Incorporating statistical clustering methods into mortality models to improve forecasting performances

by

Sihan (Echo) Cheng

B.Sc., Simon Fraser University, 2018

Project Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in the Department of Statistics and Actuarial Science Faculty of Science

©Sihan (Echo) Cheng 2020
SIMON FRASER UNIVERSITY
Spring 2020

Copyright in this work rests with the author. Please ensure that any reproduction or re-use is done in accordance with the relevant national copyright legislation.
Approval

Name: Sihan (Echo) Cheng

Degree: Master of Science (Actuarial Science)

Title: Incorporating statistical clustering methods into mortality models to improve forecasting performances

Examinning Committee:

Chair: Jinko Graham
Professor

Cary Chi-Liang Tsai
Senior Supervisor
Professor

Yi Lu
Supervisor
Professor

Liangliang Wang
Internal Examiner
Associate Professor

Date Defended: April 9, 2020
Abstract

Statistical clustering is a procedure of classifying a set of objects such that objects in the same class (called cluster) are more homogeneous, with respect to some features or characteristics, to each other than to those in other classes. In this project, we apply four clustering approaches to improving forecasting performances of the Lee-Carter and CBD models. First, each of four clustering methods (the Ward’s hierarchical clustering, the divisive hierarchical clustering, the K-means clustering, and the Gaussian mixture model clustering) are adopted to determine, based on some characteristics of mortality rates, the number and members of age subgroups from a whole group of ages 25-84. Next, we forecast 10-year and 20-year mortality rates for each of the age subgroups using the Lee-Carter and CBD models, respectively. Finally, numerical illustrations are given with R packages 'NbClust' and 'mclust' for clustering. Mortality data for both genders of the US and the UK are obtained from the Human Mortality Database, and the MAPE (mean absolute percentage error) measure is adopted to evaluate forecasting performance. Comparisons of MAPE values are made with and without clustering, which demonstrate that all the proposed clustering methods can improve forecasting performances of the Lee-Carter and CBD models.

Keywords: Lee-Carter model; CBD model; Hierarchical clustering; K-means clustering; Gaussian mixture model clustering; Bayesian model selection.
Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. Cary Tsai, for his outstanding direction, inspiration, and support over the past few years, without which this project could not be completed. I would like to appreciate him for supporting me with great kindness and continual patience.

I extend my thanks to Dr. Jinko Graham for chairing my defense, and Dr. Yi Lu and Dr. Liangliang Wang for investing their precious time in serving the examining committee, making insightful suggestions, and helping to refine the project. Also, I deeply thank all staff and faculty members of the Department of Statistics and Actuarial Science for their dedications, especially Dr. Yi Lu, Ms. Barbara Sanders, Dr. Richard Lockhart, Dr. Derek Bingham, Dr. Thomas Loughin, and Dr. Jinko Graham whose courses greatly broadened my knowledge in actuarial science and statistics. I would like to take this opportunity to thank the Graduate Student Society and the Department of Statistics and Actuarial Science at the Simon Fraser University for offering financial supports throughout my graduate study.

Furthermore, I am always grateful to my academic fellows and loving friends for sharing their experience and being with me during my studies at Simon Fraser University.
# Table of Contents

Approval ii  
Abstract iii  
Acknowledgements iv  
Table of Contents v  
List of Tables vii  
List of Figures viii  

1 Introduction 1  
1.1 Overview of mortality models and clustering 1  
1.2 Motivation 2  
1.3 Outline 2  

2 Literature review 3  
2.1 Mortality models 3  
2.2 Clustering algorithms 4  

3 Models 6  
3.1 Mortality models 6  
3.1.1 The Lee-Carter (LC) model 7  
3.1.2 The Cairns-Blake-Dowd (CBD) model 10  
3.2 Clustering approaches 11  
3.2.1 Hierarchical clustering 12  
3.2.2 K-means clustering 14  
3.2.3 Distribution-based clustering 15  
3.2.4 Clustering objects 18  

4 Numerical illustrations 21  
4.1 Visualization on classification 21  
4.2 Forecasting performance 22
4.3 Improvement illustrations ........................................ 41

5 Conclusions ...................................................... 45

Bibliography .......................................................... 47

Appendix A R Packages ............................................ 50
A.1 mclust ............................................................ 50
A.2 NbClust ........................................................... 50
A.3 cluster ............................................................ 51
## List of Tables

Table 3.1  Mixture Gaussian models implemented with the \texttt{mclust} package in R  

Table 4.1  Fitting and Forecasting year spans  

Table 4.2  $AMAPE_{[t_U+1,2016]}$ for the age span $[25, 84]$, Type I, LC  

Table 4.3  $AMAPE_{[t_U+1,2016]}$ for the age span $[25, 84]$, Type I, CBD  

Table 4.4  $AMAPE_{[t_U+1,2016]}$ for the age span $[25, 84]$, Type II, LC  

Table 4.5  $AMAPE_{[t_U+1,2016]}$ for the age span $[25, 84]$, Type II, CBD
List of Figures

Figure 3.1 \( \ln(m_{x,t}) \) against \( t \) for \( x = 30, 50 \text{ and } 70 \) ........................................ 9

Figure 3.2 Ward’s hierarchical clustering for US females with the fitting age-year window \([25, 84] \times [1954, 1996]\) ........................................ 12

Figure 3.3 K-means for UK females with the age-year fitting window \([25, 84] \times [1986, 2006]\) ........................................ 15

Figure 3.4 GMM clustering for UK males with the age-year fitting window \([25, 84] \times [1970, 2006]\) ........................................ 19

Figure 4.1 Fitting year span \([1974, 2006]\), US males, Type I .......................... 23
Figure 4.2 Fitting year span \([1984, 2006]\), US females, Type I .......................... 24
Figure 4.3 Fitting year span \([1975, 2006]\), UK males, Type I .......................... 25
Figure 4.4 Fitting year span \([1957, 2006]\), UK females, Type I .......................... 26
Figure 4.5 Fitting year span \([1982, 2006]\), US males, Type II .......................... 27
Figure 4.6 Fitting year span \([1960, 2006]\), US females, Type II .......................... 28
Figure 4.7 Fitting year span \([1995, 2006]\), UK males, Type II .......................... 29
Figure 4.8 Fitting year span \([1953, 2006]\), UK females, Type II .......................... 30
Figure 4.9 \( MAPE_{[t_L,2006]}^{[t_U,2016]} \) against \( t_L \), Type I, LC .......................... 33
Figure 4.10 \( MAPE_{[t_L,1996]}^{[t_U,1996]} \) against \( t_L \), Type I, LC .......................... 34
Figure 4.11 \( MAPE_{[t_L,2006]}^{[t_U,1996]} \) against \( t_L \), Type I, CBD .......................... 35
Figure 4.12 \( MAPE_{[t_L,1996]}^{[t_U,1996]} \) against \( t_L \), Type I, CBD .......................... 36
Figure 4.13 \( MAPE_{[t_L,2006]}^{[t_U,2006]} \) against \( t_L \), Type II, LC .......................... 37
Figure 4.14 \( MAPE_{[t_L,1996]}^{[t_U,1996]} \) against \( t_L \), Type II, LC .......................... 38
Figure 4.15 \( MAPE_{[t_L,2006]}^{[t_U,1996]} \) against \( t_L \), Type II, CBD .......................... 39
Figure 4.16 \( MAPE_{[t_L,1996]}^{[t_U,1996]} \) against \( t_L \), Type II, CBD .......................... 40
Figure 4.17 Observed and predicted \( \ln(m_{x,t})/\logit(m_{x,t}) \) against \( t = t_U + 1, \ldots, 2016 \) with age span \([25, 84]\) .......................... 44
Chapter 1

Introduction

1.1 Overview of mortality models and clustering

Mortality forecast is important to social security systems, pension providers, and life insurers. For example, life insurance companies forecast mortality rates for more accurate pricing and reserving of life insurance and annuity products. Modeling mortality rates by considering the changes through time has been a growing area of research. Graunt (1676) claimed that while the length of individual life remains unpredictable, the patterns of longevity and mortality in groups are more able to be captured. To measure the mortality time trends, we need a model that can accurately capture the patterns and trends of mortality rates, both temporal and age specific. In the last decade, more sophisticated mortality models have been developed and applied. However, these models do not consider the dissimilarities of age-specific patterns or trends among ages. Therefore, researchers start adopting modern statistical methods to resolve this issue.

Lee and Carter (1992) proposed a model which is the most widely cited and used in fitting and forecasting mortality rates in actuarial literature. The Lee-Carter (LC) model assumes that the logarithm of the central death rate consists of an age-specific constant, the product of an age-specific time-varying factor and a general time trend of mortality rates, and a model error. Another well-known mortality model is the Cairns-Blake-Dowd (CBD) model which was proposed by Cairns et al. (2006). The CBD model assumes that the logit transformation of one-year death probability for an age is driven by a time trend constant, the product of another time trend and the difference between that age and the average age over an age span, and a model error. Also, it is presumed that all the time trends in the LC and CBD models follow a random walk with drift for mortality forecasting. The model errors and the time trend errors are assumed independent.

Cluster analysis is also called data segmentation and class discovery. The purpose of cluster analysis is to classify data of unknown structure into multiple classes. It is well known as unsupervised learning and a common technique used in machine learning, pattern recognition, image analysis, and information retrieval, etc. In general, clustering algorithms
can be categorized into two basic types, hierarchical clustering and non-hierarchical clustering. Hierarchical clustering is based on the core idea of grouping similar objects into clusters. Moreover, hierarchical clustering can be agglomerative (bottom-up) or divisive (top-down). Alternatively, partitioning algorithm, also called non-hierarchical clustering, is based on iterative relocation which means that points are moved from one class to another until there is no more improvement. According to different criterion of iterative relocation, there are several methods such as K-means clustering, K-medoids clustering, silhouette, and so on. Particularly, cluster analysis can also be based on probability models. Distribution-based clustering is most closely related to statistics as it is based on distribution models, which means that objects in each cluster likely belong to the same distribution.

1.2 Motivation

Lee and Carter (1992) observed that the mortality levels differ strongly by ages and mortality rates at different ages decline at different rates. The rate of decline by age seems stable throughout the time period. At some young adult ages, the changes in mortality rates are irregular. Traditional mortality models capture most but not all information of the death rates such as the correlation structure across ages, the long-term trend of the relative rate of decrease by age, and any cohort-specific factors. In this project, we aim to incorporate clustering methods into two traditional mortality models. Several clustering algorithms are introduced and used to classify all study ages into groups such that the ages in the same group have similar features or characteristics of mortality rates. Then the mortality data for each of the age groups are fitted by the LC and CBD models, respectively. We expect that the clustering methods can improve the forecasting performances of the two traditional mortality models.

1.3 Outline

The remainder of this project is organized as follows. Section 2 provides a literature review on mortality models and clustering algorithms. Section 3 presents the Lee-Carter and CBD models and three well-known clustering algorithms including hierarchical clustering, K-means clustering, and distribution-based clustering along with four practical approaches. Moreover, we apply these four clustering approaches to the mortality data presented in two types of object sets, respectively, based on some features of mortality rates. In Section 4, we firstly illustrate the clustering results. Then we present how to apply the clustered ages for each of four populations to predict mortality rates with the Lee-Carter and CBD models. Lastly, we compare the forecasting performances of each of the two mortality models with and without the clustering methods. The numerical results with the measures of MAPE (mean absolute percentage error) and AMAPE (average of MAPEs) under multiple scenarios are exhibited. Section 5 gives the conclusion.
Chapter 2

Literature review

2.1 Mortality models

The Lee-Carter (LC) model, proposed by Lee and Carter (1992), has been heavily cited and used for mortality projection. The Singular Value Decomposition (SVD) approach is used to estimate the parameters of the LC model. Lee and Carter (1992) observed that the mortality rates have a highly distinctive age pattern even over centuries data, and it moves slowly and steadily. Even though many reasons or factors might cause changing and breaking down of the trends, they barely happened in practice. That is because the trends persist through revolutionary discoveries like medicines which are offset with obesity and smoking. Thus, the stable features make long term forecasts possible. The LC model, involving two age-specific parameters and a time-varying index, assumes that the age-specific central death rates over time have the same pattern trend. Later, Li and Lee (2005) proposed a model to forecast mortality rates for a single population by using the common patterns in a group of populations, which is called coherent forecasting. The group of populations, like males and females of a country or provinces in a country, should have similar features in morality rates. Moreover, White (2002) developed a coherent forecasting with specific regression-based approaches. Over the past ten years, actuaries have used the LC model and its variations for a wide range of applications such as prediction of mortality reduction factors (RFs) (Harberman and Renshaw (2009)), determination of the adequacy of pension (Chia and Tsui (2003)), population projections in demographic science (Booth et al. (2002)), and outlier analysis of mortality index (Li and Chan (2005)).

The CBD model, another famous stochastic mortality model, was proposed by Cairns et al. (2006). This model assumes that the logit transformation of one-year death probability can be captured by an overall time trend plus a time trend related to age, which has played an important role in predicting mortality rates for seniors. The two time-varying trends in the CBD model are jointly used as mortality indices to indicate levels of longevity risk at different time points (Chan et al. (2014)). Additionally, the two time trends are assumed to follow a bivariate random walk process with drift, while the appropriation of using this
process to forecast mortality levels has been debated by several articles (Booth et al. (2002), Tickle et al. (2006), and Sweeting (2011)).

Over the past decades, many extensions to the LC and CBD models were proposed, for example, Li and Chan (2005), Harberman and Renshaw (2009), Plat (2009), Cox et al. (2010), and Mitchell et al. (2011). Several studies propose a linear relation between two mortality sequences of equal length. For example, the linear relational (LR) model was developed by Tsai and Yang (2014) to model each of the two resulting sequences of intercept and slope parameters with a random walk with drift. Later, Tsai and Lin (2016) applied a simple linear regression to two mortality sequences with a one-year time lag for a given age to predict mortality rates for that age.

Nevertheless, a number of recent studies have found that the true underlying distribution of death rate is overdispersed naturally (Downd et al. (2009) and Downd et al. (2011)), i.e., the variance of the death rate is much greater than its mean. There are numerous articles on new models and methods for statistical forecasting of mortality rates. For instance, Li et al. (2015) proposed a local linear panel fitting method regarding the CBD model in order to free the Poisson assumption on the number of deaths.

2.2 Clustering algorithms

Since clustering is subjective, there’s plenty of ways that can be used to achieve the goal of clustering. Each clustering technique follows its own rule or criterion to describe the similarity among data points. There are in fact more than 100 known clustering algorithms in the literature (Hastie et al. (2005), and Kaufman and Rousseeuw (1990)). A wide variety of indices have also been suggested in the literature to determine the optimal number of clusters during the clustering process in partitioning a data set. In this project, we review some popular clustering algorithms below.

- Connectivity models: These models are based on the data points relative to each other in a space with more similarity. There are mainly two approaches. One approach is to classify all objects into separate clusters and then aggregate based on the distance of objects. The other approach is to label all objects as a single cluster and then partition with the increasing distance. These models have the advantage of easy interpretation while their main disadvantage is the lack of scalability of handling large data sets. A representative method is the hierarchical clustering, in which two groups are chosen to be merged by optimizing some criterion. The most popular clustering algorithms would probably be the agglomerative hierarchical methods which consider each entity as a separate cluster. Various hierarchical methods are distinguished by the criteria which determine how clusters get merged at each level. Ward (1963) proposed a well-known criterion that is the sum of within-group sums of squares. Other criteria include the complete and linkage methods (Johnson (1967)). Divisive hierarchical clustering
methods exist as another major type of hierarchical clustering, which begin with all objects in one cluster and divide it into multiple clusters (Edward and Cavalli-Sforza (1965)). For all connectivity models, it is crucial to research on how to choose the optimal cutting parameters, i.e., how to determine the number of clusters in a data set (Dudoit and Fridlyand (2002)).

- Centroid models: These models, known as iterative clustering algorithms, are based on the notion of how closely an object is toward to the centroid of the cluster. They are also called partitioning or non-hierarchical methods which tend to be used when large data sets are involved. We adopt K-means clustering (MacQueen (1967)) in this project as it is the most famous algorithm in this category. Other commonly known centroid methods are K-medoids clustering or PAM (Kaufman and Rousseeuw (1990)) and CLARA algorithm which is an extension to PAM and is adapted for large data sets. Note that, the number of clusters must be specified beforehand, which means that it is necessary to have the prior knowledge of a data set. Centroid models have the advantage of allowing objects to move from one cluster to another by iteration. Overall, the centroid-based methods are efficient but sensitive to the initial conditions and outliers.

- Distribution models: These models are based on the notion of how likely the objects belonging to the same distribution to define cluster. Finite mixture models have been proposed and studied early in the context of clustering (Edward and Cavalli-Sforza (1965), Scott and Symons (1971), Duda and Hart (1973), and Binder (1978)). More recently, it was realized that such models can offer a realistic mathematical solution to empirical problems resulting from the implementation of clustering methods (McLachlan and Basford (1988), and Fraley and Raftery (2002)). Each component probability distribution of the mixture models fits a cluster. Thus, the problems of choosing the optimal number of clusters and the appropriate clustering approach can be equivalent to a statistical model selection. Furthermore, outliers can be handled by adding one or more components which represent different distributions, while it also often causes the model to suffer from overfitting. In this project, we work on a well-known clustering method using the expectation-maximization algorithm with the multivariate normal distribution, which is called Gaussian mixture model (GMM) clustering.
Chapter 3

Models

In this chapter, we review two traditional and popular mortality models (LC and CBD) and some statistical clustering approaches (hierarchical clustering, K-means clustering, and distribution-based clustering). In particular, we give agglomerative and divisive clustering in the hierarchical algorithm, both based on the Ward’s method as the criteria of classification. For the distribution-based clustering, we consider the most widely adopted method, the Gaussian mixture model (GMM) clustering, which fits the data with multiple Gaussian distributions and each distribution belongs to a cluster.

3.1 Mortality models

Denote \( q_{x,t} \) one-year death probability of an individual aged \( x \) in year \( t \), \( p_{x,t} \) one-year survival probability, and \( \mu_{x,t} \) the corresponding force of mortality. Furthermore, let \( m_{x,t} \) be the associated central death rate, that is, the ratio of the number of deaths at age \( x \) in year \( t \) to the average number of survivors from age \( x \) in year \( t \) to age \( x+1 \) in year \( t+1 \). We assume constant force of mortality within each integer age \( x \) and year \( t \), i.e., \( \mu_{x+r,t+s} = \mu_{x,t} \) for \( r, s \in [0, 1) \), which implies that \( m_{x,t} = \mu_{x,t} = -\ln(p_{x,t}) \).

In the following, we use the Lee-Carter and CBD models to govern mortality rates in the study age-year window \([x_L, x_U] \times [T_1, T_2]\). The training data and test data are placed in the fitting window and forecasting window \([x_L, x_U] \times [t_L, t_U]\) and \([x_L, x_U] \times [t_U + 1, T_2]\), respectively, where \( t_L \geq T_1 \) and \( t_U < T_2 \). We denote \( m \left( = x_U - x_L + 1 \right) \) the number of fitting ages and \( n \left( = t_U - t_L + 1 \right) \) the number of fitting years. In Chapter 4, we will evaluate the forecasting performances of the Lee-Carter and CBD models with or without clustering by comparing the predicted values and the actual (observed) values in the forecasting window.
3.1.1 The Lee-Carter (LC) model

The Lee and Carter (1992) model, the most widely cited mortality model in actuarial literature, governs the logarithm of central death rates as

\[ \ln(m_{x,t}) = \alpha_x + \beta_x \cdot k_t + \epsilon_{x,t}, \quad x = x_L, ..., x_U, \quad t = t_L, ..., t_U, \]  

(3.1.1)

where \( \alpha_x \) describes the average age specific pattern, \( k_t \) stands for the time-varying index for the mortality at time \( t \), \( \beta_x \) represents the sensitivity of \( \ln(m_{x,t}) \) at age \( x \) toward to varying \( k_t \), and \( \epsilon_{x,t} \) is a random error. The parameters can be estimated by the singular value decomposition (SVD) subject to two constraints, \( \sum_{t=t_L}^{t_U} k_t = 0 \) and \( \sum_{x=x_L}^{x_U} \beta_x = 1 \).

Alternatively, the first constraint \( \sum_{t=t_L}^{t_U} k_t = 0 \) leads the estimate of \( \alpha_x \) to

\[ \hat{\alpha}_x = \frac{1}{t_U - t_L + 1} \sum_{t=t_L}^{t_U} \ln(m_{x,t}), \quad x = x_L, ..., x_U. \]

The second constraint \( \sum_{x=x_L}^{x_U} \beta_x = 1 \) gives the estimate of \( k_t \) as

\[ \hat{k}_t = \frac{1}{t_U - t_L + 1} \sum_{x=x_L}^{x_U} [\ln(m_{x,t}) - \hat{\alpha}_x], \quad t = t_L, ..., t_U, \]  

(3.1.2)

and \( \hat{\beta}_x \) can be obtained by regressing \( [\ln(m_{x,t}) - \hat{\alpha}_x] \) on \( \hat{k}_t \) without the constant term for each age \( x \) as

\[ \hat{\beta}_x = \frac{\sum_{t=t_L}^{t_U} [\ln(m_{x,t}) - \hat{\alpha}_x] \times \hat{k}_t}{\sum_{t=t_L}^{t_U} \hat{k}_t^2} = \frac{\sum_{t=t_L}^{t_U} \hat{k}_t \times \ln(m_{x,t})}{\sum_{t=t_L}^{t_U} \hat{k}_t^2}. \]

The time trend \( \hat{k}_t \) is assumed to follow a random walk with drift \( \theta^{LC} \), that is, \( \hat{k}_t = \hat{k}_{t-1} + \theta^{LC} + \epsilon_t \), where \( \epsilon_t \) is the error term at time \( t \). The drift can be estimated by

\[ \hat{\theta}^{LC} = \frac{1}{t_U - t_L} \sum_{t=t_L+1}^{t_U} (\hat{k}_t - \hat{k}_{t-1}) = \frac{\hat{k}_{U} - \hat{k}_{L}}{n - 1} = \frac{1}{n - 1} \sum_{x=x_L}^{x_U} [\ln(m_{x,t_U}) - \ln(m_{x,t_L})]. \]  

(3.1.3)

Thus, the logarithm of the projected central death rate for age \( x \) in year \( t_U + \tau \) is

\[ \ln(\hat{m}_{x,t_U + \tau}^{LC}) = \hat{\alpha}_x + \hat{\beta}_x \cdot \hat{k}_{t_U + \tau} = \hat{\alpha}_x + \hat{\beta}_x \cdot (\hat{k}_{t_U} + \tau \cdot \hat{\theta}^{LC}) \]

\[ = \ln(\hat{m}_{x,t_U}^{LC}) + (\hat{\beta}_x \cdot \hat{\theta}^{LC}) \cdot \tau, \quad \tau = 1, 2, ..., \]  

(3.1.4)
where \( \ln(\hat{m}_{x,t}^{LC}) = \hat{\alpha}_x + \hat{\beta}_x \cdot \hat{k}_t \). This is a linear function of \( \tau \) with intercept \( \ln(\hat{m}_{x,t}^{LC}) \) and slope \( \hat{\beta}_x \cdot \hat{\theta}^{LC} \). Note that the average of the intercepts over \([x_L, x_U]\) is

\[
\frac{1}{m} \sum_{x=x_L}^{x_U} \ln(\hat{m}_{x,t}^{LC}) = \frac{\sum_{x=x_L}^{x_U} \hat{\alpha}_x + \hat{k}_t}{m} = \frac{\sum_{x=x_L}^{x_U} \hat{\alpha}_x + \hat{k}_t}{m} = \frac{1}{m} \sum_{x=x_L}^{x_U} \ln(m_{x,t})
\]  

(3.1.5)

by (3.1.2), and the average of the slopes over \([x_L, x_U]\) by the constraint \( \sum_{x=x_L}^{x_U} \hat{\beta}_x = 1 \) is

\[
\frac{1}{m} \sum_{x=x_L}^{x_U} (\hat{\beta}_x \cdot \hat{\theta}^{LC}) = \frac{\hat{\theta}^{LC}}{m} = \frac{1}{m} \sum_{x=x_L}^{x_U} \ln(m_{x,t_L}) - \ln(m_{x,t_U})
\]  

(3.1.6)

Note that (3.1.5) demonstrates that under the LC model the average of the fitted \( \ln(m_{x,t}) \)'s equals the average of the observed \( \ln(m_{x,t}) \)'s. Since \( m_{x,t} = \mu_{x,t} = -\ln(p_{x,t}) \), the corresponding predicted one-year death probability can be easily derived as

\[
\hat{q}_{x,t_{U+\tau}}^{LC} = 1 - \exp[-\exp(\ln(\hat{m}_{x,t_{U+\tau}^{LC}}))].
\]

A special case \((m = 1)\)

Observing from Figure 3.1, we also assume that \( \ln(m_{x,t}) \) follows a random walk with drift \( \theta_x^{RW} \) for each age \( x \), which is called the random walk with drift model (RW). That is,

\[
\ln(m_{x,t}) = \ln(m_{x,t-1}) + \theta_x^{RW} + \epsilon_x^{RW}, \quad x = x_L, \ldots, x_U, \quad t = t_L + 1, t_L + 2, \ldots,
\]

where the time trend errors \( \epsilon_x^{RW} \) are assumed independent and identically normally distributed with mean zero and variance \( \sigma_x^2 \). Denote \( Y_{x,t} = \ln(m_{x,t}) - \ln(m_{x,t-1}) \) the yearly mortality decrements over \([t-1, t]\) with respect to the logarithm of central death rate; then \( Y_{x,t} = \theta_x^{RW} + \epsilon_x^{RW} \) follows a normal distribution with mean \( \theta_x^{RW} \) and variance \( \sigma_x^2 \). The drift can be estimated as

\[
\hat{\theta}_x^{RW} = \frac{1}{t_U - t_L} \sum_{t=t_L+1}^{t_U} Y_{x,t} = \frac{1}{n - 1} \sum_{x=x_L}^{x_U} \ln(m_{x,t_L}) - \ln(m_{x,t_U})
\]

(3.1.7)

where \( y_{x,t} \) is the observation or realization of \( Y_{x,t} \). Similarly, the logarithm of the predicted central death rate under the RW model is

\[
\ln(\hat{m}_{x,t_{U+\tau}^{RW}}) = \ln(m_{x,t_U}) + \hat{\theta}_x^{RW} \cdot \tau, \quad \tau = 1, 2, \ldots.
\]
It is easy to calculate that the average of the slopes \( \hat{\theta}_{RW} \) over the age span \([x_L, x_U]\) by (3.1.3) is

\[
\frac{1}{m} \sum_{x=x_L}^{x_U} \hat{\theta}_x^{RW} = \frac{\sum_{x=x_L}^{x_U} [\ln(m_{x,t_U}) - \ln(m_{x,t_L})]}{m \cdot (n - 1)} = \hat{\theta}_{LC}^m,
\]

that is, \( \hat{\theta}_{LC} = \sum_{x=x_L}^{x_U} \hat{\theta}_x^{RW} \).

Note that under the LC model, when \( x_L = x_U = x \) such that \( m = 1 \), the constraint \( \sum_{x=x_L}^{x_U} \beta_x = 1 \) reduces to \( \beta_x = 1 \), \( \hat{k}_t \) becomes \( \hat{k}_t = \ln(m_{x,t}) - \hat{\alpha}_x \) such that \( \ln(m_{x,t}) - \)
\[
\ln(m_{x,t-1}) = \hat{k}_t - \hat{k}_{t-1} = \theta^{LC} + \epsilon_t, \quad \text{and} \quad \hat{\theta}^{LC} = \frac{\hat{k}_{t_U} - \hat{k}_{t_L}}{n - 1} = \frac{\ln(m_{x,t_U}) - \ln(m_{x,t_L})}{n - 1} = \hat{\theta}^{RW},
\]

which imply that
\[
\ln(\hat{m}_{x,t_U}^{LC} + \tau) = \hat{\alpha}_x + \hat{\beta}_x(x - \bar{x}) + \epsilon_{x,t}, \quad x = x_L, \ldots, x_U, \quad t = t_L, \ldots, t_U,
\]

That is, the LC model with \(m = 1\) reduces to the RW model.

### 3.1.2 The Cairns-Blake-Dowd (CBD) model

The CBD model, proposed by Cairns et al. (2006), is also commonly cited in actuarial literature, which governs the logit function of one-year death probability with an overall time trend and an age specific time trend. The CBD model is given by

\[
\text{logit}(q_{x,t}) = \ln\left(\frac{q_{x,t}}{p_{x,t}}\right) = k_1^t + k_2^t(x - \bar{x}) + \epsilon_{x,t}, \quad x = x_L, \ldots, x_U, \quad t = t_L, \ldots, t_U,
\]

where \(\bar{x}\) is the average age over \([x_L, x_U]\), \(\epsilon_{x,t}\) is the error term, and \(k_t = (k_1^t, k_2^t)^T\) is a bivariate time trend assumed to follow a bivariate random walk with drift \(\theta = (\theta_1, \theta_2)^T\).

The time trend \(k_1^t\) represents the level of the logit-transformed mortality rate, while \(k_2^t\) stands for the slope. Both of them can be estimated by regressing \(\text{logit}(q_{x,t})\) on \((x - \bar{x})\) for each \(t\). That is,

\[
\hat{k}_1^t = \frac{1}{m} \sum_{x=x_L}^{x_U} \logit(q_{x,t}) - \frac{1}{m} \sum_{x=x_L}^{x_U} (x - \bar{x}) = \frac{1}{m} \sum_{x=x_L}^{x_U} \logit(q_{x,t}),
\]

and

\[
\hat{k}_2^t = \frac{\sum_{x=x_L}^{x_U} (\logit(q_{x,t}) - \hat{k}_1^t) \cdot (x - \bar{x})}{\sum_{x=x_L}^{x_U} (x - \bar{x})^2} = \frac{\sum_{x=x_L}^{x_U} \logit(q_{x,t}) \cdot (x - \bar{x})}{\sum_{x=x_L}^{x_U} (x - \bar{x})^2}.
\]

The drift parameter \(\theta_i\), \(i = 1, 2\), is estimated by

\[
\hat{\theta}_i = \frac{1}{t_U - t_L} \sum_{t=t_L+1}^{t_U} (\hat{k}_i^t - \hat{k}_i^{t-1}) = \frac{\hat{k}_{t_U}^i - \hat{k}_{t_L}^i}{n - 1}.
\]
Thus, the logit function of the forecasted one-year death probability for age \( x \) in year \( t_U + \tau \) is

\[
\logit(\hat{q}^{CBD}_{x,t_U + \tau}) = \hat{k}_{1,u} + \hat{k}_{2,u} \cdot (x - \bar{x})
\]

\[
= (\hat{k}_{1,u} + \tau \cdot \hat{\theta}_1) + (\hat{k}_{2,u} + \tau \cdot \hat{\theta}_2)(x - \bar{x})
\]

\[
= \logit(\hat{q}^{CBD}_{x,t_U}) + [\hat{\theta}_1 + \hat{\theta}_2(x - \bar{x})] \cdot \tau, \quad \tau = 1, 2, \ldots,
\]

where \( \logit(\hat{q}^{CBD}_{x,t_U}) = \hat{k}_{1,u} + \hat{k}_{2,u} \cdot (x - \bar{x}) \). This is a linear function of \( \tau \) with intercept \( \logit(\hat{q}^{CBD}_{x,t_U}) \) and slope \( [\hat{\theta}_1 + \hat{\theta}_2(x - \bar{x})] \). Similarly, the average of the intercepts over \([x_L, x_U]\) is

\[
\frac{1}{m} \sum_{x=x_L}^{x_U} \logit(\hat{q}^{CBD}_{x,t_U}) = \frac{1}{m} \sum_{x=x_L}^{x_U} [\hat{k}_{1,u} + \hat{k}_{2,u} \cdot (x - \bar{x})] = \hat{k}_{1,U} = \frac{1}{m} \sum_{x=x_L}^{x_U} \logit(q_{x,t_U}),
\]

and the average of the slopes over \([x_L, x_U]\) is

\[
\frac{1}{m} \sum_{x=x_L}^{x_U} [\hat{\theta}_1 + \hat{\theta}_2(x - \bar{x})] = \frac{1}{m} \sum_{x=x_L}^{x_U} \hat{\theta}_1 = \hat{\theta}_1 = \frac{\hat{k}_{1,U} - \hat{k}_{1,L}}{n - 1},
\]

\[
= \frac{1}{m \cdot (n - 1)} \sum_{x=x_L}^{x_U} [\logit(q_{x,t_U}) - \logit(q_{x,t_L})].
\]

Note that (3.1.10) shows that the average of the fitted \( \logit(q_{x,t_U}) \) equals the average of the observed \( \logit(q_{x,t_U}) \). Since \( \logit(q_{x,t}) = \ln(q_{x,t}/p_{x,t}) \), the corresponding predicted one-year death probability for age \( x \) in year \( t_U + \tau \) can be obtained as

\[
\hat{q}^{CBD}_{x,t_U + \tau} = \frac{1}{1 + \exp(-\logit(\hat{q}^{CBD}_{x,t_U + \tau}))}.
\]

### 3.2 Clustering approaches

Most clustering algorithms work by measuring the similarity among all pairs of objects. In this project, we choose the Euclidean distance as the metric to measure the dissimilarity among a set of observations. The Euclidean distance between two objects is the length of the line segment connecting them. Specifically, for an \( n \)-dimensional space, the distance (\( d \)) between vector \( \mathbf{x} = (x_1, x_2, \ldots, x_n) \) and vector \( \mathbf{y} = (y_1, y_2, \ldots, y_n) \) is defined as \( d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \). Different clustering algorithms applied to the same data set will give different clustering results. In the following, we introduce four clustering algorithms which are implemented ahead of projecting mortality rates with a mortality model. They are the Ward’s agglomerative hierarchical clustering, the divisive hierarchical clustering, the K-means clustering and the Gaussian mixture model clustering.
3.2.1 Hierarchical clustering

Hierarchical clustering results in a tree-based representation of objects, also known as dendrogram. The height of a dendrogram usually expresses the distance between each pair of objects or clusters. Objects can be broken down into classes by cutting the dendrogram to a desired level of similarity. That is to say, the number of clusters does not need to be prespecified and any number of clusters can be chosen by cutting the tree at the right level. Figure 3.1 gives an illustrative example for the dendrogram with a horizontal cutting line. The selected number of clusters can be smaller or larger by moving the horizontal dot line up and down.

Figure 3.2: Ward’s hierarchical clustering for US females with the fitting age-year window [25, 84] × [1954, 1996]

Let $d_{i,j}$ denote the distance between clusters $i$ and $j$, $n_i$ and $n_j$ represent the numbers of objects in clusters $i$ and $j$, and $D$ represent the set of all $d_{i,j}$’s. The procedure of the hierarchical clustering can be summarized as follows:

1. Find the smallest value of $d_{i,j}$ in $D$.
2. Then merge cluster $i$ and $j$ into a new cluster $k$. 
3. Compute the new distances $d_{k,m}$ with the Lance-Williams dissimilarity update formula as

$$d_{k,m} = \alpha_i \cdot d_{i,m} + \alpha_j \cdot d_{j,m} + \beta \cdot d_{i,j} + \gamma \cdot |d_{i,m} - d_{j,m}|,$$

where $m$ denotes any cluster number other than $k$ and $n_k = n_i + n_j$. The parameters $\alpha_i$, $\alpha_j$, $\beta$ and $\gamma$ are defined with reference to the used clustering criterion. These coefficients can be referred to Murtagh (1983) and Gordon (1999).

4. Repeat Steps 1-3 until a single cluster contains all objects.

In general, the hierarchical clustering contains agglomerative and divisive clustering, the former is a bottom-up (merging) approach and the latter is a top-down (splitting) approach. Agglomerative clustering works properly for identifying small clusters while divisive clustering is better for large clusters. In the category of hierarchical clustering, we only apply the Ward’s method and the divisive method in this project.

The Ward’s method, known as the Ward’s minimum variance method, is a special case of the objective function approach to minimizing the total within-cluster variance. Ward (1963) used the sum of squared errors as the objective function. To implement this method in the agglomerative hierarchical clustering procedure, the pair of clusters that leads to a minimum increase in the total variance within the merging cluster can be determined. At the initial step, each single object is formed as a cluster. The second step is to optimize the objective function as the optimal pair of clusters to merge iteratively. Note that, there is a Ward.D2 algorithm discussed in Kaufman and Rousseeuw (1990) and Murtagh and Legendre (2014). The only difference between the Ward.D2 algorithm and the Ward.D algorithm (Murtagh (1983)) is the objective function to be minimize; the Ward.D2 uses the Euclidean distance ($d$) to determine clusters while the squared Euclidean distance ($d^2$) is required in the Ward.D. Thus, in this project, we choose to use the Ward.D2 method in R’s clustering functions. The coefficients in the distance equation in Step 3 above are

$$\alpha_i = \frac{n_i + n_m}{n_k + n_m}, \quad \alpha_j = \frac{n_j + n_m}{n_k + n_m},$$

$$\beta = \frac{-n_m}{n_k + n_m}, \quad \gamma = 0.$$

The proof can be referred to Szekely and Rizzo (2005).

Divisive hierarchical clustering is a "reverse" approach of agglomerative clustering. The process starts with a single set of all objects and splits the set into the optimal clusters recursively until separating each object in its own singleton cluster or achieving a stopping criterion. At each iteration, the first step is to select the cluster with the largest dissimilarity based on the sums of squared errors of the clusters. To split the selected cluster, the algorithm assigns the most disparate object to a new group called "splinter group", which
has the largest average dissimilarity with other objects. In the subsequent steps, objects are reassigned into the "splinter group" and the old group as two new clusters.

### 3.2.2 K-means clustering

K-means is the most widely used centroid-based clustering algorithm, an alternative approach to dividing the data into a set of $K$ clusters, where $K$ is the number of clusters pre-specified by the analyst. The center or mean of the objects for each group indicates the cluster. A briefly mathematical proof is shown below. Given $N$ objects assigned to $K$ clusters, the goal is to minimize the sum of distances between objects and their centroids. We denote

$$A_{n,k} = \begin{cases} 1, & \text{the } n\text{th object is assigned to the } k\text{th cluster}, \\ 0, & \text{otherwise}, \end{cases}$$

$x_n$ the $n$th object, and $\theta_k$ the centroid of cluster $k$. To minimize

$$\min_{A,\theta} \sum_{n=1}^{N} \sum_{k=1}^{K} A_{n,k} (\theta_k - x_n)^2$$

subject to $\sum_{k=1}^{K} A_{n,k} = 1$ for all $n$, where $N$ and $K$ are the total numbers of objects and clusters, respectively, we can take the derivative with respect to $\theta_k$ and set to 0. The optimal $\theta_k$ can be easily obtained as

$$\hat{\theta}_k = \frac{\sum_{n=1}^{N} A_{n,k} x_n}{\sum_{n=1}^{N} A_{n,k}}.$$

The numerator is the number of all cluster-centroid sample distances, while the denominator is the number of objects in the $k$th cluster. Therefore, the estimate of the centroid of cluster $k$, $\hat{\theta}_k$, is the average distance in the $k$th cluster.

Conceptually, the K-means essentially treats objects as composed of a variety of roughly circular distributions, and attempts to find clusters that suit these distributions. We assume an initial number of clusters $K$, the procedure of the K-means clustering is given below:

**Step 1:** For each cluster, the algorithm randomly selects a centroid.

**Step 2:** Then each object is assigned to the closest centroid from Step 1 to obtain $K$ initial clusters.

**Step 3:** Next is to recompute the centroid for each cluster by taking the average of all objects. Once the centroids shift, these objects are reassigned to the nearest centroid.

**Step 4:** The process is repeated until convergence (that is, objects stop changing clusters). Note that in the case of large data sets, the algorithm can be interrupted before convergence using other criteria instead.
Recall that the clusters of hierarchical clustering are defined by cutting branches off the dendrogram, and K-means has to predetermine the number of clusters. Therefore, effective evaluation standards and criteria are necessary. Also, it is not confident to adopt only one validity index. In R package 'NbClust', there are 30 indices provided to select the optimal number of clusters in a data set and it offers the best clustering partition from different results. That is, each index with its own criteria enables the user to simultaneously evaluate several clustering schemes while varying the number of clusters. All of these clustering indices incorporate a lot of information such as intracluster compactness and intercluster separation, statistical properties, and dissimilarity or similarity measurements. Thus, these indices may have different choices regarding the optimal number of clusters. In this project, we choose the final optimal number of clusters based on the majority rule which indicates the most frequency of proposals among the 30 indices.

Figure 3.3 displays the suitable numbers of cluster chosen based on the highest frequency among all indices by implementing an R package 'NbClust'. Charrad et al. (2014) gave the name of each index and corresponding criterion of how to select the optimal number of clusters included in the 'NbClust' package.

3.2.3 Distribution-based clustering

This clustering approach assumes data are composed of distributions. It was discovered that cluster analysis can also be based on probability models (Bock (1996), Bock (1998)). A prominent method is the Gaussian mixture model (GMM) where the data set is fitted with a determined number of Gaussian distributions and each distribution belongs to a
cluster. The strategy of cluster analysis we apply, proposed by Fraley and Raftery (2002), consists of three cores: initialization via the hierarchical clustering, maximum likelihood estimation via the EM algorithm, and the selection of models and the number of clusters via approximate Bayes factors with the BIC. This approach is represented to work better when the objectives are well-separated.

**Mixture Models**

Given \(n\)-dimensional multivariate observations \(x = (x_1, \ldots, x_n)\), let \(f(x; \theta)\) be the probability density function of a finite mixture model, given by

\[
 f(x; \theta) = \sum_{j=1}^{J} \pi_j f_j(x; \theta_j),
\]

where \(\theta = \{\theta_1, \ldots, \theta_J\}\), \(\pi_j\) represents the \(j\)th mixing probability, and \(f_j(x; \theta_j)\) is the \(j\)th component density with parameter \(\theta_j\). Most commonly, \(f_j\) is the multivariate normal density, parameterized by mean vector \(\mu_j\) and variance-covariance matrix \(\Sigma_j\). For a mixture of multivariate normal densities, the general surface is ellipsoidal, and the geometric features (shape, volume, orientation) are determined by the variance-covariance matrix \(\Sigma_j\). Banfield and Raftery (1993) proposed a flexible framework for geometric constraints in the form,

\[
 \Sigma_j = \lambda_j D_j A_j D_j^T, \quad j = 1, \ldots, J,
\]

(3.2.1)

where \(D_j\) is the orthogonal matrix of eigenvectors which governs the orientation of the \(j\)th component of the mixture; \(A_j\) is a diagonal matrix with elements as the proportion of eigenvalues, which controls the shape; and \(\lambda_j\) is a constant of proportion, which controls the volume. This general form can construct many mixture models to be used in cluster analysis. Common instances include \(\Sigma_j = \lambda I\) (equal volume) where all clusters are spherical with the same size; \(\Sigma_j = \lambda_j I\) where clusters are spherical but have different volumes; \(\Sigma_j = \Sigma\) (equal variance) where all clusters are ellipsoidal and have the same geometry; \(\Sigma_j = \lambda D_j A_j D_j^T\) where all clusters have common shape and volume; \(\Sigma_j = \lambda_j A_j\) where all covariances are diagonal but the geometric features are allowed to vary; and unrestricted \(\Sigma_j\) where each cluster has different geometry.

**EM algorithm**

Expectation-maximization (EM) algorithm is an iterative approach to finding the maximum likelihood estimation with multivariate data, typically, with latent variables. Hence, the EM algorithm is widely used for mixture models. It is consist of two steps: the E step constructs the conditional expectation function of the complete data log-likelihood using the observed data and the current estimated parameters; the M step computes the optimal parameters to maximize the function created in the E step. The unobserved portion of the data that
contain missing values due to non-response. Denote the complete data \( x_i = (y_i, z_i) \), in which \( y_i \) is observed and \( z_i \) is unobserved. For mixture models (Fraley and Raftery (2002)), \( x_i \) follows one of the distributions in the mixture model with the probability density function \( f(x_i; \theta_i) \), where \( i = 1, ..., n \) and \( z_i = (z_{i,1}, ..., z_{i,J}) \) is the unobserved portion of the data, with

\[
    z_{i,k} = \begin{cases} 
        1, & \text{if } x_i \text{ belongs to } f_k, \\
        0, & \text{otherwise.} 
    \end{cases} 
\]

(3.2.2)

Recall that hierarchical clustering is used in the initialization of the EM algorithm. Each stage of merging in the hierarchical clustering process generates a unique number of clusters and a partition of the data. To initialize the EM algorithm, we use the transformed indicators given in (3.2.2) from the unique partition as the conditional probabilities in the M step of the EM algorithm.

**Model selection**

Model selection in cluster analysis consists of two fundamental issues: selection of the clustering method and determination of the number of clusters. There are trade-offs between these two issues. A simpler model needs more clusters to catch enough information of data while a more complex model needs fewer clusters to keep model efficient. In this project, we use an approach for model selection in clustering based on Bayesian model selection via Bayes factors and posterior model probabilities. Specifically, given \( J \) models, \( M_1, ..., M_J \) with the prior probabilities \( p(M_1), ..., p(M_J) \), respectively, the posterior probability given data \( D \) is denoted by \( p(M_j | D) \), \( j = 1, ..., J \). By the Bayes' theorem, the posterior probability of model \( M_j \) given data \( D \) is proportional to the product of the probability of the data \( D \) given model \( M_j \) and the model's prior probability,

\[
    p(M_j | D) \propto p(D | M_j) p(M_j). 
\]

Since \( p(D | M_j) = \int p(D | \theta_j, M_j) p(\theta_j | M_j) d\theta_j \), it is known as the integrated likelihood. Another important concept is the Bayes factor which defines the ratio of two integrated likelihoods, e.g., \( B_{12} = p(D | M_1) / p(D | M_2) \), the larger \( B_{12} > 1 \), the stronger evidence of favoring \( M_1 \). In practice, it is difficult to obtain the integrated likelihood. Denote \( \hat{\theta}_j \) as the estimate of \( \theta_j \) and \( \nu_j \) as the number of independent parameters to be estimated in model \( M_j \). Fortunately, the integrated likelihood can be approximated simply by the BIC (Schwarz (1978) and Haughton (1988)) as

\[
    2 \log p(D | M_j) \approx 2 \log p(D | \hat{\theta}_j, M_j) - \nu_j \log(n) = BIC_j. 
\]

Hence, the model choice can be made based on the maximum BIC.
Combining with hierarchical agglomeration, EM, and Bayes factors, the strategy proposed by Fraley and Raftery (2002) can be summarized as follows: firstly, perform the hierarchical clustering to determine a set of mixture models with the maximum classification likelihood and corresponding classifications; secondly, apply the EM to each model along with a set of numbers of clusters to estimate the parameters of the mixture model, starting from the previously obtained classification until convergence; lastly, compute the BIC for each combination case of the models and the numbers of clusters with the estimated parameters from the EM. The choice of models and the number of clusters are made corresponding to the maximum BIC.

Figure 3.4 illustrates a case of the model selection and clustering partition by fitting Gaussian distributions. In Figure 3.4 (a), the maximum BIC determines the optimal mixture model fitted by a data set with 3 ellipsoidal, equal shape but varying volume of Gaussian distributions, i.e., the VEV (ellipsoidal, equal shape). The corresponding partition of yearly decrements (see Section 3.2.4 for the definition of yearly decrements) is displayed in Figure 3.4 (b) based on the estimates of the mean and the standard deviation of yearly decrements. The abbreviations of the multivariate mixture models shown in Figure 3.4 (a) can be referred to Table 3.1.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Geometric Features of Mixture Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>EII</td>
<td>spherical, equal volume</td>
</tr>
<tr>
<td>VII</td>
<td>spherical, unequal volume</td>
</tr>
<tr>
<td>EEI</td>
<td>diagonal, equal volume and shape</td>
</tr>
<tr>
<td>VEI</td>
<td>diagonal, varying volume, equal shape</td>
</tr>
<tr>
<td>EVI</td>
<td>diagonal, equal volume, varying shape</td>
</tr>
<tr>
<td>VVI</td>
<td>diagonal, varying volume and shape</td>
</tr>
<tr>
<td>EEE</td>
<td>ellipsoidal, equal volume, shape, and orientation</td>
</tr>
<tr>
<td>EVE</td>
<td>ellipsoidal, equal volume and orientation</td>
</tr>
<tr>
<td>VEE</td>
<td>ellipsoidal, equal shape and orientation</td>
</tr>
<tr>
<td>VVE</td>
<td>ellipsoidal, equal orientation</td>
</tr>
<tr>
<td>EEV</td>
<td>ellipsoidal, equal volume and equal shape</td>
</tr>
<tr>
<td>VEV</td>
<td>ellipsoidal, equal shape</td>
</tr>
<tr>
<td>EVV</td>
<td>ellipsoidal, equal volume</td>
</tr>
<tr>
<td>VVV</td>
<td>ellipsoidal, varying volume, shape, and orientation</td>
</tr>
</tbody>
</table>

Table 3.1: Mixture Gaussian models implemented with the mclust package in R

### 3.2.4 Clustering objects

Recall from Section 3.1.1 that the LC model with \( m = 1 \) reduces to the RW model. Thus, there is a relevance among the LC model, the RW model and the clustering. The RW model is fitted for a single age \( x \), while the LC model is fitted for the whole age span \([x_L, x_U]\). Clustering is conducted, helping project mortality rates more accurately with the underlying mortality models, in both directions either by bottom-up grouping all ages with
Figure 3.4: GMM clustering for UK males with the age-year fitting window [25, 84] × [1970, 2006].
homogeneous feature(s) in the RW model, or top-down splitting the age span \([x_L, x_U]\) into multiple clusters such that the ages in the same cluster have homogeneous feature(s) in the LC model. For the forecasting performance, the RW model needs to concern the over-fitting issue and the LC model may miss some information to capture the age-specific effects. It is reasonable to infer that clustering approaches can improve the forecasting performance of the classical LC model.

With regard to the clustering objects, we measure the empirical mortality data in two ways to obtain the correlations between the mean and the standard deviation of yearly decrements regarding the RW model, and between the intercept and the slope of the simple regression of an age-specific individual time trend on the group time trend regarding the LC model. For the former, to eliminate the downward time trend over \(t\) for age \(x\) form each of \(\ln(m_{x,t})\) and \(\logit(q_{x,t})\), let \(Y_{x,t} = \ln(m_{x,t}) - \ln(m_{x,t-1})\) under the LC model and \(Y_{x,t} = \logit(q_{x,t}) - \logit(q_{x,t-1})\) under the CBD model for \(x = x_L, ..., x_U\) and \(t = t_L + 1, ..., t_U\). Since \(\ln(m_{x,t})\) follows a random walk with drift model, \(Y_{x,t}\) follows a normal distribution with mean \(\theta_x^{RW}\) and standard deviation \(\sigma_x\). We apply the clustering methods to the mean \(\theta_x^{RW}\) and the standard deviation \(\sigma_x\) of \(Y_{x,t}\) for each age \(x\), that is, \(M_x = \bar{Y}_x = \frac{1}{n+1} \sum_{t=t_L+1}^{t_U} Y_{x,t}\) and \(SD_x = \tilde{\sigma}_x = \sqrt{\frac{1}{n-2} \sum_{t=t_L+1}^{t_U} (Y_{x,t} - \bar{Y}_x)^2}\) where \(n = t_U - t_L + 1\). Hence, the clustering algorithms will be applied to all pairs of \((M_x, SD_x), x = x_L, ..., x_U\).

For the latter about the intercept and the slope of the simple linear regression of an age-specific individual time trend on the group time trend, let the group time trend \(G\) and the age-specific time trend \(S_x\) for each age \(x \in [x_L, x_U]\) be

\[
\begin{align*}
S_x &= \{S_{x,t} : t = t_L, t_L + 1, ..., t_U\}, \\
G &= \{G_t = \sum_{x=x_L}^{x_U} S_{x,t} : t = t_L, t_L + 1, ..., t_U\},
\end{align*}
\]

where \(S_{x,t} = \ln(m_{x,t})\) under the LC model and \(S_{x,t} = \logit(q_{x,t})\) under the CBD model for \(x = x_L, ..., x_U\).

To generate the objects for clustering, let \(i_x\) and \(s_x\) be the intercept and the slope of the simple linear regression model with dependence variable \(S_x\) and explanatory variable \(G\). The intercept and slope coefficients \(i_x\) and \(s_x\) for each age \(x\) can be estimated by the least-squares approach as \(\hat{i}_x = \bar{S}_x - \bar{s}_x \cdot \bar{G}\) and \(\hat{s}_x\)

\[
\hat{s}_x = \frac{\sum_{t=t_L}^{t_U} (G_t - \bar{G})(S_{x,t} - \bar{S}_x)}{\sum_{t=t_L}^{t_U} (G_t - \bar{G})^2},
\]

where \(\bar{G} = \frac{1}{n} \sum_{t=t_L}^{t_U} G_t\), and \(\bar{S}_x = \frac{1}{n} \sum_{t=t_L}^{t_U} S_{x,t}\). Then the clustering algorithms will be also applied to all pairs of \((\hat{i}_x, \hat{s}_x), x = x_L, ..., x_U\).
Chapter 4

Numerical illustrations

In this chapter, we fit the LC and CBD models with mortality data for each of the clustered age groups after applying four clustering methods, and predict mortality rates for future consecutive years. The mortality data for both genders of the US and the UK come from the Human Mortality Database (HMD, www.mortality.org). For each single population, the age-year window \([x_L, x_U] \times [T_1, T_2]\) for mortality rates is studied, where \(x_L = 25\), \(x_U = 84\), \(T_1 = 1950\) and \(T_2 = 2016\). We take a series of fitting age-year windows \([x_L, x_U] \times [t_L, t_U]\) to implement clustering methods, and evaluate the forecasting performances of the LC and CBD models with and without clustering, respectively. The fitting year spans are \([t_L, t_U] = [1950, t_U], [1951, t_U], ..., [t_U - 9, t_U]\) with the shortest one \([t_U - 9, t_U]\) of 10-year length. Moreover, we adopt two forecasting periods \([t_U + 1, T_2]\) with \(t_U = 2006\) and 1996 for 10 and 20 years wide, respectively. Table 4.1 gives a summary of the fitting and forecasting year spans.

Table 4.1: Fitting and Forecasting year spans

<table>
<thead>
<tr>
<th>Fitting year span</th>
<th>([t_L, t_U])</th>
<th>([1950, 2006])</th>
<th>([1950, 1996])</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ending year of fitting year span</td>
<td>(t_U)</td>
<td>2006</td>
<td>1996</td>
</tr>
<tr>
<td>Forecasting year span</td>
<td>([t_U + 1, T_2])</td>
<td>([2007, 2016])</td>
<td>([1997, 2016])</td>
</tr>
<tr>
<td>Width of forecasting year span</td>
<td>(T_2 - t_U)</td>
<td>10</td>
<td>20</td>
</tr>
</tbody>
</table>

4.1 Visualization on classification

Before fitting the Lee-Carter and CBD models using \(\ln(m_{x,t})\) and \(\text{logit}(q_{x,t})\) respectively, the clusters of ages are first produced by the four clustering methods which are introduced in Chapter 3. Recall that there are two measurements on the empirical mortality data with regard to the clustering objects; we denote Type I object as the bivariate of the mean and the standard deviation of yearly decrements, and Type II object as the bivariate of the intercept.
and the slope coefficients of the simple regression between the individual and group time trends. That is, we classify the entire age span \([x_L, x_U]\) into several age categories based on the following two types of objects sets:

\[
\begin{align*}
\text{Type I object set} &= \{(M_{25}, SD_{25}), (M_{26}, SD_{26}), \ldots, (M_{84}, SD_{84})\}, \\
\text{Type II object set} &= \{(\hat{i}_{25}, \hat{s}_{25}), (\hat{i}_{26}, \hat{s}_{26}), \ldots, (\hat{i}_{84}, \hat{s}_{84})\}.
\end{align*}
\]

There are two R packages, "mclust" and "NbClust", implemented in this section. These two packages and crucial functions used are introduced in Appendix A.

Figures 4.1-4.8 illustrate the clustering results of applying the four algorithms to Type I and Type II objects for both genders of the US and the UK, respectively. Clusters are distinguished by symbol and colour. The four clustering approaches generate the same clustering results with some of \(t_L\)'s \((t_L = 1950, 1951, \ldots, t_U - 9)\). Therefore, in order to illustrate the comparisons among the clustering approaches, we choose a representative \(t_L\) for each population, which results in quite different components of clusters from each algorithm. From Figures 4.1-4.8, we observe that different \(t_L\)'s lead to different results and the two types of objects yield dissimilar distributions. However, there is a common pattern for all populations and types. As long as the \(t_L\) sets closer to \(t_U\) (i.e., a shorter fitting year span), the outline of the correlation between \(M_x\) and \(SD_x\) or between \(\hat{i}_x\) and \(\hat{s}_x\) switches from negative to positive. For a longer fitting year span with a small value of \(t_L\), the standard deviation of \(Y_x\) decreases as the mean of \(Y_x\) increases, whereas the standard deviation of \(Y_x\) increases as the mean of \(Y_x\) increases for a shorter fitting year span. Thus, the common pattern for Type I is that ages with low mean and low standard deviation of \(Y_x\) or high mean and high standard deviation of \(Y_x\) tend to be classified to a cluster. The scatters for Type II look similar to those for Type I but different scales. Moreover, the three clustering methods, Ward.D2, divisive and K-means, classify objects to clusters mainly based on \(\hat{i}_x\), the estimate of the intercept of simple linear regression of the age-specific time trend on the group time trend, whereas the GMM clustering groups up the bivariate objects fitted by a mixture of Gaussian distributions, which may explain why the GMM has a better forecasting performance than the other three methods shown in Section 4.2.

### 4.2 Forecasting performance

In this section, we evaluate the forecasting performances of the LC and CBD models with and without incorporating the clustering approaches.

After applying each of the four clustering methods to decompose the age span \([x_L, x_U]\) into multiple age clusters, we then apply the LC and CBD model, respectively, to each of the resulting age clusters separately to obtain the forecasted \(\ln(\hat{m}_{x,t_U+\tau}^{LC})\) (see (3.1.4)) and \(\logit(\hat{q}_{x,t_U+\tau}^{CBD})\) (see (3.1.9)) for age \(x\) in year \(t_U + \tau\). Since the ages classified into the same cluster have similar features/characteristics and are treated as having homogeneous risks,
Figure 4.1: Fitting year span \([1974, 2006]\), US males, Type I

(a) Ward.D2

(b) Divisive

(c) K-means

(d) GMM
Figure 4.2: Fitting year span [1984, 2006], US females, Type I

(a) Ward.D2
(b) Divisive
(c) K-means
(d) GMM
Figure 4.3: Fitting year span [1975, 2006], UK males, Type I

(a) Ward.D2

(b) Divisive

(c) K-means

(d) GMM
Figure 4.4: Fitting year span [1957, 2006], UK females, Type I

(a) Ward.D2

(b) Divisive

(c) K-means

(d) GMM
Figure 4.5: Fitting year span [1982, 2006], US males, Type II

(a) Ward.D2

(b) Divisive

(c) K-means

(d) GMM
Figure 4.6: Fitting year span [1960, 2006], US females, Type II

(a) Ward.D2

(b) Divisive

(c) K-means

(d) GMM
Figure 4.7: Fitting year span [1995, 2006], UK males, Type II

(a) Ward.D2

(b) Divisive

(c) K-means

(d) GMM
Figure 4.8: Fitting year span [1953, 2006], UK females, Type II

(a) Ward.D2

(b) Divisive

(c) K-means

(d) GMM
the slope for each of the ages in the same cluster is set to equal the average of the slopes over all ages in that cluster. Specifically, let $C_i$ be the $i$th age cluster with cluster size $m_i$.

Then the modified slope for age $x \in C_i$,

$$slope_x^M = \frac{1}{m_i} \sum_{x \in C_i} slope_x = \begin{cases} \frac{1}{m_i} \sum_{x \in C_i} \hat{\beta}_x \cdot \hat{\theta}^{LC_i} = \frac{\hat{\theta}^{LC_i}}{m_i}, & \text{LC model}, \\ \frac{1}{m_i} \sum_{x \in C_i} [\hat{\theta}_1^i + \hat{\theta}_2^i \cdot (x - \bar{x}_i)] = \hat{\theta}_1^i, & \text{CBD model}, \end{cases}$$

where superscript $i$ is attached to $\hat{\theta}^{LC}$, $\hat{\theta}_1$ and $\hat{\theta}_2$ to indicate cluster $i$, and $\bar{x}_i$ is the average of the ages in cluster $i$.

To measure the prediction accuracy between the observed one-year death probability $q$ and the predicted one $\hat{q}$, we adopt a common statistical quantity, the $MAPE$ $[t_L,t_U]$ $\{t_{U+1},T_2\}$ (mean absolute percentage error), which is defined below as the average of the absolute percentage errors of $\hat{q}s$ over the forecasting age-year window $[x_L,x_U] \times [t_U+1,T_2]$ based on the fitting year span $[t_L, t_U]$,

$$MAPE^{[t_L,t_U]}_{[t_U+1,T_2]} = \frac{1}{(T_2 - t_U)(x_U - x_L + 1)} \sum_{t=1}^{T_2-t_U} \sum_{x=x_L}^{x_U} \frac{|\hat{q}_{x,t_U+t} - q_{x,t_U+t}|}{q_{x,t_U+t}},$$

where $\hat{q}_{x,t_U+t}$ is projected by the LC or CBD model with or without clustering.

In order to evaluate the overall prediction performances, we use the measure of AMAPE which is the average of MAPEs over $t_L = T_1, T_1 + 1, \ldots, t_U - 9$. That is,

$$AMAPE^{[t_U+1,T_2]}_{[t_U+1,T_2]} = \frac{1}{t_U - 9 - T_1 + 1} \sum_{t_{L=1}}^{t_{L=9}} MAPE^{[t_L,t_U]}_{[t_U+1,T_2]},$$

where $T_1 = 1950$ and $T_2 = 2016$.

Figures 4.9-4.16 display the MAPEs produced from the LC and CBD mortality models with and without clustering, under Type I and Type II objects for both genders of the US and the UK populations, against $t_L = 1950, 1951, \ldots, t_U - 9$, $t_U = 2006$ and 1996, respectively. The AMAPE values are shown at the top of each figure. The goal is to compare the forecasting performances against $t_L$ among models, fitting year spans, and types of clustering objects. We have the following observations from Figures 4.9-4.16.

- It is obviously to see the four MAPE curves of Ward.D2, divisive, K-means and GMM methods produce lower MAPEs for most of $t_L$ values than the LC and CBD models without clustering. This observation demonstrates that forecasting performance can be improved by incorporating clustering methods into a mortality model.

- There are no common trends in the MAPE curves against $t_L$ for all populations, both types of objects, and both models with or without clustering.

31
• Most of the MAPEs for the 10-year forecasting period are smaller than those for 20-year one for each population under both models and both types. In particular, the US female shows a close scale of MAPEs for the two forecasting periods.

• Under the CBD model, all clustering methods for males perform better than females for both types (Type I and Type II) and countries (the US and the UK).

• Comparing to the MAPE curves for the LC and CBD models without clustering, the curves of MAPEs for the models with clustering are less smooth. The reason could be that clustering algorithms generally are sensitive to outliers. Even if just one outlier occurred, two adjacent values of $t_L$ may produce two completely different clustering results, which lead to a jump-up or jump-down MAPE value.

• The GMM clustering works remarkably better in forecasting than the other three clustering methods for Type II object, and the GMM for Type I object also performs well.

Tables 4.2-4.5 display the AMAPEs for mortality models with and without clustering under Type I and Type II objects for both genders of the US and the UK populations with $t_U = 2006$ and 1996, respectively. These tables give an overview of the comparisons of forecasting performances between the LC/CBD models without and with incorporating clustering approaches. There are Panels A and B in each table, representing 10 and 20 years forecasting periods, [2007, 2016] and [1997, 2016], respectively. Each panel also shows the average of the AMAPEs over the four populations under each of the five methods, and highlights the smallest value in bold, representing the overall best method in forecasting performance. We have the following observations from Tables 4.2-4.5.

• All the AMAPEs for the LC/CBD models without clustering are larger than the AMAPEs for both models with each of the four clustering methods. Comparing the average of AMAPEs, we observe that the LC and CBD models without clustering perform worse than incorporating with clustering methods for all populations and cases.

• The LC model without clustering has a better forecasting performance than the CBD model for males of both countries and the 10-year forecasting period [2007, 2016] (Panel A), whereas the CBD model without clustering predicts the mortality rates better than the LC model for females of both countries and the 20-year forecasting period [1997, 2016] (Panel B).

• For Type I object (Table 4.2 and Table 4.3), females of the US and the UK, and both forecasting periods, all clustering methods improve forecasting performance of the LC model more significantly than that of the CBD model.
Figure 4.9: $MAPE_{[t_L, 2006]}^{[2007, 2016]}$ against $t_L$, Type I, LC

(a) US males

(b) UK males

(c) US females

(d) UK females

<table>
<thead>
<tr>
<th>Method</th>
<th>LC</th>
<th>Ward D2</th>
<th>Divisive</th>
<th>Kmeans</th>
<th>GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMAPE</td>
<td>8.69%</td>
<td>7.63%</td>
<td>7.55%</td>
<td>7.53%</td>
<td>7.69%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>LC</th>
<th>Ward D2</th>
<th>Divisive</th>
<th>Kmeans</th>
<th>GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMAPE</td>
<td>10.12%</td>
<td>9.14%</td>
<td>9.31%</td>
<td>9.28%</td>
<td>8.99%</td>
</tr>
</tbody>
</table>
Figure 4.10: $MAPE^{[t_{L,1996}]}_{[1997,2016]}$ against $t_{L}$, Type I, LC

(a) US males

(b) UK males

(c) US females

(d) UK females

<table>
<thead>
<tr>
<th>Method</th>
<th>LC</th>
<th>Ward D2</th>
<th>Divisive</th>
<th>Kmeans</th>
<th>GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMAPE</td>
<td>15.08%</td>
<td>10.31%</td>
<td>10.44%</td>
<td>10.39%</td>
<td>11.21%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>LC</th>
<th>Ward D2</th>
<th>Divisive</th>
<th>Kmeans</th>
<th>GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMAPE</td>
<td>17.63%</td>
<td>15.18%</td>
<td>15%</td>
<td>14.98%</td>
<td>15.17%</td>
</tr>
</tbody>
</table>
Figure 4.11: $MAPE_{[t_L,2006]}^{[2007,2016]}$ against $t_L$, Type I, CBD

(a) US males

(b) UK males

(c) US females

(d) UK females
Figure 4.12: $MAPE_{[t_L,1996]}^{[1997,2016]}$ against $t_L$, Type I, CBD

(a) US males

(b) UK males

(c) US females

(d) UK females
Figure 4.13: $\text{MAPE}^{[t_L,2006]}_{[2007,2016]}$ against $t_L$, Type II, LC

(a) US males
(b) UK males
(c) US females
(d) UK females

<table>
<thead>
<tr>
<th>Method</th>
<th>LC</th>
<th>Ward D2</th>
<th>Divisive</th>
<th>Kmeans</th>
<th>GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMAPE</td>
<td>8.69%</td>
<td>7.11%</td>
<td>7.21%</td>
<td>7.18%</td>
<td>5.13%</td>
</tr>
<tr>
<td></td>
<td>10.12%</td>
<td>8.46%</td>
<td>8.63%</td>
<td>8.7%</td>
<td>7%</td>
</tr>
</tbody>
</table>

37
Figure 4.14: $MAPE_{[t_L,1996]}^{[1997,2016]}$ against $t_L$, Type II, LC

(a) US males

(b) UK males

(c) US females

(d) UK females
Figure 4.15: $MAPE_{[t_L,2006]}^{[t_L,2006]}$ against $t_L$, Type II, CBD

(a) US males

(b) UK males

(c) US females

(d) UK females
Figure 4.16: $\text{MAPE}^{[t_L, 1996]}_{[1997, 2016]}$ against $t_L$, Type II, CBD

(a) US males
(b) UK males
(c) US females
(d) UK females
Particularly, the GMM clustering method works outstandingly on Type II object (see both Panel A and Panel B in Tables 4.4 and 4.5). For instance, the AMAPE for the US males and the forecasting period \[2007, 2016\] is reduced to 5.55% from 11.45% by applying the GMM under the CBD model.

Among the four clustering approaches applied to Type I object, the GMM and the K-means produce the lowest averages of AMAPEs over four populations for 10-year (see Panel A in Tables 4.2 and 4.3) and 20-year (see Panel B in Tables 4.2 and 4.3) forecasting periods, respectively, under both the LC and CBD models.

Among the four clustering approaches applied to Type II object, the GMM dominates the others in terms of the averages of AMAPEs over four populations except for the UK and the 20-year forecasting period under the CBD model (see Panel B in Table 4.5) where the Ward.D2 produces a slightly remarkable forecasting improvement than the other three, compared to the original CBD mortality model.

Comparing the AMAPE values in Tables 4.2-4.5, we find that the GMM applied to Type II object produces lower AMAPE values than Type I object for all combinations of populations, mortality models, and forecasting periods. Therefore, Type II object are more effective for the GMM to improve the forecasting performances of both the LC and CBD models.

4.3 Improvement illustrations

In this section, we investigate the forecasting improvement of the LC and CBD mortality models incorporating with clustering. Recall that \(\ln(\hat{m}_{x,t_U+\tau})\) and \(\text{logit}(\hat{q}_{x,t_U+\tau})\) for age \(x\) in year \(t_U + \tau\) under the LC and CBD models are linear functions of \(\tau\) with different intercepts and slopes (see (3.1.4) for the LC model and (3.1.9) for the CBD model). The intercept determines the location where mortality forecasts originate with the slope, and the slope is to capture the downward time trend of the future mortality rates. The effects of the slope and intercept of \(\ln(\hat{m}_{x,t})\) or \(\text{logit}(\hat{q}_{x,t})\) are illustrated in Figure 4.17 for forecasting \(\ln(m_{x,t}), t = t_U + 1, ..., 2016\), with the mortality rates for US males and fitting age-year window \([25, 84] \times [1950, t_U]\) where \(t_U = 1996\) and 2006. Comparing the observed mortality rates and the forecasted mortality rates under the LC or CBD model with and without clustering, we have the following observations from Figure 4.17.

- For all the cases shown in Figure 4.17, incorporating the clustering methods into the LC or CBD model produces closer forecasted mortality rates to the observed ones.
- Applying the clustering methods lead the LC and CBD mortality models to a better starting location and a more accurate slope to capture the time trend of the observed mortality rates.
Table 4.2: $AMAPE_{[t_U+1,2016]}$ for the age span [25, 84], Type I, LC

<table>
<thead>
<tr>
<th>Population/Method</th>
<th>LC</th>
<th>Ward.D2</th>
<th>Divisive</th>
<th>Kmeans</th>
<th>GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A:</strong> $t_U = 2006$ for forecasting year span [2007, 2016]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US</td>
<td>Male</td>
<td>8.69%</td>
<td>7.63%</td>
<td>7.55%</td>
<td>7.53%</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>8.66%</td>
<td>6.65%</td>
<td>6.62%</td>
<td>6.71%</td>
</tr>
<tr>
<td>UK</td>
<td>Male</td>
<td>10.12%</td>
<td>9.14%</td>
<td>9.31%</td>
<td>9.28%</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>8.02%</td>
<td>6.80%</td>
<td>7.03%</td>
<td>6.81%</td>
</tr>
<tr>
<td></td>
<td><strong>Average</strong></td>
<td>8.87%</td>
<td>7.56%</td>
<td>7.63%</td>
<td>7.58%</td>
</tr>
<tr>
<td><strong>Panel B:</strong> $t_U = 1996$ for forecasting year span [1997, 2016]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US</td>
<td>Male</td>
<td>15.08%</td>
<td>10.31%</td>
<td>10.44%</td>
<td>10.39%</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>9.12%</td>
<td>7.15%</td>
<td>7.06%</td>
<td>7.03%</td>
</tr>
<tr>
<td>UK</td>
<td>Male</td>
<td>17.63%</td>
<td>15.18%</td>
<td>15.00%</td>
<td>14.98%</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>12.79%</td>
<td>9.32%</td>
<td>9.39%</td>
<td>9.33%</td>
</tr>
<tr>
<td></td>
<td><strong>Average</strong></td>
<td>13.66%</td>
<td>10.49%</td>
<td>10.47%</td>
<td><strong>10.43%</strong></td>
</tr>
</tbody>
</table>

Table 4.3: $AMAPE_{[t_U+1,2016]}$ for the age span [25, 84], Type I, CBD

<table>
<thead>
<tr>
<th>Population/Method</th>
<th>CBD</th>
<th>Ward.D2</th>
<th>Divisive</th>
<th>Kmeans</th>
<th>GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A:</strong> $t_U = 2006$ for forecasting year span [2007, 2016]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US</td>
<td>Male</td>
<td>11.45%</td>
<td>7.00%</td>
<td>6.30%</td>
<td>6.30%</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>7.82%</td>
<td>7.02%</td>
<td>7.07%</td>
<td>7.14%</td>
</tr>
<tr>
<td>UK</td>
<td>Male</td>
<td>13.38%</td>
<td>8.57%</td>
<td>8.74%</td>
<td>8.65%</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>8.33%</td>
<td>7.98%</td>
<td>8.06%</td>
<td>7.91%</td>
</tr>
<tr>
<td></td>
<td><strong>Average</strong></td>
<td>10.24%</td>
<td>7.64%</td>
<td>7.54%</td>
<td>7.50%</td>
</tr>
<tr>
<td><strong>Panel B:</strong> $t_U = 1996$ for forecasting year span [1997, 2016]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US</td>
<td>Male</td>
<td>14.55%</td>
<td>11.06%</td>
<td>11.23%</td>
<td>11.20%</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>7.91%</td>
<td>7.44%</td>
<td>7.31%</td>
<td>7.27%</td>
</tr>
<tr>
<td>UK</td>
<td>Male</td>
<td>18.59%</td>
<td>13.71%</td>
<td>13.88%</td>
<td>13.69%</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>10.63%</td>
<td>9.22%</td>
<td>9.25%</td>
<td>9.18%</td>
</tr>
<tr>
<td></td>
<td><strong>Average</strong></td>
<td>12.92%</td>
<td>10.36%</td>
<td>10.42%</td>
<td><strong>10.34%</strong></td>
</tr>
</tbody>
</table>
Table 4.4: $AMAPE_{tU+1,2016}$ for the age span $[25, 84]$, Type II, LC

<table>
<thead>
<tr>
<th>Population/Method</th>
<th>LC</th>
<th>Ward.D2</th>
<th>Divisive</th>
<th>Kmeans</th>
<th>GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$t_U = 2006$ for forecasting year span</td>
<td>2007, 2016</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>8.69%</td>
<td>7.73%</td>
<td>7.89%</td>
<td>7.91%</td>
<td>5.84%</td>
</tr>
<tr>
<td>Female</td>
<td>8.66%</td>
<td>7.11%</td>
<td>7.21%</td>
<td>7.18%</td>
<td>5.13%</td>
</tr>
<tr>
<td>Male</td>
<td>10.12%</td>
<td>8.46%</td>
<td>8.63%</td>
<td>8.70%</td>
<td>7.00%</td>
</tr>
<tr>
<td>Female</td>
<td>8.02%</td>
<td>6.57%</td>
<td>6.25%</td>
<td>6.48%</td>
<td>5.45%</td>
</tr>
<tr>
<td>Average</td>
<td>8.87%</td>
<td>7.47%</td>
<td>7.49%</td>
<td>7.57%</td>
<td>5.85%</td>
</tr>
<tr>
<td>UK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>15.08%</td>
<td>10.80%</td>
<td>10.79%</td>
<td>10.78%</td>
<td>10.56%</td>
</tr>
<tr>
<td>Female</td>
<td>9.12%</td>
<td>7.15%</td>
<td>7.06%</td>
<td>7.03%</td>
<td>7.47%</td>
</tr>
<tr>
<td>Male</td>
<td>17.63%</td>
<td>14.38%</td>
<td>14.77%</td>
<td>14.82%</td>
<td>14.46%</td>
</tr>
<tr>
<td>Female</td>
<td>12.79%</td>
<td>10.11%</td>
<td>10.01%</td>
<td>10.07%</td>
<td>9.52%</td>
</tr>
<tr>
<td>Average</td>
<td>13.66%</td>
<td>10.72%</td>
<td>10.78%</td>
<td>10.81%</td>
<td>10.42%</td>
</tr>
</tbody>
</table>

Table 4.5: $AMAPE_{tU+1,2016}$ for the age span $[25, 84]$, Type II, CBD

<table>
<thead>
<tr>
<th>Population/Method</th>
<th>CBD</th>
<th>Ward.D2</th>
<th>Divisive</th>
<th>Kmeans</th>
<th>GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$t_U = 2006$ for forecasting year span</td>
<td>2007, 2016</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>11.45%</td>
<td>8.33%</td>
<td>8.16%</td>
<td>8.47%</td>
<td>5.55%</td>
</tr>
<tr>
<td>Female</td>
<td>7.82%</td>
<td>6.16%</td>
<td>6.30%</td>
<td>6.26%</td>
<td>4.95%</td>
</tr>
<tr>
<td>Male</td>
<td>13.38%</td>
<td>7.29%</td>
<td>7.68%</td>
<td>7.64%</td>
<td>6.50%</td>
</tr>
<tr>
<td>Female</td>
<td>8.33%</td>
<td>6.03%</td>
<td>5.87%</td>
<td>5.97%</td>
<td>4.91%</td>
</tr>
<tr>
<td>Average</td>
<td>10.24%</td>
<td>6.95%</td>
<td>7.00%</td>
<td>7.08%</td>
<td>5.47%</td>
</tr>
<tr>
<td>UK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>14.55%</td>
<td>11.80%</td>
<td>11.78%</td>
<td>11.74%</td>
<td>10.93%</td>
</tr>
<tr>
<td>Female</td>
<td>7.91%</td>
<td>7.06%</td>
<td>7.29%</td>
<td>7.26%</td>
<td>6.94%</td>
</tr>
<tr>
<td>Male</td>
<td>18.59%</td>
<td>12.87%</td>
<td>13.20%</td>
<td>13.10%</td>
<td>13.75%</td>
</tr>
<tr>
<td>Female</td>
<td>10.63%</td>
<td>8.30%</td>
<td>8.38%</td>
<td>8.30%</td>
<td>8.70%</td>
</tr>
<tr>
<td>Average</td>
<td>12.92%</td>
<td>10.01%</td>
<td>10.16%</td>
<td>10.10%</td>
<td>10.08%</td>
</tr>
</tbody>
</table>
• Figure (a) shows that the LC model with the GMM clustering can provide a more accurate time trend.

• Figures (c) and (d) present that the CBD model with the GMM clustering can locate the starting point much better than the original model without clustering, both under closer slope values, i.e., it can produce a closer fitted value \( \ln(\hat{m}_{x,t_U}) \) to the observed value \( \ln(m_{x,t_U}) \).

Figure 4.17: Observed and predicted \( \ln(m_{x,t})/\logit(m_{x,t}) \) against \( t = t_U + 1, \ldots, 2016 \) with age span [25, 84]

(a) Fitting year span [1981, 1996], \( x = 35 \)
(b) Fitting year span [1985, 2006], \( x = 25 \)

(c) Fitting year span [1951, 2006], \( x = 50 \)
(d) Fitting year span [1950, 2006], \( x = 25 \)
Chapter 5

Conclusions

This project studies incorporation of statistical clustering methods into the Lee-Carter and CBD models to improve their performances in forecasting mortality rates. Statistical clustering in the sense of modeling mortality rates is a procedure of classifying all study ages to classes such that ages in the same class are more homogeneous, with respect to some key features or characteristics in the underlying mortality models, each other than to those in other classes. We apply four popular clustering approaches: the Ward’s hierarchical clustering, the divisive hierarchical clustering, the K-means clustering, and the Gaussian mixture model clustering. They are adopted to first determine the number and partition of age subgroups from a whole group of ages \([25, 84]\). The mortality data we used are both gender of the US and the UK from the Human Mortality Database, and the clustering approaches are implemented with R packages "NbClust" and "mclust". Afterwards, we forecast 10-year and 20-year mortality rates for each of the age subgroups using the Lee-Carter and CBD models, respectively. The MAPE (mean absolute percentage error) measure and the average of the MAPEs over both genders of the US and the UK are adopted to evaluate forecasting performance. By comparing the MAPE values with and without clustering, we demonstrate that all the proposed clustering methods can improve forecasting performances of the Lee-Carter and CBD models. The Gaussian mixture model clustering overall outperforms the other three clustering methods and results in a remarkable mortality improvement over the traditional Lee-Carter and CBD models without clustering.

However, it is not easy to conclude which clustering method performs the best in mortality forecasting for all cases. There is no objectively the most appropriate clustering algorithm. The most appropriate clustering algorithm for a particular problem often needs to be chosen experimentally, unless there is theoretical support to favor a particular cluster model. It can be seen in this project that various combinations of populations, objects, clustering methods, and study data sets lead to different forecasting results. The reason could be that each clustering algorithm has its own limitation and some of them heavily rely on the initialization which is usually set randomly in iterative approaches. For instance, although the model-based clustering has been applied successfully in this project, the clus-
tering methods based on multivariate normal mixture models indeed have limitations for non-Gaussian, high-dimensional, and large data sets.

For the future work, we will develop more clustering applications in other modern mortality models. It may be considerable to incorporate some other statistical or actuarial approaches, such as Bayesian and Bühlmann credibility, into mortality models with clustering. Moreover, we will study the relevance between clustering methods and mortality models so that mathematical expressions can be derived for the mortality models with clustering.
Bibliography


Appendix A

R Packages

A.1 mclust

'mclust' is an R package developed for model-based clustering, classification, and density estimation based on finite multivariate Gaussian mixture models. It relies on the EM algorithm for model-based clustering and classification, and the Bayesian Information Criterion (BIC) in comprehensive strategies for model selection. Model-based hierarchical clustering is also implemented in 'mclust' and is used to initialize the EM algorithm. Additional functions are available for displaying and visualizing fitted models along with clustering results. Its flexibility, availability and relatively frequent good performance make this package one of the most popular. More details can be referred to https://mclust-org.github.io/mclust/.

A.2 NbClust

The R package 'NbClust' provides 30 indices for determining the best clustering scheme with the optimal number of clusters from the various results obtained by comparing all combinations of cluster sizes, distance measurements, and clustering methods to be used. The indices can be adopted to outputs of two clustering algorithms, K-means and hierarchical agglomerative clustering. Distance measures available in the 'NbClust' package are: Euclidean distance, maximum distance, Manhattan distance, Canberra distance, binary distance and Minkowski distance. Several agglomeration methods are also provided by the ‘NbClust’ package, namely: Ward, single, complete, average, McQuitty, median and centroid. All of these methods and distance measures can be referred to Charrad et al. (2014). Also, the package is available from the Comprehensive R Archive Network (CRAN) at http://CRAN.R-project.org/package=NbClust.
A.3 cluster

The R package 'cluster' implements the methods for cluster analysis based on the article 'Finding Groups in Data' (Kaufman and Rousseeuw (1990)). The package is available at https://svn.r-project.org/R-packages/trunk/cluster.