On Clustering Comparisons Using Data from a Seroprevalence Study

ON CLUSTERING COMPARISONS USING DATA FROM A SEROPREVALENCE STUDY

BY

Chandra Grewal, B.Sc.

A THESIS
SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS
AND THE SCHOOL OF GRADUATE STUDIES
OF MCMASTER UNIVERSITY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Master of Science (2021) McMaster University
(Statistics) Hamilton, Ontario, Canada

TITLE:	On Clustering Comparisons Using Data From A Seroprevalence Study
AUTHOR:	Chandra Grewal B.Sc. (Mathematics), Queen's University, Kingston, Ontario, Canada
SUPERVISOR:	Dr. Paul D. McNicholas
NUMBER OF PAGES:	x, 51

To my mother, father and brother for their endless support and love.

Abstract

Various longitudinal clustering approaches are discussed and compared on an application to a seroprevalence study. The data contains information about the behaviours of individuals throughout the course of the COVID-19 pandemic. First, a review of the various longitudinal clustering methods compared throughout this thesis is discussed. Longitudinal k-means, growth mixture models, latent class growth analysis and a two-step approach involving growth curve models and k-means are reviewed. Longitudinal model-based clustering based on a modified Cholesky decomposition of a Gaussian mixture and Gaussian linear means are also reviewed. The BIC is used as the primary criterion to determine the number of components, and the ARI is used to determine cluster similarity between models. The various clustering approaches are then compared as they attempt to identify gathering patterns within the population of the seroprevalence dataset.

Acknowledgements

First, I would like to thank my supervisor Dr. Paul D. McNicholas for his continued guidance and support throughout my graduate studies at McMaster University. I am grateful for his advice and direction even as we navigated through these unfamiliar learning circumstances. Secondly, I would like to thank my examining committee, Dr. Dawn Bowdish and Dr. Anas Abdallah for their time. Finally, to my family and friends for their love and support the entire way.

Contents

List of Figures **ix**

List of Tables **x**

1 Introduction **1**

2 Methodology **4**

2.1 Clustering 4

2.2 Mixture-Model Based Methods 5

2.2.1 Finite Mixture Models 5

2.2.2 Modified Cholesky Decomposition 6

2.2.3 Gaussian Mixture Modelling of Longitudinal Data 6

2.2.4 Linear Means 8

2.2.5 EM Algorithm for Model-based Clustering 8

2.2.6 Convergence Criterion for EM Algorithm 9

2.3 Growth Curve Models 10

2.4 Growth Mixture Modelling 11

2.5 Latent Class Growth Analysis 12

2.6 K-means Clustering 13

2.7 Two-Step Clustering 14

2.8 Model Selection 15

2.9 Measuring Similarity 16

3 Seroprevalence Data **17**

3.1 Research Ethics 17

3.2 Background 17
3.3 Preliminary Analysis of Data 19

4 Analysis **27**
4.1 Introduction 27
4.2 Missing values 28
4.3 Model Comparisons 28
4.4 Further Exploration 41

5 Conclusions **45**

Bibliography **47**

List of Figures

3.1 Boxplot illustrating social distancing score by stage. 20
3.2 Six histograms illustrating the distribution of observed social distancing scores within each stage. 21
3.3 Number of gatherings attended per two weeks by stage. 22
3.4 Six histograms illustrating the distribution of the number of gatherings attended per two weeks by stage. 23
3.5 Boxplot illustrating social distancing score by group. 24
3.6 Six histograms illustrating the distribution of observed social distancing scores between groups made based on age and health status. 24
3.7 Number of gatherings attended per 2 weeks by group. 25
3.8 Six histograms illustrating the distribution of the number of gatherings attended based on group. 26

4.1 Selected model as per the BIC and ICL from CDGMM family. . . . 29
4.2 Selected model as per BIC and ICL from the CDGMM family using linear means. 30
4.3 Selected model as per the BIC and AIC produced by LCGA. 32
4.4 AIC for models from 2 to 8 components produced by GMM. 33
4.5 Selected model as per the BIC and AIC produced by GMM. 33
4.6 Model with 7 components produced by KML, selected as per the BIC and AIC. 34
4.7 BIC by number of components for each method. 36
4.8 Models produced by CDGMM family for varying component numbers. 37
4.9 CDGMM family model with 3 components. 42

4.10 Boxplot of age of inidividuals by component assigned by CDGMM model with 3 components. 43
4.11 Boxplot of number of conditions per inidividual by component assigned by CDGMM model with 3 components. 44

List of Tables

2.1 The different covariance structures along with the number of free covariance parameters for each member of the CDGMM family. . . 8

3.1 Breakdown of point system used to calculate social distancing score by category. 19
3.2 Average social distancing scores by stage. 20
3.3 Average gatherings per two weeks by stage. 22
3.4 Average social distancing score by group. 23
3.5 Average gatherings attended per 2 weeks by group. 25

4.1 CDGMM family BIC and ICL scores for each model. 29
4.2 CDGMM Linear Means family BIC and ICL scores for each model. 30
4.3 BIC and AIC results for the LCGA model. 31
4.4 BIC and AIC results for the GMM. 32
4.5 BIC and AIC results for the KML model. 34
4.6 BIC and AIC results for the two-step model. 35
4.7 Best BIC for each component number across methods. 35
4.8 Cross-tabulation of CDGMM (1–5) and KML (A–G) clusters. . . . 38
4.9 Cross-tabulation of GMM (1–4) and KML (A–G) clusters. 38
4.10 Cross-tabulation of GMM (1–4) and CDGMM (A–E) clusters. . . . 38
4.11 Cross-tabulation of GMM (1–5) and KML (A–E) clusters. 39
4.12 Cross-tabulation of CDGMM (1–5) and KML (A–E) clusters. . . . 39
4.13 Cross-tabulation of CDGMM (1–5) and GMM (A–E) clusters. . . . 39
4.14 Cross-tabulation of GMM (1–4) and KML (A–D) clusters. 40
4.15 Cross-tabulation of CDGMM (1–4) and KML (A–D) clusters. . . . 40

4.16 Cross-tabulation of CDGMM (1–4) and GMM(A–D) clusters. . . . 40
4.17 Cross-tabulation of GMM (1–3) and KML (A–C) clusters. 40
4.18 Cross-tabulation of CDGMM (1–3) and KML (A–C) clusters. . . . 41
4.19 Cross-tabulation of CDGMM (1–3) and GMM (A–C) clusters. . . . 41
4.20 Statistical Summary of ages of individuals by component. 42
4.21 Statistical Summary of # of conditions of individuals by component. 44

Chapter 1

Introduction

Longitudinal data, sometimes referred to as panel data, can be described as collecting observations from the same subjects over multiple time measures. This data can be very useful in tracking changes over time, or trends/behavioural patterns, depending on the type of study conducted. This type of data is popular in a wide variety of fields such as economics, finance, sociology, epidemiology and/or medical studies and more. These studies can range from tracking stock market trends to evaluating the survival rate of a disease. A more comprehensive review of longitudinal experimental design, data collection and analysis can be found in Menard (2007) and Lynn (2009).

Clustering, involves the task of grouping observations so that the observations in any such grouping are more similar than observations in other groupings. This is also referred to as unsupervised learning, as there are no labels on the observations we attempt to cluster. A more detailed review of clustering can be found in Everitt et al. (2011).

In recent years, there have been an increasing number of methods that can be used to analyze longitudinal data. Our focus herein will be on clustering methods for longitudinal data. While there are a multitude of approaches available, some more common than others, there is still little information on which methods are most effective under certain conditions.

For the past year-and-a-half the world has been battling the unprecedented COVID-19 pandemic. In this thesis we will use data collected regarding the be-

haviours of individuals throughout the course of the pandemic. From this data we hope to gain an understanding of which individuals were most adherent to the social distancing protocols along with those who were the least adherent.

The data used throughout this thesis comes from a study conducted by the research group of Dr. Dawn Bowdish. A study was conducted of approximately 300 people over the course of the pandemic. Data used in this thesis were collected from May 2020 to January 2021. The dataset includes the observations from a monthly survey sent to participants asking them questions surrounding their social distancing behaviours. These behaviours include things like frequency of hand-washing, gatherings attended, hours spent in a workplace, etc.

When the study was originally conducted, the hypothesis was that age and health could affect individuals adherence to safety protocols and hence their social distancing behaviours. This could ultimately lead to over/under estimates of which age groups/health conditions make people more susceptible to contracting COVID-19.

To test this hypothesis, the goal of the analysis was to determine if there was a statistically significant difference in the total social distancing score for individuals of different ages and health conditions. We also hoped to determine if there was a difference in the behaviour of individuals throughout different stages of the pandemic (i.e., as there were differences in government mandates and case levels).

Using longitudinal clustering algorithms we attempt to analyze the different gathering patterns of individuals throughout the course of the pandemic. Through the process of identifying these patterns we are also given the opportunity to compare and contrast the different clustering algorithms and determine which algorithms perform well under the conditions of our data. With this work we hope to add to the existing knowledge by illustrating how each method is able to perform on the given data.

The longitudinal clustering methods explored throughout this thesis include mixture-model based methods, k-means, growth mixture models and latent growth class analysis, along with a two-step clustering method.

In Chapter 2 a background and description of methodology used is introduced. Each of the longitudinal clustering methods used is described. In Chapter 3 the data is introduced and described, and a preliminary analysis is performed. In Chapter 4 the clustering methods are applied to the data, models are compared and contrasted. Then analysis pertaining to the type of individual (i.e., age, health status) in each cluster is conducted. Finally, Chapter 5 includes concluding statements and discussions of future work.

Chapter 2

Methodology

2.1 Clustering

Clustering can also be regarded as the unsupervised sub-species of classification. Classification can be generally defined as assigning unlabelled observations to a group. The idea is that those observations within a group are more similar than to those outside of the group. The three types of classification include supervised, semi-supervised and unsupervised (clustering). With n observations and k labelled observations, where $k < n$ we can define each type of classification as follows. Supervised classification uses k labelled observations to determine the $n - k$ unlabelled observations. Semi-supervised classification uses all n observations to determine the $n-k$ unlabelled observations. On the other hand, unsupervised classification (i.e., clustering) assigns group labels with no prior knowledge of group membership (i.e., data is completely unlabelled).

Clustering broadly includes a multitude of methods that work to solve the same problem. The process in which this problem is solved and similarity within groups is decided varies across clustering approaches and will be discussed throughout this chapter.

2.2 Mixture-Model Based Methods

2.2.1 Finite Mixture Models

The first methods we will discuss are referred to as model-based clustering methods. Before we define and explain these methods, we first discuss the connection between clustering and mixture models. McNicholas (2016b) states that the idea of defining a cluster in terms of a component in a mixture model was given in a paper by Tiedeman (1955) who built on previous works by Pearson (1894) and Rao (1952). This work was used to encourage what we now know as clustering. McNicholas (2016b) makes the mixture-model based clustering definition given by Tiedeman (1955) more specific by stating:

> A cluster is a unimodal component within an appropriate finite mixture model.

McNicholas (2016b) clarifies that the appropriate finite mixture model here is the one that is appropriate for the data, i.e. the model with the flexibility to fit the data. With this established, we now move to discuss model-based clustering. Model-based clustering is the use of finite mixture models to perform clustering. We can define a finite mixture model as follows: for a random variable $\mathbf{X}$ and p-dimensional data the probability density function for the model with G components is

$$f(\boldsymbol{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(\boldsymbol{x}|\boldsymbol{\theta}_g),$$

where we have π_g as the gth mixing proportion with $\pi_g > 0$ and $\sum_{g=1}^{G} = 1$, $f_g(\boldsymbol{x}|\boldsymbol{\theta}_g)$ as the gth component density and $\boldsymbol{\vartheta} = (\pi_1, \ldots, \pi_g, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g)$ as the vector of parameters. It is common for the component densities $f_1(\boldsymbol{x}|\boldsymbol{\theta}_1), \ldots, f_g(\boldsymbol{x}|\boldsymbol{\theta}_g)$ to all be of the same type. The most frequently used component density is the Gaussian. This will be discussed more later on. A more in depth review of finite mixture models can be found in McNicholas (2016a,b).

2.2.2 Modified Cholesky Decomposition

Benoît (1924) founded the Cholesky decomposition which is used to decompose a matrix into the product of a lower triangular matrix and its transpose. Pourahmadi (1999, 2000) used a modified version of this Cholesky decomposition which was applied to the covariance matrix $\mathbf{\Sigma}$ of a random variable. This obtains,

$$\mathbf{T\Sigma T'} = \mathbf{D} \iff \mathbf{\Sigma}^{-1} = \mathbf{T'D}^{-1}\mathbf{T}, \tag{2.1}$$

where we have $\mathbf{T}$ as a unique unit lower triangular matrix and $\mathbf{D}$ as a unique diagonal matrix containing only positive entries. As per Pourahmadi (1999), $\mathbf{T}$ and $\mathbf{D}$ can be interpreted as generalized linear autoregressive parameters and innovation variances respectively. Due to this, we are able to predict a measurement taken at time t using those measurements taken at previous time points. So, we see that the linear least squares predictor of X_t based on $X_{t-1}, \ldots, X_1$ is as below:

$$X_t = \mu_t + \sum_{s=1}^{t-1}(-\varphi)(X_s - \mu_s) + \sqrt{d_t}\varepsilon_t,$$

where φ are the lower triangular elements of $\mathbf{T}$, d_t are the diagonal elements of $\mathbf{D}$ and $\varepsilon \sim \mathrm{N}(0,1)$.

This modified Cholesky decomposition has come to be very useful for a multitude of longitudinal methods, as we will see later with our Gaussian mixture modelling of longitudinal data (McNicholas and Murphy, 2010).

2.2.3 Gaussian Mixture Modelling of Longitudinal Data

McNicholas and Murphy (2010) provided a Gaussian mixture model with a modified Cholesky decomposed covariance structure for each component in order to model longitudinal data. For our purposes in this thesis, will we focus on the model-based clustering applications of this model. To illustrate, we consider a

Gaussian mixture model, which can be defined as

$$f(\boldsymbol{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \phi(\boldsymbol{x}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g),$$

where we have the probability density function of a multivariate Gaussian distribution

$$\phi(\boldsymbol{x}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}_g|}} \exp -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}_g^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_g),$$

and we have $\boldsymbol{\mu}_g$ as the mean, $\boldsymbol{\Sigma}_g$ as the covariance matrix. As per McNicholas and Murphy (2010), utilizing the decomposition in (2.1) we can rewrite this as

$$\phi(\boldsymbol{x}|\boldsymbol{\mu}_g, (\mathbf{T}_g' \mathbf{D}_g^{-1} \mathbf{T}_g)^{-1}) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}_g|}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_g)' \mathbf{T}_g' \mathbf{D}_g^{-1} \mathbf{T}_g (\boldsymbol{x} - \boldsymbol{\mu}_g) \right\},$$

where $\mathbf{T}_g$ and $\mathbf{D}_g$ are the $p \times p$ unit lower triangular matrix and $p \times p$ diagonal matrix that follow from the modified decomposition of $\boldsymbol{\Sigma}_g$.

We find that 8 Gaussian mixture models result from the different constraints imposed. There is the option to constrain one of or both of $\mathbf{T}_g$ and $\mathbf{D}_g$ to be equal across components along with the option to impose the isotropic constraint given by $\mathbf{D}_g = \delta_g \mathbf{I}_p$. Throughout this thesis we will denote this family of Gaussian mixture models as CDGMM (Cholesky-decomposed Gaussian mixture models) as was done in McNicholas (2016a). We note these CDGMM's seem to fit the longitudinal data very easily as we notice the following patterns. Constraining $\mathbf{T}_g = \mathbf{T}$ results in the autoregressive relationship between all time points being the same among components. This means that the correlation structure of the longitudinally recorded data values is the same for all the clusters. By constraining $\mathbf{D}_g = \mathbf{D}$ we get that the variability at each time point is the same across all components. And finally, we see that the isotropic constraint $\mathbf{D}_g = \delta_g \mathbf{I}_p$ gives that the variability is the same at each time point within the component, i.e., the noise is the same at all time points. The eight models and their corresponding constraints can be seen in Table 2.1.

Table 2.1: The different covariance structures along with the number of free covariance parameters for each member of the CDGMM family.

Model	$\mathbf{T}_g$	$\mathbf{D}_g$	$\mathbf{D}_g$	Free Covariance Parameters
EEA	Equal	Equal	Anisotropic	$p(p-1)/2+p$
VVA	Variable	Variable	Anisotropic	$G[p(p-1)/2]+Gp$
VEA	Variable	Equal	Anisotropic	$G[p(p-1)/2]+p$
EVA	Equal	Variable	Anisotropic	$p(p-1)/2+Gp$
VVI	Variable	Variable	Isotropic	$G[p(p-1)/2]+G$
VEI	Equal	Equal	Isotropic	$G[p(p-1)/2]+1$
EVI	Equal	Variable	Isotropic	$p(p-1)/2+G$
EEI	Equal	Equal	Isotropic	$p(p-1)/2+1$

2.2.4 Linear Means

McNicholas and Subedi (2012) showed that the CDGMM family can also model the means $\boldsymbol{\mu}_g$ using a linear combination. They defined

$$\boldsymbol{\mu}_g = \boldsymbol{Q}\boldsymbol{\beta}_g = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ t_1 & t_2 & t_3 & \cdots & t_p \end{bmatrix}' \begin{bmatrix} a_g \\ b_g \end{bmatrix},$$

where b_g is the slope and a_g is the intercept. This gives us the following likelihood:

$$\boldsymbol{L}(\boldsymbol{\vartheta}) = \prod_{i=1}^{n}\prod_{g=1}^{G}[\pi_g \phi(\boldsymbol{x}_i|\boldsymbol{Q}\boldsymbol{\beta}_g, (\mathbf{T}_g'\mathbf{D}_g^{-1}\mathbf{T}_g)^{-1})].$$

This method models the mean using a line of best fit. In general, this is not always the most desirable approach as most data does not necessarily follow a linear pattern. We note that the modelling of the mean can be extended to other situations such as quadratic curves or other polynomial combinations.

2.2.5 EM Algorithm for Model-based Clustering

The EM algorithm (Dempster et al., 1977) is an iterative process that works to find the maximum likelihood estimate in the case of missing or incomplete data. The algorithm consists of two steps. The expectation (E) step and the maximization (M) step. The E-step is used to compute the expected value of the complete

data log-likelihood. The M-step maximizes the expected value of complete data log-likelihood. These two steps are repeated until we reach a convergence point.

Here, complete data refers to the combination of the observed and unobserved data. When looking at a clustering problem, we see that the observed data are given by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ while the unobserved data is given by the unknown labels, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$. We use $\boldsymbol{z}_i$ to denote the group membership of the ith observation and z_{ig} is an indicator variable that is used to represent whether x_i belongs to group g, as shown below. We note that $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ig})$, where

$$z_{ig} = \begin{cases} 1 & \text{if } x_i \text{ belongs to component } g, \\ 0 & \text{otherwise.} \end{cases}$$

In model-based clustering, our main goal is to estimate z_{ig}.

2.2.6 Convergence Criterion for EM Algorithm

There are several approaches when it comes to a stopping criterion for the EM algorithm, all mainly centered around determining when there is a lack of progress in the log-likelihood. More explicitly we can write this as

$$l^{(k+1)} - l^{(k)} < \epsilon, \tag{2.2}$$

where we have $l^{(k+1)}$ and $l^{(k)}$ as the likelihood at the $k+1$ and kth iteration and ϵ to be some very small value. This method is generally very effective where the log-likelihood continues to increase until it reaches a plateau point, as that plateau would be easily identified. However, this is not always the case, as there are instances where the log-likelihood can increase in more of a staircase fashion which would make this method potentially ineffective (McNicholas et al., 2010).

As a result, we may want to consider another approach. Following McNicholas et al. (2010) we can consider another convergence method based on Aitken's acceleration (Aitken, 1926). This approach works to estimate the asymptotic maximum log-likelihood at each iteration of the EM algorithm and decide on whether con-

vergence has been achieved. Aitken's acceleration at iteration k can be written as

$$a^{(k)} = \frac{l^{(k+1)} - l^{(k)}}{l^{(k)} - l^{(k-1)}},$$

where $l^{(k+1)}$, $l^{(k)}$ and $l^{(k-1)}$ are the log-likelihoods from iterations $k+1$, k and $k-1$ respectively.

We see that the asymptotic log-likelihood at iteration $k+1$ (Böhning et al., 1994) is given by

$$l_\infty^{(k+1)} = l^{(k)} + \frac{1}{1 - a^{(k)}}(l^{(k+1)} - l^{(k)}).$$

There are several methods that can be used to determine convergence (stopping criterions). The first is given by Böhning et al. (1994): $|l_\infty^{(k+1)} - l_\infty^{(k)}| < \epsilon$. The next is given by Lindsay (1995): $l_\infty^{(k)} - l^{(k)} < \epsilon$. In both, ϵ is a very small value.

Alternatively, McNicholas et al. (2010) proposes that the stopping criterion is met when

$$l_\infty^{(k+1)} - l^{(k)} < \epsilon, \tag{2.3}$$

for a small value for ϵ, provided this difference is positive. This criterion was shown to be just as strict as the criterion given by Lindsay (1995) by McNicholas et al. (2010) since $l^{(k+1)} \geq l^{(k)}$. McNicholas et al. (2010) also showed the that criterion in (2.3) was just as strict as the criterion in (2.2). The criterion in (2.3) will be used in this thesis.

2.3 Growth Curve Models

Growth curve modelling, a term that covers many other similar or identical approaches such as mixed effect models, latent curve modeling, etc. is used to test hypotheses based on differences in individual trajectories between persons. Conventional growth modeling assumes that individuals come from a single population and therefore a single growth trajectory is able to accurately depict the entire population. They look to model the relationship between an explanatory variable and a repeatedly measured outcome. The literature on these models is extensive and more detail can be found in Burchinal et al. (2006) or Preacher et al.

(2008).

We define $k = 1, \dots K$ as the number of classes, $t = 1, \dots, T$ as the time points, and $i = 1, \dots, N$ as the number of subjects. The equation for a single trajectory growth curve model is

$$y_{it} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})X_{it} + (\beta_2 + b_{2i})X_{it}^2 + \epsilon_{it}, \tag{2.4}$$

where y_{it} is the measured outcome for individual or subject i at time t, X_{it} is the predictor value for subject i at time t, β_0, β_1 and β_2 are our fixed effects, b_{0i}, b_{1i} and b_{2i} are the random effects (allowing for differences between individuals and the average trend), and ϵ_{it} represent the error, i.e., individual variability. In the case of equidistant values of X (the outcome is measure at the same value of X across all time points) this can simply be denoted by t_i. The random effects and errors follow a normality assumption, i.e., $b_{ji} \sim N(0, \sigma_{b_j})$, $\epsilon_{it} \sim N(0, \sigma_{\epsilon_t})$. The sum of the random effects and the error terms give us the difference between an individual and the average trend. In (2.4), we are making the assumption that the effect of X on y is given by a second order polynomial over time. However, in real life applications this is not always the case and this model can be adapted to fit higher order polynomials or constrained to fit a linear trend.

2.4 Growth Mixture Modelling

While conventional growth curve modelling can only model a single population trajectory, growth mixture modelling (GMM) is able to consider different groups within a larger population. It does this by allowing for multiple latent classes with each having its own growth curve model (GCM). To do this, GMM relaxes the assumption that all individuals are drawn from a single population with common parameters. It instead allows for differences in growth parameters across unobserved subpopulations. This results in separate growth models for each latent class. Each group follows a different mean trajectory, of possibly different forms (i.e., linear vs. higher order polynomial) and each with unique estimates

of variances and covariate influences. The flexibility in this model is what makes the GMM unique (Muthen and Asparouhov, 2006). Random effects are used to capture individual differences in trajectories within a class (Muthén and Muthén, 2000), since the outcome at the start (the intercept) and the rate of change (the slope) may vary between individuals within a class. Therefore, the distance between the class-specific mean trend and the individual belonging to that class is given by the sum of the random effect and the random error. Furthermore, the random effects and errors follow the same assumptions as in GCM, but now per latent class. The formula for GMMs is given by:

$$y_{it}^k = (\beta_0^k + b_{0i}^k) + (\beta_1^k + b_{1i}^k)X_{it} + (\beta_2^k + b_{2i}^k)X_{it}^2 + \epsilon_{it}^k, \tag{2.5}$$

where all the parameters are the same as defined in the previous section for GCMs; however, each parameter is class-specific to class k.

Since the number of classes is unknown, estimation is carried out conditionally on a pre-specified number of classes. Estimates are found using maximum likelihood with the EM algorithm as was described in a previous section. As there is a possibility of several local maxima for the likelihood (especially with more complex models) it is beneficial to test several starting points before determining that maximization has been reached.

2.5 Latent Class Growth Analysis

Latent class growth analysis (LCGA) models are a special type of GMM. They assume no individual level random variation within each class. This means that all individuals within a class share the same trajectory, i.e., are homogeneous. They describe a longitudinal dataset in terms of a mixture of group trajectories, without having regard for within-group variability (Nagin and Land, 1993; Nagin and Odgers 2010). Individual deviations from the class specific trend are treated as residual error, random effects are not factored in. Instead, it allows for discrete individual differences by letting fixed effects (given by the trend) differ between

classes (Pennoni and Romeo, 2017). We note that k represents the class, and therefore all parameters are class specific. Each subject or individual within the class is expected to follow the group trajectory, although this trajectory is different between classes. The formula for LCGA is given by

$$y_{it}^k = \beta_0^k + \beta_1^k X_{it} + \beta_2^k X_{it}^2 + \epsilon_{it}^k, \tag{2.6}$$

where we again have $\epsilon_{it} \sim N(0, \sigma_{\epsilon_t})$, our error term under the normality assumption. The group-based trajectory model (GBTM) is a popular special case of the LCGA in which the error variance is assumed to be the same for all classes and all time points (Nagin, 2005; Nagin and Land, 1993).

As LCGA exhibits no between-subject variability within a class, far fewer parameters need to be estimated. Therefore, it may be useful in cases of smaller sample sizes or in the presence of more complex models that fail to converge, produce out of range estimates, or it may be used as an initial modelling step before specifying a GMM (Jung and Wickrama, 2008).

2.6 K-means Clustering

Distance based methods optimize a global criteria based on the distance between patterns. A suitable distance function is used to measure the dissimilarity between two subjects and then a clustering algorithm is applied to those distance measurements. K-means clustering is an example of a distance-based clustering method. Euclidean distance (Golay et al., 1998; Košmelj and Batagelj, 1990; Policker and Geva, 2000) is used when performing K-means clustering. The algorithm works as follows (MacQueen 1967; Genolini and Falissard 2010):

- Step 1: We choose K data points as the centers of each of our K clusters.
- Step 2: Next, the Euclidean distance between every data point and each of the K cluster centres is computed. Every data point is assigned to the cluster in which it is the smallest distance from the centre.

- Step 3: Once this is complete for every data point, we re-calculate the mean of each cluster, and the data point closest to that value is reassigned as the center value.
- Step 4: The process in step 2 is then repeated.
- Step 5: Steps 2–4 are repeated until there is no longer any changes (movement between clusters).

This process is similar when we refer to longitudinal K-means clustering. The main difference is that in the longitudinal version the cluster centres are representative of a group trajectory of a cluster. Subjects that are assigned to this cluster are assumed to also follow this trajectory, which we may refer to as a vector $\boldsymbol{\mu}_K$.

2.7 Two-Step Clustering

The final approach considered is a two-step clustering approach that utilizes both growth curve models (Laird and Ware, 1982) and k-means (MacQueen et al., 1967), methods we have discussed earlier in this chapter. This method has been described and reviewed in Twisk and Hoekstra (2012) along with Den Teuling et al. (2020).

To start, the dataset is expressed as a single group trajectory by the growth curve model. This is known as the fixed effects. Each individual within the dataset is described by their deviation from this trajectory (the random effects) using a polynomial of the kth order (Nagin and Odgers, 2010). This trajectory and those random effects are given by

$$y_{ij} = \sum_{k=0}^{K} \beta_{ki} t_{ij}^{k} + \epsilon_{ij}, \tag{2.7}$$

where

$$\beta_{ki} = \alpha_k + \xi_{ki},$$

α_k represents the kth order coefficient of the polynomial trajectory, ξ_{ki} represents the between-subject variability (random effect) for subject i for the kth coefficient,

ϵ_{ij} denotes our error term (i.e., within subject variability). Our error terms are assumed to be independent and normally distributed, $N \sim (0, \sigma)$. The random effects are also assumed to be normally distributed with mean 0, uncorrelated, but with the possibility of an unstructured (i.e., unconstrained) variance-covariance matrix.

Once the growth curve model and random effects are established, we can commence the second step. In this step the outcome can be predicted for each subject over time and passed into the longitudinal K-means algorithm. These predicted outcomes are passed into the longitudinal K-means algorithms as subject trajectories. The longitudinal K-means algorithm is then carried out as described in the previous section.

2.8 Model Selection

Throughout this thesis, all models in a family will be fitted through a range of values of groups. There are several ways to determine which model size is deemed "best". Here, we will consider the BIC, short for Bayesian Information Criterion (Schwarz et al., 1978). This is the criterion of choice in most model-based clustering applications. The BIC will be used to select the "best" model (e.g. covariance or scale decomposition) and the number of components. The BIC is given by:

$$\text{BIC} = 2l(\boldsymbol{x}, \hat{\boldsymbol{\vartheta}}) - p \log n,$$

where $l(\boldsymbol{x}, \hat{\boldsymbol{\vartheta}})$ is the maximized log-likelihood, $\hat{\boldsymbol{\vartheta}}$ is the log-likelihood estimate of ϑ, p is the number of free parameters and n is the number of observations.

The model with the largest BIC is selected as the best model. It is worthy of mention to note that the BIC does not always select the best model from the point of view of the classification performance. There have been many comparative studies performed (for example, Steele and Raftery, 2010), but there has not been a method that has emerged and shown to do better than the BIC across model-based clustering applications.

For completeness we will also consider the AIC (Akaike's information criterion; Akaike, 1998) and the ICL (integrated completed likelihood; Biernacki et al., 2000). The model with the largest AIC or ICL is selected as the best model. Their formulas are given below:

$$\text{AIC} = 2l(\boldsymbol{x}, \hat{\boldsymbol{\vartheta}}) - 2p,$$

$$\text{ICL} \approx \text{BIC} + 2\sum_{i=1}^{n}\sum_{g=1}^{G}\text{MAP}(\hat{z}_{ig})\log(\hat{z}_{ig}),$$

where

$$\text{MAP}(z_{ig}) = \begin{cases} 1 & \text{if } g = \arg\max_h(\hat{z}_{ih}) \\ 0 & \text{otherwise.} \end{cases}$$

2.9 Measuring Similarity

The adjusted Rand index, often referred to as the ARI, (Hubert and Arabie, 1985) is used to compare the clustering results from the different longitudinal clustering methods used throughout the thesis. The ARI can measure the similarity between true classes and predicted classes. However in our case it will be used to measure similarity between two clustering results. This measurement adjusts the Rand index (Rand, 1971) to consider randomness. Here an ARI of 1 indicates perfect class agreement while an ARI of 0 means that the clustering result is similar to the result of a random class assignment. We can also achieve a negative ARI value which indicates that the clustering result is worse than the expected performance of randomly classifying results.

Chapter 3

Seroprevalence Data

3.1 Research Ethics

Participants from the Greater Hamilton Area who had previously provided consent to contact were asked if they would be interested in participating in the study. Informed consent was obtained and participants provided demographic data and filled out questionnaires by email. Participants were provided with a study number and only the research coordinator had access to de-identifying information. All analysis was done using de-identified data. The protocol was approved by the Hamilton Integrated Research Ethics Board (HiREB 10757).

3.2 Background

In this chapter we introduce the data collected from the study conducted by the research group of Dr. Dawn Bowdish. We have just over 300 individuals who responded to surveys from May 2020 to January 2021 on many categories of behaviour throughout this time period. Participants also completed a baseline survey providing information such as age, race, health conditions, prior working situations, etc. The response rate for all of the monthly surveys was approximately 87%.

A social distancing score was derived by considering 7 different categories of behaviour: gathering, hand-washing, care, volunteering, visiting, public transit,

and workplace. The data contains responses from individuals regarding their behaviour within the past two weeks in each category used for the social distancing score. Individuals were then given a score based on their response for each of the seven categories and those scores were summed together to obtain the total social distancing score for that two-week period.

Each category can be described as follows:

- Gathering: the number of gatherings the individual attended in the past two weeks. Includes things like visits with family and friends, trips to the mall or movies, sports games or practices, school, etc.

- Hand-washing: the individual reported on average how many times per day they washed their hands throughout the past two weeks.

- Care: how many times in the past two weeks the individual received any sort of care in which they were in contact with another person not from their own household.

- Volunteering: the number of hours an individual spent volunteering (not including virtual hours) within the past two weeks.

- Visiting: the amount of times within the last two weeks the individual visited a long-term care or retirement facility.

- Public transit: the number of times within the last two-weeks an individual used public transportation.

- Workplace: the number of hours within the past two weeks an individual spent physically at their workplace.

The score is on a scale from 0–14, with 14 being the highest attainable score. The higher the score, the better social distancing practices individuals have. A breakdown of the points earned from each category is shown in Table 3.1.

Table 3.1: Breakdown of point system used to calculate social distancing score by category, where the first column indicates the criterion and the second column indicates points earned.

Breakdown	**Points**
Gather	
Gathering events = 0 times/week	2
Gathering events ≥ 1 & < 5 times/week	1
Gathering events ≥ 5 times/week	0
Hand-washing	
Hand-washing ≥ 3 times/day = 0 times/day	2
Hand-washing ≥ 1 & < 3 times/day	1
Hand-washing = 0 times/day	0
Care	
Receiving care = 0 times/week	2
Receiving care ≥ 1 & < 3 times/week	1
Receiving care ≥ 3 times/week	0
Volunteering	
Volunteering = 0 hours/week	2
Volunteering > 0 & ≤ 10 hours/week	1
Volunteering > 10	0
Visiting	
Visiting = 0 hours/week	2
Visiting > 0 & ≤ 10 hours/week	1
Visiting > 10	0
Public Transportation	
Public transportation = 0 times/week	2
Public transportation ≥ 1 & ≤ 4 times/week	1
Public transportation > 4 times/week	0
Work Away From Home (WAFH)	
WAFH = 0 hours/week	2
WAFH ≥ 1 & ≤ 31 hours/week	1
WAFH 31 hours/week	0

3.3 Preliminary Analysis of Data

First, we are looking to determine if the varying government restrictions or case levels had an impact on individuals social distancing score. To do so, we will look at social distancing score through different "stages" of the pandemic. Each stage is defined as a period of time where government rules/regulations were different from the previous time period. These stages were defined as follows:

Stage 1: week of May 11th – week of June 15th

Stage 2: week of June 22nd – week of July 20th

Stage 3: week of July 27th – week of September 28th

Stage 4: week of October 5th – week of December 21st

Stage 5: week of December 28th – week of January 18th

Again centered around when public guidelines changed surrounding gathering limits, restaurants and other businesses opening and closing, etc.

Table 3.2: Average social distancing scores by stage.

Stage	1	2	3	4	5
Average Score	10.666667	10.142857	9.621622	9.563492	10.206897

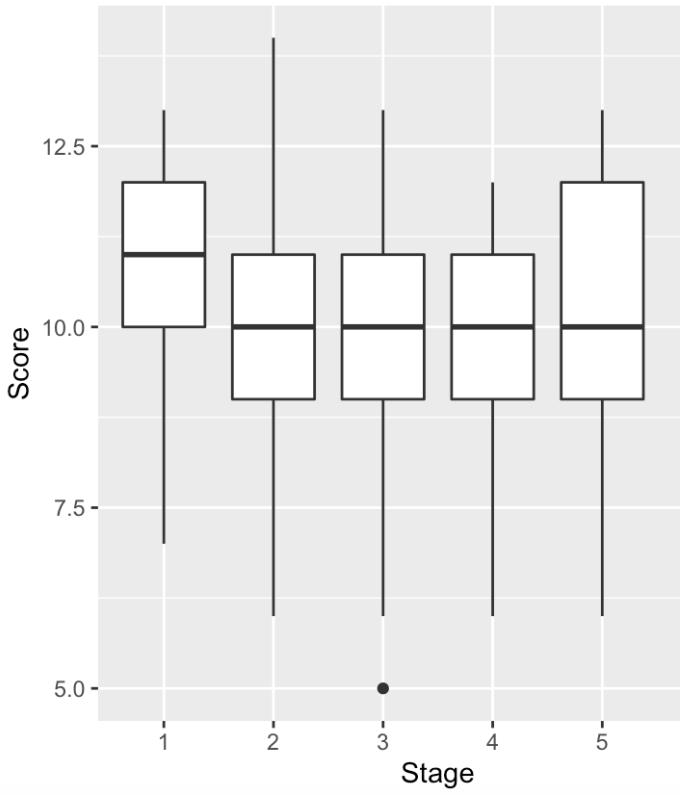


Figure 3.1: Boxplot illustrating social distancing score by stage.

Table 3.2 shows the average social distancing scores in each stage. As we can see the scores between stages are all quite similar, with a slight dip in the middle stages. Based on the boxplots shown in Figure 3.1 it appears that stages 2–4 appear to have the same distribution while stages 1 and 5 are different. This is

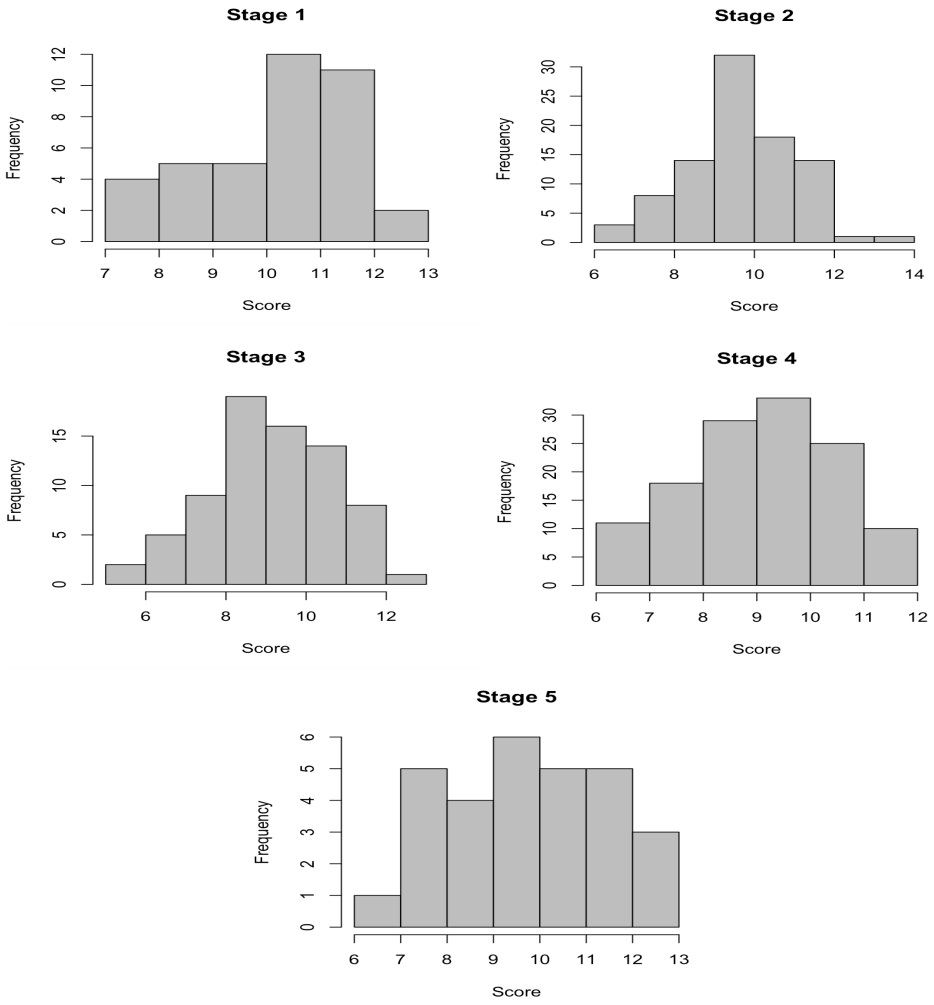


Figure 3.2: Six histograms illustrating the distribution of observed social distancing scores within each stage.

confirmed in Figure 3.2 where we see the individual histograms breaking down the distribution for each stage. Since all 5 stages do not follow the same distribution we cannot safely use a test such as the Kruskal-Wallis. However, pairwise comparisons between stages 2, 3 and 4 can be made using Dunn's test. This tells us that there is no statistically significant differences between social distancing scores in any of the three stages. Overall, we notice that social distancing scores are highest in stages 1 and 5 when restrictions were tightest, and lowest in stages 3 and 4 when restrictions were loosest.

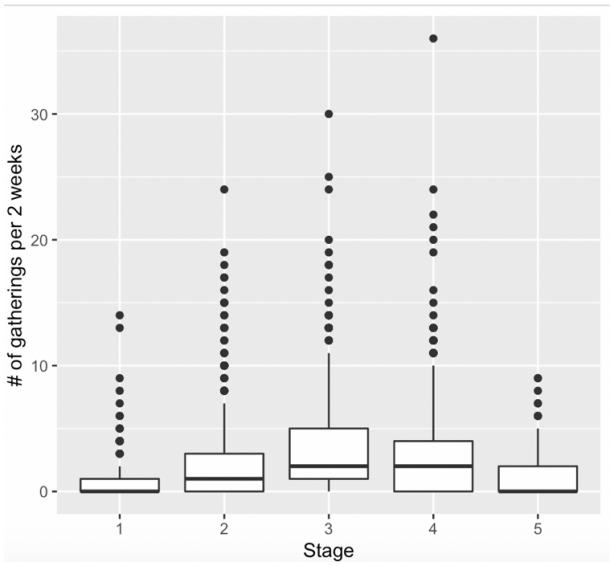


Figure 3.3: Number of gatherings attended per two weeks by stage.

Table 3.3: Average gatherings per two weeks by stage.

Stage	1	2	3	4	5
Average # gatherings	0.6811594	2.1493056	3.4925094	2.5218477	1.0833333

Table 3.3 and Figure 3.3 show that more gatherings were attended in stages 2–4, when there were looser restrictions imposed by the government, along with lower case numbers. Stages 1 and 5 average the least number of gatherings attended. This is when government restrictions were strongest. Individuals attended somewhere between 1.5–2.5 more gatherings per two weeks on average in stages 2–4 than they did in stages 1 and 5. Figure 3.4 gives histograms illustrating the distribution of each stages social distancing scores. The distributions between stages are not the same, and we cannot safely perform a non-parametric test to determine if there is a significant difference between means. The histograms provide another visualization of the differences between gathering numbers in the middle stages and the outer ones. Specifically, in stage 3 it is clear how much more gathering occurred as compared to the other stages.

Next, we grouped the individuals within the study into 6 groups: “Young and sick”, “young and healthy”, “middle-aged and sick”, “middle-aged and healthy”, “elderly and sick”, “elderly and healthy”. Young people were considered to be

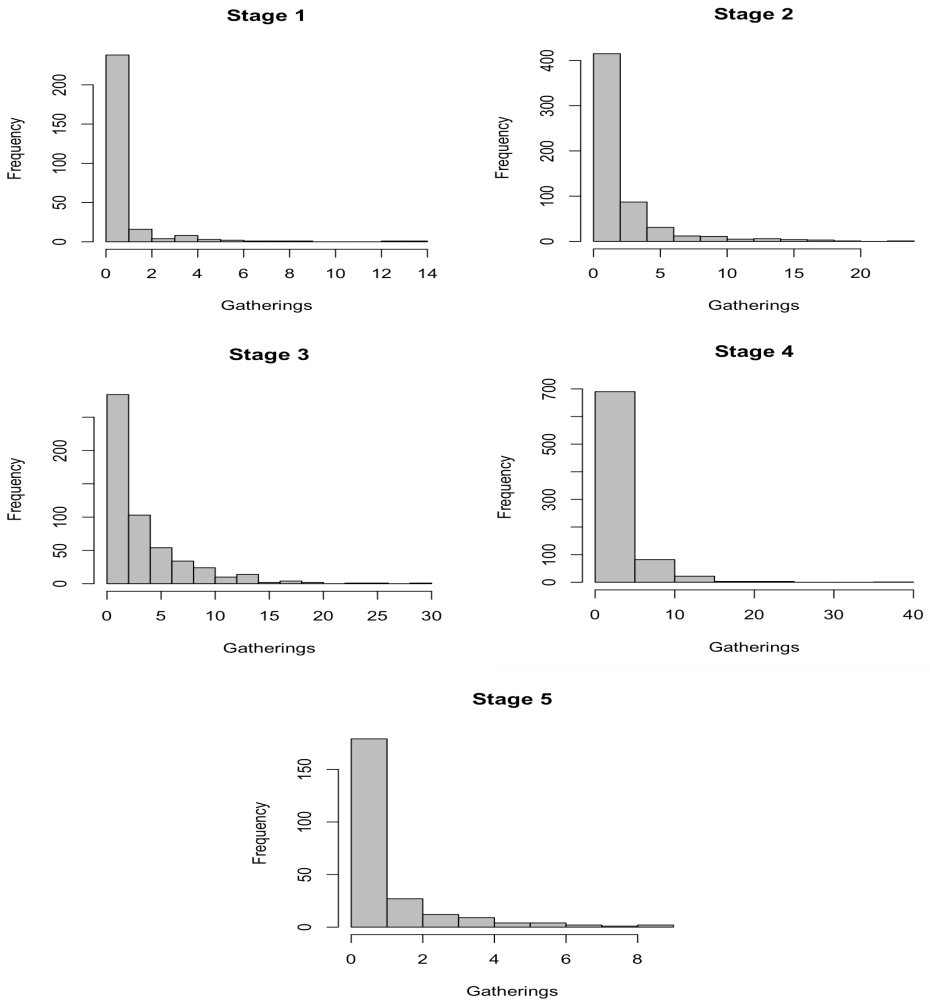


Figure 3.4: Six histograms illustrating the distribution of the number of gatherings attended per two weeks by stage.

those younger than 40. Middle-aged were those between the ages of 40 to 65. The elderly were those over 65. A person was deemed "sick" if they had at least one pre-existing or current condition from the following list: asthma, COPD, ILD, lung disease, diabetes, high blood pressure, cancer, organ failure, autoimmune disorder, pneumonia, or any other "long-term" health condition. The mean social distancing scores for each group are shown in the Table 3.4.

Table 3.4: Average social distancing score by group.

Group	Y&H	Y&S	M&H	M&S	E&H	E&S
Average Score	9.738636	9.333333	9.990291	9.940000	11.384615	9.714286

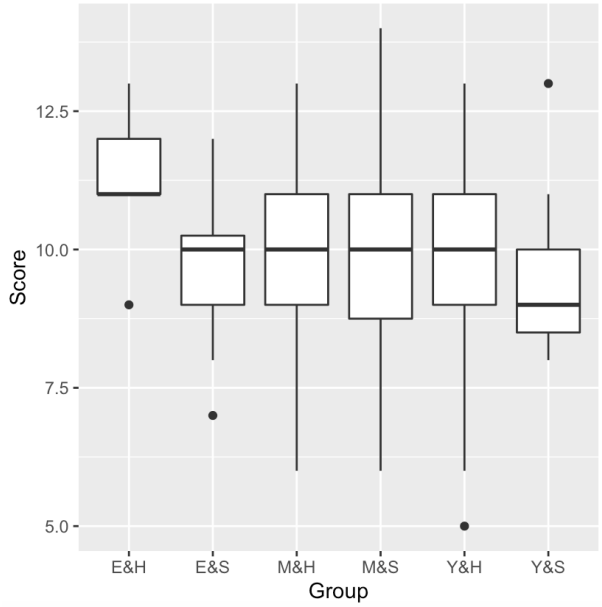


Figure 3.5: Boxplot illustrating social distancing score by group.

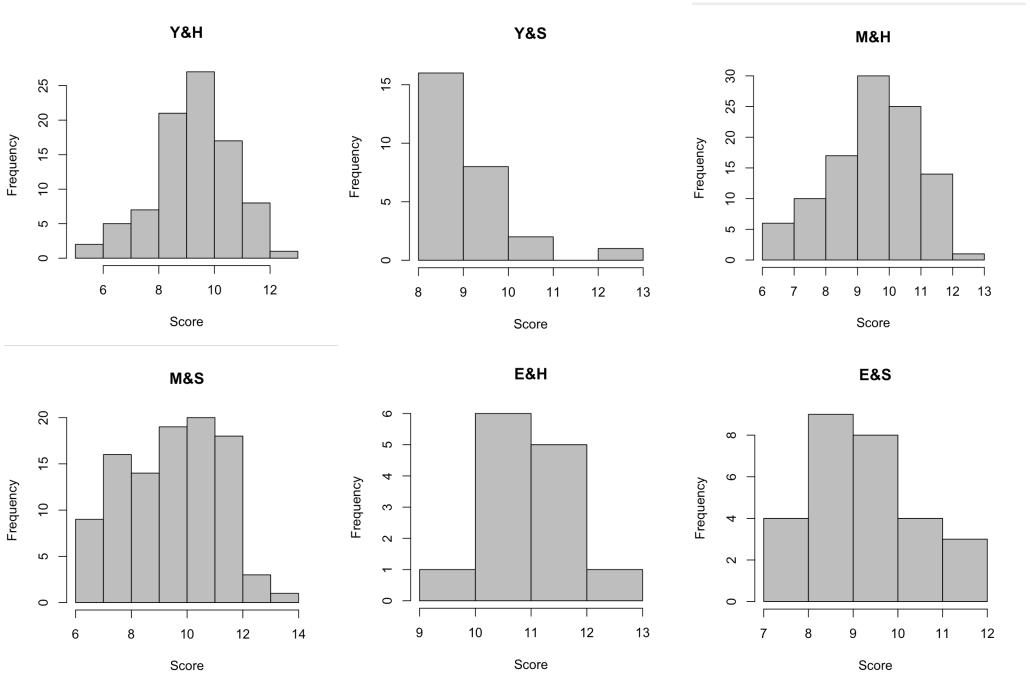


Figure 3.6: Six histograms illustrating the distribution of observed social distancing scores between groups made based on age and health status.

In Table 3.4 we see that the elderly have the highest social distancing scores, followed by the middle-aged and then young. The elderly and healthy have the highest average social distancing score by approximately a point and a half. We

then look to see if the distributions are the same between groups using Figure 3.6. From the histograms in Figure 3.6 it appears we can only safely conclude that Y&H and M&H have the same distribution. This allows us to perform a pairwise Dunn Test between the two groups where we found that there was no significant difference in score. We should note that it was difficult to interpret the young and sick group, as it had by far the smallest sample size. We then compared the average number of gatherings attended per week for each group. The results are shown in both Table 3.5 and Figure 3.7.

Table 3.5: Average gatherings attended per 2 weeks by group.

Group	Y&H	Y&S	M&H	M&S	E&H	E&S
Average # gatherings	2.876126	2.734043	2.555556	2.398589	1.542169	1.553734

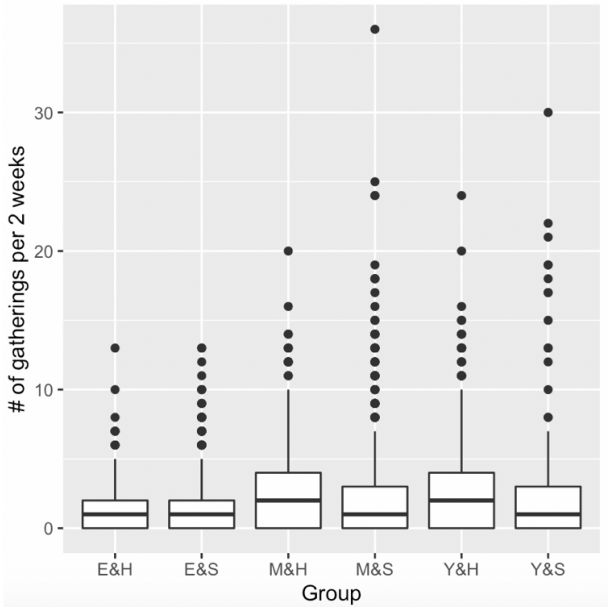


Figure 3.7: Number of gatherings attended per 2 weeks by group.

Right away we notice that the number of gatherings attended seems to be similar for the E&H and E&S group. The numbers for the remaining four groups are also similar. In Figure 3.8 we see the histograms illustrating the distribution of observations for each group. Based on these histograms it is fair to conclude that E&H and E&S follow the same distribution. It also appears that Y&H and

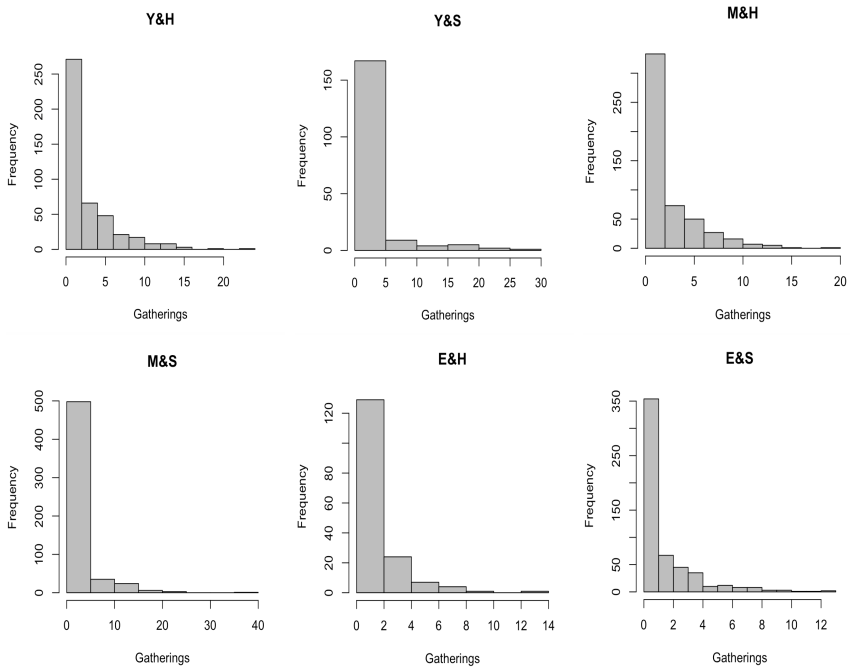


Figure 3.8: Six histograms illustrating the distribution of the number of gatherings attended based on group.

M&H also follow the same distribution. Since all distributions are not the same we cannot safely perform a non-parametric test.

After running a pairwise test to determine if there is a significant difference in the mean gathering numbers of E&H vs. E&S and Y&H and M&H, we found that there was no significant difference in both cases. We note that using Table 3.5 and Figure 3.7 we can observe that the E&S and E&H averaged the lowest number of gatherings per two weeks. On the other hand, the M&H and Y&H groups gathered most frequently per two weeks. M&S and Y&S appear to fall in the middle of the pack, although do have more outliers, as compared to the other groups.

Chapter 4

Analysis

4.1 Introduction

An analysis of the performance of several longitudinal clustering methods will be performed on the same seroprevalence data described in the previous chapter. We will apply longitudinal k-means, growth mixture models, latent class growth analysis, two-step clustering and a variety of mixture-based longitudinal clustering approaches. These include both the Gaussian method and the Gaussian method with linear means.

From the seroprevalence data we will be specifically looking at gathering numbers throughout the pandemic. We hope to be able to cluster groups based on differences in activity levels over the observed time period. For example, it may be natural to assume there were people who rarely/never attend gatherings, while others disregard guidelines and gather more than one may find favorable.

By clustering subjects into these behavioural groups, we can then hope to uncover more information on what type of individual is most common in each setting. For example: average age, number of medical conditions, etc.

Participants provided the gathering information on a monthly basis. Within the monthly survey, individuals were asked to disclose how many gatherings they attended within the past two weeks. These questions were broken up into 11 parts with individuals inputting the number of each type of gathering they attended. Some of categories included family gatherings, sports and recreation, shopping

and movies, religious gatherings, meetings (work or otherwise), restaurants, etc. There was also an "other" category. Responses from each category were then added together to obtain the total number of gatherings per two weeks for each individual.

4.2 Missing values

It is important to note that there are instances of missing data as sometimes individuals missed their monthly questionnaire once or twice. To handle said missing data a carry forward method was used, where the number of gatherings attended listed in the previous month for that individual was slotted into that missing months slot. If the individual had a missing value for the first month of data collection (May 2020), a 0 was inputted to replace the missing NA value. This was done as 0 was the most common response in May 2020.

4.3 Model Comparisons

Throughout this analysis we will use the Gaussian and Gaussian linear means mixture-model based methods. We will also look at longitudinal k-means, growth mixture models, latent class growth analysis and a two-step method using growth curve models and longitudinal k-means to analyze our seroprevalence data. We will start by looking at the mixture-model based methods. The `longclustEM` function from the R package `longclust` was applied to each method for component size 2-8. We can compare the resulting BIC and ICL scores for each of the 8 resulting models of these two methods (16 total). Table 4.1 provides the BIC and ICL values for each model in the CDGMM family. NA is used to represent the models that could not be fit as their likelihood tended towards negative infinity. We see that the CDGMM family selects the EVI model with 5 components as the selected model in terms of both the BIC and the ICL. The EVI model has equal autoregressive structure and variable, isotropic noise across groups. The selected model is shown in Figure 4.1.

Table 4.1: CDGMM family BIC and ICL scores for each model.

BIC								
	VVI	VVA	EEI	EEA	VEI	VEA	EVA	EVI
2	-11607	NA	-12823	-12626	-12690	-12516	NA	-11474
3	-11265	NA	-12618	-12689	-12750	-12535	NA	-10973
4	-11341	NA	-12515	-12752	-12714	-12469	NA	-10921
5	-11573	NA	-12740	-12816	-12955	-12743	NA	-10912
6	-11570	NA	-12792	-12879	-13014	-12702	NA	NA
7	-11782	NA	-12703	-12942	-13121	-12980	NA	NA
8	-11980	NA	-13003	-13006	-13275	-13013	NA	NA
ICL								
	VVI	VVA	EEI	EEA	VEI	VEA	EVA	EVI
2	-11626	NA	-13239	-13030	-12692	-12646	NA	-11495
3	-11277	NA	-12634	-13334	-12761	-12667	NA	-10993
4	-11365	NA	-12528	-13572	-12718	-12483	NA	-10951
5	-11613	NA	-13321	-13790	-12969	-12842	NA	-10948
6	-11594	NA	-13486	-13972	-13026	-12748	NA	NA
7	-11824	NA	-13433	-14130	-13128	-13090	NA	NA
8	-12005	NA	-14038	-14266	-13326	-13142	NA	NA

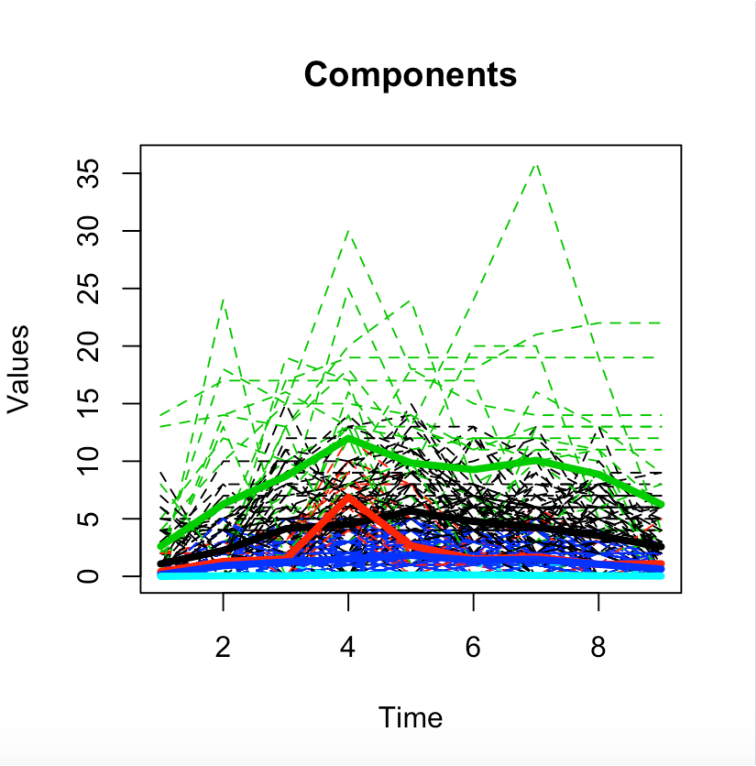


Figure 4.1: Selected model as per the BIC and ICL from CDGMM family. This is the EVI model with 5 components where light blue represents component 1, dark blue represents component 2, red represents component 3, black represents component 4 and green represents component 5.

Table 4.2: CDGMM Linear Means family BIC and ICL scores for each model.

BIC								
	VVI	VVA	EEI	EEA	VEI	VEA	EVA	EVI
2	-11782	-11749	-12968	-12857	-12936	NA	-11616	-11637
3	-11778	-11849	-12991	-12880	-12882	-12783	-11592	-11464
4	-11475	-11830	-13014	-12903	NA	-12760	NA	-11004
5	-11602	-12060	-13060	-12926	NA	NA	-11103	-10991
6	-11787	-12175	-13083	-12949	NA	-12890	-11148	-11020
7	-11940	-12301	-13106	-12972	NA	NA	NA	NA
8	-12019	NA	-13003	-12995	NA	NA	-11130	NA
ICL								
	VVI	VVA	EEI	EEA	VEI	VEA	EVA	EVI
2	-11801	-11762	-13403	-13235	-12943	NA	-11629	-11661
3	-11822	-11877	-13624	-13460	-12886	-12789	-11637	-11499
4	-11498	-11854	-13867	-13721	NA	-12813	NA	-11031
5	-11630	-12089	-14035	-13902	NA	NA	-11132	-11061
6	-11812	-12203	-14143	-14028	NA	-12915	-11199	-11083
7	-11998	-12329	-14271	-14141	NA	NA	NA	NA
8	-12064	NA	-14374	-14226	NA	NA	-11200	NA

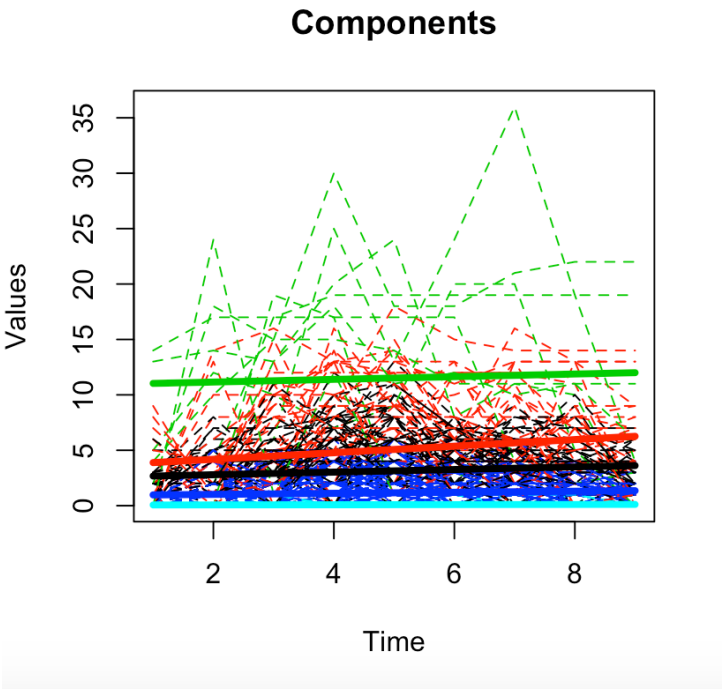


Figure 4.2: Selected model as per BIC and ICL from the CDGMM family using linear means. The EVI model with 5 components where light blue represents component 1, dark blue represents component 2, black represents component 3, red represents component 4, and green represents component 5.

Next, we see Table 4.2 which provides the BIC and ICL values for the models produced from the CDGMM family with linear means. Linear means selects the EVI model with 5 components as the optimal model in terms of both the BIC and the ICL. We see the optimal model in Figure 4.2 where we notice the main difference in linear means is that our trajectories for each component are in fact linear.

The CDGMM family performs the best in terms of both the BIC and ICL across every component number and almost every model. We see that the linear means applied to the CDGMM family performs similarly to the CDGMM family, however most BIC and ICL values are slightly worse than those from the regular CDGMM family. As we notice in Figure 4.3, this can be attributed to the fact that linear means provides linear trajectories and the data we observe does not appear to follow a linear path for the most part. There are a few instances (for example the VVA model) where linear means does better than non-linear means. The CDGMM family selects the EVI model with 5 components as the optimal model in terms of both the BIC and the ICL. The CDGMM family with linear means selects the same model.

Next we will consider the results found from models produced using the `LCMM` package. Using this package we were able to produce models using latent class growth analysis and growth mixture models. We will start by looking at the results from the BIC and the AIC given for the models with 2-8 components by latent class growth analysis as shown in Table 4.3.

Table 4.3: BIC and AIC results for the LCGA model.

	BIC	AIC
2	-14237.59	-14215.05
3	-13844.72	-133810.92
4	-13726.26	-13681.19
5	-13826.83	-13770.50
6	-13844.10	-13776.50
7	-13861.37	-13782.50
8	-13878.64	-13788.50

Table 4.3 shows that LCGA selects the model with 4 components as per both

the BIC and the AIC. The optimal model with the class assignments highlighted by colour are shown in Figure 4.3.

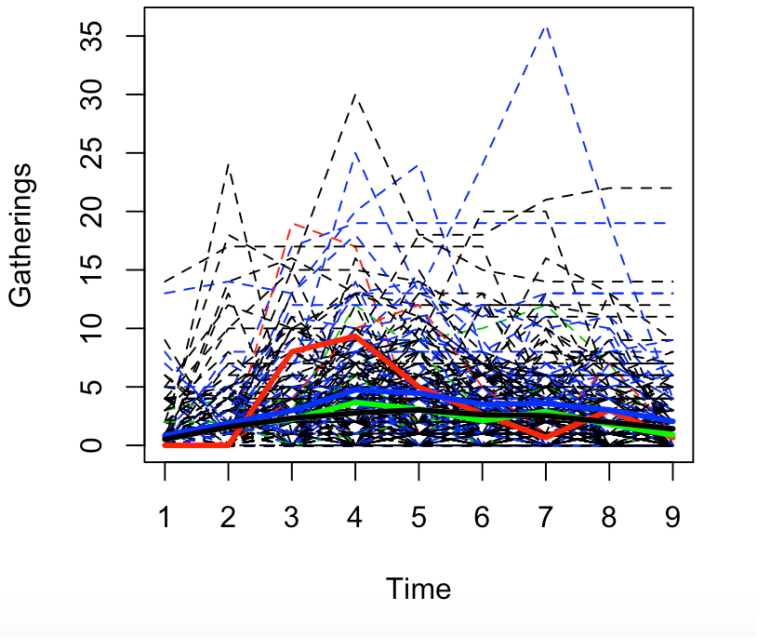


Figure 4.3: Selected model as per the BIC and AIC produced by LCGA. The model has 4 components where black represents component 1, red represents component 2, green represents component 3 and blue represents component 4.

Table 4.4 shows both the BIC and AIC values given by the growth mixture models from 2 to 8 components. We see that the BIC selects the model with 4 components while AIC actually selects the model with 8 components.

Table 4.4: BIC and AIC results for the GMM.

	BIC	AIC
2	-13786.26	-13756.22
3	-13673.46	-13628.39
4	-13625.19	-13565.09
5	-13633.75	-13558.64
6	-13655.47	-13565.33
7	-13678.50	-13573.34
8	-13674.56	-13554.38

Table 4.4 shows that the BIC selects the model with 4 components while AIC actually selects the model with 8 components. However, in Figure 4.4 we see that

at 4 components the increase in AIC that comes with an increase in components is not significant. This supports the BIC results in that the model with 4 components produced by the GMM is the model that optimizes both the AIC and BIC. This model with the highlighted classes can be shown in Figure 4.5.

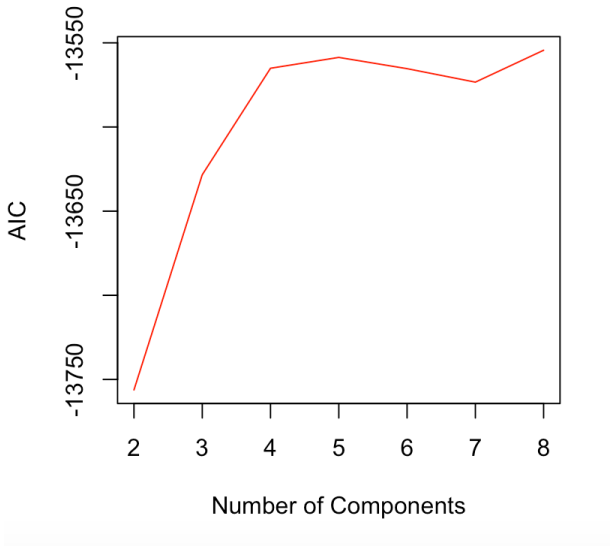


Figure 4.4: AIC for models from 2 to 8 components produced by GMM.

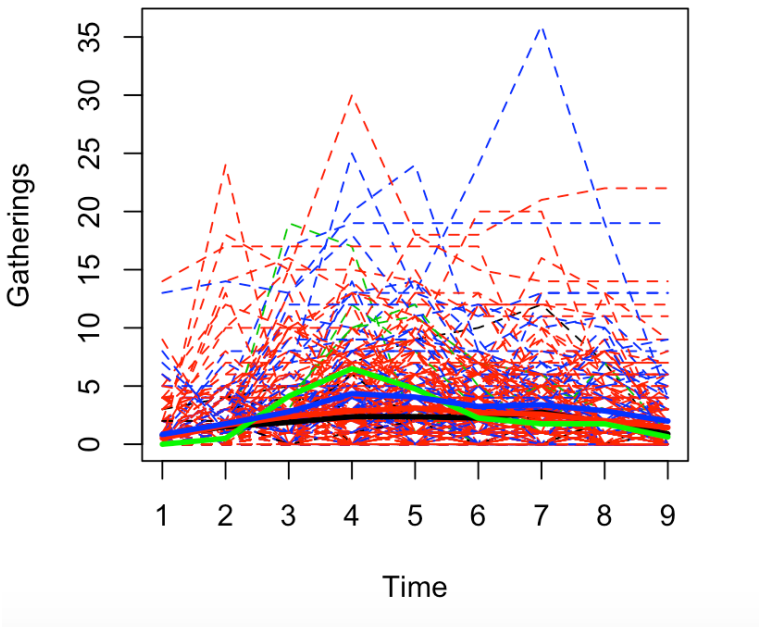


Figure 4.5: Selected model as per the BIC and AIC produced by GMM. The model has 4 components, where black represents component 1, red represents component 2, green represents component 3 and blue represents component 4.

Next we consider our longitudinal k-means approach using the `KML` package. We will analyze the models produced for clusters from size 2-8 using both the BIC and the AIC as is provided by the `KML` package.

Table 4.5: BIC and AIC results for the KML model.

	BIC	AIC
2	-14084.34	-14012.87
3	-13442.49	-13337.15
4	-13419.96	-13280.76
5	-13379.17	-13206.11
6	-13341.16	-13134.24
7	-13080.22	-12839.45
8	-13416.65	-13142.02

Table 4.5 shows that the BIC values given by the KML model. We also notice that KML selects the model with 7 components as per the BIC and the AIC. The selected model with its mean trajectories highlighted are shown in Figure 4.6.

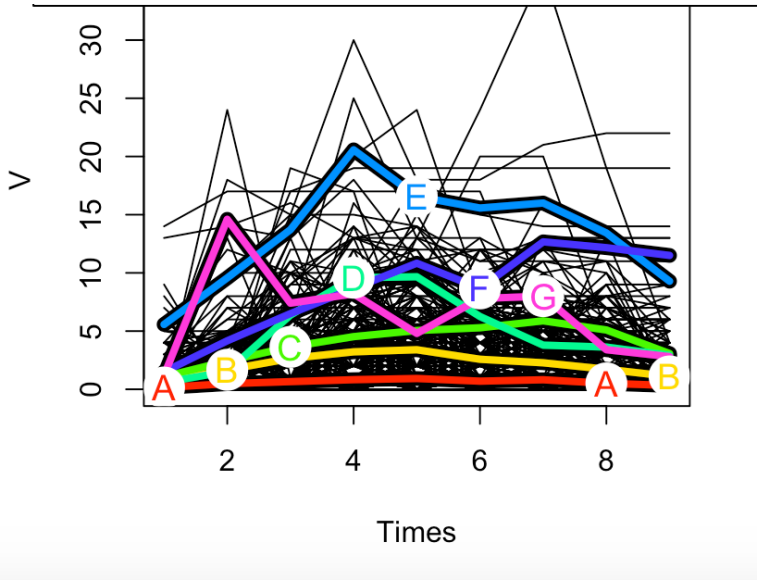


Figure 4.6: Model with 7 components produced by KML, selected as per the BIC and AIC. The components are labelled A through G with the mean trajectories highlighted and labelled respectively.

Lastly, we see the results for our final approach, the two-step clustering method. The growth curve model was fit using R package `lme4` and k-means was performed using `KML`. Table 4.6 shows the BIC and AIC for the models produced from 2-8 clusters.

Table 4.6: BIC and AIC results for the two-step model.

	BIC	AIC
2	-14357.66	-14316.35
3	-14028.10	-13968.01
4	-13656.80	-13577.93
5	-13603.97	-13506.32
6	-13428.00	-13311.57
7	-13302.59	-13167.38
8	-13083.38	-12929.40

From the results shown in Table 4.6 we see that both the BIC and the AIC continue to improve as the number of clusters increase. This means that both metrics select the model with 8 clusters. This continued significant decrease in BIC and AIC as the number of clusters increase is different from the results we have seen from the other methods. Ultimately, this method was not very effective in identifying those with similar gathering patterns.

Because the BIC and ICL or BIC and AIC gave the same conclusions in the tables above, we have compared all models by their BIC values in Table 4.7. We see a variation in the number of components selected by the BIC from each method. KML selects the model with 7 components, CDGMM and CDGMM with linear means selects the model with 5 components, LCGA and GMM select the model with 4 components, and the two-step method selects the model with 8 components. The model with the best BIC overall is from the CDGMM family, the EVI model with 5 components.

Table 4.7: Best BIC for each component number across methods.

	BIC					
	KML	LCGA	GMM	Two-Step	CDGMM	CDGMM-LM
2	-14084.34	-14237.59	-13786.26	-14357.66	-11474	-11616
3	-13442.49	-13844.72	-13673.46	-14028.10	-10973	-11464
4	-13419.96	-13726.26	-13625.19	-13656.80	-10921	-11004
5	-13379.17	-13826.83	-13633.75	-13603.97	-10912	-10991
6	-13341.16	-13844.10	-13655.47	-13428.00	-11570	-11020
7	-13080.22	-13861.37	-13678.50	-13302.59	-11782	-11940
8	-13416.65	-13878.64	-13674.56	-13083.38	-11980	-11130

However, both Table 4.7 and Figure 4.7 show that the difference in BIC for

the CDGMM family from 3–5 components is very minimal. Same can be said for the linear means model from 4–6 components. We also observe that the GMM has a relatively steady BIC across the number of components, while the LCGA has a slight peak at 4, with minimal difference between the BIC at 3 or 5 components. Even KML has a relatively steady BIC until the slight spike at 7. When determining which model is optimal we often choose the model that provides us with the greatest leap of improvement from the previous model. Due to this, it is reasonably to assume the CDGMM model with 3 or 4 components could perform just as well as the model with 5 components when judging based on the BIC.

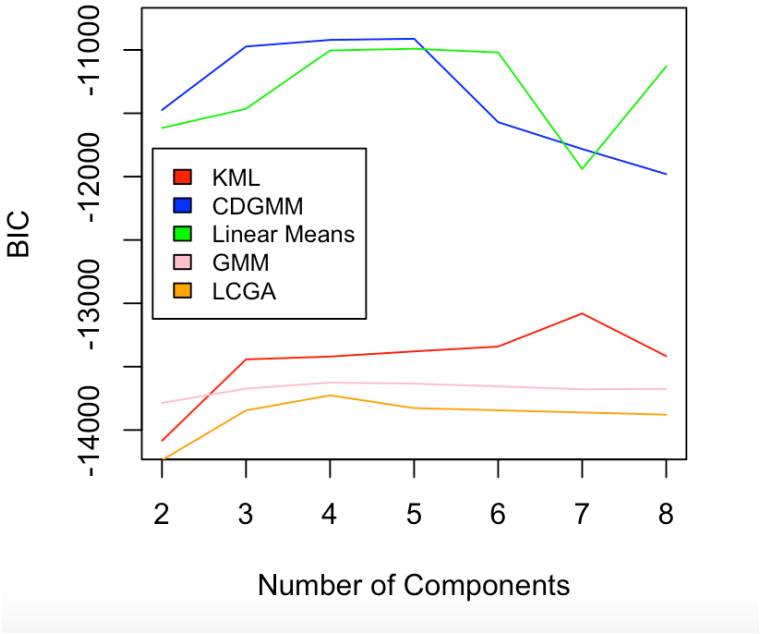
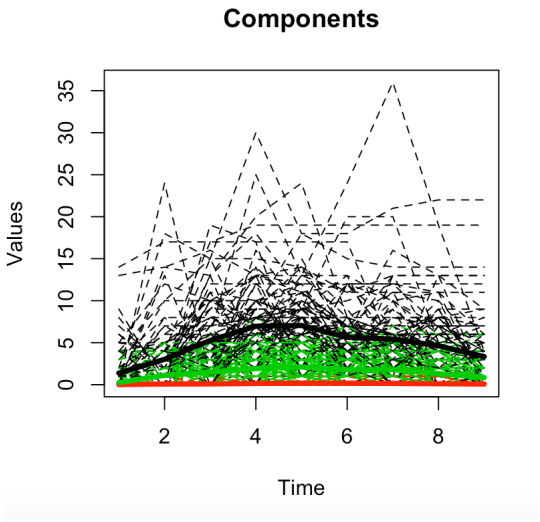
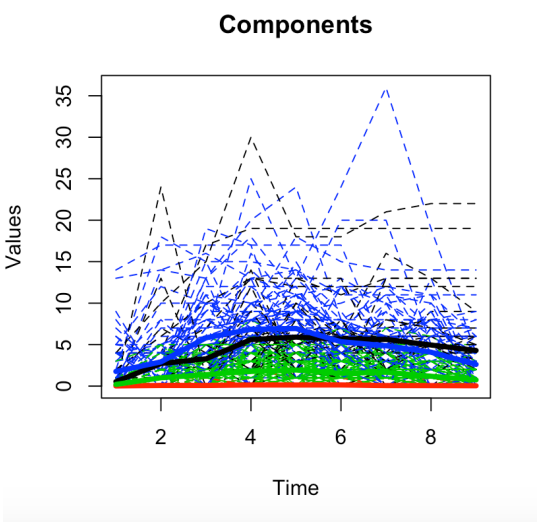


Figure 4.7: BIC by number of components for each method.

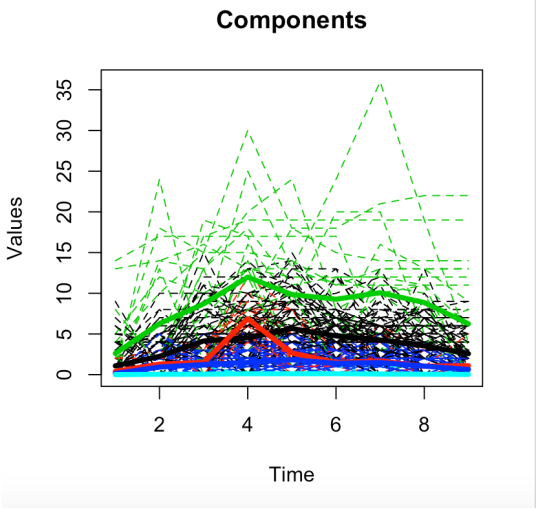
Figure 4.8 shows the models produced by the CDGMM family with 3,4 and 5 components. The models with 3 and 4 components suggest that each group follows an almost linear trajectory, whereas the model with 5 components has 2 components that show a spike around time 4.



(a) 3 Components



(b) 4 Components



(c) 5 components

Figure 4.8: Models produced by CDGMM family for varying component numbers.

Beyond analyzing models using the BIC, we can also contrast models using ARI and cross-tabulation tables. These will be used to evaluate the similarities between clusters. We will start by looking at the cross-tabulation tables and between model ARI's for the models with the best BIC from the following categories: CDGMM family, growth models and KML. This means we will compare the CDGMM model with 5 components, the GMM model with 4 components and the KML model with 7 components to start. This is given in Tables 4.8, 4.9 and 4.10.

Table 4.8: Cross-tabulation of CDGMM (1–5) and KML (A–G) clusters, this gives an ARI of 0.2881164.

No. of clusters	A	B	C	D	E	F	G
1	2	40	43	5	1	0	0
2	6	5	1	4	0	0	0
3	0	0	5	3	6	5	3
4	96	42	0	7	0	0	0
5	42	0	0	0	0	0	0

Table 4.9: Cross-tabulation of GMM (1–4) and KML (A–G) clusters, this gives an ARI of 0.02691652.

No. of clusters	A	B	C	D	E	F	G
1	4	1	1	2	0	0	0
2	106	62	31	11	3	4	1
3	4	0	4	0	0	0	0
4	32	24	13	6	4	1	2

Table 4.10: Cross-tabulation of GMM (1–4) and CDGMM (A–E) clusters, this gives an ARI of 0.003258861.

No. of clusters	A	B	C	D	E
1	2	1	0	5	0
2	62	9	13	101	33
3	2	1	1	3	1
4	25	5	8	36	8

In Tables 4.8 and 4.9 we notice that there are very few observations in clusters 5, 6 and 7 for our KML model. We also notice that according to our ARI the similarity between clusters is very lacking. To confirm this statement we will compare each model at 3, 4, and 5 components. This is because we see in Table 4.7 these components numbers average the lowest BIC.

First we will compare the CDGMM, GMM and KML models when each have 5 components. Both the ARI and cross-tabulation tables will give us an insight into the similarity (or lack there-of) between clusters. These results are given in Tables 4.11. 4.12 and 4.13. Even though we are now comparing each model type with an equal number of components we do not see any improvement in our cluster similarity (ARI values), as from when we were simply comparing the models with

the best BIC values. We still notice that there are very few observations in the 4th and 5th cluster for KML. However, the comparison between CDGMM and GMM produces the worst results in terms of similarity with an ARI of -0.00457875. This tells us the the results are even worse than simple randomization could obtain (in terms of similarity between clusters).

Table 4.11: Cross-tabulation of GMM (1–5) and KML (A–E) clusters, this gives an ARI of 0.03682972.

No. of clusters	A	B	C	D	E
1	152	48	7	5	2
2	4	4	0	0	0
3	14	12	0	1	2
4	41	11	2	1	1
5	7	1	1	1	0

Table 4.12: Cross-tabulation of CDGMM (1–5) and KML (A–E) clusters, this gives an ARI of 0.24931.

No. of clusters	A	B	C	D	E
1	79	0	0	0	0
2	0	1	2	8	5
3	14	65	12	0	0
4	89	10	6	0	0
5	36	0	0	0	0

Table 4.13: Cross-tabulation of CDGMM (1–5) and GMM (A–E) clusters, this gives an ARI of -0.004578785.

No. of clusters	A	B	C	D	E
1	56	3	6	10	4
2	9	1	3	2	0
3	54	3	11	11	2
4	67	0	8	27	3
5	28	1	1	6	0

Secondly, we see our comparisons when each type of model has 4 components in Tables 4.14, 4.15 and 4.16. The results show poor ARI scores, meaning there is little similarity between clusters in each model comparison. We found the highest ARI between the CDGMM and KML model, however it is still quite low around 0.36.

Table 4.14: Cross-tabulation of GMM (1–4) and KML (A–D) clusters, this gives an ARI of 0.0235861.

No. of clusters	A	B	C	D
1	7	1	0	0
2	156	49	6	7
3	4	4	0	0
4	53	21	5	3

Table 4.15: Cross-tabulation of CDGMM (1–4) and KML (A–D) clusters, this gives an ARI of 0.3633604.

No. of clusters	A	B	C	D
1	35	62	0	6
2	148	1	0	2
3	36	0	0	0
4	1	12	11	2

Table 4.16: Cross-tabulation of CDGMM (1–4) and GMM(A–D) clusters, this gives an ARI of -0.01074832.

No. of clusters	A	B	C	D
1	3	68	3	29
2	5	104	3	39
3	0	29	1	6
4	0	17	1	8

Finally we run our comparison at 3 components. The results are shown in Tables 4.17, 4.18 and 4.19.

Table 4.17: Cross-tabulation of GMM (1–3) and KML (A–C) clusters, this gives an ARI of 0.02771571.

No. of clusters	A	B	C
1	4	4	0
2	160	52	6
3	62	23	5

Table 4.18: Cross-tabulation of CDGMM (1–3) and KML (A–C) clusters, this gives an ARI of 0.3517941.

No. of clusters	A	B	C
1	17	65	11
2	49	0	0
3	160	14	0

Table 4.19: Cross-tabulation of CDGMM (1–3) and GMM (A–C) clusters, this gives an ARI of -0.01115127.

No. of clusters	A	B	C
1	4	63	26
2	2	39	8
3	2	116	56

Of the models with 3 components the KML and CDGMM had the most similar clusters with an ARI of around 0.35. However, this is still a fairly low ARI and lets us know the clusters produced by the two models are still largely dissimilar. Throughout our entire ARI analysis we noticed that the CDGMM and KML models produced the most similar clusters, while the CDGMM and GMM seemed to produce the least similar clusters. Overall, we see a huge lack of similarity between the clusters found by the differing methods. None of the methods produced any sort of truly similar clusters.

4.4 Further Exploration

The goal of our analysis to start was to see if we could extrapolate differing gathering patterns found within our data. Based on the analysis we have done in the previous section we have found evidence that models of different types and sizes have performed "best" in terms of different metrics such as the BIC, AIC, ICL or through our cross-tabulation tables.

In this section we hope to look further into the models we have produced. Using a selected model we can analyze the individuals in each component, to determine if there is any noticeable difference in the age or health conditions between the different behavioural trajectories identified by our model.

Here, we will analyze the CDGMM model with 3 components. For reference, Figure 4.9 shows the highlighted components of our chosen model. Component 1 is black, component 2 is red, component 3 is green.

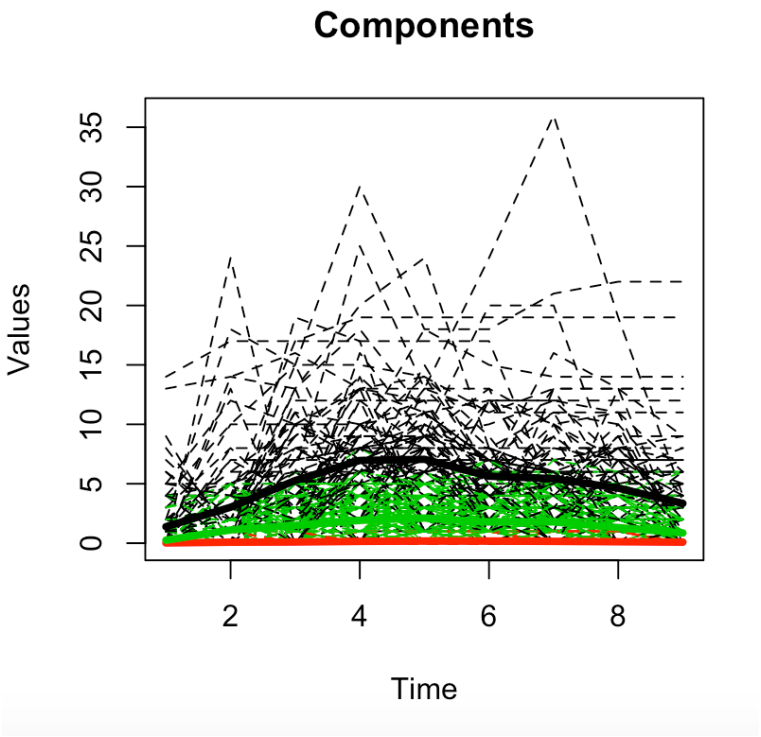


Figure 4.9: CDGMM family model with 3 components.

Table 4.20: Statistical Summary of ages of individuals by component.

	1	2	3
Minimum	23	20	20
1st Quantile	37.00	52	40.25
Mean	50.17391	60.73469	53.05747
Median	48	65	56.5
3rd Quantile	63.25	71	65.00
Maximum	86	81	81
Std Deviation	14.95973	15.39315	14.69447

Based on Table 4.20 above we can see the descriptive statistics of the ages of individuals in each component. The descriptive statistics mostly line up with what one would assume to be true about each component. Component 1 has the youngest average age at 48 and is the component where individuals gather the

most often. Component 2 has the highest average age at 65 and is the component where individuals gather the least, averaging between 0 and 1 for the entirety of the pandemic. Component 3 is right in the middle in terms of gatherings attended and average age at 56.5. We also notice that 50% of individuals in component 1 are between the ages of 37 and 63.25 while 50% of the individuals in component 3 are between the ages of 52 and 71. This information reinforces our original analysis, younger individuals partook in voluntary gatherings much more frequently than older individuals throughout this part of the pandemic. A visual representation of these statistics are given in Figure 4.10.

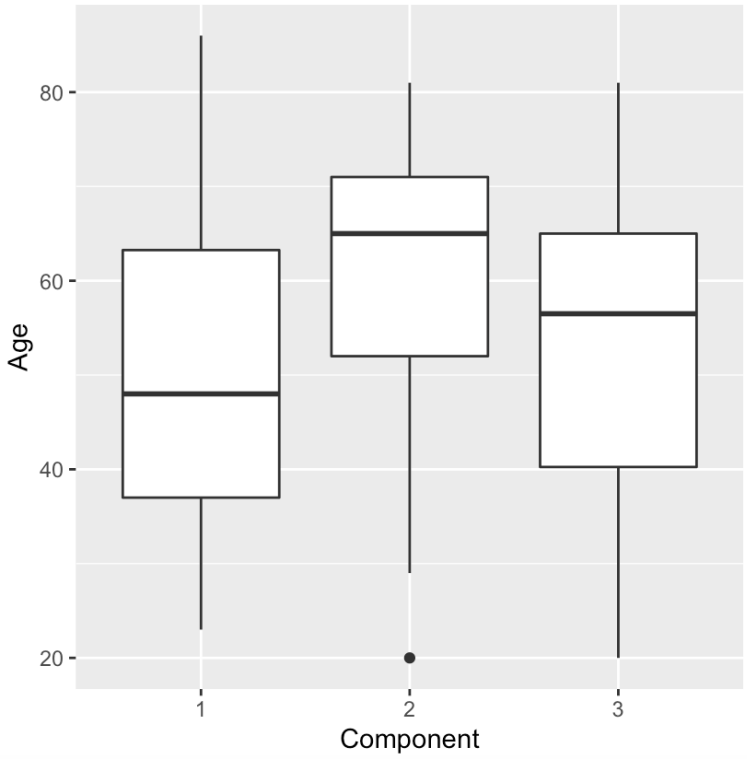


Figure 4.10: Boxplot of age of inidividuals by component assigned by CDGMM model with 3 components.

Next, we see the descriptive statistics of the number of conditions had by individuals within each component in Table 4.21.

Table 4.21: Statistical Summary of # of conditions of individuals by component.

	1	2	3
Minimum	0	0	0
1st Quantile	0	0	0
Mean	0.9239130	1.3061224	0.9655172
Median	0.5	1.0	1.0
3rd Quantile	2	2	1
Maximum	4	5	6
Std Deviation	1.160084	1.417217	1.196800

Those in component 1, who gather the most frequently average the lowest number of conditions at 0.9239130. Those in component 2, who gather the least frequently, average the most number of conditions at 1.3061224. Component 3 again falls in the middle in terms of gathering attended on average and number of conditions at 0.9655172. A visual representation of this is given in Figure 4.11.

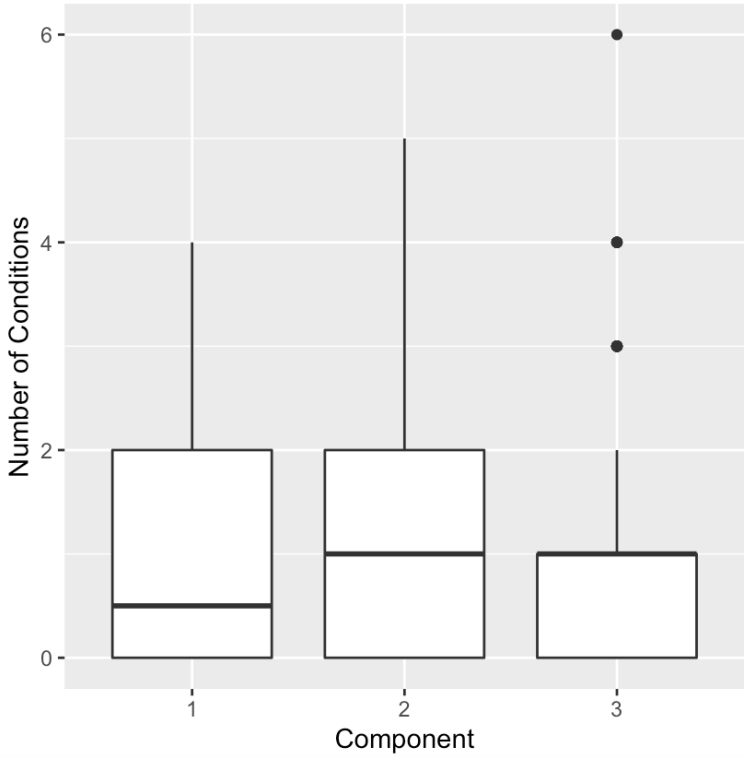


Figure 4.11: Boxplot of number of conditions per inidividual by component assigned by CDGMM model with 3 components.

Chapter 5

Conclusions

Throughout this thesis we worked with a seroprevalence dataset looking to uncover information about the behaviours of individuals of different age and heath. A preliminary analysis of our dataset confirmed that those who are younger and healthier were worse social distancers than those who were older and sicker. It also confirmed that people were better social distancers when there were tighter government restrictions in place as opposed to when restrictions were looser.

We compared a multitude of longitudinal clustering methods on our dataset, specifically looking to determine if we could find groups of different gathering patterns among surveyed individuals. Gaussian and Gaussian linear means mixture model-based methods were both considered. Also, growth mixture models, latent class growth analysis, longitudinal k-means, and a two-step clustering method using growth curve models and k-means. We note that mixtures of t-distributions (the CDtMM family; McNicholas and Subedi, 2012) were also attempted, although convergence was not able to be achieved. The models were incredibly difficult to fit despite best efforts and, therefore, were not included in this thesis.

We first looked at the BIC as a primary method of selecting the number of components for each model. We found that most methods including the Gaussian mixture model based methods, GMM and LCGA agreed that the selected model had somewhere from 3—5 components based on the BIC. Longitudinal k-means and the two-step method found the selected model had somewhere from 7—8 components.

We then looked at the ARI between the models that produced the best BIC within three groups: KML, the mixture models and the growth models. This led us to compare the CDGMM model with 5 components, the KML model with 7 components and the GMM with 4 components using the ARI and cross-tabulation tables. Ultimately, we found ARI scores close to 0, indicating there is almost no similarity in clusters between the different methods.

Since each type of model had a different number of components give the minimized BIC we then compared the clusters for the models of each type with 3, 4 and 5 components. In each comparison we found poor ARI results, in terms of similarity. Even with the same number of clusters there was no significant increase in cluster similarity among methods. The CDGMM and KML had the highest ARI (most similarity between clusters) hovering around 0.35 when the models had 3 or 4 components. The CDGMM and GMM models were the least similar, producing ARIs closest to zero. ARI comparisons ultimately proved that the differing methods were producing very dissimilar clusters. This leads us to conclude that there may be no definite groupings within this dataset. If this is the case, the reason is likely sample size. It is possible that the dataset is simply too small, i.e. we do not have enough data points to definitively identify clusters.

A further exploration was conducted on the CDGMM model with 3 components. This model separated the participants into groups that gather rarely, sometimes and often consistently throughout the time of the year we considered. Unsurprisingly, the average age of those in the components who gathered most often was the youngest while the average age of those who gathered the least was the highest.

In future work, it is still important to continue to perform longitudinal clustering comparisons in a wider range of situations. This could be through datasets of different types, or a large simulation involving multiple scenarios. This would help to provide a clearer picture on the growing number of longitudinal clustering methods and the situations in which they are most effective.

Bibliography

Aitken, A. (1926). A series formula for the roots of algebraic and transcendental equations. *Proceedings of the Royal Society of Edinburgh 45*(1), 14–22.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pp. 199–213. Springer.

Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *Transactions on Pattern Analysis and Machine Intelligence 22*(7), 719–725.

Böhning, D., E. Dietz, R. Schaub, P. Schlattmann, and B. G. Lindsay (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics 46*(2), 373–388.

Burchinal, M. R., L. Nelson, and M. Poe (2006). Growth curve analysis: An introduction to various methods for analyzing longitudinal data. *Monographs of the Society for Research in Child Development 71*(3), 65–87.

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

Den Teuling, N., S. Pauws, and E. van den Heuvel (2020). A comparison of methods for clustering longitudinal data with slowly changing trends. *Communications in Statistics-Simulation and Computation*, 1–28.

Everitt, B. S., S. Landau, M. Leese, and D. Stahl (2011). Cluster analysis 5th ed.

Genolini, C., X. Alacoque, M. Sentenac, C. Arnaud, et al. (2015). kml and kml3d: R packages to cluster longitudinal data. *Journal of Statistical Software 65*, 1–34.

Golay, X., S. Kollias, G. Stoll, D. Meier, A. Valavanis, and P. Boesiger (1998). A new correlation-based fuzzy logic clustering algorithm for FMRI. *Magnetic Resonance in Medicine 40*(2), 249–260.

Herle, M., N. Micali, M. Abdulkadir, R. Loos, R. Bryant-Waugh, C. Hübel, C. M. Bulik, and B. L. De Stavola (2020). Identifying typical trajectories in longitudinal data: modelling strategies and interpretations. *European Journal of Epidemiology 35*(3), 205–222.

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of classification 2*(1), 193–218.

Jung, T. and K. Wickrama (2008). An introduction to latent class growth analysis and growth mixture modeling.

Košmelj, K. and V. Batagelj (1990). Cross-sectional approach for clustering time varying data. *Journal of Classification 7*(1), 99–109.

Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics*, 963–974.

Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 5*. Hayward, CA: Institute of Mathematical Statistics.

Lynn, P. (2009). *Methodology of longitudinal surveys*. John Wiley & Sons.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Volume 1, pp. 281–297. Oakland, CA, USA.

McLachlan, G. J. and D. Peel (1998). Robust cluster analysis via mixtures of multivariate t-distributions. In *Joint IAPR International Workshops on Statistical*

Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), pp. 658–666. Springer.

McNeish, D. and T. Matta (2018). Differentiating between mixed-effects and latent-curve approaches to growth modeling. *Behavior Research Methods 50*(4), 1398–1414.

McNicholas, P., K. Jampani, and S. Subedi (2012). longclust: Model-based clustering and classification for longitudinal data. *R package version 1*.

McNicholas, P. D. (2016a). *Mixture Model-Based Classification*. Boca Raton: Chapman & Hall/CRC Press.

McNicholas, P. D. (2016b). Model-based clustering. *Journal of Classification 33*(3), 331–373.

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

McNicholas, P. D., T. B. Murphy, A. F. McDaid, and D. Frost (2010). Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics & Data Analysis 54*(3), 711–723.

McNicholas, P. D. and S. Subedi (2012). Clustering gene expression time course data using mixtures of multivariate t-distributions. *Journal of Statistical Planning and Inference 142*(5), 1114–1127.

Menard, S. (2007). *Handbook of longitudinal research: Design, measurement, and analysis*. Elsevier.

Muthen, B. and T. Asparouhov (2006). Item response mixture modeling: Application to tobacco dependence criteria. *Addictive Behaviors 31*(6), 1050–1066.

Muthén, B. and L. K. Muthén (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research 24*(6), 882–891.

Nagin, D. S. and K. C. Land (1993). Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed poisson model. *Criminology 31*(3), 327–362.

Nagin, D. S. and C. L. Odgers (2010). Group-based trajectory modeling in clinical research. *Annual Review of Clinical Psychology 6*, 109–138.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A 185*, 71–110.

Pennoni, F. and I. Romeo (2017). Latent markov and growth mixture models for ordinal individual responses with covariates: a comparison. *Statistical Analysis and Data Mining: The ASA Data Science Journal 10*(1), 29–39.

Policker, S. and A. B. Geva (2000). Nonstationary time series analysis by temporal clustering. *Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 30*(2), 339–343.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika 86*(3), 677–690.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika 87*(2), 425–435.

Preacher, K. J., A. L. Wichman, R. C. MacCallum, and N. E. Briggs (2008). *Latent growth curve modeling.* Number 157. Sage.

R Core Team (2019). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Ram, N. and K. J. Grimm (2009). Methods and measures: Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. *International Journal of Behavioral Development 33*(6), 565–576.

Rao, C. R. (1952). Advanced statistical methods in biometric research.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *Annals of Statistics 6*(2), 461–464.

Student (1908). The probable error of a mean. *Biometrika*, 1–25.

Tiedeman, D. (1955). On the study of types. In *Symposium on pattern analysis*, pp. 1–14. Air University, USAF School of Aviation Medicine Randolph Field, TX.

Twisk, J. and T. Hoekstra (2012). Classifying developmental trajectories over time should be done with great caution: a comparison between methods. *Journal of Clinical Epidemiology 65*(10), 1078–1087.

van der Nest, G., V. L. Passos, M. J. Candel, and G. J. van Breukelen (2020). An overview of mixture modelling for latent evolutions in longitudinal data: Modelling approaches, fit statistics and software. *Advances in Life Course Research 43*, 100323.