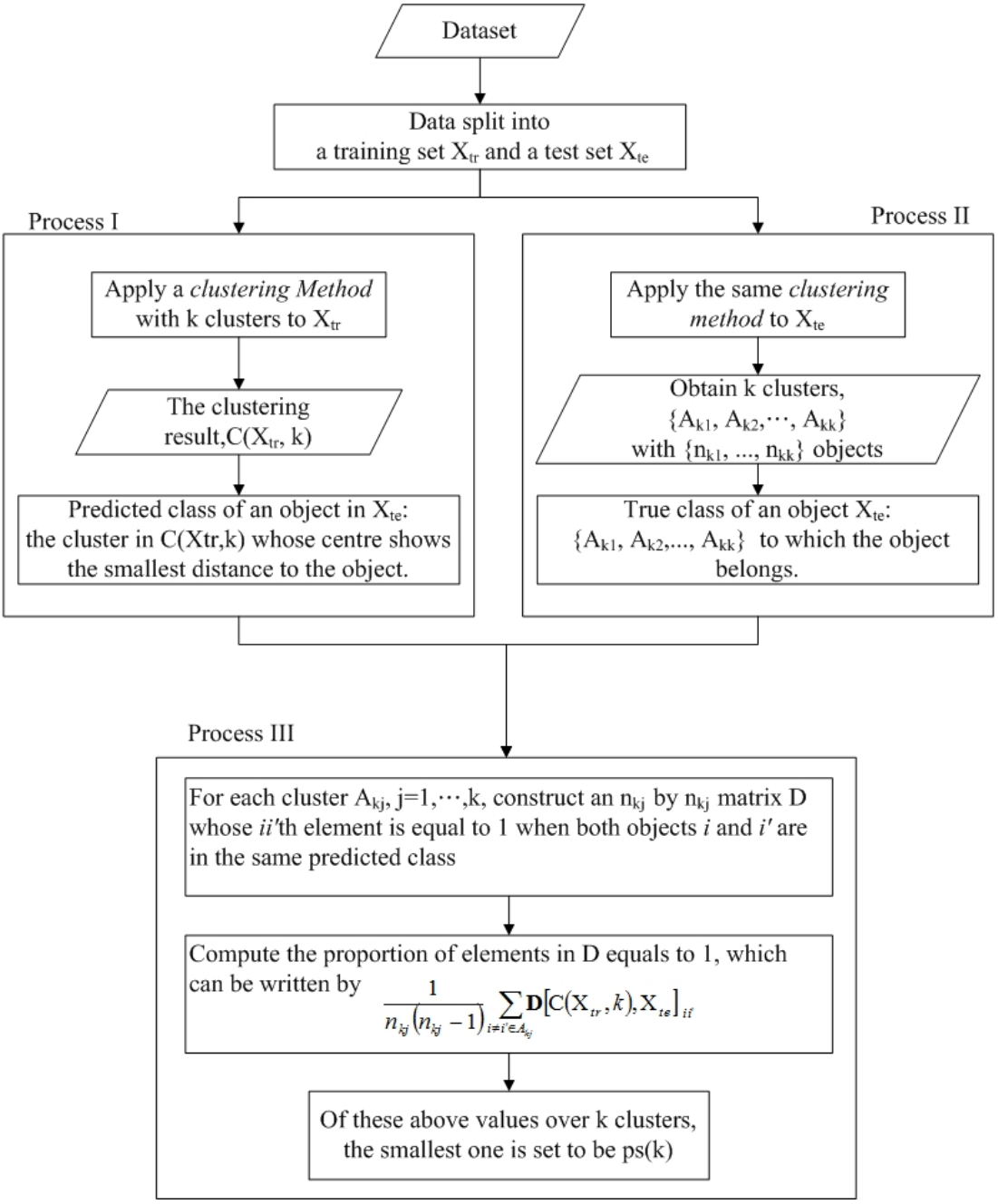


Figure 5.1: Flowchart of the Prediction Strength. - This flowchart illustrates the calculation of the Prediction Strength for all $k = 2, 3, ...$ The clustering method can be one of the hierarchical clustering methods, the K-Means and the PAM method. Of the objects in X_{te} with the same true class, their predicted classes are compared and presented by a matrix **D**.

"training centres" is used to refer to the mean vectors of the training clusters. Next, The "predicted" class of objects in $\mathrm{X_{te}}$ is defined to be the training cluster with nearest training centre.

The outcome of the process II is the true class for the objects in $\mathrm{X_{te}}$. By true class, the authors mean that it is defined as the true class. The clustering *Method* with k clusters is applied to $\mathrm{X_{te}}$. Denote the obtained k clusters by $\{\mathrm{A}_{k1}, \mathrm{A}_{k2}, \ldots, \mathrm{A}_{kk}\}$. Denote the number of objects in these clusters by $n_{k1}, n_{k2}, \ldots, n_{kk}$, respectively. The "true" class of objects is defined to be the cluster in $\{\mathrm{A}_{k1}, \mathrm{A}_{k2}, \ldots, \mathrm{A}_{kk}\}$ to which they belong.

The outcome of the process III is the value of the Prediction Strength for each cluster $\mathrm{A}_{kj}, j = 1, \ldots, k$. Also, the minimum of these values is defined as the Prediction Strength with k cluster. The process III works as follows. First of all, for cluster $\mathrm{A}_{kj}, j = 1, \ldots, k$, the "true" and the "predicted" classes of its objects are compared. This is done by using an matrix $\mathbf{D}$ whose ii'th element is the indicator of whether two objects i and i' are assigned to the same cluster by the training centres of $C(\mathrm{X_{tr}}, k)$. Next, the probability of any two objects i and i', $i \neq i' \in A_{kj}$, having the same "predicted" class is estimated, that is, the frequency of 1 of $\mathbf{D}_{n_{ki} \times n_{ki}}$. The Prediction Strength for the selected *Method* with k clusters is defined as the minimum of these estimated probabilities of $\{\mathrm{A}_{k1}, \mathrm{A}_{k2}, \ldots, \mathrm{A}_{kk}\}$. It can be written by

$$ps(k) = \min_{1 \leq j \leq k} \frac{1}{n_{kj}(n_{kj} - 1)} \sum_{i \neq i' \in A_{kj}} \mathbf{D}[C(\mathrm{X_{tr}}, k), \mathrm{X_{te}}]_{ii'}. \tag{5.3}$$

Hence, the average Prediction Strength is the average of m repetitions of $ps(k)$. Tibshirani and Walther [2005] suggested to use the largest k in which the average Prediction Strength is greater or equal to 0.8 or above a user-specified threshold. The average Prediction Strength finds the number of clusters through cluster validation. The advantage is that the proportions of clusters can be used as a measurement for the stability of each cluster. A similar method is proposed by Fang and Wang [2012], their idea is that objects in a training set and a test set are drawn from the dataset with replacement. The R-package fpc developed by Christian Hennig has implemented these two indexes, one being *prediction.strength* and the other being *nselectboot*, respectively.

5.3.1 New rules for modifying the Prediction Strength

In some circumstances, the hierarchical clustering methods are preferred to the K-Means method (Tibshirani et al. [2001]), so one might use the hierarchical clustering methods to obtain the clustering result $C(\mathrm{X_{tr}}, k)$. However, in the step 4 of the algorithm of the PS, the predicted class is always determined by the training centres of $C(\mathrm{X_{tr}}, k)$ regardless which of the clustering methods is used. This step is not appropriate to the linkage methods and the PAM method that are not build on the basis of the mean vectors of clusters. For instance, the Single Linkage method uses the nearest neighbour rule to obtain $C(\mathrm{X_{tr}}, k)$. This method works in a substantially different way from the K-Means. But labelling the class with closest training centre as the predicted class ignores the nearest neighbour rule. Moreover, there are dissimilarity functions in which mean is not defined, such as the p-dissimilarity. The step 4 needs to be modified. Therefore, we consider three linkage methods and the PAM method, and propose the corresponding solutions for modifying the Prediction Strength. Each of the solutions is a new rule to determine the predicted class. In step 4, labelling the class with closest training centre is called "original PS" and labelling the class with the new rule is called "modified PS". The new rules are as follows.

Notation

Denote the number of clusters by k. Denote k training clusters obtained from $C(\mathrm{X_{tr}}, k)$ by $\{G_1, G_2, \ldots, G_k\}$ where $G_i \cap G_j = \phi$, for all $i \neq j$. Denote data for an object r by $\mathbf{x}_r = [x_{r1}, \ldots, x_{rT}]$. Denote the dissimilarity between variables by $d(.,.)$. Denote the dissimilarity between a training cluster $G_{k'}$ and an object i in $\mathrm{X_{te}}$ by $D(G_{k'}, i)$.

Single Linkage

The "predicted" class of an object i in $\mathrm{X_{te}}$ is $G_{k'}$ when an object r in $G_{k'}$ has the shortest dissimilarity from the object i among all objects in the training set, that is

$$d(\mathbf{x}_r, \mathbf{x}_i) < d(\mathbf{x}_j, \mathbf{x}_i),$$

for all $j \in \mathrm{X_{tr}} \setminus \{r\}$ and $r \in G_{k'}$.

Complete Linkage

The "predicted" class of an object i in $\mathrm{X_{te}}$ is $G_{k'}$ if the object i has a shortest dissimilarity to $G_{k'}$ among all k training clusters. The dissimilarity between $G_{k'}$ and the object i is defined to be the greatest distance from the object i to objects in $G_{k'}$. In other words, the object i will be assigned to cluster $G_{k'}$ if

$$D(G_{k'}, i) < D(G_j, i), j \in \Theta = \{1, 2, \ldots, k\} \setminus \{k'\}.$$

The dissimilarity between cluster $G_{k'}$ and object i is defined by

$$D(G_{k'}, i) = \max_{\mathbf{x}_r \in G_{k'}} d(\mathbf{x}_r, \mathbf{x}_i). \tag{5.4}$$

Average Linkage

The "predicted" class of an object i in $\mathrm{X_{te}}$ is $G_{k'}$ if object i has a shortest distance to $G_{k'}$ among all k training clusters. The dissimilarity between $G_{k'}$ and the object i is defined by the average of all distances from the object i to all objects in $G_{k'}$. This means that the object i will be assigned to cluster $G_{k'}$ if

$$D(G_{k'}, i) < D(G_j, i), j \in \Theta = \{1, 2, \ldots, k\} \setminus \{k'\}.$$

The dissimilarity between cluster $G_{k'}$ and $\mathbf{y}_i$ is defined by

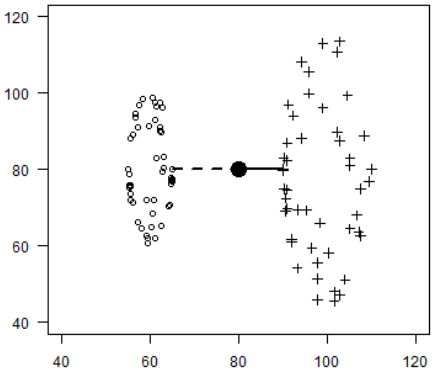
$$D(G_{k'}, i) = \frac{1}{n_{kk'}} \sum_{x_r \in G_{k'}} d(x_r, \mathbf{x}_i), \tag{5.5}$$

where $n_{kk'}$ is the number of objects in $G_{k'}$.

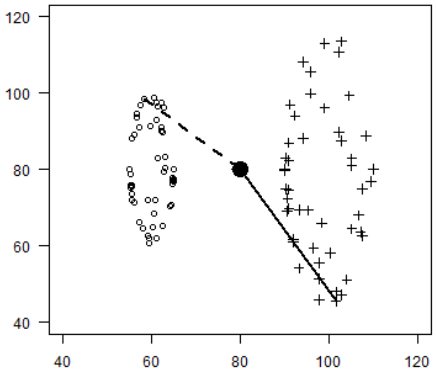
PAM method

The PAM method partitions objects in $\mathrm{X_{tr}}$ into k clusters in which each object is assigned to the cluster with the closest medoid. The clustering result for $\mathrm{X_{tr}}$ includes k clusters and k medoids. The "predicted" class of an object i in $\mathrm{X_{te}}$ is $G_{k'}$ if the object i has a shortest dissimilarity to the medoid of $G_{k'}$.

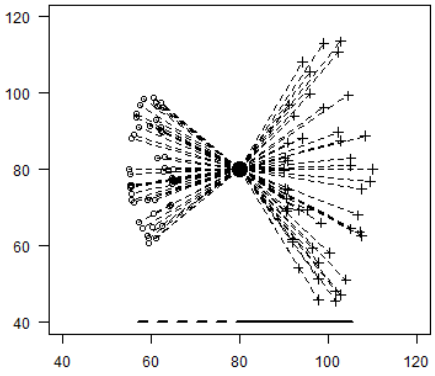
The above rules are for the modification of the process of generating the predicted class of objects in $\mathrm{X_{te}}$. Figure 5.2 illustrates an example of the application of the new rules to generate the predicted class for an object in a test set when linkage methods are applied. In the figure, the circles and crosses represent the data for objects in two



(a)



(b)



(c)

Figure 5.2: Application of the new rules on $C(\mathbf{X_{tr}}, 2)$ in the three linkage methods. - The data, appearing in two symbols, is randomly selected from points within two ellipses. For the ellipse with the circles, the diameters on the x-axis and the y-axis are 5 cm and 20 cm respectively, and the centre is (60, 80), whereas those for the ellipse with the crosses are 10 cm, 35 cm and (100, 80). An object in $\mathrm{X_{te}}$ is located at (80, 80). Dash lines show the Euclidean distance between the new object and circles, whereas solid lines show the Euclidean distance between the new object and crosses with respect to the Single Linkage method shown in (a), the Complete Linkage method shown in (b) and the Average Linkage method shown in (c). Note that, regarding Average Linkage method, the distance between the new object and a cluster is the average of all lines. The "predicted" class for this object will be cross, circle and circle regarding the three linkage methods, respectively.

training clusters obtained from $C(\mathrm{X_{tr}}, 2)$. For demonstration purposes, these two clusters are, in fact, generated from two ellipses. The points, appearing in circle, are from an ellipse with diameter on the x-axis 5 unit, that on the y-axis 20 unit and the centre (60, 80). In contrast, the points, appearing in cross, are from an ellipse with diameter on the x-axis 10 unit, that on the y-axis 35 unit and the centre (100, 80). Assume the object in $\mathrm{X_{te}}$ whose data is located at (80, 80). The lines in the three graphs show the distances between the object and the two clusters, circle in dashed lines and cross in solid lines, for the three linkage methods. In Figure 5.2(c), the distance in the Average Linkage method is the average of all distances between the object and all objects in circle. Likewise, the distance between the object and the other cluster is calculated based on all objects. Finally, the "predicted" class of this object will be cross, when the Single Linkage method is applied as shown in Figure 5.2(a). Both of the "predicted" classes of the Complete Linkage and the Average Linkage are circle as shown in Figure 5.2(b) and Figure 5.2(c).

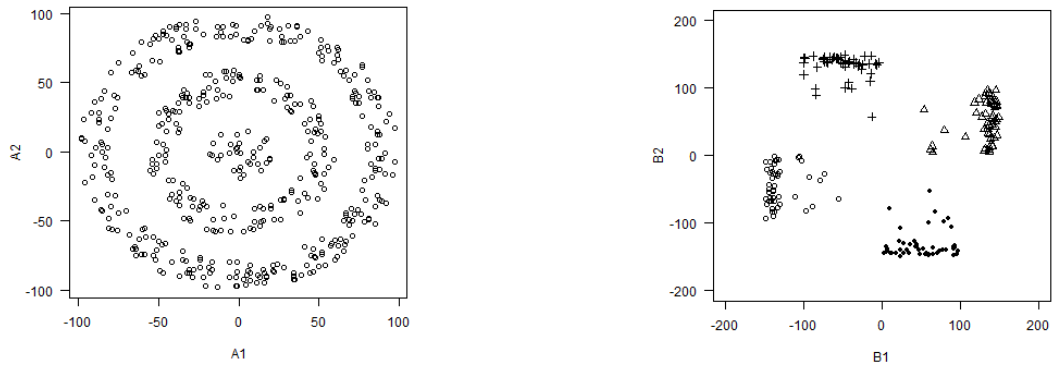
Here is an illustration of how the proposed three rules have a positive contribution to the Prediction Strength. Two datasets are simulated, denoted by dataset A and dataset B. We evaluate the efficiency of the rules by comparing the original average Prediction Strength with 100 repetitions for 2 to 9 clusters and the modified Prediction Strength with 100 repetitions for 2 to 9 clusters for the three linkage methods for the simulated dataset.

Dataset A

The dataset in which objects are randomly selected from three circles with radius 100, 60 and 20, respectively, consists of a total of 473 objects as shown in Figure 5.3(a).

Dataset B

The dataset contains four clusters. Each of which has 50 objects where they are randomly selected from a rectangle of which area equals to 100 times 100. Let two random variables B_1 and B_2 be the data for objects. The distributions for B_1 and B_2 with respect to each of the four clusters are as follows. For convenience, by $B \sim U(u_1, u_2)$, we mean that the random variable B follows a uniform distribution with minimum u_1



(a) (b)

Figure 5.3: Simulated datasets - (a)Dataset A contains 473 objects. Data for objects are generated from three ellipses with radius 100, 60 and 20, respectively. (b)Dataset B contains four clusters. Each cluster includes 50 objects, generated from a square.

and maximum u_2 from which a number b is generated.

$$
\begin{aligned}
\text{Cluster 1:} \quad & 40 \text{ objects with } B_1 \sim U(-150, -130.001), B_2 \sim U(-100, 0); \\
& 10 \text{ objects with } B_1 \sim U(-130, -50), B_2 \sim U(-100, 0) \\
\text{Cluster 2:} \quad & 10 \text{ objects with } B_1 \sim U(50, 130), B_2 \sim U(0, 100); \\
& 40 \text{ objects with } B_1 \sim U(130.001, 150), B_2 \sim U(0, 100) \\
\text{Cluster 3:} \quad & 10 \text{ objects with } B_1 \sim U(-100, 0), B_2 \sim U(50, 130); \\
& 40 \text{ objects with } B_1 \sim U(-100, 0), B_2 \sim U(130.001, 150) \\
\text{Cluster 4:} \quad & 40 \text{ objects with } B_1 \sim U(0, 100), B_2 \sim U(-150, -130.001); \\
& 10 \text{ objects with } B_1 \sim U(0, 100), B_2 \sim U(-130, -50)
\end{aligned}
$$

The data for the 200 objects is shown in Figure 5.3(b).

Both original PS and modified PS are applied on the dataset A. As seen from Figure 5.3(a), there are three circular arcs referring to the three subsets. Each of which is covered by another but three arcs are isolated. The Single Linkage is an ideal method for capturing elongated shapes, so we are most interested in the result of PS for the Single Linkage. Table 5.1 shows the original average PS and modified average PS with 100 repetitions for 2 to 9 clusters for the three linkage methods for dataset A. The first column is the numbers of clusters, followed by every two columns the values of the original average PS and modified PS for the Single Linkage, the Complete Linkage and the Average Linkage, respectively. With regards the Single Linkage, the original

average PS gives lower values than the modified average PS. Also, the original average PS and modified average PS for Single Linkage with three clusters is 0.35 and 0.86, respectively. According to the suggestion of Tibshirani and Walther [2005], namely to use the largest k in which the average PS is greater or equal to 0.8, the potential number of clusters for dataset A is three. The values of the Complete Linkage and that of the Average Linkage are all smaller than 0.7.

Both original PS and modified PS are applied on the dataset B of which the four subsets are separated from each other as shown in Figure 5.3(b). Table 5.2 shows the original average PS and modified average PS with 100 repetitions for 2 to 9 clusters for the three linkage methods. As seen, for each column, $k = 4$ has the highest score.

To sum up, the ASW and PS were proposed for determining the number of clusters by measuring the cluster stability and cluster coherence. We took into account the logic behind each clustering method and modified the PS. In the following chapters we use the term PS to refer to the modified PS. In addition, in the next chapter we use ASW and PS to compare the performance of the clustering methods on stability and coherence and let our data to decide which clustering method to use.

Table 5.1: The average Prediction Strength of dataset A - dataset A has three subsets. Each of which is covered by another but three arcs are isolated. The Single Linkage is an ideal method for capturing elongate shape, so we are most interested in the result of PS for the Single Linkage. For the Single Linkage method, the modified PS gives higher values than the original PS. Moreover, the modified average PS for the Single Linkage method for three clusters is 0.86, whereas the original average PS gives a value of 0.35.

Number of Clusters	Single Linkage		Complete Linkage		Average Linkage	
	Original	Modified	Original	Modified	Original	Modified
2	0.489	0.982	0.603	0.553	0.632	0.619
3	0.354	0.858	0.504	0.518	0.486	0.504
4	0.276	0.644	0.469	0.467	0.458	0.460
5	0.229	0.494	0.430	0.408	0.411	0.409
6	0.182	0.384	0.381	0.362	0.373	0.389
7	0.150	0.286	0.376	0.341	0.373	0.359
8	0.153	0.247	0.368	0.391	0.348	0.360
9	0.100	0.193	0.349	0.380	0.340	0.334

Table 5.2: The average Prediction Strength of dataset B - The dataset B contains four clusters, each of which has 50 objects generated from a square. The original average PS and modified average PS with 100 repetitions for 2 to 9 clusters for the three linkage methods are shown below. As seen, both original PS and modified PS for the three linkage methods show highest score at $k = 4$. The modified PS is preferred as it takes into account the logic behind each clustering method.

Number of Clusters	Single Linkage		Complete Linkage		Average Linkage	
	Original	Modified	Original	Modified	Original	Modified
2	0.568	0.584	0.713	0.733	0.638	0.638
3	0.542	0.543	0.577	0.493	0.529	0.556
4	0.943	0.917	0.980	0.977	0.976	0.981
5	0.239	0.291	0.536	0.551	0.319	0.339
6	0.053	0.062	0.392	0.391	0.159	0.127
7	0.004	0.006	0.323	0.322	0.081	0.087
8	0.000	0.005	0.294	0.282	0.062	0.052
9	0.000	0.000	0.242	0.215	0.031	0.036

Chapter 6

The clustering method and number of clusters for CO_{314}

In this Chapter we select values for the parameters of the p-dissimilarity (β and p), clustering method, and the number of clusters. Figure 6.1 shows the association between the decision making and the structure of this chapter. Section 6.1 discusses how to determine β, p and clustering method. The PAM method and the p-dissimilarity with $p = 0.6$ and $\beta = 1.42$ are selected. In Section 6.2 we propose a null model test. The null model test uses the null model and parametric bootstrap to investigate whether the clusters found according to PAM and the value of the indexes can be explained by random variation. Section 6.3.3 shows an application of the null model test for CO_{314}. Note that the term "average PS" refers to the "average modified PS" in this chapter.

6.1 Determination of β, p, and clustering method

6.1.1 Determination of β

The p-dissimilarity function is used to construct a proximity matrix to which the clustering methods will apply. It includes δ, p, α, and β (see Section 4.3.2). The p-dissimilarity between two participants is the sum of $(1 - p^{\alpha(t)})$ and $(1 - p^{\beta})$ over time. To use the p-dissimilarity, the values of β and the p need to be decided.

The parameter β relates to missing values. Suppose data for a participant r contains T records, denoted by $\mathbf{x}_r = [x_{r1}, \ldots, x_{rT}]$. For any two participants i and j, there

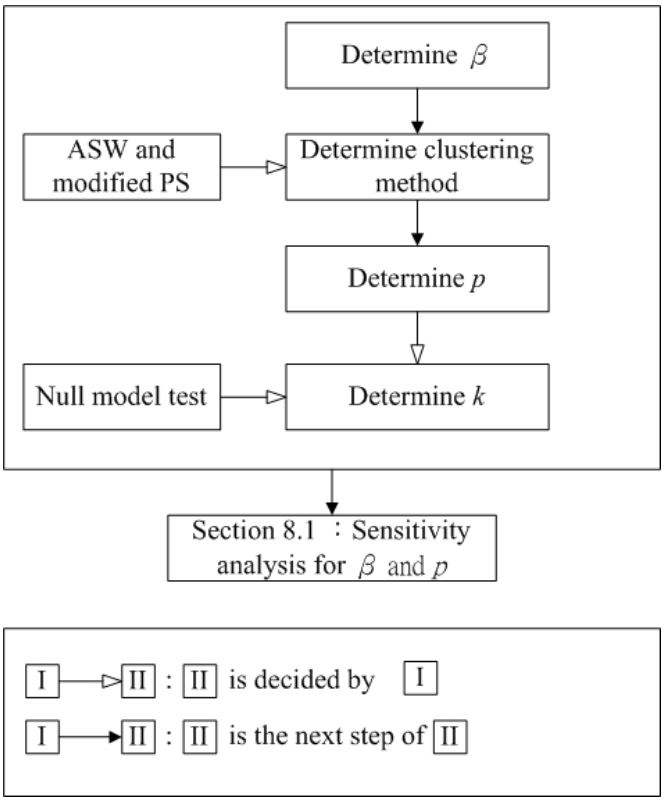


Figure 6.1: Process of decision making. -

Table 6.1: The frequency of α. - The dosage value is partitioned into six categories, so that the possible outcomes of α are between 0 and 5 as shown in the first column. The average of the values of α is 1.42.

α	Frequency	Relative frequency (%)
0	1641493	23.12
1	2590795	36.53
2	1619155	22.79
3	829515	11.65
4	338222	4.79
5	78915	1.11

are T pair differences in terms of their daily records, that is, $\{|x_{it} - x_{jt}| : t = 1, \ldots, T\}$. We categorized the T pair differences into "observed differences" α in which the t^{th} records of both participants are non-missing, and "semi-observed differences" in which at least one of the participants has their t^{th} record missing. The p-dissimilarity for an observed difference is $(1 - p^{\alpha(t)})$, while that for a semi-observed difference is $(1 - p^{\beta})$.

The process of selecting β was based on an assumption that if missing values were observed, values of the observed differences and values of the semi-observed differences followed the same distribution. So, for CO_{314}, we set the β to the average of all observed differences occurring in the dataset between different participants on the same day, that is, missing values were treated as "in average distance to everything".

Each pair of participants produced 180 differences according to their daily records. There were $\binom{314}{2}$ pairs of participants. One of which produced 90 observed differences, which was also the minimum numbers of observed differences. Around 80% of the $\binom{314}{2}$ times 180 differences were observed differences. Table 6.1 shows the frequency of α. The possible values for α range from 0 to 5 as the categories 1 to 6 represent observed dosages. The third column shows the relative frequency. The observed difference that is equal to 1 has the highest relative frequency, 36.53%, whereas the relative frequency of 2 is 22.79%. We set β equal to the mean of α, 1.42. $\beta = 1.42$ is applied to category-ordered data for 314 participants.

6.1.2 Determination of the clustering methods

We compared the clustering methods, namely the Single Linkage, the Complete Linkage, the Average Linkage and the PAM method, by their values of the modified PS (see Section 5.3.1) and their values of the ASW (see Section 5.2). The clustering method with higher values for 2 to 20 clusters would be used for CO_{314}. Note that the values of the modified PS and the values of the ASW are higher for the lower numbers of clusters. The number of repetitions of the modified PS was set to 30.

The seven graphs in Figure 6.2 show the average PS for the four clustering methods from 2 to 20 clusters with the combination of β=1.42 and $p = 0.2, 0.3, \ldots, 0.8$. The y-axes represent the value of the average PS. The x-axes represent the number of clusters. What can be observed is that, by and large, the PAM method has the higher values of the average PS up to 20 clusters. Similarly, the seven graphs in Figure 6.3 show the ASW for the four clustering methods from 2 to 20 clusters with the combination of β=1.42 and $p = 0.2, 0.3, \ldots, 0.8$. The y-axes represent the values of the ASW and the x-axes the number of clusters. Among the three linkage methods, the values of ASW for the Complete Linkage are higher. Comparing the values between the Complete Linkage and the PAM, the values of the ASW for the Complete Linkage are the highest for $k = 2$ and $k = 3$ and those for the PAM are higher in large numbers of clusters. Overall, PAM has higher values of the average PS and higher values of the ASW, the PAM method therefore was selected.

6.1.3 Determination of p

A large p means to consider data as approximately ordinal, while a small p means to consider data as approximately categorical. It was clear that $p = 0.9$ and $p = 0.1$ were not suitable for CO_{314} for the following reasons. The values for $(1 - 0.9^{\alpha})$ where $\alpha = 1, 2, 3, 4$ were 0.1, 0.19, 0.271, 0.3439. The dissimilarity between neighbouring categories did not seem capable to capture movements of dosage levels from one stage to another. The consequence of using $p = 0.9$ would be similar to the clustering result of using the Euclidean distance (See Section 4.4). Also, the values for $(1 - 0.1^{\alpha})$ where $\alpha = 1, 2, 3, 4$ were 0.9, 0.99, 0.999, 0.9999. The ordinality for categories does not seem to be well represented by these dissimilarities. Likewise, we were sceptical about $p = 0.8$

and $p = 0.2$. Suggestion from expertise in selecting a value of p given how sensitive the p-dissimilarity needs to be to separate categories, and the importance of whether CO_{314} should be considered as more ordinal or as more categorical would be helpful in practice. In our study, we determined p based on the result of previous section. We were aware of the issue that the determination of the β and p were subjective, so once the values for β and p were decided, we would then perform a sensitivity analysis with various values of p and β.

In previous section, we considered all combinations of the remaining choices of p and clustering methods. With regards the choice of p, the decision was made based on the values for the PAM method. Of the values of the average PS, the values with $p = 0.6$ were higher than the values with lower p; but, the values with $p = 0.6$ were similar to the values with higher p. Similar result for the values of the ASW with respect to $p = 0.6$ was also observed. We decided to use $p = 0.6$ because the values of indexes for $p = 0.6$ were similar to higher p and it meant to consider CO_{314} not to be fully categorical or ordinal.

Finally, the PAM method and the p-dissimilarity with $p = 0.6$ and $\beta = 1.42$ would be used for CO_{314}. Details on the stability analysis can be found in Section 8.1 from which we observe that p and β do not have a strong impact on performing a cluster analysis of the real data.

6.2 Null model test

6.2.1 Motivation

Let k denote the number of clusters. Most of the indexes for finding the number of clusters produce values for every $k > 1$, and yet only one k will be used. Which k to use is determined by the values of the index. For some indexes, the k which scores the highest value is used, while for other indexes, the first k with a value above a threshold is used and, for other indexes, the k for which there is a gap between its value and that of $(k + 1)$ is used (Milligan and Cooper [1985]). However, there is no systematic research about how the value changes when k changes. By and large, values for indexes are lower for the larger k. A larger k means that the homogeneity within clusters gets

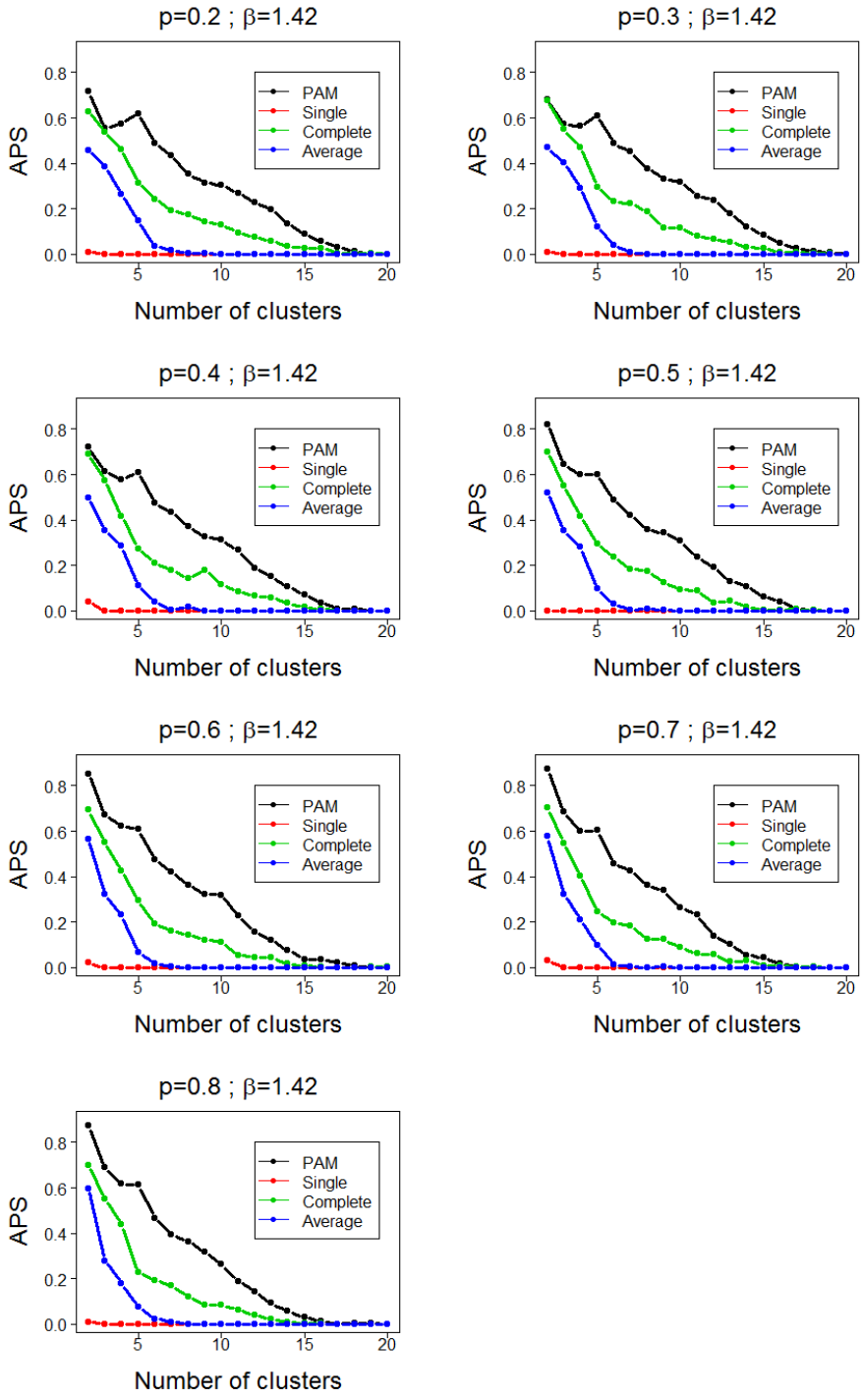


Figure 6.2: The average Prediction Strength for the four clustering methods for 2 to 20 clusters - The graphs show the average PS for the four clustering methods for 2 to 20 clusters with $\beta = 1.42$ and $p = 0.2, 0.3 \ldots, 0.8$. The y-axes represent the average PS with 30 random partitions. The x-axes represent the number of clusters. By and large, the PAM method has the higher average PS for 2 to 20 clusters.

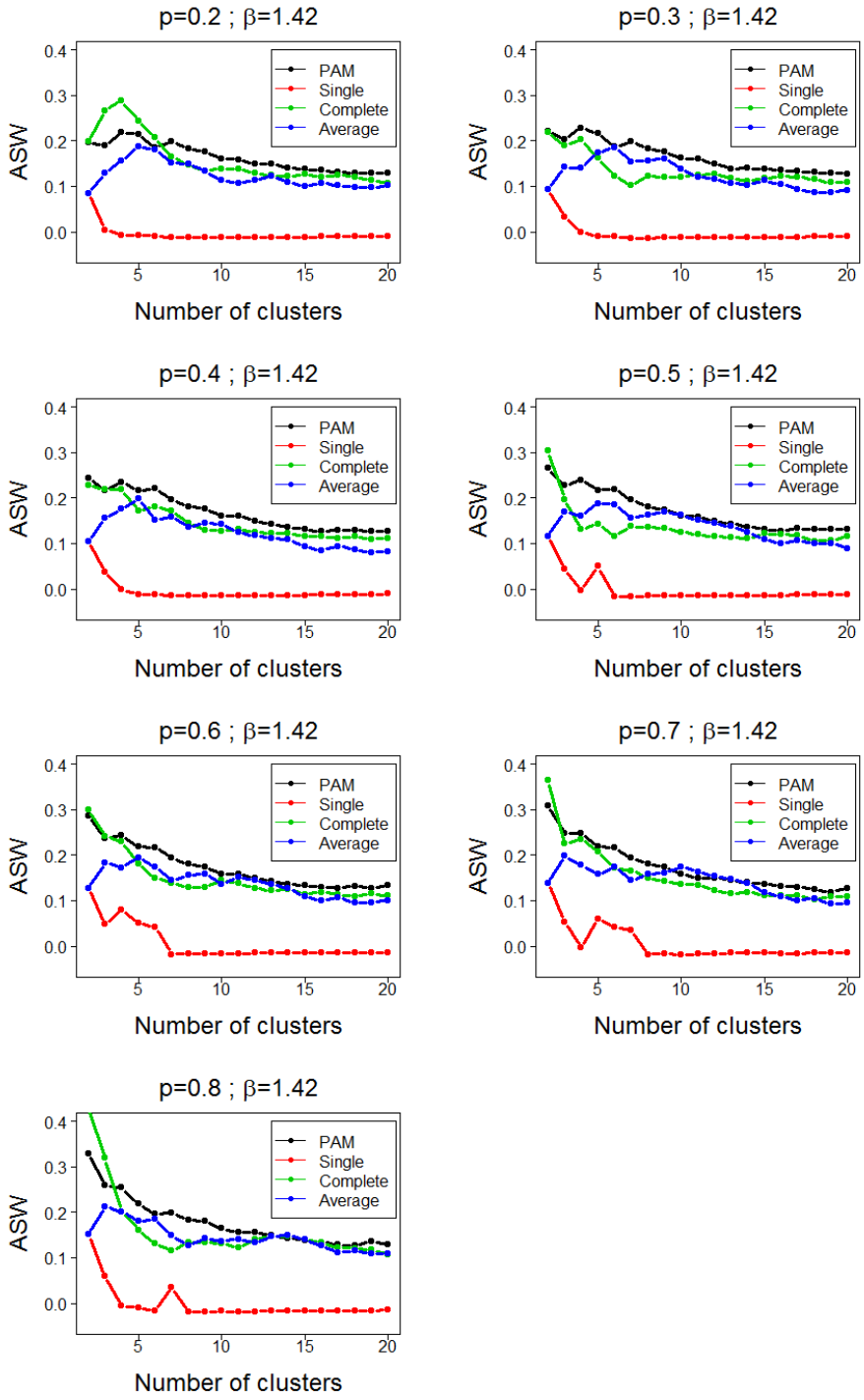


Figure 6.3: The Average Silhouette Width for the four clustering methods for 2 to 20 clusters. - The graphs show the Average Silhouette Width for the four clustering methods for 2 to 20 clusters with $\beta = 1.42$ and $p = 0.2, 0.3 \ldots, 0.8$. The y-axes represent the Average Silhouette Width and the x-axes the number of clusters. As can be seen, the values of the ASW of the PAM method are higher for large numbers of clusters and those of the Complete Linkage are the higher for $k = 2$ and $k = 3$.

better but the separation of clusters gets worse. The maximum k above a threshold does not indicate that a dataset indeed has k clusters.

Also, Tibshirani and Walther [2005] suggested to use the largest k in which the average Prediction Strength is greater or equal to 0.8 or above a user-specified threshold. But we notice that there is no combination of clustering methods and p for which the value of the average PS is higher than 0.8 when the number of clusters is greater than or equal to 3. To choose any value of k is, in a way, to assume that there are clusters in a dataset. Because the values of the average PS and the values of the ASW are rather low in our case, we wonder our dataset cannot achieve 0.8 because there is no clear cluster or the 0.8 is too high. We wonder if there is a way of using the values of the indexes, which can be backed up with a rationale, which can be used to determine the number of clusters and test the existence of clustering structure.

Jain and Dubes [1988] discussed validation of hierarchical structure obtained from a hierarchical clustering method. Indexes listed in the book used rank correlations to compare a given proximity matrix of hierarchical structure and a proximity matrix of random partitions. Note that most indexes depend on the data type, the number of clusters, the type of hierarchical clustering method used. Also, Bock [1985, 1996] studied several significance tests for homogeneous population and an alternative involving clustering or heterogeneity. However, all of these do not take into account structure in the data that is not from clustering, such as time series/Markov structure, missing values, etc. Hennig and Liao [2013] took account of data structure and concerned about selecting k with the highest value of the ASW. An idea of null assumption in their study was in line with that used by Buja et al. [2009]. In the research of Buja et al. [2009], the authors illustrated a comparison between real data and reference datasets that were simulated by a null assumption. In the research of Hennig and Liao [2013], under a null assumption that there were no clusters, a model was built to represent real data; they called this model a null model. The null model included features that were the same as those of the real data. But the structure of clusters in the null model was absent and it was unknown whether there existed clusters in the real data. Their purpose was to test the homogeneity of the real data by comparing it to the null model. They generated reference datasets from the null model and applied the ASW to the

real data and to all simulated reference datasets. At each k, the distribution of the values of the ASW of the null model without any clustering structure was constructed by the values of the ASW of the reference datasets. They found that the k that scored the highest ASW in the real data did not necessarily have a higher value than most of the values of the same k under the null model, whereas some other k had their values of ASW higher than those of the null model.

Therefore, our next objective is to build on the average PS and the ASW, and offer a rationale for determining k by considering the existence of a clustering structure in a dataset. We propose a null model test for investigating whether the found number of clusters can be explained by random variation and to use it as a rationale for determining the number of clusters.

6.2.2 Proposed null model test

We attempt to construct a null model test to test if the dataset is homogeneous, that is, there is no real clusters exist. Because the PS and ASW are used to measure the quality of clustering, we use them as test statistics for the hypothesis test. Also, the value of an index depends on k and clustering methods *Method*. The distribution of the test statistic, therefore, depends on a fixed k and a specific *Method*. The observed test statistic of the real dataset is the value of the index of the real dataset and the distribution of the test statistic is the values of the index under the null assumption. The null model test for every fixed k of interest is performed by comparing the observed test statistic with the distribution of the test statistic. It is used to explore k. Moreover, we define a single test of the homogeneity hypothesis against a clustering alternative by aggregating the test results for different k.

To construct an exact distribution of the test statistic is sometimes impossible. Alternatively, we can build a model from which a set of reference datasets are drawn. Then, the values of average PS or the values of the ASW for a set of reference datasets can be used to explore the distribution of a statistic. Regarding sampling methods, we consider the following. Firstly, the non-parametric boostrap, in which a sample with the same size as in a real dataset is drawn from an empirical distribution with replacement; secondly, the parametric boostrap, in which a sample is drawn from a

model with estimated parameters constructed from the real dataset. As for the Monte Carlo method, the sample is drawn from a model with fixed values of the parameters (Davison and Hinkley [1997]).

The use of the null model in the test is that it fits all non-clustering aspects of the real dataset, such as relationships between variables, time dependency, marginal distributions and etc. The null model should relate to the real dataset. The non-parametric and parametric boostrap are thus considered. The null assumption is that there are no clusters. The non-parametric boostrap gives reference datasets in which the clustering structure remains the same as the real dataset, so the distribution of the test statistic cannot be used to compare with the observed statistic. The parametric boostrap which gives reference datasets without clustering structure is thus considered.

The null model test uses the null model and the parametric boostrap to obtain the distribution of a statistic on the basis of an index. We define the test statistic and the p-value for every fixed k and *Method* as follows.

Test statistic

Under the null assumption, the distribution of the test statistic for a fixed k and a specific clustering method (*Method*) is estimated as follows. Denote the test statistic for a real dataset for a fixed k and a specific clustering method *Method* by s. It is the value for an index. Denote the distribution of the test statistic under the null hypothesis by S. It is estimated from the values of the index for a set of reference datasets drawn from a null model. Assume R reference datasets are simulated. Their values of the index for k and *Method* are denoted by $s_1, s_2, \ldots, s_R$, which are used to estimate S.

P-value

The p-value is defined as a probability that is used to measure the level of evidence against the null hypothesis, that is, under the null assumption, the probability of a new dataset having its test statistic greater than s. In the null model test, the observed test statistic s is compared with $s_1, s_2, \ldots, s_R$. If exactly a of the simulated values are greater than s, then the approximate p-value of the significance test is defined by

(Davison and Hinkley [1997])

$$\text{p-value}_k = \frac{a+1}{R+1}. \tag{6.1}$$

Because the test statistics $s_1, s_2, \ldots, s_R$ are discrete, a value 1 is added to the denominator and the numerator in order to avoid obtaining a zero probability.

The above test uses the value of index to explore k by testing whether a dataset has a specific number of clusters. The k for which p-value$_k$ is less than a significance level is identified to be a potential number of clusters.

Test for homogeneity

Another purpose for performing the null model test is to test the existence of clustering structure. We attempt to carry out this test by using a p-value$_H$ which summarizes the information on the results of several tests for different k. The null model test for homogeneity works as follows.

Step 1: denote the test statistic for the i^{th} simulated dataset for j cluster by s_{ij}. Denote the observed statistic of the real data for j cluster by s_j. We summarize the test statistics for the simulated R datasets and the observed statistic for 1 to k clusters by the following matrix,

$$\begin{bmatrix} s_{11} & \ldots & s_{1k} \\ \vdots & \ddots & \vdots \\ s_{R1} & \ldots & s_{Rk} \\ s_1 & \ldots & s_k \end{bmatrix}.$$

Step 2: for j clusters in columns, assign ranks to the values $\{s_{(1j)}, \ldots, s_{(Rj)}, s_i\}$. Denote the ranks by $\{r_{1j}, \ldots, r_{Rj}, r_i\}$. The reason for taking rank transformation is to alleviate effects caused by large values. Then, the above matrix becomes

$$\begin{bmatrix} r_{11} & \ldots & r_{1k} \\ \vdots & \ddots & \vdots \\ r_{R1} & \ldots & r_{Rk} \\ r_1 & \ldots & r_k \end{bmatrix}.$$

Step 3: for dataset i in rows, the average of $\{r_{i1}, \ldots, r_{ik}\}$ is computed, denoted by $\overline{r_{i.}}$, $i = 1, \ldots, R$. Also, the average of the rank values for the real dataset $\{r_1, \ldots, r_k\}$

are computed, denoted by $\overline{r}^*$.

Step 4: the distribution of the test statistic for the null model test for homogeneity is estimated by $\overline{r_{1.}}, \overline{r_{2.}}, \ldots, \overline{r_{R.}}$. The observed statistic is $\overline{r}^*$. If exactly a of $\overline{r_{1.}}, \overline{r_{2.}}, \ldots, \overline{r_{R.}}$ are greater than $\overline{r}^*$, then the approximate p-value of this test is defined by

$$\text{p-value}_H = \frac{a+1}{R+1}. \tag{6.2}$$

There are more possibilities of defining p-value$_H$, such as, the grade for each simulated dataset can be the average of the values produced by the indexes. But the p-value$_H$ might be dominated by some k in which the s_k are relatively small. Or, defining p-value$_H$ as the p-value$_k$ which scores the smallest, but an adjustment of p-values for multiple comparisons needs to be taken into account. Also, the disadvantage of using the smallest p-value$_k$ is that more reference datasets are required in order to reach a good accuracy. By accuracy, we mean the following. The critical value with Bonferroni correction is approximated by $\frac{\alpha}{k}$ where α is the significant level. In a null model test with $\alpha = 0.01$ and $k = 2, \ldots, 20$, the critical value for each k is $\frac{0.01}{19} = 0.0005$. We will need a lot of reference datasets in order to possibly have a k with p-value$_k = 0.0005$.

6.3 Application of the null model test to CO_{314}

There are some useful models that can be used to model a random variable that changes over time (Shumway and Stoffer [2010]), such as the Autoregressive model (AR), the Moving Average model (MA) and the Markov model. In the AR model, the current state can be estimated by a linear weighted sum of previous states. The weights are the auto regression coefficients. In the MA model, the current state can be estimated by a linear weighted sum of current and previous errors. A first-order Markov model follows a Markov property in which the next state depends on only the current state but not on the sequence of previous states. Let the random variable C be the state and Θ be the state space, so that $C \in \Theta$. For the first-order Markov model, the production of any sequence can be described by transition probabilities. The transition probability of C_t being in state j, given that C_{t-1} is in state i, can be written as

$$P_{ij(t)} = P(C_t = j | C_1 = c_1, C_2 = c_2, \ldots, C_{t-1} = i) = P(C_t = j | C_{t-1} = i). \tag{6.3}$$

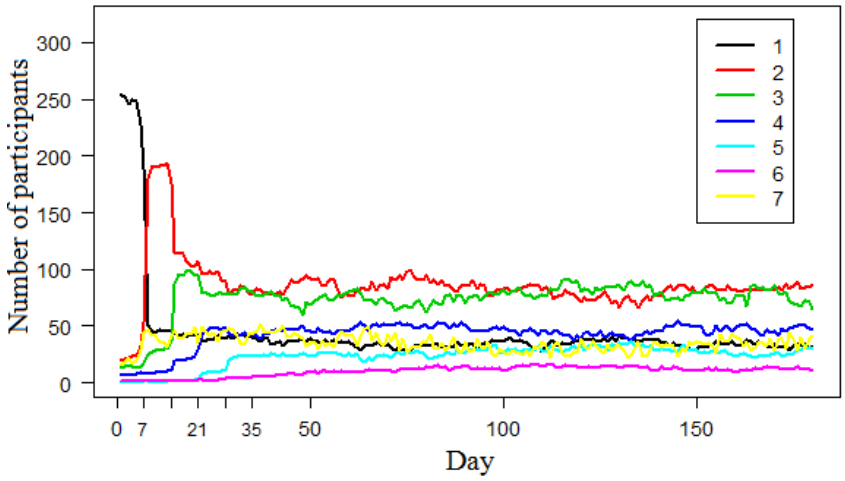


Figure 6.4: Distributions of the number of participants in the seven categories. - The y-axis represents the number of participants and the x-axis the days. The colours designate the seven categories. The numbers of participants slightly change within periods of seven days but dramatically change at the beginning of the next 7-day period.

We attempt to model the distributions for categories in CO_{314} by a Markov chain because of

- the first weekly prescription: as a result of 20 mg for participants who had no experience on methadone. There are more changes in the first week than later.
- weekly prescriptions: dosage level is more stable within a 7 days. Most of the sudden changes in dosage level happen at the beginning of a weekly prescriptions.

Also, we notice that most of the participants stayed in the same category on the next day within 7 days. Of those who moved to other categories, most of them moved to the neighbouring categories. With the limited number of participants in CO_{314}, we will estimate the parameters for the Markov model by the relative frequency in CO_{314}. Also, we will treat the beginning of prescription and a period of 6 days separately.

6.3.1 Exploration of movements of categories in CO_{314}

Figure 6.4 shows the distributions of the numbers of participants of the seven categories over 180 days. The vertical axis are the number of participants and the horizontal axis the days. The colours designate the seven categories, corresponding to the aforementioned dosage values. What can be observed is that the number of participants in category 1 remains steady over the first seven days, with an average of 238.29 participants. Subsequently, on day 8, the number of participants in category 1 plunges to 51, while the number in category 2 rockets, reaching to 180, with an average of 27.71 participants in the first week. For both categories, the number of participants during the second week is relatively stable, on day 15, the numbers drop to 44 and 115 for categories 1 and 2, respectively. In contrast, on the same day, the number of participants in category 3 sees a sudden increase proceeded by a marginally incremental trend in the first two weeks. As for category 4, the number climbs steadily during the first three weeks, but virtually doubles on day 22. There is also an upward trend in categories 5 and 6, while, for category 7, the number fluctuates within a narrow margin.

We discovered associations between medical decisions and the distributions of these numbers. The initial prescription dosage for participants, most of whom had no previous experience of MMT, was 20 mg (category 1). This resulted in most of the 314 participants having their dosage in category 1 from day 1 to day 7. Also, in all categories except category 7, the numbers of participants changed slightly within each seven day periods but changed dramatically at the beginning of the next 7-day period. This is consistent with weekly prescriptions. Regarding the number of participants in category 7, those who were determined to quit heroin would try to reduce the methadone dosage, and then try to quit methadone, so their records of methadone appeared to be 0 mg. On the other hand, those who abused heroin would not need to take methadone to accommodate their addiction, and their records of methadone remained 0 mg as well. These made to model records of participants in category 7 far more complicated than to model records of participants in categories 1 to 6. Therefore, we focused on the categories 1 to 6 first. Given that category 7 was excluded on day t, we estimated the probability of transitioning from category i to category j on day t by calculating the relative frequency, defined by

$$RF_{ij}^{(t)} = \frac{\#\{\text{category } j \text{ on day } t, \text{ given that category } i \text{ on day } (t-1)\ \}}{\#\{\text{category } i \text{ on day } (t-1)\} - \#\{\text{category 7 on day } t\}} \tag{6.4}$$

where $i, j = 1, 2, \ldots, 6$; $t = 2, \ldots, 180$, and $\#\{\text{category } i \text{ on day } (t-1)\}$ denotes the number of participants whose records are category i on day $(t-1)$.

Table 6.2 shows the RF from category 1 to categories 1 to 6 in a single day from day 2 to day 23. Of $\{RF_{11}^{(t)} : t = 2, \ldots, 23\}$, four are below 90%, namely $RF_{11}^{(8)} = 26.09\%$, $RF_{11}^{(9)} = 85.71\%$, $RF_{11}^{(15)} = 86.05\%$ and $RF_{11}^{(22)} = 87.18\%$. As seen, of those whose records are in category 1 on day 7, 26.09% stay in the category 1, while 70.65% move to the category 2. Similarly, Table 6.3 shows $\{RF_{2j}^{(t)} : j = 1, \ldots, 6; t = 2, \ldots, 23\}$. The RF for category 2 to itself which are below 90 % are $RF_{22}^{(8)} = 88.37\%$, $RF_{22}^{(15)} = 63.03\%$ and $RF_{22}^{(22)} = 88.78\%$. Of those whose records are in category 2 on day 14, 63.03% stay in the category 2, whereas 31.52% increase their dosage to category 3. Likewise, there is a similar pattern in $\{RF_{33}^{(t)} : t = 2, \ldots, 23\}$, of which values smaller than 90% occur on day 8, 15 and 22. In addition, most of participants change their dosage to either category 2 or category 4, which are the neighbouring categories of the category 3. Evidences are $RF_{33}^{(8)} = 82.36\%$, $RF_{32}^{(8)} = 11.76\%$, $RF_{33}^{(15)} = 81.08\%$, $RF_{34}^{(15)} = 18.92\%$, $RF_{33}^{(22)} = 74.45\%$, and $RF_{34}^{(22)} = 23.33\%$. In comparison to the RF from category 3 to category 4 on other days, $RF_{34}^{(22)} = 23.33\%$ is high.

To sum up, many participants changed their dosages to another level between days 7 and 8; between days 14 and 15; between days 21 and 22. The transition probability fluctuates on days 8, 15 and 22 and it is approximately constant on day $t \in \Theta = \{1, 2, \ldots, 23\} \setminus \{8, 15, 22\}$.

Next, we attempt to formalise and then simplify the relative frequencies. Note that only categories 1 to 6 are considered for the moment. We knew that the prescription dosage constrained movements between categories within any seven day period and most of the noticeable changes happened on the first day of a new prescription. Those first days were the common multiple of the integer 7 plus 1 day. Therefore, these seven days of a prescription can be divided into two parts, one being the beginning day on which the noticeable changes happened, and the other being the other six days in which

the relative frequencies are approximately constant. We defined two sets of days, ψ_1 and ψ_2 as follows.

- ψ_1: we use a subscript d to refer to the first day of a new prescription plus 1, $T_d = \{\text{days } d, (d+1), (d+2), \ldots, (d+5) : d = 2, 9, 16, \ldots, 170\}$ and $T_{177} = \{\text{days } 177, 178, 179, 180\}$, so $\psi_1 = \{T_d : d = 2, 9, \ldots, 177\}$.
- ψ_2: the set of the first days of new prescriptions. $\psi_2 = \{\text{days } 8, 15, \ldots, 169, 176\}$

Because the relative frequencies for every six days in T_d were approximately constant, we assumed that the relative frequencies for days in T_d were equal and defined the weekly average relative frequency by

$$ARP_{ij}^{(T_d)} = \frac{1}{\text{number of days in } T_d} \sum_{t \in T_d} RF_{ij}^{(t)}$$

where $i, j = 1, 2, \ldots, 6$ and t is the day of observations. For instance, $T_2 = \{\text{days } 2, 3, \ldots, 7\}$, $ARP_{12}^{(T_2)} = \frac{1}{6}(RF_{12}^{(2)} + RF_{12}^{(3)} + RF_{12}^{(4)} + RF_{12}^{(5)} + RF_{12}^{(6)} + RF_{12}^{(7)})$.

The figure on the top left in Figure 6.5 illustrates the ARP from category 1 to all six categories over ψ_1. The vertical axes represent the ARP and the horizontal axes represent ψ_1. The colours indicate the next possible states, where transitions from category 1 to categories 1, 2, 3, 4, 5 and 6, appear as black, red, green, blue, cyan and purple, respectively. The ARP from category 1 to itself, ARP_{11}, appearing in black, fluctuates around an average of 98.44% whereas that from category 1 to category 2, ARP_{12}, appearing in red, remains in the margin of 1%. The ARP from category 1 to categories 3, 4, 5 and 6 are also plotted, but those in those cases the lines overlap. Consequently, only one line appears in cyan. The figure on the top right represents the ARP from category 2 to all six categories. ARP_{22}, appearing in red, maintains a high level at 98.2%, while there are negligible changes in those from category 2 to the other five categories. Similarly, the four graphs from the middle left to the bottom right show the ARP from categories 3, 4, 5 and 6 to all six categories, respectively. It can be seen that, by and large, the ARP from categories to themselves, ARP_{ii}, $i = 1, 2, \ldots, 6$, fluctuate within a narrow margin around 97%. In ψ_1, participants are more likely to have their dosages in the same category.

As for ψ_2, because there were dramatic changes in terms of their RF on days in ψ_2, we computed the relative frequencies for every single day. In total, there were twenty-five days. Each of the six graphs in Figure 6.6 illustrates the relative frequencies from a category to categories on days in ψ_2. The vertical axes represent the relative frequencies and the horizontal axes represent days in ψ_2. The colours black, red, green, blue, cyan and purple indicate the next possible states. The graph on the left in the first row in Figure 6.6 shows the relative frequency from category 1 to the six categories. We can see that the relative frequency from category 1 to itself, appearing in black, soars from $RF_{11}^{(8)} = 0.26$ to $RF_{11}^{(15)} = 0.86$ and continues with an upward trend. In contrast, the relative frequency from category 1 to category 2, appearing in red, plunges from $RF_{12}^{(8)} = 0.71$ to $RF_{12}^{(15)} = 0.14$ and carries on downward, hitting a low of 0. The relative frequencies from category 1 to categories 3, 4, 5 and 6, appearing in cyan are also plotted, but, once again, the lines overlap. Next, the graph on the right in the first row shows the relative frequencies of from category 2 to all six categories on days in ψ_2. The relative frequencies from category 2 to category 1 fluctuate around 2%. However, those from category 2 to itself and to category 3 fluctuate widely in the opposite direction, for example, while the $RF_{22}^{(15)}$ falls to 0.63, the $RF_{23}^{(15)}$ peaks at 0.31 on day 15. Similarly, the four graphs from the graph on the left in the second row to that on the right in the third row show the relative frequencies from categories 3, 4, 5 and 6 to all six categories, respectively. Note that less than ten participants have their dosages in category 5 or category 6 in days $1, 2, \ldots 28$.

The pattern found in the Figure 6.6 for ψ_1 can be summarized as follows: firstly, on day 8, many participants have their dosage moved from categories 1 to 2. Then, on day 15, many participants have their dosage moved from categories 2 to 3. Next, on day 22, the dosage of a group of people move from categories 3 to 4. On day 29, some of the participants increase their dosage from categories 4 to 5 and on days 78, 85, 92 and 99, of those participants who move from category 4, most move to category 3. A similar phenomenon can be observed in the graphs of the relative frequencies from categories 5 and 6, where, of those participants who move from categories 5 and 6 respectively, most move to categories 4 and 5 respectively. This suggests that after three months treatment, some participants start to show some positive outcome of the decrease of their methadone dosage.

Table 6.2: Relative frequencies (%) from category 1 to categories 1, 2, 3, 4, 5 and 6 over 22 days.-Of these participants who have their dosage changed, most increase their dosage to category 2. The relative frequencies $RF_{11}^{(t)}$, $t = 2, \ldots, 23$ are rather stable except on days 8, 15 and 22.

Day \ category	1	2	3	4	5	6
2	100	0	0	0	0	0
3	99.15	0.85	0	0	0	0
4	99.57	0.43	0	0	0	0
5	99.58	0.42	0	0	0	0
6	97.37	1.75	0.88	0	0	0
7	92.2	7.31	0.49	0	0	0
8	26.09	70.65	3.26	0	0	0
9	85.71	14.29	0	0	0	0
10	100	0	0	0	0	0
11	100	0	0	0	0	0
12	100	0	0	0	0	0
13	100	0	0	0	0	0
14	95.45	4.55	0	0	0	0
15	86.05	13.95	0	0	0	0
16	97.62	2.38	0	0	0	0
17	100	0	0	0	0	0
18	100	0	0	0	0	0
19	100	0	0	0	0	0
20	100	0	0	0	0	0
21	97.22	2.78	0	0	0	0
22	87.18	12.82	0	0	0	0
23	100	0	0	0	0	0

Table 6.3: Relative frequencies (%) from category 2 to categories 1, 2, 3, 4, 5 and 6 over 22 days.-The relative frequencies $RF_{22}^{(t)}$, $t = 2, \ldots, 23$ are higher than 90 %, except on days 8, 15 and 22. Of those participants who have their dosage changed, most of them move to either category 1 or category 3.

Day \ category	1	2	3	4	5	6
2	0	100	0	0	0	0
3	0	100	0	0	0	0
4	0	100	0	0	0	0
5	0	100	0	0	0	0
6	0	100	0	0	0	0
7	0	96.67	3.33	0	0	0
8	2.33	88.37	9.3	0	0	0
9	0.58	99.42	0	0	0	0
10	0	98.37	1.63	0	0	0
11	0	100	0	0	0	0
12	0	99.46	0.54	0	0	0
13	0	98.92	0.54	0.54	0	0
14	0	94.89	5.11	0	0	0
15	4.24	63.03	31.52	1.21	0	0
16	0	92.66	7.34	0	0	0
17	0	98.17	1.83	0	0	0
18	0	100	0	0	0	0
19	0	98.04	1.96	0	0	0
20	0	100	0	0	0	0
21	1.06	95.74	3.2	0	0	0
22	1.02	88.78	10.2	0	0	0
23	1.09	97.82	1.09	0	0	0

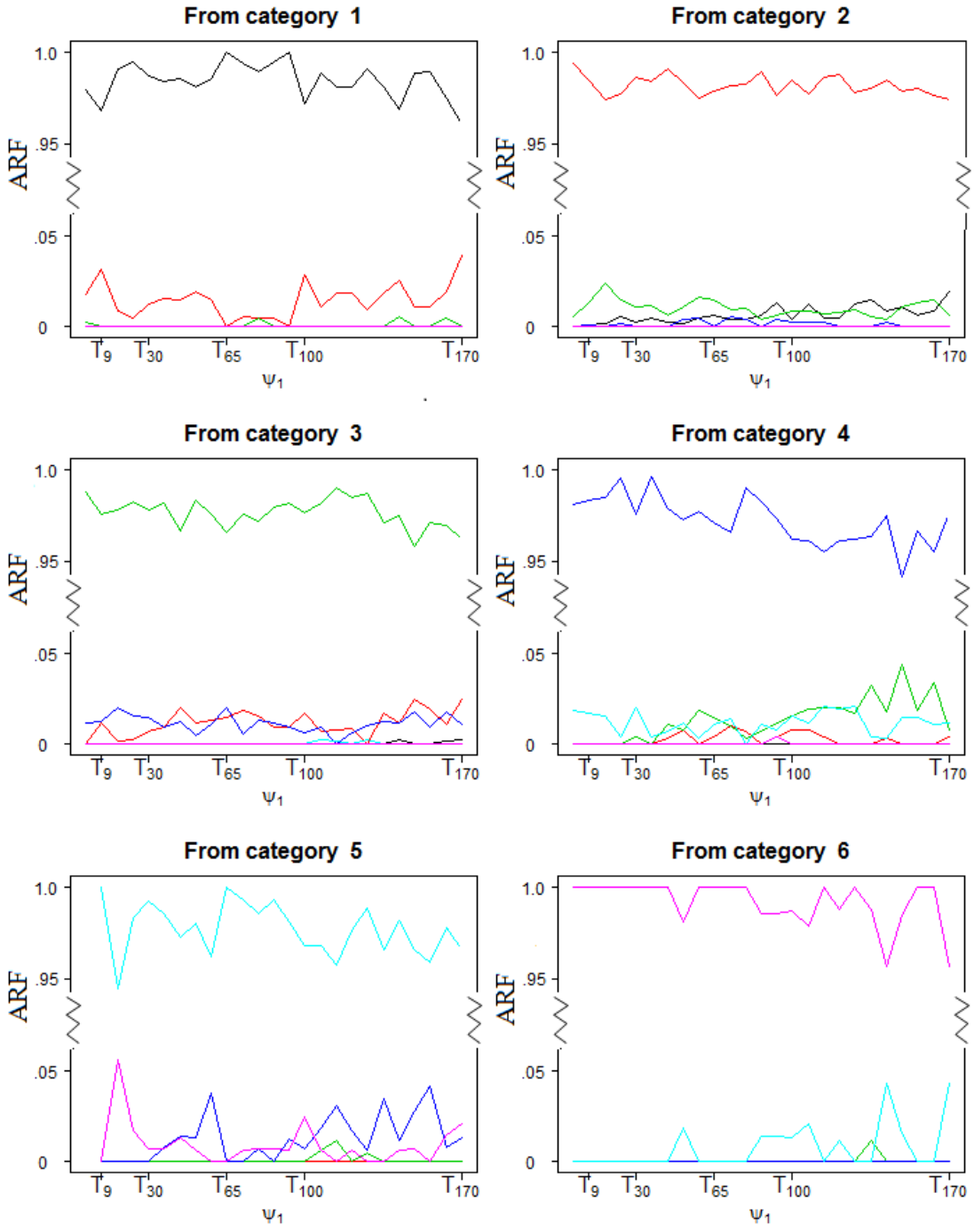


Figure 6.5: The average relative frequencies of ψ_1 - $\psi_1 = \{T_2, T_9, \ldots, T_{177}\}$ where $T_d = \{\text{days } d, (d+1), (d+2), \ldots, (d+5) : d = 2, 9, 16, \ldots, 170\}$ and $T_{177} = \{\text{days } 177, 178, 179, 180\}$. The six graphs from the top left to the bottom right show the average relative frequencies from categories 1, 2, 3, 4, 5 and 6 to all six categories, coloured in black, red, green, blue, cyan and purple, respectively. The $ARF_{ij}^{(T_d)}$ is the average of the relative frequencies from category i to category j on days in T_d. Overall, the ARF from categories to themselves, ARF_{ii}, fluctuates within a narrow margin, around 98 %.

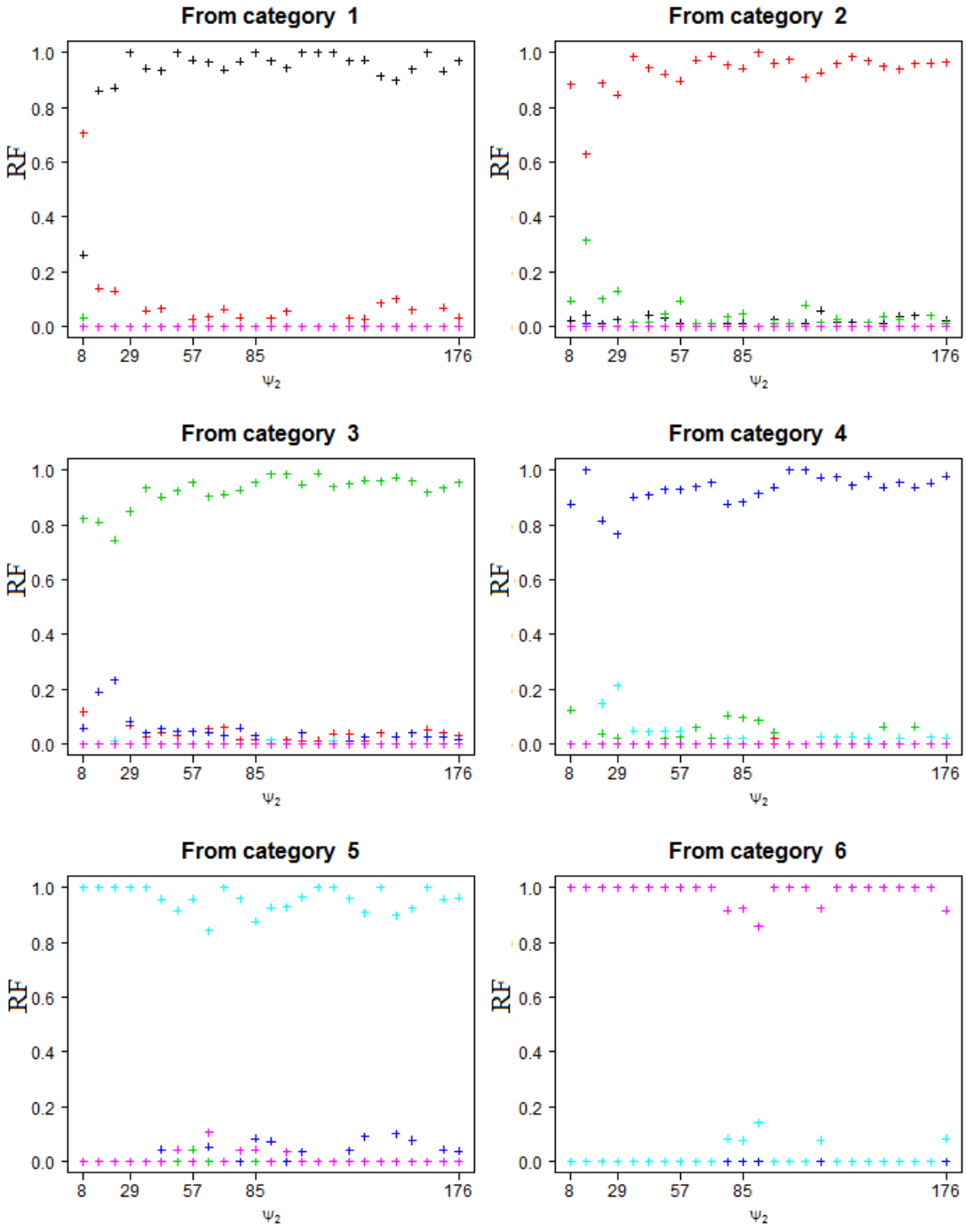


Figure 6.6: The relative frequency of ψ_2. - The six graphs from the top left to the bottom right show the relative frequencies from categories 1, 2, 3, 4, 5 and 6 to all six categories. The y-axes represent relative frequencies and the x-axes represent $\psi_2 = \{\text{days } 8, 15, 22, \ldots, 176\}$. The colours indicate the categories, black, red, green, blue, cyan and purple, respectively.

6.3.2 The null model for CO_{314}

The parameters for the Markov model were estimated by relative frequencies. As for category 7, it did not appear to be representable by relative frequency. Figure 6.7 shows the relative frequencies from category 7 to categories 1 to 7. As can be seen, because the prescribed dosages for the first prescriptions of most of the participants were in category 1, the relative frequencies for category 1 for the first seven days are higher. Modeling category 7 with other six categories will result in participants has dosage in category i on day t but has dosage in category j, which is a distant-neighbouring category, on day $(t+1)$. This contradicted the fact that the valid prescribed dosage might be below category j. The category 7 should not be included in the Markov model. A simplified solution for ensuring that the distribution of category 7 in the reference datasets would be the same as that in CO_{314} was

- to generate a reference dataset from the model. Note that participants in the reference dataset were in a random order.
- to order participants in CO_{314} by the date they commenced MMT.
- to plug in the patterns of category 7 of CO_{314} into the reference dataset.

This meant that we identified the days on which category 7 appeared. Then the t^{th} record of the $i^{th}, i = 1, \ldots, 314$ participant in the reference datasets would be replaced by category 7 if the t^{th} record of the i^{th} participant in CO_{314} was category 7.

We considered the Markov model for categories 1 to 6 for ψ_1 and ψ_2 separately. For days in ψ_1, the relative frequencies from category i to category j fluctuated in a very narrow margin. For this reason, we assumed that the transition probability from category i to category j was a constant from day 2 to day 180 except for days in ψ_2. So, the values of categories for days in ψ_1 would be generated from a stationary Markov model with the estimated transition probabilities obtained from aggregating all the relative frequencies and dividing by the total observing days. The estimated transition probabilities for days in ψ_1 is defined by

$$ETP_{ij}^{(\psi_1)} = \frac{1}{\text{number of days in } \psi_1} \sum_{t \in \psi_1} RF_{ij}^{(t)}$$

Table 6.4: The estimated transition probabilities matrix of ψ_1.-The first column displays the current state. The remaining columns show the estimated transition probabilities from the current state to the next states of ψ_1. Note that the probability is estimated by using Eq 6.4. The estimated transition probabilities from categories to themselves are all above 97 %.

From Category \ To Category	1	2	3	4	5	6
1	98.47	1.46	0.07	0	0	0
2	0.65	98.21	1.01	0.13	0	0
3	0.03	1.16	97.65	1.14	0.02	0
4	0	0.25	1.33	97.25	1.16	0.01
5	0	0	0.1	1.27	97.72	0.91
6	0	0	0.048	0	0.772	99.18

where $i, j = 1, 2, \ldots, 6$. Table 6.4 displays the estimated transition probabilities matrix for ψ_1. The first column stands for the current state. The remaining columns show the estimated transition probabilities from the given state to the next states. As seen, the estimated transition probabilities from categories to themselves are all above 97%.

The relative frequencies for ψ_2 varied from day to day (see Figure 6.6). Therefore, the category for the days in ψ_2 would be generated from a Markov model with the transition probability estimated by $RF_{ij}^{(t)}$, $t \in \psi_2$.

The number of participants of CO_{314} was 314 and the proportions of categories 1, 2, 3, 4, 5 and 6 of CO_{314} on day 1 were 85.5%, 6.7%, 4.3%, 2.7%, 0.1% and 0.7%, respectively, from which the initial states of 314 observations on day 1 were generated. Then, the states from day 2 to day 180 were generated from the aforementioned Markov model. Next, the pattern of category 7 in CO_{314} was plugged into the reference dataset. Finally, the marginal distributions of the categories 1 to 6 and the distribution of the missing values in the reference datasets were the same as in CO_{314}.

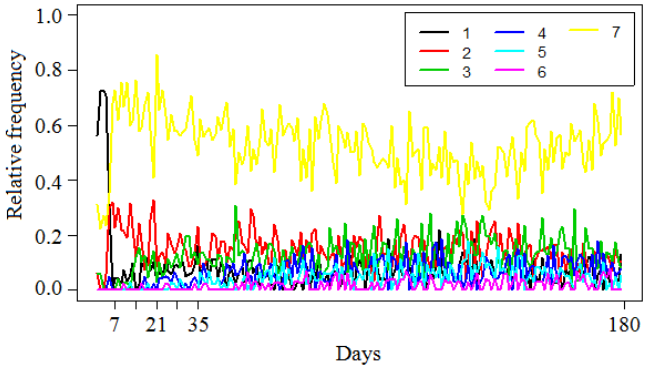


Figure 6.7: Relative frequencies from category 7 to categories 1 to 7. - The y-axis represents the relative frequencies and the x-axis represents the day. The colour designates the seven categories.

6.3.3 Determination of the number of clusters

We performed null model tests for CO_{314}. A total of 1000 reference datasets was generated from the null model and then the p-dissimilarity with β=1.42 and p=0.6 were applied to these reference datasets. We used the PS and ASW for 2 to 20 clusters for a specific clustering method. Because method of PS was computationally heavy and the PAM was selected for CO_{314}, we then used PS for PAM, and we used ASW for PAM, the Complete Linkage and the Average Linkage methods.

We applied the average PS with 50 repetitions for the PAM method with 2 to 20 clusters. Figure 6.8 shows the values of the average PS. The red line refers to the observed statistic and the black lines refer to the values for the 1000 reference datasets. A number of observations can be made about the values of the average PS of CO_{314}. Firstly, they decrease from 2 to 4 clusters. Secondly, they are above all the values of the reference datasets from 5 to 10 clusters but below most of them for higher numbers of clusters. Thirdly, they are zero for k= 17, 19 and 20.

The p-value$_k$ for CO_{314} for between 2 to 20 clusters is shown in Figure 6.9. To make the p-value for between 2 to 20 clusters easier to read from the figure, we plot the $-log_{10}$(p-value). The $-log_{10}$(p-value) is greater than 2 means that the p-value is smaller than 0.01. Also, $-log_{10}(0.05)$=1.3. What can be observed with respect of the

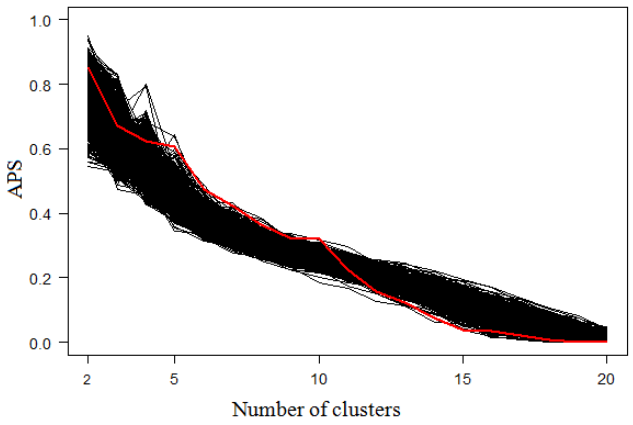


Figure 6.8: Test of each number of clusters for CO_{314} for the PAM method with the average Prediction Strength. - The figure shows the average PS with 50 repetitions for 2 to 20 clusters for the PAM method. The y-axis represents the average PS and the x-axis represents the number of clusters. The colours of the lines indicate the datasets, black for each of the reference datasets and red for CO_{314}.

PAM method is that $-log_{10}$(p-value) is greater than 1.3 for values of k ranging from 5 to 10. This suggests that the potential numbers of clusters for the PAM method is between 5 and 10.

Similarly, we performed another null model test with ASW. Figure 6.10 shows the values of the ASW for the PAM method with the p-dissimilarity. The y-axis represents the values of the ASW and the x-axis represents the number of clusters. The black lines represent the ASW values for each of the reference datasets, and the red line represents the values for CO_{314}. As can be seen, there is a drop from $k = 2$ to $k = 3$ since ASW is more likely to be higher for a lower number of clusters. Also, the null model test for Complete Linkage and the Average Linkage were performed. The results are shown in Figure 6.11(a) and Figure 6.11(b). The y-axes and x-axes represent the values of the ASW and the number of clusters, respectively. The black lines represent the values for each of the reference datasets for 2 to 20 clusters, and the red line represents the values for CO_{314}. As can be seen, none of them is significant. Figure 6.12 shows the result for the p-value$_k$ for between 2 to 20 clusters with the ASW. What can be observed is that for the PAM method, when $k = 3$, 4, 5 or 6, the value of the ASW for CO_{314} is higher

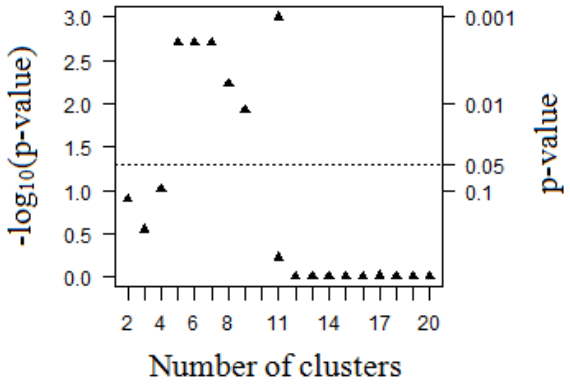


Figure 6.9: The null model test with average Prediction Strength. - The figure shows the $-log_{10}$(p-value) for numbers of clusters between 2 and 20 for three clustering methods. The vertical axis on the left represents the $-log_{10}$(p-value), while that on the right represents the p-value. Note that since $-log_{10}(0.05) = 1.3$, a p-value smaller than 0.05 is equivalent to a value of $-log_{10}$(p-value) greater than 1.3.

than 95 % of the corresponding values for the reference datasets. This suggests that the potential number of clusters for the PAM method with the p-dissimilarity is 3, 4, 5 or 6.

Test for homogeneity

The p-value$_H$ for CO_{314} with the average PS was 0.4748, while that for CO_{314} with ASW was 0.1392. This suggests that there is not enough evidence to conclude that there exists a structure of clustering in CO_{314}.

To sum up, a set of the potential number of clusters is identified but there is no enough evidence to conclude about clustering structure. However, the clusters are still useful in our study regardless of whether or not there is a clear clustering structure. The clusters can be used to explore patterns in the daily dosage taken by participants. According to the clusters, participants with similar 180 records of dosages can be grouped together and the variations of daily dosages among participants narrowed down to those among participants within a cluster. The average PS and the ASW can be used to determine the number of clusters, and the null model test can be used to back up these indexes. By the results of the null model test for average PS and ASW,

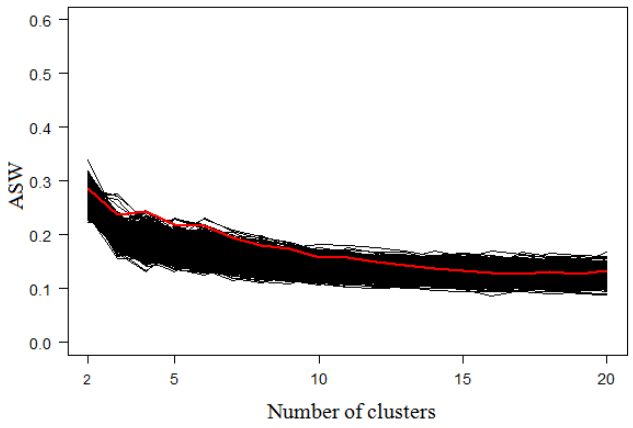
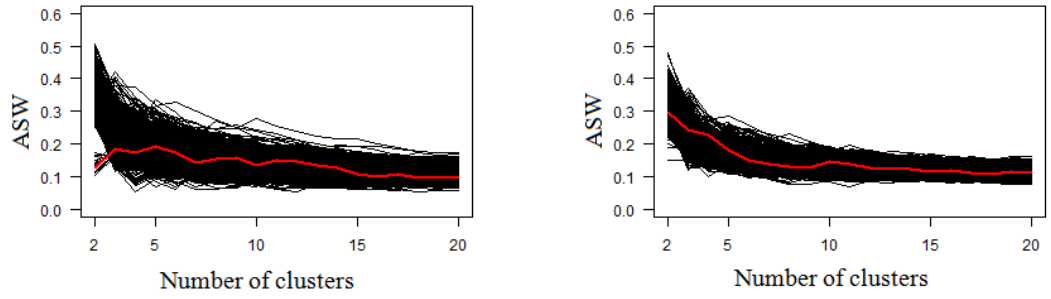


Figure 6.10: Test of the homogeneity between the null model and CO_{314} for the ASW. - The graph shows the values of the ASW for the PAM method for between 2 to 20 clusters. The y-axis represents the values of the ASW and the x-axis represents the number of clusters. The values of the ASW for each of the 1000 reference datasets are displayed as in black lines and those for CO_{314} are displayed as a red line.

only a few k are identified to be potential number of clusters. Since both indexes picked up the value 5 and $k = 5$ has the highest value of the ASW among these identified potential k, later we will use the PAM clustering with five clusters.

In the next chapter we move to the topic of assessing the quality of clustering. We propose algorithms for information visualisation via heatplots and show their applications on CO_{314} with the PAM clustering and five clusters.



(a) (b)

Figure 6.11: Test of the homogeneity between the null model and CO_{314} with ASW. - The graphs (a) and (b) show the values of the ASW of the Average Linkage and those of the Complete Linkage for between 2 and 20 clusters. The values of the ASW for each of the 1000 reference datasets are displayed as black lines, and those for CO_{314} are displayed as in a red line.

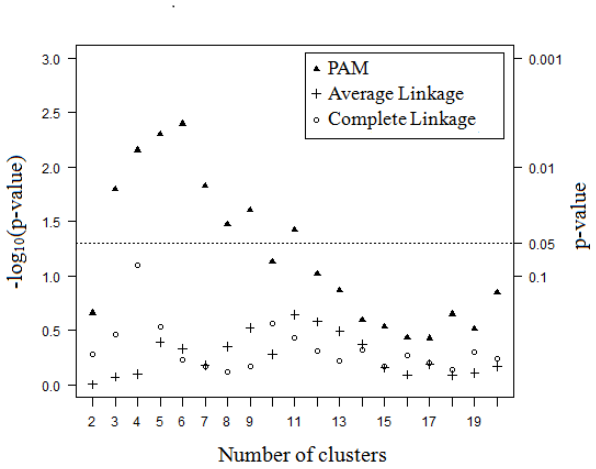


Figure 6.12: The null model test with ASW. - It shows the $-log_{10}$(p-value) for 2 to 20 clusters with respect to three clustering methods. The vertical axis on the left represents the $-log_{10}$(p-value), while that on the right represents the p-value. Note that since $-log_{10}(0.05) = 1.3$, a p-value smaller than 0.05 is equivalent to a value of $-log_{10}$(p-value) greater than 1.3.

Chapter 7

Visualisation of the PAM results

Two algorithms for obtaining orders for objects based on clusters to which objects belong and dissimilarity between objects are proposed. These algorithms are useful for visualizing clustering results to assess the quality of the clustering. Also, the ordering algorithm with the heatplot will be used for visual significance test.

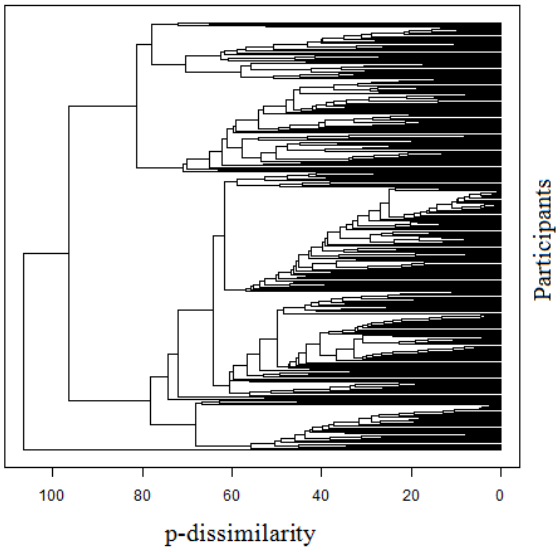
7.1 Motivation

A number of summary statistics, such as mean, median, quartile, IQR, variance, etc., and some graphics, such as histograms, boxplots, scatter plots and heatplots can be used to represent the data for clusters (Leisch [2008]). Among these, we focus on heatplots. A heatplot is a graph that represents data by colour. It consists of horizontal lines representing the data for objects. It is particularly useful for visualizing relationship between objects when clustering method is used. However, the interpretability of a heatplot strongly depends on the order of the objects. In this chapter we show two heatplots for a dataset. Their purposes and layout are as follows.

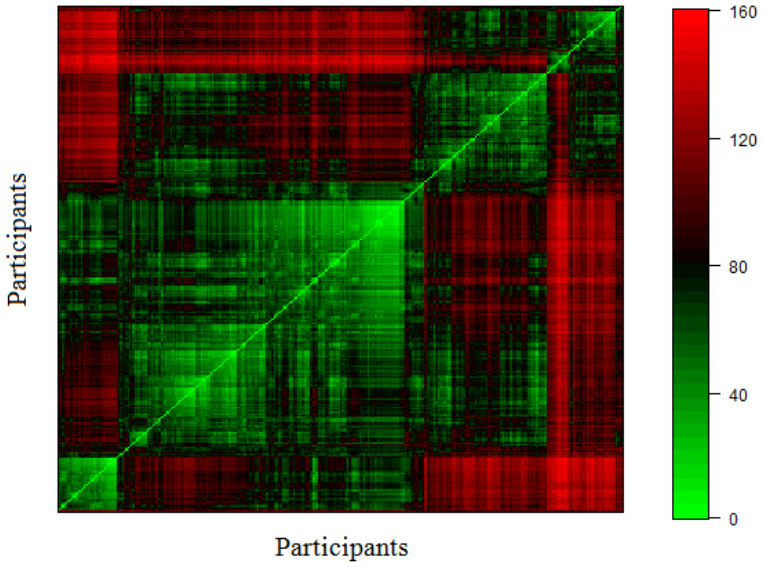
- heaplot of the dosage data is for observing the dosage patterns over time. The record of the 180 days of each participant will be plotted along the x-axis. The colour designates dosage.
- heatplot of the p-dissimilarity matrix is for observing the relationships between participants/clusters. Both the x- and y-axes represent participants. Note that the order of the participants on the x-axis mirrors the order of the participants on the y-axis.

For a clustering result obtained by hierarchical clustering methods whose clustering process can be displayed by a dendrogram, all horizontal lines in the heatplot that represent the data for objects can be re-ordered by their locations in the dendrogram as shown in Figure 7.1. Figure 7.1(a) is a dendrogram obtained from applying the Average Linkage method and the p-dissimilarity to CO_{314}. The x-axis represents the dissimilarity. On the y-axis, the right end of each node indicates one participant. Assume this y-axis is a 0-1 scale from the bottom to the top, and the nodes indicate locations for the participants on the y-axis. We observe that, of these participants who are close to each other, most are linked together. This suggest that the location can be used to represent the relationship between participants, that is, participants show are close to each other have a smaller dissimilarity than participants who are further away. Next, the p-dissimilarity matrix for CO_{314} is re-arranged according to the locations. Figure 7.1(b) shows the p-dissimilarity matrix for participants with a heatplot. Both the x- and y-axes represent participants. Note that the order of the participants on the x-axis as well as that on y-axis mirrors the locations established by the dendrogram. The colour designates the p-dissimilarity, ranging from 0 to 159, appearing in a sequence of green, black and red. The p-dissimilarity matrix is symmetric about its diagonal, so the heatplot is symmetric. Also, the diagonal line shows as green, because the dissimilarities between participants and themselves are zero. Moreover, there are some green squares along the diagonal line. They represent the dissimilarity matrix within clusters. As seen, some green squares have colour gradients – green, black, red – from the diagonal line to the border of the figure. The colour gradients within clusters themselves can be improved by flipping nodes in the tree in Figure 7.1(a) without changing the structure of the hierarchy. However, flipping trees is beyond the scope of this study, readers are referred to research on optimal leaf ordering for hierarchical clustering (Gale, Halperin, and Costanzo [1984] Bar-Joseph, Gifford, and Jaakkola [2001]).

As for a result obtained by the PAM method, Figure 7.2 shows the p-dissimilarity matrix with the heatplots. The cluster information is preserved by plotting each of the five clusters one by one. In each cluster, participants are organized in terms of the date they joined the MMT. Both the x- and y-axes represent participants. Also, the numbers 1 to 5 on the y-axis indicate the five clusters. The colour indicates the dissimilarity, ranging from 0 to 160. What can be observed is that most of the participants



(a)



(b)

Figure 7.1: The dendrogram of the Average Linkage and the heatplot of the p-dissimilarity matrix of CO_{314} - (a) shows the dendrogram obtained from the Average Linkage method with the p-dissimilarity. On the y-axis, the right end of each node indicates each of the 314 participants. The x-axis represents the dissimilarity. (b) shows the p-dissimilarity matrix. Both the x- and y-axes represent the 314 participants. The order of the participants mirrors the order found by the dendrogram of (a).

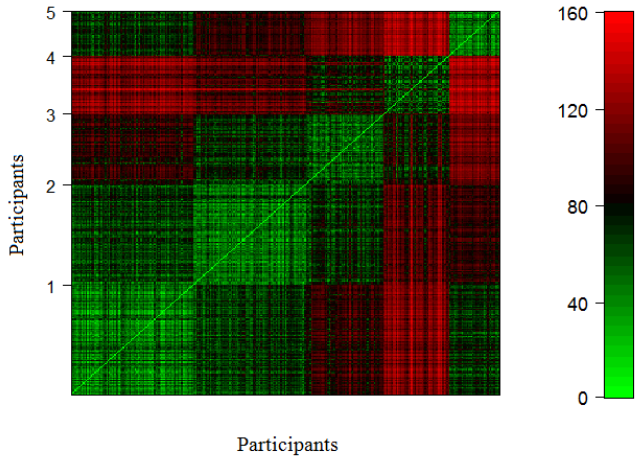


Figure 7.2: Heatplot of p-dissimilarity matrix of CO_{314} with random orders within clusters. - The heatplot of the p-dissimilarity matrix for participants. Both the x- and y-axes represent participants. Note that the five clusters are obtained by the PAM method and then plotted into heatplot separately. The participants within a cluster are organized in terms of the date they joined the MMT. The colour indicates the p-dissimilarity, ranging from 0 to 160.

within a cluster have small p-dissimilarity, appearing in green. The performance of PAM p-dissimilarity on the category-ordered data is good as highlighted by the green squares along the diagonal line. However, this does not indicate that there is a clear cluster structure in the data. The random order of participants within clusters makes it unnecessary for neighbouring participants to be regarded as most similar. The random order generates a rather artificial order which might lead to an overoptimized view of the heatplot regardless of whether a cluster structure exists or not. The visible cluster structure might be misleading.

Some research has been done on information visualization via heatplots of row data matrices and proximity matrices (Chen [2002]; Hahsler and Hornik [2011]; Hahsler, Hornik, and Buchta [2008]; Tien, Lee, Wu, and Chen [2008]; Wu, Tien, and Chen [2010]). They constructed an order for objects that preserved the clustering structure and used the order with heatplots to illustrate and to assess the quality of clustering results, which is what we are interested in. Hahsler and Hornik [2011] developed the R package *seriation*, for visualizing the dissimilarity of the partitioning methods. An

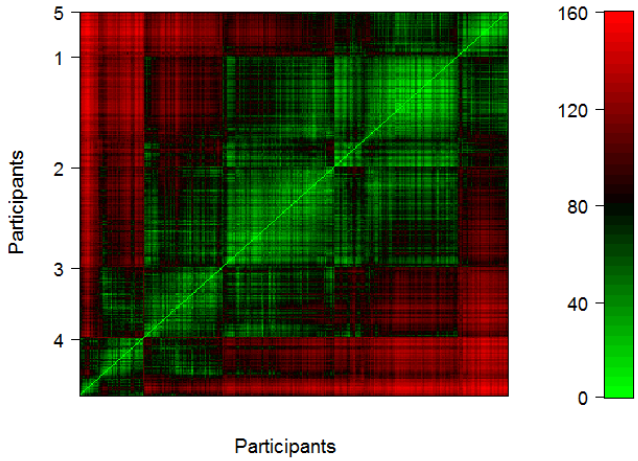


Figure 7.3: Heatplot of p-dissimilarity matrix of CO_{314} by seriation. - The heatplot of the p-dissimilarity matrix among participants. Both the x- and y-axes represent participants. Note that the five clusters are obtained by the PAM method and labeled on the y-axis. The order of the participants within a cluster is obtained from the algorithm of Chen [2002] by *seriation*. The colour represents the p-dissimilarity, ranging from 0 to 160.

ordering algorithm proposed by Chen [2002] was implemented. The ordering algorithm aimed at placing minimally dissimilar participants within a cluster close to the diagonal of the heatplot. Figure 7.3 shows the heatplot of the p-dissimilarity matrix with $p = 0.6$ of CO_{314} with the order of the participants obtained from the algorithm of Chen [2002] by *seriation*. Both the x- and y-axes represent participants. As can be seen, the order preserves the clustering structure and the heatplot shows that the color gradient from the diagonal line to the borders of the figure is much smoother in comparison to that in Figure 7.2. However, this ordering method does not bear on clustering methods. The relationship between medoids has been neglected.

With regards the PAM method, medoids play important roles because each object is assigned to the cluster with the closest medoid. To visualize the clustering result for the PAM method, we take into account the following: preservation of the clustering structure, the information of the selected medoids, the information about the distance between each object and their medoid and that between them and their neighbouring

medoids. We attempt to develop an ordering rule which indicates the similarity structure of clusters and similarity structure of participants. We attempt to decide where to locate an object so that its location reflects the dissimilarity between the object and the medoid in the same cluster as well as the dissimilarity between the object and the most closest medoid in a different cluster. An object which is close to the border line between two medoids should be considered to be less similar to its medoid in comparison to objects belonging to the same cluster, so we conclude that such an object should be plotted distant from its medoid. This in a heatplot would look like a smooth colour gradient. Our intention is to make as smooth as possible the colour gradient of the heatplot representing the transition to a neighbouring cluster. Then, one can make statements about whether there really is some clustering which is visible by looking at the border regions of the clusters on the heatplot.

7.2 Multidimensional scaling

Before getting into the ordering rules, we would like to introduce a method called multidimensional scaling (MDS) (Cox and Cox [1990]). We would use MDS to construct ordering rules for the PAM result in Chapter 7 and we would use MDS to produce a map for monitoring the movement of the final five clusters over time in Chapter 8.

The concept of the MDS is to represent a dissimilarity matrix in a multidimensional space so that information in the high dimensional data can be reflected in the lower dimensional space. Some information is lost in the process of dimension reduction. The lost information is measured by a loss function called *stress* and is defined by

$$stress = \sqrt{\frac{\sum_{ij}(f(d_{ij}) - d_{ij}^q)^2}{\sum_{ij}(d_{ij}^q)^2}}, \tag{7.1}$$

where d_{ij}^q is the spatial distance between objects i and j, which is computed using the Euclidean distance, and the value of $f(d_{ij})$ depends on whether metric or non-metric MDS is used. In metric MDS, $f(d_{ij})$ is the original dissimilarity between objects i and j. In non-metric MDS, $f(d_{ij})$ represents the rank of the dissimilarity between objects i and j. A dataset consisting of n objects has $n(n-1)/2$ dissimilarities for $\binom{n}{2}$ pair objects. $f(d_{ij})$ is the value which is mapped from the original dissimilarity and best

preserves the rank order (Kruskal [1964]).

In our study, we will use the non-metric MDS because we are aiming at ordering participants instead of focusing on the numerical values of the dissimilarities. Also, only one dimension is needed for generating orders for objects in the heatplot. A function called isoMDS{MASS} in R which produces the non-metric MDS will be used throughout this study.

7.3 Order of clusters

In Chapter 6 we have selected the PAM method and five to be the clustering method and the number of clusters for CO_{314}. The following sections dealing with the ordering rules are organized as follows. First of all, we introduce ordering rules in a general situation, that is, for k clusters. Afterwards, we show a heatplot of CO_{314} and a heatplot of the p-dissimilarity with the order of participants obtained by the proposed ordering algorithms.

The ordering algorithm starts from preserving the clusters in a heatplot. This can be done by plotting clusters separately in a heatplot. We chose to determine the location of the clusters in a heatplot first.

We propose to locate the clusters based on their objects. For each cluster, one object is selected as the medoid by the PAM method. The medoids is thus used. The order of the medoids on the one dimensional MDS is used to locate the clusters. There are two ways to order the medoids. One is to apply MDS to the dissimilarity matrix of the dataset and then obtain the order of the medoids. The orders are the locations of the medoids on the one dimensional MDS. The other is to apply MDS to the dissimilarity matrix of the k medoids and then to obtain the locations of the medoids. The former uses all objects in the dataset, while the latter uses k objects.

In our case, the 314 participants were partitioned into 5 clusters. The MDS order for the five medoids was generated from 1-dimensional MDS. We tried both approaches for $k = 5$ and observed that the orders of the medoids were the same. Likewise, we

tried both approaches for $k = 7, 9, 11, 13$. We obtained the same results. In general, this might not always be the case. Since the one-dimensional MDS returned a similar order of the medoids irrespective of whether it was applied to all participants or only to the selected participants in our case, the first approach was used to generate an order for clusters.

7.4 Order of objects within clusters

In the previous section the position of each of the k clusters referring to the k medoids in a heatplot is determined. In this section we introduce ordering algorithms that aim at preserving the similarity structure of objects by which the order of objects within a cluster are obtained.

We introduce two algorithms for ordering objects within clusters. The algorithm 1 is to use MDS (see Section 7.4.1) and the algorithm 2 is to use projection vectors (see Section 7.4.2). The description of each of them is organized as follows: first of all, we introduce the process in a general situation, that is, for k clusters. Secondly, we apply the ordering algorithm to cluster result of the PAM method with $k = 5$ for CO_{314} to obtain an order of the participants. Then, with the obtained order, we show a heatplot of Dosage_{314} and a heatplot of the p-dissimilarity matrix in order to assess the quality of the clustering.

7.4.1 Ordering by multidimensional scaling

The first algorithm is to utilize MDS. Assume that the PAM partitions data into k clusters $\{G_i : i = 1, \ldots, k\}$ and selects k objects as the k medoids. Then, these medoids are ordered by the MDS (see the previous section). Denote the ordered medoids by $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(k)}$. Let $G_{(1)}, G_{(2)}, \ldots, G_{(k)}$ be the corresponding clusters.

The concept of the ordering algorithm by MDS is to order of objects in $G_{(i)}$ by using the information on objects in the neighbouring clusters. By neighbouring cluster, we mean that cluster whose medoid is next to $\mathbf{x}_{(i)}$ according to its location on the 1-dimensional MDS. The algorithm 1 works as follows. Step 1: the one-dimensional MDS is applied to all objects in $G_{(j-1)}$, $G_{(j)}$ and $G_{(j+1)}$, $j = 1, \ldots, k$. Step 2: the

objects in $G_{(j)}$ are organized in terms of their locations on the one-dimensional MDS. Then, this algorithm repeats Step 1 and Step 2 until order for all objects are obtained.

We applied this algorithm to the clustering result of PAM p-dissimilarity with $k = 5$. Figure 7.4 displays the heatplot of Dosage_{314} and that of the p-dissimilarity matrix of CO_{314}. The heatplot of Dosage_{314} displays a much smoother colour gradient from clusters 1 to 5. Also, the heatplot of p-dissimilarity matrix shows green, black and red from the diagonal to the borders of the figure. What can be observed is that both figures present a better colour gradient in comparison to the aforementioned figures without ordering. What can be inferred from this is that the algorithm has improved the use of the heatplot. Also, this algorithm generates an order that preserves clusters, and shows similarity structure of clusters and similarity structure of participants. What can be concluded from Figure 7.4 is that there does not seem to be a clear separation between clusters. Also, there are two red lines that cross in cluster 4 in the graph on the right. At the same position in the graph on the left where the 180 records of this participant is plotted, it can be seen that, at the beginning of the MMT, this participant has a higher dosage than the other participants in cluster 4. This suggests that this participant might be an outlier in cluster 4.

7.4.2 Ordering by projection vectors

Medoids are important for the PAM method because each object is assigned to the cluster with the closest medoid, so we want to have a method that arranges objects based on medoids. Following the notations that are defined in the Section 7.4.1, $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(k)}$ are the ordered k medoids, which are also k objects, $G_{(1)}, G_{(2)}, \ldots, G_{(k)}$ are the corresponding clusters, and $d(.,.)$ is the dissimilarity between objects. For convenience, $d_{i,j}$ represents the dissimilarity between objects i and j, and $d_{(i),(j)}$ represents the dissimilarity between ordered medoids i and j. The aim is to generate order for objects in which the order is capable of displaying the relationships between objects and medoids, in other words, $d_{r,(i)}$ and $d_{r,(j)}$ in relation to $d_{(i),(j)}$, where object r and medoid $\mathbf{x}_{(i)}$ are in the same clusters, and $\mathbf{x}_{(j)}$ is the neighbouring medoid of $\mathbf{x}_{(i)}$. In order to achieve the purpose, we attempt to transform $d_{r,(i)}$ and $d_{r,(j)}$ into a new value which is a value in proportion to $d_{(i),(j)}$. The idea is that, on the basis of medoids, we create a two dimensional space into which dissimilarities between an object and

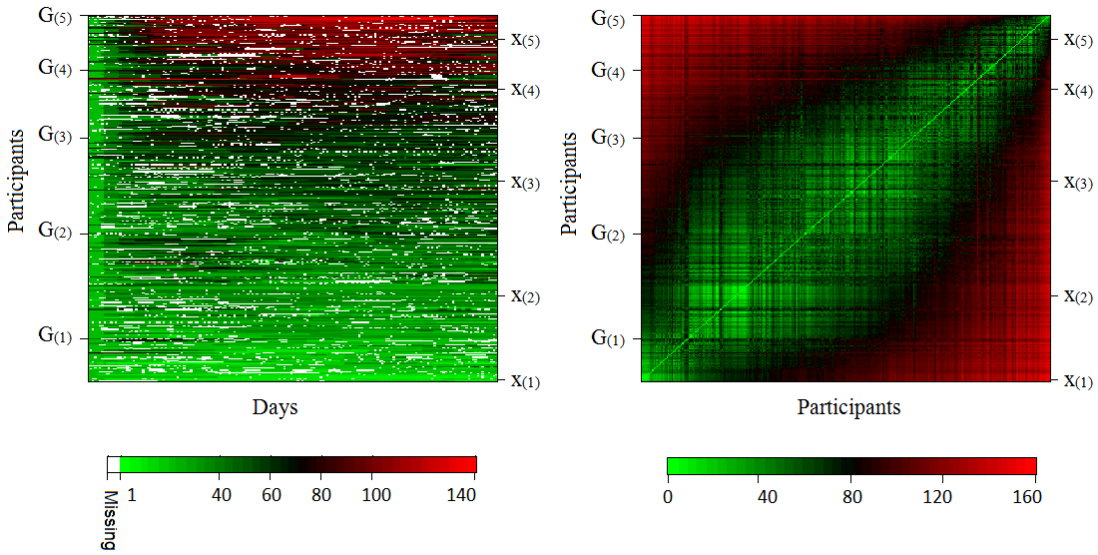


Figure 7.4: Heatplot of Dosage$_{314}$ and heatplot of the p-dissimilarity matrix of CO$_{314}$ with the order of participants generated from MDS on all participants belonging to the same and the neighbouring clusters. - The graph on the left displays Dosage$_{314}$. The record of the 180 days of each participant is plotted along the x-axis. The colour designates dosage. The graph on the right shows the p-dissimilarity matrix among participants. Both the x- and y-axes represent participants. The colour represents the dissimilarity. Note that the order of the participants within clusters is generated by applying MDS on all participants belonging to the same and the neighbouring clusters.

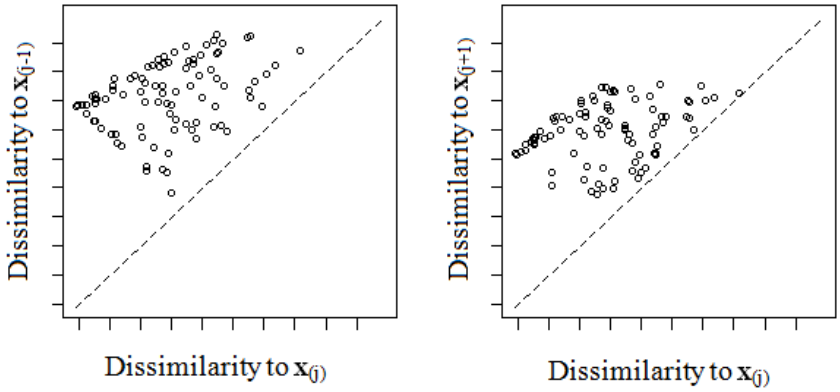


Figure 7.5: Illustration of the planes - The plane on the left is created by using $\mathbf{x}_{(j)}$ and $\mathbf{x}_{(j-1)}$, while that on the right is created by using $\mathbf{x}_{(j)}$ and $\mathbf{x}_{(j+1)}$. The figures show the scatter plot of dissimilarity. The points represent objects in $G_{(j)}$ in terms of their dissimilarity to $\mathbf{x}_{(j)}$ and to the neighbouring medoids. Also, a dashed line at 45 degrees is drawn on both of the graphs.

medoids will be transformed. Then, the transformed dissimilarity will be used to order the objects.

The ordering algorithm is divided into two cases, denoted process I and process II. Process I is designed for medoids which have neighbouring medoids on both sides. It is used to generate orders for objects in $G_{(2)}, \ldots, G_{(k-1)}$. Process II is for the first and last medoids. It is used to generate orders for objects in $G_{(1)}$ and $G_{(k)}$.

The algorithm of process I works as follows. Step 1 is to create spaces with medoids. For any $j = 2, \ldots, k-1$, the neighbouring medoids of $\mathbf{x}_{(j)}$ are $\mathbf{x}_{(j-1)}$ and $\mathbf{x}_{(j+1)}$. One space is created by using a pair of medoids $\mathbf{x}_{(j)}$ and $\mathbf{x}_{(j-1)}$. The x-axis is labelled dissimilarity to $\mathbf{x}_{(j)}$ and the y-axis is labelled dissimilarity to $\mathbf{x}_{(j-1)}$. Mark objects in $G_{(j)}$ on the space for these two axes according to their dissimilarities as points shown in the graph on the left in Figure 7.5. The other space is created by using $\mathbf{x}_{(j)}$ and $\mathbf{x}_{(j+1)}$. Also, mark objects on the plane according to their dissimilarities as points shown in the graph on the right in Figure 7.5.

Step 2 is to transform dissimilarities into one dissimilarity using projection vectors.

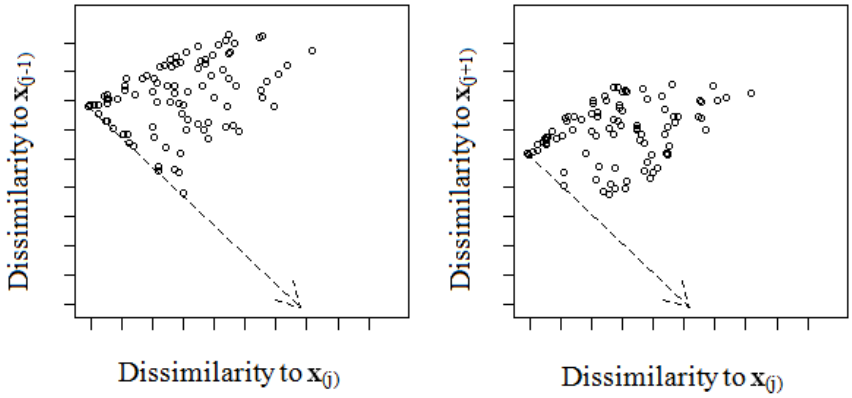


Figure 7.6: Illustration of representing the dissimilaty between $\mathbf{x}_{(j)}$ and $\mathbf{x}_{(j-1)}$ by a vector - Each point in the figures represents the dissimilarities between an object and two medoids labelled on the axes. The dashed lines with arrows at the end refer to the vectors $\overrightarrow{v_{(j)(j-1)}}$ and $\overrightarrow{v_{(j)(j+1)}}$.

On the plane of $\mathbf{x}_{(j)}$ and $\mathbf{x}_{(j-1)}$, each point represents the dissimilarities between an object in $G_{(j)}$ and two medoids. Also, the dissimilarities between $\mathbf{x}_{(j)}$ and $\mathbf{x}_{(j-1)}$ can be represented by the length of the vector from point $(0,\ d_{(j),(j-1)})$ to point $(d_{(j),(j-1)},0)$. Denote this vector by $\overrightarrow{v_{(i)(j-1)}}$. Next, transforming two dissimilarities into one is done by projecting the vector from point $(0, d_{(j),(j-1)})$ to point $(d_{i,(j)},\ d_{i,(j-1)})$ onto $\overrightarrow{v_{(j)(j-1)}}$. We call the length of the projection vector "standardized projection". Likewise, the dissimilarities between an object and medoids, $\mathbf{x}_{(j)}$ and $\mathbf{x}_{(j+1)}$, are transformed into another standardized projection. The standardized projection of object i with respect to medoids $\mathbf{x}_{(j)}$ and $\mathbf{x}_{(j')}$ is defined by

$$\text{Proj}_{(j)(j\prime)}i = \frac{\overrightarrow{v_{i,(j')}} \bullet \overrightarrow{v_{(j)(j')}}}{\|\overrightarrow{v_{(j)(j')}}\|} \tag{7.2}$$

where $\bullet$ is the inner product of the vectors, and $\|\cdot\|$ is the Euclidean norm. Figure 7.6 shows an example for $\overrightarrow{v_{(i)(j-1)}}$ and $\overrightarrow{v_{(i)(j+1)}}$. Also, Figure 7.7 shows the standardized projections of an object i in $G_{(j)}$ with respect to medoids $\mathbf{x}_{(j-1)}$, $\mathbf{x}_{(j)}$ and $\mathbf{x}_{(j+1)}$. The two standardized projections of the object i are shown as bold lines.

Step 3 is to determine whether an object should be plotted between $\mathbf{x}_{(j)}$ and $\mathbf{x}_{(j-1)}$ or it should be plotted between $\mathbf{x}_{(j)}$ and $\mathbf{x}_{(j+1)}$. It is determined by comparing the standardized projections. Those whose $\text{Proj}_{(j)(j-1)}i$ is greater than $\text{Proj}_{(j)(j+1)}i$ are

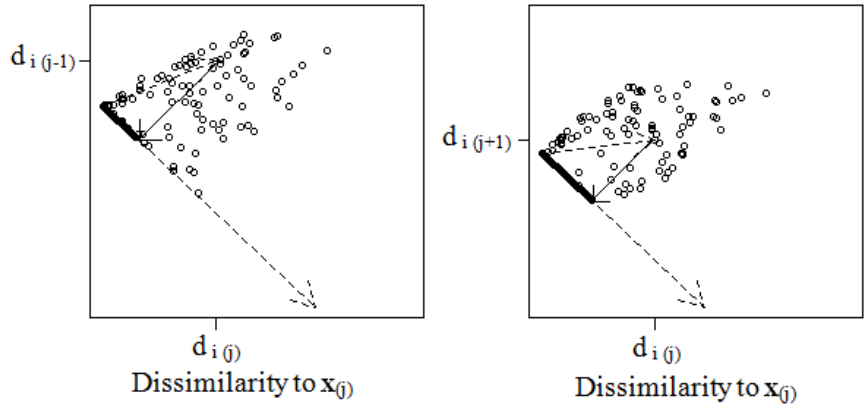


Figure 7.7: Illustration of standardized projection of an object - Each point in the figures represents the dissimilarities between an object in $G_{(j)}$ and two medoids labelled on the axes. Assume an object i in $G_{(j)}$ has its dissimilarities from three medoids, $\mathbf{x}_{(j-1)}$, $\mathbf{x}_{(j)}$, and $\mathbf{x}_{(j+1)}$ be $d_{i(j-1)}$, $d_{i(j)}$, and $d_{i(j+1)}$, respectively. The figures show the standardized projections of the object i with respect to medoids $\mathbf{x}_{(j)}$ and $\mathbf{x}_{(j-1)}$, and medoids $\mathbf{x}_{(j)}$ and $\mathbf{x}_{(j+1)}$, displayed as bold lines.

plotted between $\mathbf{x}_{(j)}$ and $\mathbf{x}_{(j-1)}$, while those with a larger $\text{Proj}_{(j)(j+1)}i$ are plotted between $\mathbf{x}_{(j)}$ and $\mathbf{x}_{(j+1)}$.

Step 4 is to order the objects. Those plotted between $\mathbf{x}_{(j)}$ and $\mathbf{x}_{(j-1)}$ are organized by the value of $\text{Proj}_{(j)(j-1)}i$, whereas those plotted between $\mathbf{x}_{(j)}$ and $\mathbf{x}_{(j+1)}$ are ordered by $\text{Proj}_{(j)(j+1)}i$. To sum up, we first identify between which two boundary medoids objects should be located, and then, for all objects located between the given two boundary medoids, we establish the order.

Process II is for the cases of the first and last medoids because they have only one neighbouring medoid. For an object i in $G_{(1)}$, whether to plot it between $\mathbf{x}_{(1)}$ and $\mathbf{x}_{(2)}$ or between $\mathbf{x}_{(1)}$ and the border of the heatplot is determined by comparing two dissimilarities, namely $d_{(1),(2)}$ and $d_{i,(2)}$. If $d_{i,(2)}$ is smaller than $d_{(1),(2)}$ then the object i is more similar to $\mathbf{x}_{(2)}$ than $\mathbf{x}_{(1)}$ is to $\mathbf{x}_{(2)}$. In order to address this relationship, objects with a smaller dissimilarity to $\mathbf{x}_{(2)}$ are plotted between $\mathbf{x}_{(1)}$ and $\mathbf{x}_{(2)}$. These are then organized in terms of $\text{Proj}_{(1)(2)}i$. The remaining objects in $G_{(1)}$ are plotted on the other side of $\mathbf{x}_{(1)}$ in order to deliver the fact that they are distant from $\mathbf{x}_{(2)}$, and

they are ordered by the dissimilarity between them and $\mathbf{x}_{(1)}$, that is, $d_{(1),i}$. As for the objects in $G_{(k)}$, their dissimilarity to $\mathbf{x}_{(k-1)}$ is compared to the dissimilarity between $\mathbf{x}_{(k)}$ and $\mathbf{x}_{(k-1)}$, and those with a smaller dissimilarity to $\mathbf{x}_{(k-1)}$ are plotted between $\mathbf{x}_{(k)}$ and $\mathbf{x}_{(k-1)}$ and ordered by $\text{Proj}_{(k)(k-1)}i$, while the rest are plotted on the other side and ordered by $d_{(k),i}$. To sum up, we first identify the location of objects, and then all objects located between the given two medoids are organized by the standardized projection, and the remaining objects organized by their dissimilarities to the medoid.

We applied this algorithm to the clustering result of PAM p-dissimilarity with $k = 5$. The PAM method is based on medoids. Figure 7.8 shows the heatplot of Dosage_{314}, and the heatplot of the p-dissimilarity matrix of CO_{314}. What can be observed is that both graphs presents a good colour gradient. Also, Figure 7.8 provides the information on the location of medoids and the relationship of dissimilarities between the participants. This figure indicates how far apart the medoids are, and the colour gradient around the medoids indicates the density of the clusters. In the heatplot of the p-dissimilarity, in cluster 4, there is a participant who has p-dissimilarities between themself and others represented by red coloured values, which creates two red lines cross at a point in cluster 4. The dosage data for this participant start from high dosages, followed by some missing dosages. In comparison to records for other clusters, this participant has record that should be considered to be more similar to records of participants in cluster 4. Despite the fact, the heatplot of Dosage_{314} shows that this participant behaves differently, so this participant might be considered as an outlier in cluster 4.

We proposed two ordering algorithms. One uses MDS and the other uses projections. Both methods smooth heatplots. The MDS quantifies information on one dimension and consequently information is lost. The projections quantifies information on one dimension on the basis of medoids, so that information of medoids and the density of clusters are more visible. Figure 7.4 shows the heatplots with ordering algorithm of MDS, while Figure 7.8 shows the heatplots with ordering algorithm of projections. Assume that we are interested in $G_{(2)}$ and $G_{(3)}$, so we look at the heatplots of the p-dissimilarity for $G_{(2)}$. We notice from Figure 7.4 that the separation between these two cluster might not be clear. With Figure 7.8, we obtain more information. We

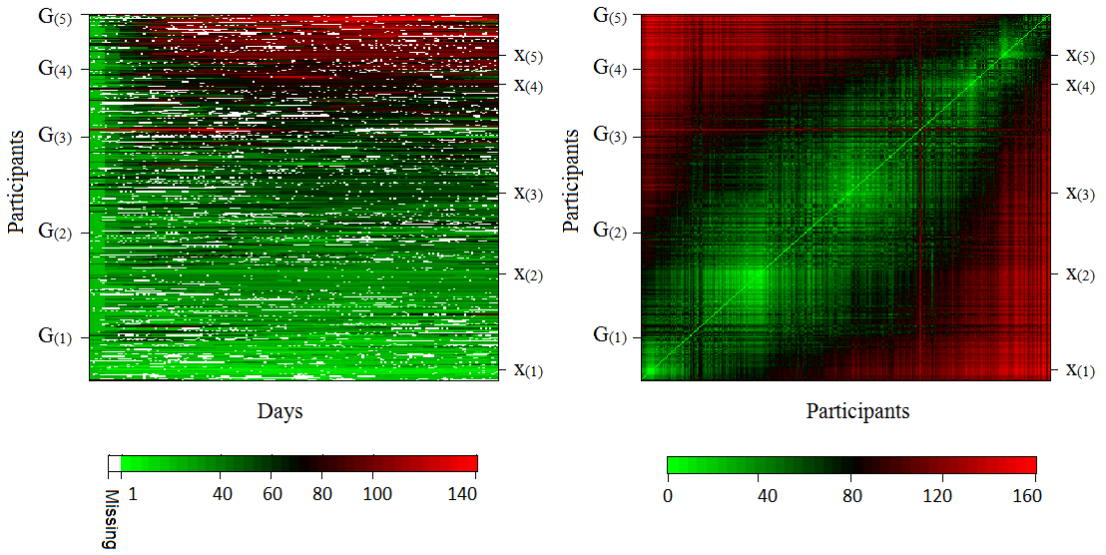


Figure 7.8: Heatplot of $Dosage_{314}$ and heatplot of the p-dissimilarity matrix with the order of the participants obtained by using projection vectors. - The graph on the left displays $Dosage_{314}$. The record of the 180 days of each participant is plotted along the x-axis. The participants belong to the same cluster are organized in terms of the standardized dissimilarity. The graph on the right shows the p-dissimilarity matrix among participants. Both the x- and y-axes represent participants. There is a potential outlier indicated by a point in cluster 4 at which two red lines cross.

notice that there is a coherence of participants around $x_{(2)}$. On the other hand, participants in $G_{(3)}$ tend to be scattered. The algorithm of the projections is very useful for visualizing the clustering result obtained from the PAM method.

7.5 Comparison of CO_{314} and the reference datasets

Graphs are often used to visualize more complex patterns in datasets for exploratory data analysis. However, there is no standard method for how to read graphs. Reading graphs depends on human viewers. Buja, Cook, Hofmann, Lawrence, Lee, Swayne, and Wickham [2009] attempted to develop graphical statistics that could be used for statistical inference. Their idea was whether the graph was, they were generated under the null assumption. They were interested in whether the real dataset looked anything like the datasets generated from the model. As a solution, they suggested a test by

human viewer compare a plot of the real dataset with plots of simulated datasets.

In Section 6.3 we proposed a Markov model which fitted the marginal distributions for categories but it did not model the clustering. We attempted to test how well the null model fitted the CO_{314} by using the graphical statistic. The test was done by human viewers comparing the heatplot of CO_{314} with heatplots of simulated CO_{314}, and comparing the heatplot of p-dissimilarity matrix of CO_{314} with heatplots of p-dissimilarity matrices of simulated CO_{314}. We considered

- a probability of the real dataset being picked up might be unstable if the graph statistic was determined by one person .
- interviewees might have difficulty to read graph, so they picked up a graph at random.
- interviewees might want to pick up more than one graphs.

Therefore, we surveyed 19 participants, 2 of which have PhD in Statistics, 5 of which are the PhD students in the Statistics department, 2 of which are senior statisticians, and the remaining have at least an undergraduate degree. We asked them to answer questions based on the figures they received.

Critical region

In order to deal with the result obtained from more one person and answers for more than one question, the critical region in our study is defined based on "sum". We sum up the frequency of datasets being nominated. A conclusion is drew according to whether the real dataset is nominated the most or not. The real dataset can not get the highest score means that there is no enough evidence to reject the null hypothesis.

P-value

Denote the frequencies of the R simulated datasets being nominated by $s_1, s_2, \ldots, s_R$. Denote the frequency of the real dataset being nominated by s. If exactly a of $\{s_1, s_2, \ldots, s_R\}$ are greater than s, then the approximate p-value of this test is defined by

$$\text{p-value} = \frac{a+1}{R+1}. \tag{7.3}$$

Here is how the figures were generated. For CO_{314}, the order of the participants were obtained by applying algorithm 2 to the clustering result of PAM method with $k = 5$ and the p-dissimilarity with $p = 0.6$ and $\beta = 1.42$. We randomly generated 24 reference datasets and applied the same process to obtain the order of the participants. The twenty-five graphs in Figure 7.9 show the heatplots of the simulated CO_{314}. In addition, the heatplot of the real CO_{314} is embedded in the plots with a random position. In all twenty-five graphs, the y-axes represent the participants of five clusters and the x-axes represent the 180 days. The seven categories are represented by seven colours, black, red, green, blue, cyan, purple and white. The order of participants is obtained by the algorithm of the projections. On the other hand, the twenty-five graphs in Figure 7.10 show the heatplots of the p-dissimilarity matrix. All x- and y-axes represent the participants. The colour designates the dissimilarity, ranging from 0 to 159, appearing in a sequence of green, black and red. The order of the participants in each graph mirrors the order of the participants on the y-axes in Figure 7.9. The questions that participants received are as follows. Note that Fig 1 and Fig 2 refer to Figure 7.9 and Figure 7.10, respectively.

This is a homogeneity test between a real dataset and reference datasets, simulated from a null model, by human viewer comparing plots. Enclosed are two figures, namely Fig 1 and Fig 2. Both of them consist of 25 graphs. The 25 graphs in Fig 1 and those in Fig 2 represent the same dataset and the real data is embedded among the plots with a random position. Each horizontal line in Fig 1 represents daily dosage of a participant from day 1 to day 180, while that in Fig 2 indicates distances (eg. Euclidean distance) between this participant and other participants.

(Q1) Pick up three graphs in Fig 1 that you find different from the rest.
(Q2) Among those three, what is the most different one?
(Q3) Pick up three graphs in Fig 2 that you find different from the rest.
(Q4) Among those three, what is the most different one?

Define the t^{th}-paired-graph being the t^{th} graph in Fig 1 and that in Fig 2.
(Q5) Pick up three paired-graph that you find different from the rest.
(Q6) Among those three, what is the most different one?

Table 7.1 shows the frequencies of the various answers to the questions. The first column indicates the 25 graphs. The columns 2 to 7 show the frequency of the graphs being picked up with respect to the six questions in the survey. The last column is the sum of the number of times that the graph has been picked. The p-value for this test is 0.0769. We asked participants who picked up graph 21 and graph 9 about the likely reasons. Their responses to our addition question was that they made choices based on the shape of the black area toward the bottom and right, which was largest in graph 21 and smallest in graph 9.

The null model seems to model the distributions for categories in CO_{314} well because the heatplots of the simulated CO_{314} show similar dosage patterns to the heatplot of the real CO_{314}. Also, comparing the heatplots of the p-dissimilarity matrix, that of the real CO_{314} seems to be one of the possible heatplots of the p-dissimilarity of the null model. Moreover, graph 9, the real dataset, is not identified as a dataset that is significantly different from the null model. Therefore, we concluded that CO_{314} does not seem to have a clustering structure.

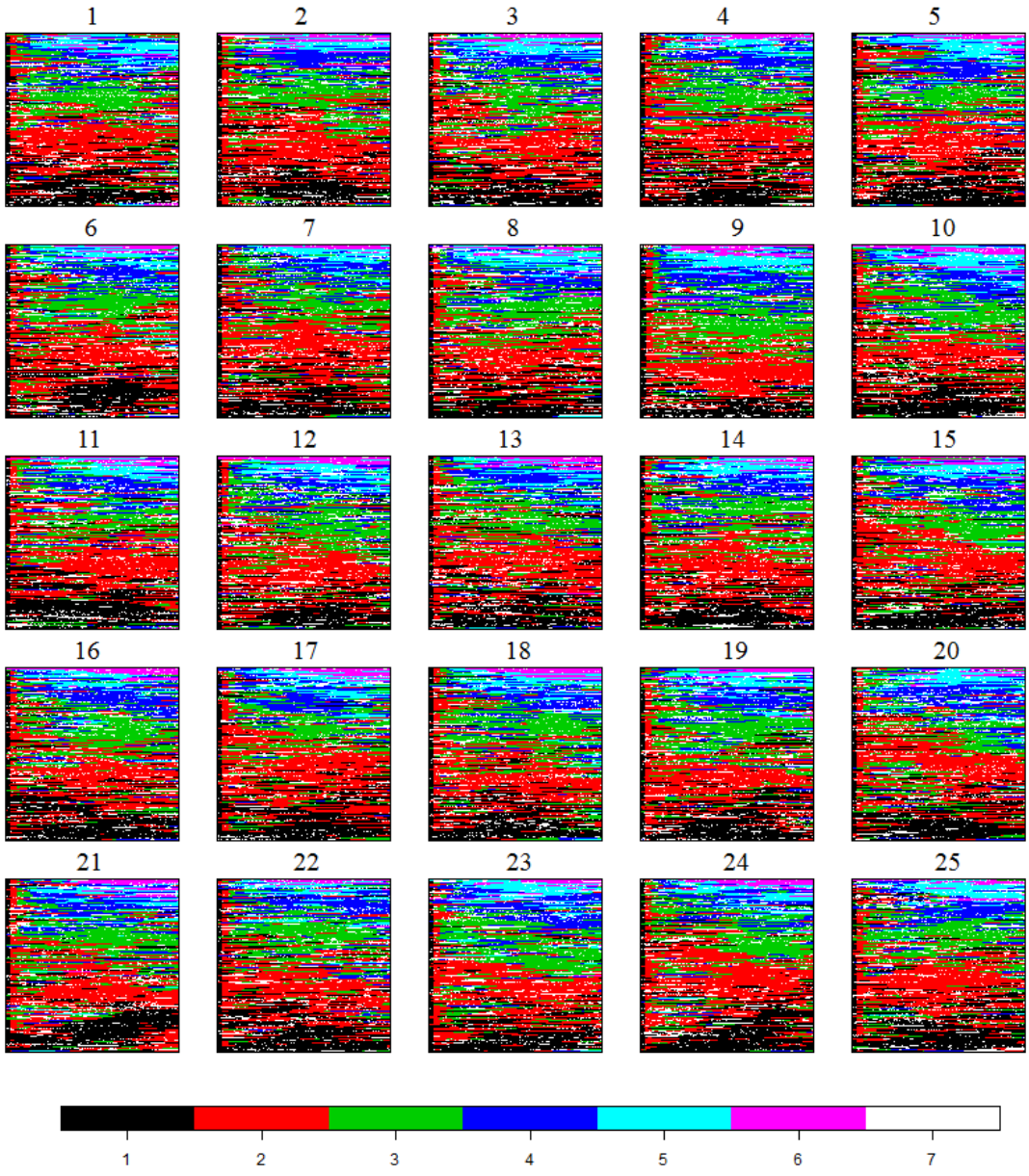


Figure 7.9: The heatplots of simulated CO_{314}. - All y-axes represent the participants of five clusters, while all x-axes represent the 180 days. The seven categories are represented by seven colours, black, red, green, blue, cyan, purple and white. The heatplot of the real CO_{314} is embedded in the simulated reference datasets.

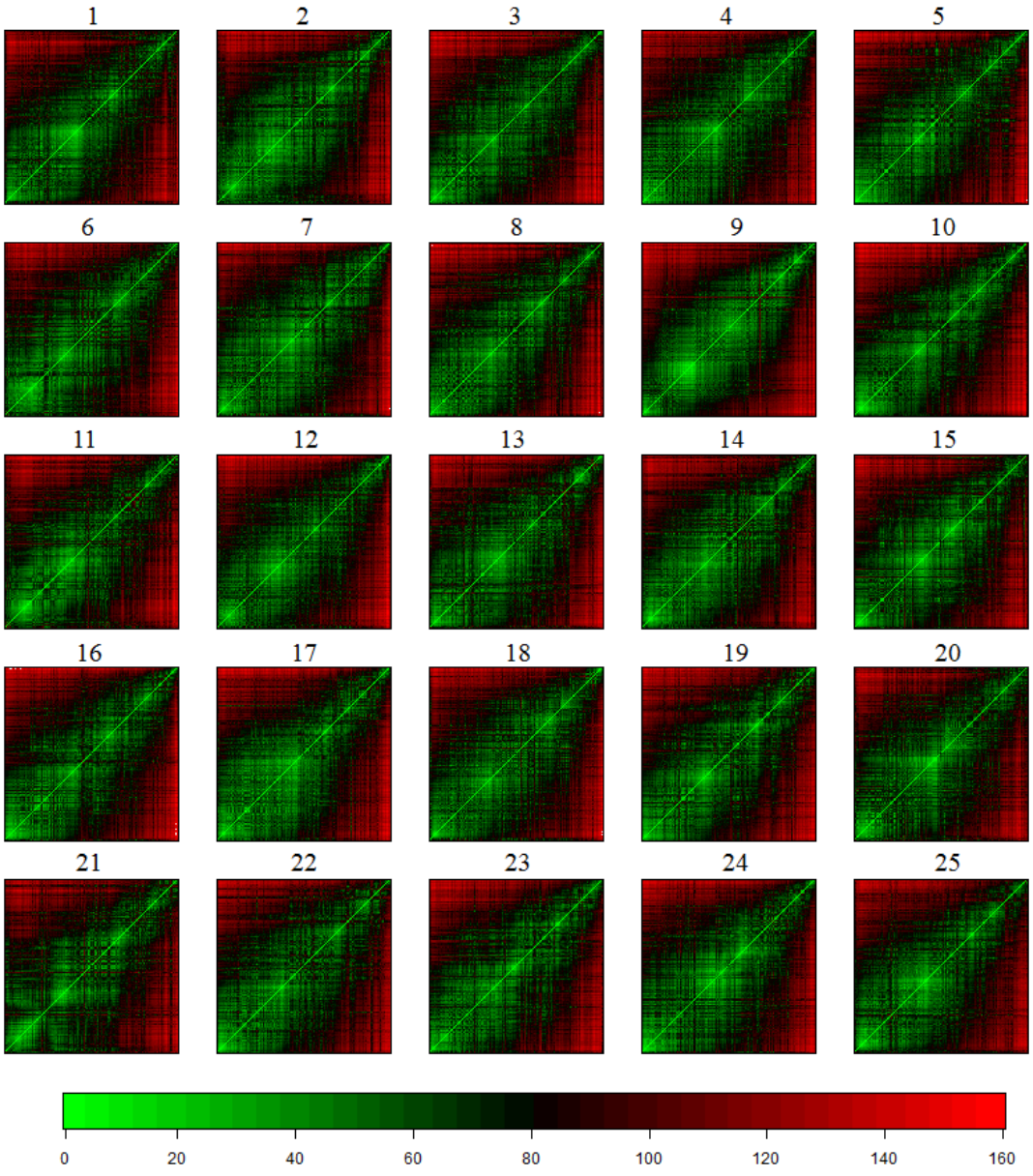


Figure 7.10: The heatplots of the p-dissimilarity matrix of simulated CO_{314}. - All x- and y-axes represent the participants of five clusters. The colour indicates the dissimilarity, ranging from 0 to 159, displayed in a sequence of green, black and red. The heatplot of the p-dissimilarity matrix of the real CO_{314} is embedded in the simulated reference datasets.

Table 7.1: The frequencies of answers to each question in the survey. The answers to the six questions of 18 participants who have at least an undergraduate degree. Note that the empty cell refers to a frequency of 0, and only 9 participants have answered the questions 5 and 6. Graph 21 is considered to be the real dataset, with a p-value=0.0769.

Label of the dataset	Q1	Q2	Q3	Q4	Q5	Q6	sum
1	1	1	2		3	1	8
2	3	1	3	1	1		9
3	2	1	2	2	2		9
4			3				3
5			5	1	1	1	8
6	7	4			1		12
7	1		1				2
8	1		1				2
9	5	4	4		2		15
10			1		1		2
11	4		3	2	2		11
13			1	1			2
14			2				2
15	2	1	1		1		5
16	1		1				2
17	1		2	1	1		5
18	3		1				4
19	1		2	1	1		5
20	5	2	4	1	1	1	14
21	7	3	9	5	5	4	33
22	2						2
23	1		1	1			3
24	2		2		3	1	8
25	5	1	3	1	2	1	13

Chapter 8

Sensitivity analysis, stability analysis and features of the final five clusters

In this chapter we investigate the stability of the final clustering by comparing clustering results for different settings of p and β. This is done by using the Adjusted Rand Index (Rand [1971]). We use the bootstrap distribution of the Jaccard coefficient (Hennig [2007]) to explore the stability of the clustering solution. We use the Adjusted Rand Index to measure the agreement of the final clustering of CO_{314} and the clustering result of the imputed CO_{314}. Afterwards, we display the demographical information related to the found five clusters.

8.1 Sensitivity analysis

To assess the sensitivity of the clustering, we evaluate how clustering results are affected by changes in the two parameters p and β in the p-dissimilarity. To explore the stability of the clustering, the Adjusted Rand index proposed by Hubert and Arabie [1985] is used. The concept of the Rand index (Rand [1971]) was to calculate the proportion of the agreement of two clustering solutions. By "agreement", we mean how similar two clustering solutions are when comparing any two objects. There are two cases of agreement. One of which is that the two objects are assigned to the same cluster in respect of one clustering solution, and to the same cluster in respect of the other

clustering solution. Another case of "agreement" is that two objects are in different clusters in respect of one clustering solution, and in different clusters in respect of the other clustering solution. Let X be the set of all n objects. Assume that two clustering methods are applied to X separately. Let k be the number of clusters, $(k \leq n)$. Let $\{G_i : i = 1, \ldots, k\}$ be the collection of the k clusters obtained from one of the clustering methods. Let $\{H_i : i = 1, \ldots, k\}$ be the k clusters obtained from the other clustering method. The two clustering results can be listed by a contingency table (See Table 8.1). The first column and the first row indicate the clusters. The n_{ij} shows the number of participants that are clustered in G_i and in H_j. Considering all possible combinations of paired participants $\binom{n}{2}$, Table 8.1 can be summarized by 4 numbers, a, b, c and d, where the value of a is the number of pairs of observations that are in the same cluster of $\{G_i, i = 1, \ldots, k\}$ and in the same cluster of $\{H_j, j = 1, \ldots, k\}$; d is the number of pairs of observations that are in different clusters whichever one of the two compared clustering solutions is selected; b is the number of pairs of observations that are in the same cluster G_i but in the different clusters H_j and $H_{j'}$; and c is the number of pairs of observations that are in different clusters G_i and $G_{i'}$ but in the same cluster H_j. The Rand Index is defined by

$$RI = \frac{a + d}{a + b + c + d}. \tag{8.1}$$

Thus, RI indicates how similar two clustering solutions are. Note that the expected value of RI for two random clusterings is not 0. The Adjusted Rand Index (Hubert and Arabie [1985]) is an improvement of the Rand Index and is defined by

$$ARI = \frac{\binom{n}{2}(a + d) - [(a + b)(a + c) + (c + d)(b + d)]}{\binom{n}{2}^2 - [(a + b)(a + c) + (c + d)(b + d)]}. \tag{8.2}$$

The expected value of the Adjusted Rand Index for two random clusterings is 0. The value of the Adjusted Rand Index can be negative and its maximum value, indicating strong agreement, is 1.

Stability of p and β

The final five clusters are obtained by partitioning the category-ordered data with the PAM clustering method and the p-dissimilarity where $p = 0.6$ and $\beta = 1.42$. Table 8.2 shows the agreement of the final five clusters and other clusterings obtained from

Table 8.1: Contingency table of two clustering results.- A total of n participants are partitioned into k clusters by two clustering methods. The results, which are the two k clusters, are represented by $\{G_i : i = 1, \ldots, k\}$ and $\{H_i : i = 1, \ldots, k\}$.

	G_1	G_2	$\ldots$	G_k	subtotal
H_1	n_{11}	n_{12}	$\ldots$	n_{1k}	$n_{1.}$
H_2	n_{21}	n_{22}	$\ldots$	n_{2k}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
H_k	n_{k1}	n_{k2}	$\ldots$	n_{kk}	$n_{k.}$
subtotal	$n_{.1}$	$n_{.1}$	$\ldots$	$n_{.1}$	n

various other choices of p. What can be observed is that the clustering results for $p = 0.4$, 0.5 and 0.7 show high agreement with the final clustering result ($p = 0.6$). This suggests that the p-dissimilarity with $p = 0.4, \ldots, 0.7$ generate very similar clustering results. The choice $p = 0.9$ scores the smallest Adjusted Rand Index 0.746, but even this is considered as representing a high agreement with the final cluster result. So, we can conclude that the choice of p does not influence the clustering result too strongly. Also, the values of the Adjusted Rand Index between the clustering result for $p = 0.6$ and $\beta = 1.42$ and those for $p = 0.6$ and $\beta = 0.5, 1, \ldots, 6$ are computed. All these values of the Adjusted Rand Index are 1. To sum up, the p which uses to distinguish categories and the β which represents the missing record in the p-dissimilarity do not have a strong impact on the performance of a cluster analysis of the real data.

Stability of the clustering

Several decisions have been made to perform a cluster analysis of the real data, such as the choice of the clustering method (PAM), the number of clusters (5), the dissimilarity function (the p-dissimilarity), and its parameters ($p = 0.6$, $\beta = 1.42$). The question arising now is how stable is the clustering of the real data? In the paper of Hennig [2007], in order to assess cluster stability, the author used the bootstrap distribution of the Jaccard coefficient, which gave the stability of every single clusters of a clustering. The method worked as follows. First of all, a bootstrap dataset was generated from real data. Next, a clustering method was applied both to the real data and the bootstrapped dataset. Two clustering results were obtained, one being the clusters obtained

Table 8.2: Stability of the p-dissimilarity of $p = 0.6$ to the found clusters.- The final five clusters are obtained by PAM clustering method on the p-dissimilarity of $p = 0.6$ and $\beta = 1.42$. This table shows the stability of $p = 0.6$ by employing Adjusted Rand Index (ARI) to compare the clustering result of $p = 0.6$ and $\beta = 1.42$ to that of various p and $\beta = 1.42$.

p	ARI	p	ARI
0.1	0.876	0.5	0.984
0.2	0.918	0.7	1.000
0.3	0.931	0.8	0.766
0.4	0.975	0.9	0.746

by applying the clustering method to the real data, and the other being the clusters constructed from the bootstrapped dataset. The two clustering results for the bootstrapped points could then be compared. This was done by a Jaccard coefficient. It computed every single clusters and its most similar bootstrapped cluster. The Jaccard coefficient of clusters G and H is defined by

$$\gamma(G, H) = \frac{\{\text{number of objects that belong to cluster } G \text{ and also belong to cluster } H\}}{\{\text{all objects in either clusters } G \text{ or } H\}}$$

A total of M bootstrapped datasets were generated and the average Jaccard coefficients for every single cluster over M replications, denoted by $\overline{\gamma}$, were used to evaluate the cluster stability. $\overline{\gamma} > 0.75$ indicates good recoveries (Hennig [2007]). We used their method to run $M = 30$ bootstrap replications for CO_{314}. The result of the Jaccard coefficients for the five clusters in our final clustering result are 0.852, 0.786, 0.764, 0.853 and 0.966. These figures show that the five clusters are very stable and no cluster is extremely unstable.

8.2 Comparison between CO_{314} and the imputed datasets

In Section 2.4.2 three datasets were created by applying the imputation methods to CO_{314} and $Dosage_{314}$. They are 1) $ImpCO_{314}$ in which participants continuously lack of 14 days records were not imputed, 2) $ImpCO^{7}_{314}$ in which participants continuously lack of 7 days records were not imputed, 3) $ImpDosage_{314}$ in which the dosage dataset $Dosage_{314}$ was imputed by a linear interpolation. They are used to see the effect of

treating the missing dosages in CO_{314} the same.

For category-ordered data, the PAM with $k = 5$ and the p-dissimilarity with $p = 0.6$ and $\beta = 1.42$ are applied. For dosage data, the PAM with $k = 5$ and the Euclidean distance are applied. The ARI for CO_{314} and the three imputed datasets, $ImpCO_{314}$, $ImpCO^{7}_{314}$ and $Dosage_{314}$, are 0.718, 0.726 and 0.54. Both clustering solutions of $ImpCO_{314}$, $ImpCO^{7}_{314}$ show an agreement with the final clustering solution of CO_{314}, and the clustering solutions of $Dosage_{314}$ and CO_{314} were fairly similar.

Table 8.3 shows the crosstable of the clustering solutions between CO_{314} and $ImpCO_{314}$. Of those participants assigned to the same cluster in respective of the cluster solution of CO_{314}, most of them are assigned to the same cluster in respective of the cluster solution of $ImpCO_{314}$. We know from Figure 7.8 that participants in $G_{(3)}$ tend to be scattered. Also, the third row in the Table 8.3 shows that there are 82 participants in $G_{(3)}$ by CO_{314}. Of these 82 participants, 15 are assign to $G_{(2)}$ and $G_{(4)}$ by $ImpCO_{314}$. Around 18% of the participants are assigned to different clusters.

We applied the algorithm of the projections to the clustering result of $ImpCO_{314}$. Figure 8.1 shows the heatplot of $Dosage_{314}$ and that of the p-dissimilarity matrix of $ImpCO_{314}$. What can be observed is that the heatplot of $Dosage_{314}$ looks similar to that in Figure 7.8.

Based on the result and values of ARI, we conclude that treating all the missing dosages the same, regardless of their length, will not influence the clustering result too strongly.

8.3 Result of dosage patterns

One month is often regarded as the minimum length of receiving methadone treatment. Of those participants who complete the treatment for a month, some of them will continue on receiving MMT. Moreover, if they stay in MMT for three months, the possibility of overcoming their addictions becomes higher. Also, participants who stay in MMT for six months are considered to be candidates who can achieve abstinence, so

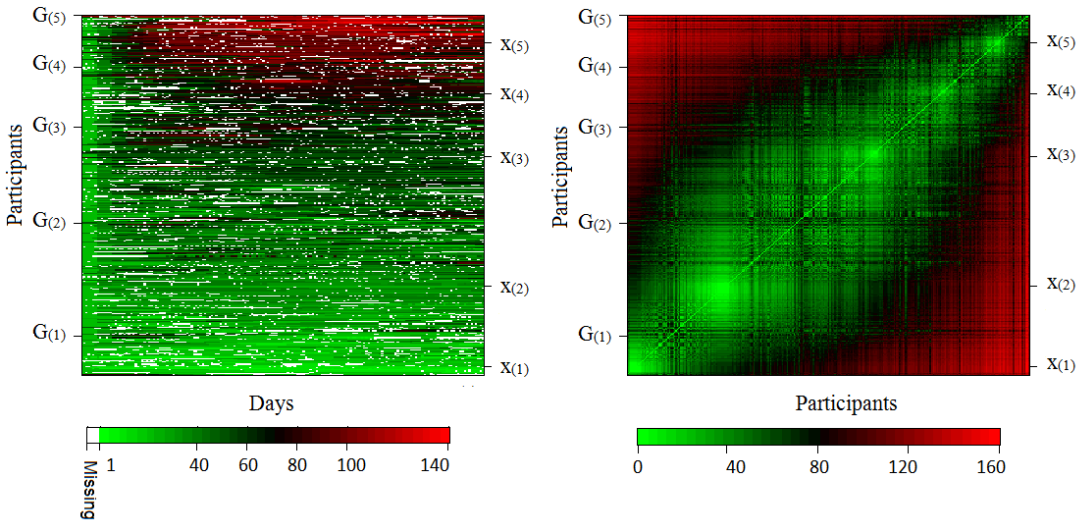


Figure 8.1: Heatplot of Dosage$_{314}$ and of the p-dissimilarity matrix of ImpCO$_{314}$ with the order obtained by the algorithm of the projections. - The graph on the left displays Dosage$_{314}$. The record of the 180 days of a participant is plotted horizontally. The colour indicates the dosage. The missing and 0 mg are shown in white, while the dosages between 1 and 140 mg are represented by a sequence of green, black and red. The participants belonging to the same cluster are organized in terms of the standardized dissimilarity. The graph on the right shows the p-dissimilarity matrix of ImpCO$_{314}$. Both the x- and y-axes represent participants.

Table 8.3: The crosstable of the clustering solution of CO$_{314}$ and that of ImpCO$_{314}$ The clusters obtained from CO$_{314}$ are shown in the first column. The remaining columns show the cluster obtained from ImpCO$_{314}$. Of those participants assigned to the same cluster by the cluster solution of CO$_{314}$, most of them are assigned to the same cluster by the clustering solution of ImpCO$_{314}$.

CO$_{314}$ \ ImpCO$_{314}$	1	2	3	4	5	Row Sum
1	35	2	0	0	0	37
2	0	84	12	0	0	90
3	0	6	69	9	0	82
4	0	0	1	47	4	58
5	0	0	0	2	43	47
Col Sum	35	98	84	52	45	

Table 8.4: The mean and standard deviation of dosage of the found five clusters over three time intervals.- The first column show three time intervals for which the mean (standard deviation) of dosages are calculated. Columns 2 to 6 show the mean (standard deviation) of dosages of the final five clusters.

	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5
Mean (SD)					
Day 1-30	24 (15)	32 (13)	38 (15)	45 (22)	53 (25)
Day 31-90	20 (11)	37 (13)	48 (14)	69 (17)	88 (21)
Day 91-180	19 (9)	35 (11)	53 (11)	69 (15)	98 (18)

the physicians are interested in those who are active during the 1st month and carry on MMT for three months, and those who stay in MMT for six months. Since one month, three months and six months are meaningful time intervals, we monitor the changes of dosage of the five clusters in relation to three time intervals, namely day 1 to 30, day 31 to 90 and day 91 to 180.

Table 8.4 shows the mean and standard deviation of dosage for the five clusters for the three intervals. We observe that all clusters have their mean dosage go up from the first month to the third month. This show a process of detoxification. The clusters 1 and 2 have their mean dosage go down at the sixth month. This might reflect that participants are trying to quit methadone. The clusters 3,4,5 have their mean dosage go up. This suggest that participants who are highly addicted to heroin might take longer to finish the process detoxification.

Figure 8.2 shows the frequency of the categories from day 1 to day 30 for the five clusters. The y-axis indicates category and the x-axis indicates days. The colour designates frequency. We observe the followings: (1) There is an upward trend in categories over time, particular in cluster 5. (2) Cluster 1 seems to have more category 7 than cluster 5. Similarly, Figure 8.3 shows the frequency of the categories from day 31 to day 180 for the five clusters. In this figure, frequency is defined as the number of categories for a 7-day period. For convenience, we define pattern of detoxification in three stages. Stage I represents methadone dosage goes up, stage II represents dosage stays

stable, and stage III represents dosage goes down. We observe the followings: (1) In cluster 1, the participant who continuously has dosages in category 3 moves to category 2 on day 85 and moves to category 1 on day 109. Another participant has dosages in category 2 on days 107-119, category 3 on days 120-133, category 2 on days 134-147, category 3 on days 148-180. Most of the participants stay in low dosages. (2) In cluster 2, category 3, the frequency increases at the beginning and then goes down after day 50. This suggests that some participants with low addictions go into the stage III in approximately 2 months. Also, around day 120, the frequency of category 3 goes up again. This suggests that some participants take around 3 months to go into stage II. (3) In cluster 3, many participants stay in either category 2 or category 3. After about day 50, many participants start to move to category 3 from category 2. After day 140, many participants in category 3 move to category 2. (4) In cluster 4, category 5, the frequency slightly goes down after day 150. Also, for category 3, the frequency is fairly stable on days 100-150. Few participants in higher categories move to category 3 after day 150. (5) In cluster 5, participants contiguously move to higher categories from lower categories. The frequency of category 4 is fairly stable on days 100-150. The frequencies of categories 3-4 go up after day 150, that is, few participants move to lower categories after day 150.

We attempt to summarize the pattern of detoxification for each cluster by the date on which the three stages are observed: Cluster 2 (day 1-40-100), Cluster 3 (day 1-80-140), cluster 4 (1-100-150), cluster 5 (1-100-150). Note that these dates are roughly numbers. Also, majority in cluster 4 and cluster 5 have their dosages in high categories. This means that participants who are highly addicted to heroin might take longer to finish the detoxification process.

Next, we used MDS plot to monitor the movement of the five clusters from one time interval to the next. We applied a three dimensional MDS to the p-dissimilarity matrix of the category-ordered data at each of the three time intervals. First of all, the three dimensional MDS was applied to the p-dissimilarity matrix of the records of days 1 to 30, the records of days 31 to 90 and records of days 91 to 180 separately. The stress for those models were 0.1, 0.09 and 0.08, respectively. Figure 8.4 illustrates how the clusters moved apart depending on the dosage of their participants. What can be

observed from Figure 8.4(a) is that most of the points of cluster 1 stay close to each other and that there is a tail that is formed by a few points in clusters 4 and 5. Figure 8.4(b) shows the MDS result for the second interval. As can be seen, cluster 5 is distant from clusters 1, 2 and 3. There is a mixing of points in clusters 2 and 3. Figure 8.4(c) presents the results for the third interval. This graph displays very clearly each of the five clusters. What can be concluded is that the dosage patterns of the clusters in the first month overlap, and that they begin to show some difference in the following three months. They are also clearly distinguishable from the third month to the sixth month.

8.4 Demographical information relating to the five clusters

Table 8.5 shows age, age of the onset of heroin, gender, education, marital status and occupation for the five clusters. The one-way ANOVA is used to compare mean ages of the five clusters and the Chisq-test is performed to identify variables associated with clusters. The result of ANOVA shows that there is not enough evidence to conclude that the mean ages of the five clusters are different (p-value=0.084). The result of ANOVA test the mean ages of heroin onset is borderline significant (p-value=0.062), there could be some effect of age of heroin onset. The result of Chisq-test shows that there is not enough evidence to conclude that there exists a relationship between clusters with respect to gender (p-value=0.377), education (p-value0.996), marital status (p-value=0.429) and occupation (p-value=0.310).There are fewer females than males in each of the five clusters, with an average of proportion of females 17.76 %. The highest female-male ratio is 23.40 % in cluster 5, while the smallest is 10.98 % in cluster 3. With regards education, about half of the participants have received basic education (elementary and junior high school) in all five clusters. With regards marital status, an average proportion of the single and divorced participants in the five clusters is 76.18%. With regards occupation, more than half of the participants have jobs, with an average proportion of 63.27%.

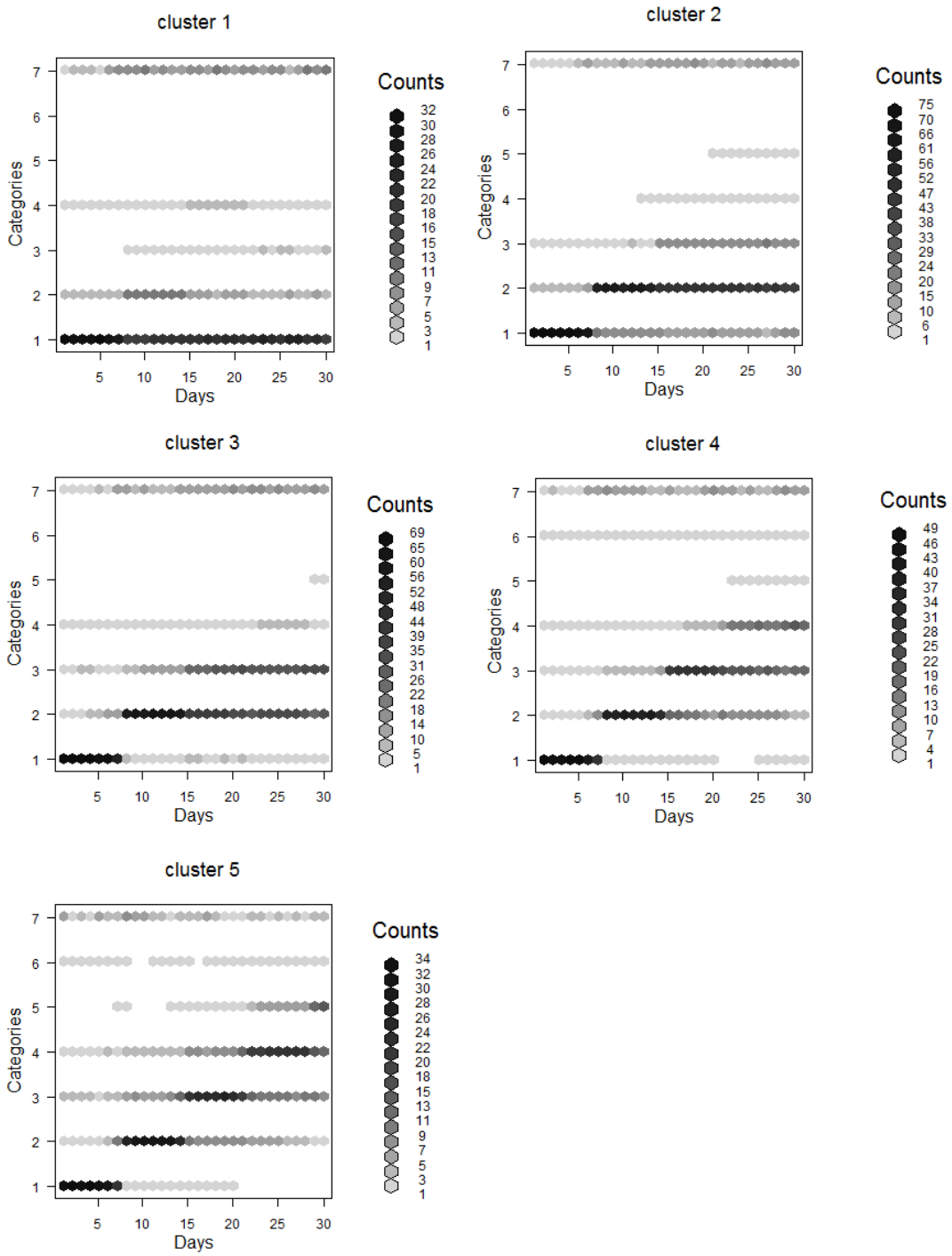


Figure 8.2: Frequency of the categories from day 1 to day 30 for the five clusters. - The y-axis indicates category and the x-axis indicates days. The colour designates frequency.

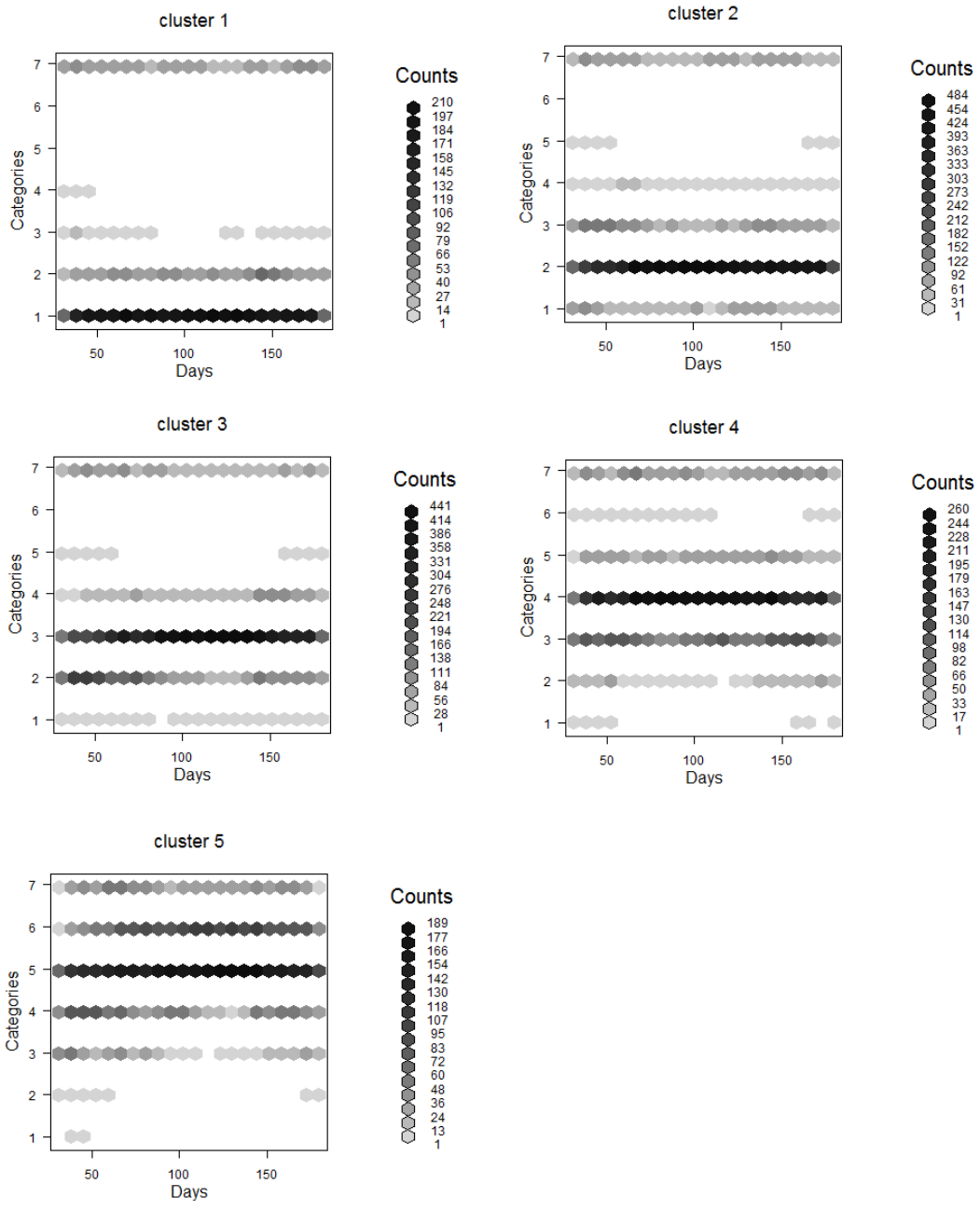


Figure 8.3: Frequency of the categories from day 31 to day 180 for the five clusters. - The y-axis indicates category nd the x-axis indicates days. The colour designates frequency. The frequency is calculated based on a 7-day period.

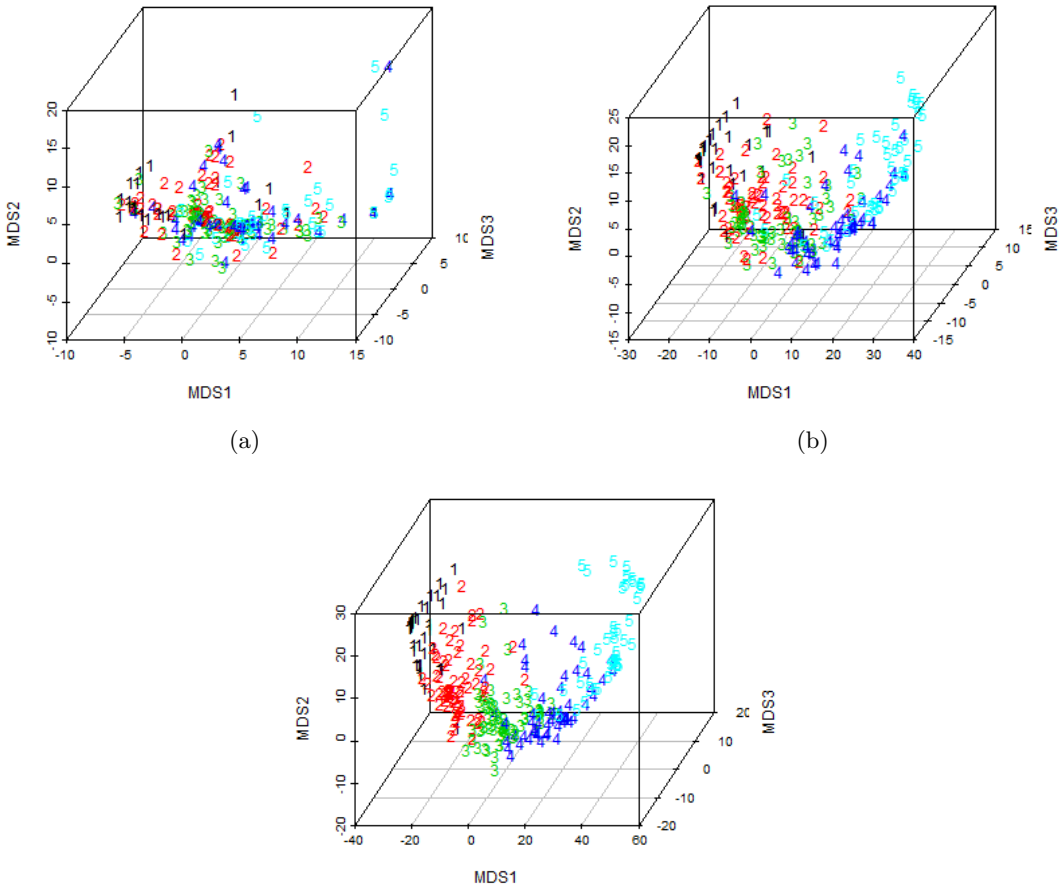


Figure 8.4: The movement of clusters over time. - The three dimensional MDS is used to illustrate the movements of the five clusters for three intervals. The graphs (a), (b) and (c) show the three dimensional MDS of the p-dissimilarity matrix of the category-ordered records of days 1 to 30, days 31 to 90 and days 91 to 180, respectively. The dosage patterns of the clusters in the first month overlap, and they begin to show some difference in the following three months. They are distinguishable from the third month to the sixth month.

Table 8.5: Demographical information relating to the five clusters.-Age is expressed as mean standard deviation. The one-way ANOVA is used to compare mean ages of the five clusters and the Chisq-test is performed to identify categorical variables associated with clusters. The Chisq-test shows that there is not enough evidence to conclude that there exists a relationship between clusters with regards any of the following features: gender, education, marital and occupation.

		Clusters					
		1	2	3	4	5	p-value
Number of participants		37	90	82	58	47	
Age	mean	35.14	39.41	38.12	34.81	36.24	0.0839
	SD	6.45	7.43	7.61	6.10	6.80	
Age of heroin onset	mean	25.58	26.24	26.79	23.84	24.29	0.0621
	SD	7.37	7.07	7.22	5.67	5.98	
Gender	female	8	14	9	10	11	0.3777
	male	29	76	73	48	36	
Education	elementary	4	6	6	4	3	0.9963
	junior high	14	39	35	23	21	
	high school	19	44	38	30	19	
	undergraduate	0	1	2	0	3	
Marital	single	24	45	39	38	25	0.4294^{+}
	married	5	26	21	13	9	
	divorced	8	18	20	6	11	
	windowed	0	1	0	0	0	
	living with partner	0	0	1	1	1	
Occupation	Yes	23	56	56	31	33	0.3102
	No	14	34	26	27	13	

+: (single, divorced, windowed) and (married, living with partner)

Chapter 9

Conclusion and discussion

This thesis covers the various stages of clustering analysis, including data transformation, selection of dissimilarity functions, selection of clustering methods, determination of number of clusters, test of homogeneity, quality check for clustering results and interpretation. Also, we define the category-ordered data and propose the following: the p-dissimilarity, the modification of the Prediction Strength, the null model test of homogeneity, the null model test for determination of numbers of cluster, the Markov model for the category-ordered data and two ordering algorithms for information visualisation via heatplots.

Study design, data collection, data quality, etc., have a huge impact on the findings. Sadly, many study plans do not involve statistician from the beginning of the research resulting in statisticians having to spend great effort on understanding data and on data structuring. Also, it might result in research limitations on the data or worse if the data is not able to answer the proposed research questions.

Quality check of data and understanding data are the first steps of performing an analysis. Having analysed the MMT data, we witnessed some problems. (1) The datasets in the MMT database were not synchronized. This reduced the number of participants with full records. This could be improved by making staff aware of the importance of having a completed dataset. (2) The difference between prescribed dosages and dosages taken by participants reflected whether prescriptions were appropriate and whether participants had followed physicians' instructions. However, there was no in-

dication for which one of the multiple prescriptions was used. This problem could be avoided by listing the full research objectives and then creating a recording system with the flexibility to accommodate the research objectives and future research purposes. (3) The coding for dosage taken records was not unique. This problem could be avoided by designing a coding book for the MMT recording system. Despite these facts, we selected a meaningful sample with 314 participants. We took account of the weekly prescriptions, fluctuations in dosage taken records and patterns of missing dosages, and defined the category-ordered data CO_{314} by transforming the plain dosage data Dosage_{314} into categories. The final clusters were obtained by using the PAM method with five clusters and the p-dissimilarity with $p = 0.6$ and $\beta = 1.42$. Although none of the five clusters could very easily be distinguished in terms of, say, their demographics, the sequences of categories for the five clusters were clinically useful. The sequences of categories indicated detoxification. We found the heroin onset age might have an influence on the patterns of detoxification. Participants with low addictions reduced the use of heroin by addicting to methadone at the first month and attempted to reduce/quit the use of methadone at the third month. As for participants with high addictions, few attempted to reduce the use of methadone at the fifth month and most required more time to finish the detoxification process.

Regarding developing methodologies, our first contribution was to propose the p-dissimilarity. The p-dissimilarity was based on assessing the interpretative dissimilarity between categories and focused more on sequence of constancy and less on sudden changes in categories. This was used to measure dissimilarity between the 180-day time series of the participants. Also, it implemented concepts of variables having information on categorical and ordinal, and thus can be used for incomplete data. The p-dissimilarity uses p as the switch between data being categorical and ordinal, and uses β to deal with missing values. Further, we showed that values of p and β do not have as strong an impact on clustering as those measured by the Adjusted Rand Index. Moreover, the p-dissimilarity quantifies the structure of the categories which are partly categorical, partly ordinal and also contains quantitative information. The principle behind the measure can be used in a wider field of applications, in which there is more information about the meaning of categories than just those that are "ordinal" or "categorical", such as survey studies. These studies use questionnaires with choices

on Likert scales and a don't know-category and the researchers have a quantitative idea about the interpretative distance between categories.

The Prediction Strength determines the number of clusters by measuring cluster stability. It favours the K-Means method. We proposed rules to modify the use of the Prediction Strength so that it could be fully applied to the hierarchal clustering methods and the PAM method. Additionally, instead of preselecting the clustering method, we let data to decide on the basis of cluster stability and cluster coherence, which were measured by the Prediction Strength and the Average Silhouette Width.

We proposed the null model test for determining the number of clusters and for testing homogeneous population. This method took account of data structure not arising from clustering and constructed the distribution of the statistic so that a hypothesis test for each number of clusters could be performed. Moreover, this allowed us to investigate the existence of a clustering structure. In this study, we constructed the Markov null model to represent the category-ordered data with no clustering structure. Also, we carried out a graphical test to validate the Markov model. The result showed that the Markov model seemed to be a good model and there were no significant clusters.

We proposed two ordering algorithms to visualise information via heatplots. The first general use algorithm employed multidimensional scaling (MDS). It aimed at preserving cluster structures, similarity structure of clusters and that of objects in a cluster. The second PAM method algorithm used projection vector. This second algorithm aimed at preserving the aforementioned information, locating medoids, displaying how far apart the medoids were, and displayed density of the clusters by way of a colour gradient around the medoids. We used the algorithms and heatplot to access the quality of clustering. For CO_{314}, the result of the Jaccard coefficients for the five clusters in our final clustering result were 0.852, 0.786, 0.764, 0.853 and 0.966, which showed that the five clusters were very stable. From the heatplot of the clustering result, we observed that the participants in cluster 1 had similar dosage patterns. So do cluster 4 and cluster 5. Participants in clusters 2 and 3 tended to be scattered.

Future work will focus on the development of the p-dissimilarity and explore the null model test. In some applications, variables are correlated where researchers have a quantitative idea about the interpretative distance between grouped variables. Therefore, we will take account of correlation between variables in order to improve the use of the p-dissimilarity and the null model test for the existence of clustering structures. This method is limited on the MMT data, which was found to have no clustering structure. As such, its performance on data with clustering is unknown. Therefore, we will simulate data with clustering structure. Then, we will apply the null model test to determine the number of clusters and test the absence of clustering in order to explore and improve the null model test. Apart from methodology, we are also interested in comparing the patterns found in our study with the clustering results of recent data and then aim to construct intervals of prescribed dosages in relation to clusters.

Bibliography

O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46:243–256, 2013.

J.C. Ball and A. Ross. *The Effectiveness of Methadone Maintenance Treatment*. Springer-Verlag, 1991.

Z. Bar-Joseph, D. K. Gifford, and T. S. Jaakkola. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, 17:22–29, 2001.

E. Bellin, J. Wesson, V. Tomasino, J. Nolan, A.J. Glick, and S. Oquendo. High dose methadone reduces criminal recidivism in ppiate addicts. *Addiction Research*, 7: 19–29, 1999.

D. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. *AAAI-94 workshop on knowledge discovery in databases*, pages 229–248, 1994.

H. H. Bock. On some significance tests in cluster analysis. *Classification*, 2:77–108, 1985.

H. H. Bock. Probabilistic models in cluster analysis. *Computational Statistics & Data Analysis*, 23:5–28, 1996.

A. Buja, D. Cook, H. Hofmann, M. Lawrence, E. Lee, D. F. Swayne, and H. Wickham. Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society*, 367:4361–4383, 2009.

R.B. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3:1–27, 1974.

C. H. Chen. Generalized association plots: information visualization via iterative generated correlation matrices. *Statistica Sinica*, 12:7–29, 2002.

T.F. Cox and M.A.A. Cox. *Multidimensional Scaling.* Chapman & Hall, 1990.

A.P.M. Coxon and P.M. Davies. *The user's guide to multidimensional scaling.* Heinemann Educational Books, 1982.

T. D'Aunno and H. A. Pollack. Changes in methadone treatment practices: results from a national panel study, 1988-2000. *American Medical Association*, 288:850–856, 2002.

A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application.* Cambridge University press, 1997.

A. Douzal-Chouakria and P. Nagabhushan. Adaptive dissimilarity index for measuring time series proximity. *Advances in Data Analysis and Classification*, 1:5–21, 2007.

A. Douzal-Chouakria, A. Diallob, and F. Giroudb. Adaptive clustering for time series: application for identifying cell cycle expressed genes. *Computational Ststistics and Data Analysis*, 53:1414–1426, 2009.

R. Duda, P. Hart, and D. Stork. *Pattern Classification.* Wiley-Interscience, 2001.

B. Efron and R. Tibshirani. *An introduction to the bootstrap.* Chapman and Hal, 1993.

Y. Fang and J. Wang. Selection of the number of clusters via the bootstrap method. *Computational Statistics & Data Analysis*, 56:468–477, 2012.

C. Fraley and A. Raftery. Model-based clustering, discriminat analysis, and density stimation. *Journal of the American Statistical Association*, 97:611–631, 2002.

C. Fraley and A. Raftery. Mclust version 3 for r: normal mixture modeling and model-based clustering. *Technical Report No.504*, 2010.

N. Gale, W.C. Halperin, and C.M. Costanzo. Unclassed matrix shading and optimal ordering in hierarchical cluster analysis. *Classification*, 1, 1984.

A.D. Gordon. Constructing dissimilarity measures. *Classification*, 7:257–269, 1990.

A.D. Gordon. *Classification*. Chapman and Hall, 1999.

M. Gossop, J. Marsden, D. Stewart, and A. Rolfe. Patterns of improvement after methadone treatment: 1 year follow-up results from the national treatment outcome research study. *Drug Alcohol Abuse*, 60:275–286, 2000.

J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27:857–874, 1971.

J. C. Gower and P. Legendre. Matrics and euclidean properties of dissimilarity coefficients. *Classification*, 3:5–48, 1986.

M. Hahsler and K. Hornik. Dissimilarity plots: a visual exploration tool for partitional clustering. *Computational and Graphical Statistics*, 10:335–354, 2011.

M. Hahsler, K. Hornik, and C. Buchta. Getting things in order: an introduction to the r package seriation. *Statistical Software*, 25:1–34, 2008.

J. A. Hartigan. *Clustering algorithms*. Wiley, 1975.

J.A. Hartigan and M.A. Wong. A k-means clustering algorithm. *Applied Statistics*, 28: 100–108, 1979.

C. Hennig. Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, 52:258–271, 2007.

C. Hennig and B. Hausdorf. Design of dissimilarity measures: a new dissimilarity measure between species distribution ranges. *Data Science and Classification, Springer, Berlin*, pages 26–38, 2006.

C. Hennig and T.F. Liao. How to find an appropriate clustering for mixed type variables with application to socio-economic stratification. *Royal Statistical Society Series C*, 62:1–25, 2013.

L. Hubert and P. Arabie. Comparing partitions. *Classification*, 2:193–218, 1985.

A. K. Jain and R. C. Dubes. *Algorithms for clustering data*, chapter 4. Prentice Hall, 1988.

R. E. Johnson, M. A. Chutuape, E. C. Strain, S. L. Walsh, M. L. Stitzer, and G. E. Bigelow. A comparison of levomethadyl acetate, buprenorphine, and methadone for opioid dependence. *The New England Journal of Medicine*, 343:1290–1297, 2000.

L. Kaufman and P.J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley and Sons, 1990.

J.B. Kruskal. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29:115–129, 1964.

W. Krzanowski and Y. Lai. A criterion for determing the number of groups in a data set using sum-of-squares clustering. *Biometrics*, 44:23–34, 1988.

M. Langendam, H. Van Haastrecht, G. Van Brussel, A. Van den Hoek, R. Coutinho, and E. Van Ameijden. Research report differentiation in the amsterdam dispensing circuit: determinants of methadone dosage and site of methadone prescription. *Addiction*, 93:61–72, 1998.

Friedrich Leisch. Visualizing cluster analysis and finite mixture models. In C.H. Chen, W. Härdle, and A. Unwin, editors, *Handbook of data visualization*, pages 561–587. Springer Berlin Heidelberg, 2008.

B. Li. A new approach to cluster analysis: The clustering-function-based method. *Royal Statistical Society Series B*, 68:457–476, 2006.

T.F. Liao. Measuring and analyzing class inequality with the gini index informed by model-based clustering. *Sociological Methodology*, 36:201–224, 2006.

G. De Luca and P. Zuccolotto. A tail dependence-based dissimilarity measure for financial time series clustering. *Advances in Data Analysis and Classification*, 5: 323–340, 2011.

I. Maremmani, M. Pacini, S. Lubrano, and M. Lovrecic. When enough is still not enough: effectiveness of high-dose methadone in the treatment of heroin addiction. *Heroin Addiction and Related Clinical Problems*, 5, 2003.

L. A. Marsch. The efficacy of methadone maintenance interventions in reducing illicit opiate use, HIV risk behavior and criminality: a meta-analysis. *Addiction*, 93, 1998.

C. Masson, P. Barnett, K Sees, K. Delucchi, A. Rosen, W. Wong, and S. Hall. Cost and cost-effectiveness of standard methadone maintenance treatment compared to enriched 180-day methadone detoxification. *Addiction*, 99:718–126, 2004.

R. P. Mattick, J. Kimber, C. Breen, and M. Davoli. Buprenorphine maintenance versus placebo or methadone maintenance for opioid dependence. *Cochrane Database of Systematic Reviews*, 3, 2009.

S. Maxwell and M. Shinderman. Optimizing long-term response to methadone maintenance treatment. *Addictive Diseases*, 21:3:1–12, 2002.

G. McLachlan and K. Basford. *Mixture models : inference and applications to clustering*. Marcel Dekker Incorporated, 1988.

G. McLachlan and D. Peel. *Finite mixture models*. John Wiley and Sons, 2000.

A. T. McLellan, L. Luborsky, J. Cacciola, J. Griffith, F. Evans, H. Barr, and C. O'Brien. New data from the addiction severity index: reliability and validity in three centers. *Nervous and Mental Disorders*, 173:412–423, 1985.

G. Milligan and M. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 59:159–179, 1985.

H. Murray, R. McHugh, E. Behar, and E. Pratt. Personality factors associated with methadone maintenance dose. *Drug and Alcohol Abuse*, 34, 2008.

E. Peles, S. Schreibera, Y. Naumovskya, and M. Adelsona. Depression in methadone maintenance treatment patients: Rate and risk factors. *Affective Disorders*, 99:213–220, 2007.

K. I. Powers and M. D. Anglin. Cumulative versus stabilizing effects of methadone maintenance. *Evaluation Review*, 17:243–270, 1993.

D. Pud, C. Zlotnick, and E. Lawental. Pain depression and sleep disorders among methadone maintenance treatment patients. *Addictive Behaviors*, 37:1205–1210, 2012.

W. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, pages 846–850, 1971.

C. Ratanamahatana and E. Keogh. Making time-series classification more accurate using learned constraints. *In proceedings of SIAM International Conference on Data Mining*, pages 22–24, 2004.

P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics*, 20:53–65, 1987.

R. H. Shumway and D. S. Stoffer. *Time Series analysis and its applications: with R Examples*. Springer, 2010.

S. S. Stevens. On the theory of ccales of measurements. *Science*, 103:677–680, 1946.

E. C. Strain, M. L. Stitzer, I. A. Liebson, and G. E. Bigelow. Dose-response effects of methadone in the treatment of opioid dependence. *Annals of Internal Medicine*, 119: 23–27, 1993a.

E. C. Strain, M. L. Stitzer, I. A. Liebson, and G. E. Bigelow. Methadone dose and treatment outcome. *Drug and Alcohol Dependence*, 33:105–117, 1993b.

F. Termorshuizen, A. Krol, M. Prins, and E. van Ameijden. Long-term outcome of chronic drug use: the amsterdam cohort study among drug users. *American Journal of Epidemiology*, 161:271 –279, 2005.

R. Tibshirani and G. Walther. Cluster validation by prediction strength. *Computational and Graphical Statistics*, 14:511–528, 2005.

R. Tibshirani, G. Walther, D. Botstein, and P. Brown. Cluster validation by prediction strength. Technical report, 2001.

Y.J. Tien, Y.S. Lee, H.M. Wu, and C.H. Chen. Methods for simultaneously identifying coherent local clusters with smooth global patterns in gene expression profiles. *BMC Bioinformatics*, 9:1–16, 2008.

J. Ward, W. Hall, and R.P. Mattick. Role of maintenance treatment in opioid dependence. *Lancet*, 353:221–226, 1999.

C. Winick. Maturing out of narcotic addiction. *Bull Narc*, 14:1–7, 1962.

H.M. Wu, Y.J. Tien, and C.H. Chen. Gap: a graphical environment for matrix visualization and cluster analysis. *Computational Statistics & Data Analysis*, 54:767–778, 2010.