Estimating Veterans' Health Benefit Grants Using the Generalized Linear Mixed Cluster-Weighted Model with Incomplete Data

ESTIMATING VETERANS' HEALTH BENEFIT GRANTS USING THE GENERALIZED LINEAR MIXED CLUSTER-WEIGHTED MODEL WITH INCOMPLETE DATA

BY

XIAOYING DENG, 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 (2017) McMaster University
(Mathematics & Statistics) Hamilton, Ontario, Canada

TITLE: Estimating Veterans' Health Benefit Grants Using the Generalized Linear Mixed Cluster-Weighted Model with Incomplete Data

AUTHOR: Xiaoying Deng
B.Sc., (Mathematics and Statistics)
St. Francis Xavier University, Antigonish, NS, Canada

SUPERVISOR: Dr. Paul McNicholas

CO-SUPERVISOR: Dr. Petar Jevtic

NUMBER OF PAGES: ix, 37

To my beloved parents, who have always told me "where there is life, there is hope".

Abstract

The poverty rate among veterans in US has increased over the past decade, according to the U.S. Department of Veterans Affairs (2015). Thus, it is crucial to veterans who live below the poverty level to get sufficient benefit grants. A study on prudently managing health benefit grants for veterans may be helpful for government and policy-makers making appropriate decisions and investments. The purpose of this research is to find an underlying group structure for the veterans' benefit grants dataset and then estimate veterans' benefit grants sought using incomplete data. The generalized linear mixed cluster-weighted model based on mixture models is carried out by grouping similar observations to the same cluster. Finally, the estimates of veterans' benefit grants sought will provide reference for future public policies.

Acknowledgements

Firstly, I would like to express sincere appreciation to my supervisor, Dr. Paul McNicholas, for his continuous encouragements, guidance and support, without which this work at McMaster University would not be possible. It is my pleasure to study and work under his supervision.

Secondly, I would like to thank my co-supervisor, Dr. Petar Jevtic, for his support and suggestions towards the completion of my thesis, and Dr. Tatjana Milijkovic, for suggesting this interesting data on veterans' health benefit grants.

I also would like to thank my committee members Dr. Shui Feng and Dr. Peter Macdonald for accepting to be my thesis examiners and for their invaluable discussion, ideas, and feedback.

Then I would like to thank Dr. Sunny Wang, for her recommendation, encouragement and support all the time, with which I gain more confidence to keep challenging and improving myself.

I would like to thank my friends, classmates and professors I met at McMaster University and my roommates in Hamilton for their encouragement and assistance.

Finally, I want to thank my family, especially my parents, for all their love, care, and understanding. Without their support, I am not able to have this valuable experience studying in Canada.

Contents

Abstract **iv**

Acknowledgements **v**

1 Introduction **1**

2 Data Structure **3**

2.1 Background 3

2.2 Monthly Income across Different Groups 6

2.3 Benefit Grants with Monthly Income across Different Groups 7

3 Methodology **10**

3.1 Finite Mixture Models 10

3.2 Cluster-Weighted Model (CWM) 11

3.3 Generalized Linear Mixed CWMs 12

3.4 Likelihood and Parameter Estimation 14

3.4.1 EM Initialization 19

3.4.2 EM Convergence Criterion 19

3.5 Model Selection 20

4 Analysis of Veterans' Benefits Data **22**

4.1 Generalized Linear Mixed CWMs 22

4.2 Model Evaluation 26

5 Conclusions and Future Work **31**

Bibliography **33**

List of Figures

2.1 Boxplots illustrating the monthly income for each value of the variables Spouse, Widow and Gender, respectively. 6

2.2 Boxplot illustrating the monthly income in each age group. 7

2.3 Boxplots illustrating the benefit grants for each value of the variables Spouse, Widow and Gender, respectively. 8

2.4 Boxplot illustrating the benefit grants in each age group. 9

4.1 Boxplots illustrating the difference between clusters in age, monthly income and benefit grants, respectively. 24

4.2 Boxplots illustrating the difference between clusters in monthly income (left) and benefit grants (right) with respect to the variable Gender. . 25

4.3 Boxplots illustrating the difference between clusters in monthly income (left) and benefit grants (right) with respect to the variables Spouse and Widow. 26

4.4 Boxplots illustrating the comparisons between true (left) and expected (right) benefit grants for the uncensored and censored groups. 27

4.5 Boxplots illustrating the comparisons between true (left) and expected (right) benefit grants in each cluster. 27

4.6 Boxplot illustrating the comparisons between expected benefit grants for censored observations in each cluster. 28

4.7 Boxplots illustrating the comparisons between true (left) and expected (right) benefit grants with respect to the variable Gender in each cluster. 29

4.8 Boxplots illustrating the comparisons between true (left) and expected (right) benefit grants with respect to the variable Spouse and Widow in each cluster. 30

Chapter 1

Introduction

As of the end of 2010, the US Census Bureau recorded that the percentage of living veterans has increased to 7.3% of the total population (U.S. Department of Veterans Affairs, 2010). The poverty rate among veterans has also increased over the past decade. This is evidenced by Essenburg and Hanson (2014), where they stated that veterans poverty increased nearly 1 percentage point from 2007 to 2009, to 6.3 percent.

There are a number of federal and state benefit programs and services offered by the Department of Veterans Affairs (DVA) for eligible veterans who may be disabled or have low income after serving in the active military service. With an increasing number of benefit programs and services implemented, the number of uninsured veterans has decreased in recent years. Thus, appropriate benefit programs are important for veterans, especially for those who live below the poverty level.

In ND, benefit programs are carried out under the supervision of the Administrative Committee of Veterans Affairs (ACOVA). Limited funds are available to eligible veterans and their families. To manage the funds well, thereby helping government make reasonable decisions, the problem of evaluating the level of benefit grants for veterans in ND becomes very crucial. This thesis extends a statistical approach, the generalized linear mixed cluster-weighted model (CWM), to deal with the missingness (right-censoring) in the data. The aim of this research is to estimate the benefit needs for veterans in ND, so that more accurate assessment of future needs is available.

The data collected on health benefit grants of applications include right censored observations. More details about how censoring is defined and its related topics can be found in Miller (1976); Koul *et al.* (1981); Le (1997); Guiahi (2001); Miller Jr (2011). In the financial assessments of benefit grants, the amount of payments are defined to be right censored if the benefit grants required exceed the amount limited. Some works that has used expectation-maximization (EM) algorithm (Dempster *et al.*, 1977) to deal with censoring can be found in Chauveau (1995); Qin *et al.* (2011); Park and Lee (2012).

This thesis is organized as follows. Chapter 2 gives a brief overview of the data set. In Chapter 3, a family of CWMs and algorithms for parameter estimation as well as model selection is presented. Model analysis and discussions are given in Chapter 4. This thesis ends by conclusions and future work in Chapter 5.

Chapter 2

Data Structure

2.1 Background

The data were provided by ND DVA on dental benefits grants from the year 2000 to 2010. Besides dental benefits, there are other categories of medical benefits offered: dentures, hearing, optical, and special. However, only a small amount of data were collected for these four categories and hence they are not considered herein.

Benefit grants are only available for a capped or a limited amount, which has been changed at different times. From the year 2000 to 2010, the dental benefit cap experienced several adjustments. The benefit cap was $500 by the end of 2004, while it was increased to $750 as of early January of 2006. By November of 2007, it was increased to $1000.

For any application, if the amount claimed exceeds the present benefit limit, then the applicant will only be reimbursed an amount corresponding to the benefit limit. In this situation, the benefit grants for this application will be recorded as capped, i.e., right-censored. The amounts of benefits granted for each application were adjusted for inflation using the Consumer Price Index published by the Bureau of Labor Statistics, US Department of Labor (Miljkovic and Barabanov, 2015).

The data set consists of 575 observations with 311 uncensored and 264 censored. There are 9 variables considered in total, with their descriptions listed in Table 2.1.

Table 2.1: Description of variables in the ND DVA data set with the corresponding field values in the database.

Variable Name	Description
ApplicantTBLPK	Applicant's unique non-identifiable ID
InfPaid	The amount of benefits granted adjusted by inflation
CencID	Uncensored benefit grants (0) / Censored benefit grants (1)
AppYear	Application year (2000-2010)
Age	Applicant's age
Gender	Male (0) / Female (1)
IncomeLevel	Applicant's monthly income
Spouse	Primary beneficiary-veteran (0) / Spouse of a living veteran (1)
Widow	Not widowed (0) / Widowed-widow or widower (1)

Each applicant can submit more than one application if they have several dental appointments. There are 368 different applicants. The variables Spouse and Widow both describe the status of an individual receiving benefit grants. The spouse of a veteran can only use the grant if the veteran is alive.

Of a total of 368 applicants, 98.0% have monthly income less than $1400 and the monthly income of 46.5% applicants is below $800. The highest monthly income is $2600. In all, one female and 33 male applicants are reported to have zero monthly income.

The range of applicants' age is from 23 to 94 years, and divided into five age groups. Table 2.2 gives the number of applications classified in each of the age groups.

Table 2.2: The number of applications classified in each of the age groups.

Age	[23, 50)	[50, 60)	[60, 70)	[70, 80)	[80, 94]
Frequency	74	210	118	117	56

Notably, 87.1% applications came from applicants whose age is greater than or equal to 50. There are only three applicants over 90 years old and six applicants under 40.

The applicants include 287 males and 81 females. Table 2.3 gives the partition of applicants with respect to the variables Spouse, Widow, and Gender.

Table 2.3: The number of applications are partitioned with respect to the variables Spouse, Widow, and Gender.

	Not widowed (0)			Widowed (1)		
	Total	Male (0)	Female (1)	Total	Male (0)	Female (1)
Veteran (0)	309	283	26	33	2	31
Spouse (1)	26	2	24	0	0	0

Almost all of the applicants indicated as either spouses of living veterans or widowed are female. There are 24 females among 26 applicants who are spouses of living veterans and 31 widows among 33 widowed applicants.

2.2 Monthly Income across Different Groups

In Figure 2.1, it is observed that the monthly income of applicants who are spouses of living veterans is higher than that of the veterans. The monthly income of widowed applicants is lower than those are not widowed. Further, female applicants have a higher average monthly income, whereas monthly income of the a few male applicants is relatively high.

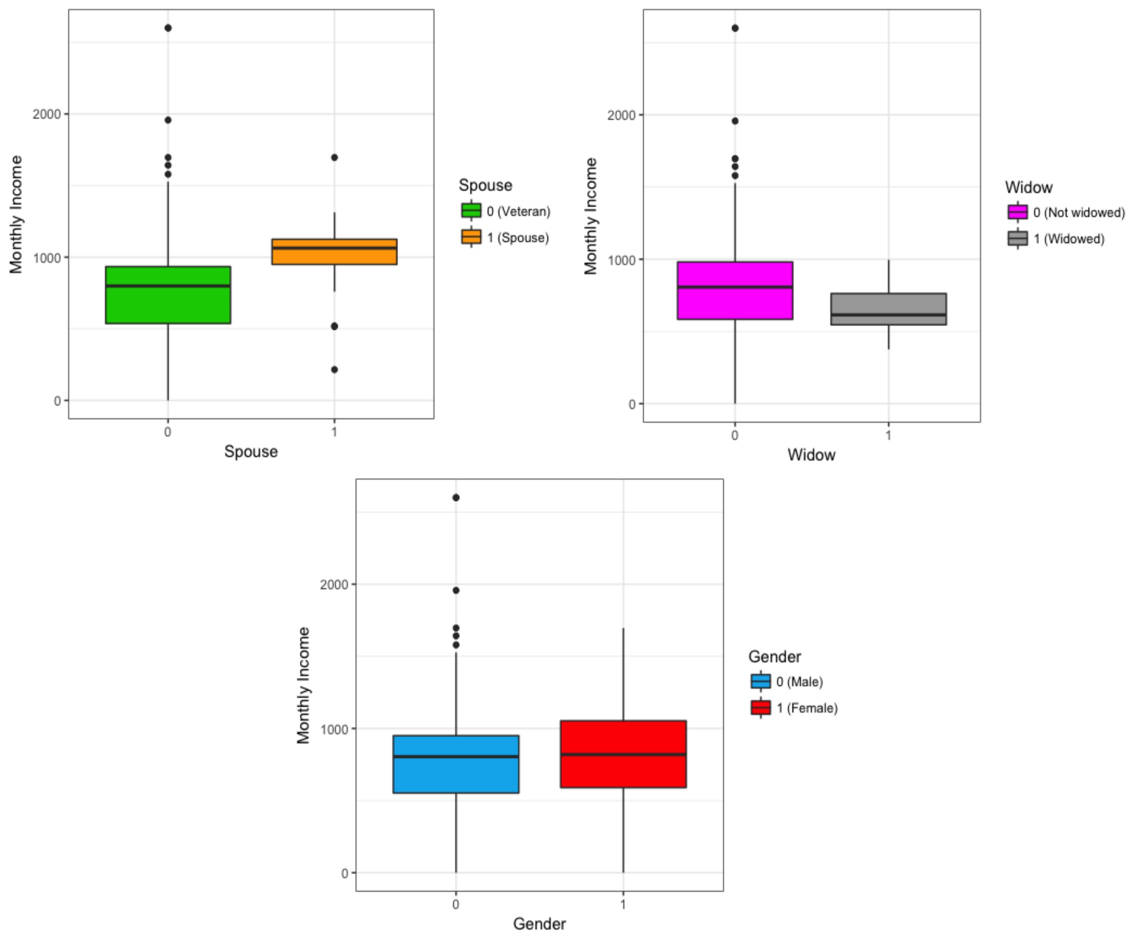

Figure 2.1: Boxplots illustrating the monthly income for each value of the variables Spouse, Widow and Gender, respectively.

The plot of Figure 2.2 shows that applicants whose age is greater or equal to 80 achieve the highest average as well as the highest median monthly income. Applicants in the age group between 60 to 70 have the second highest average and also median monthly income.

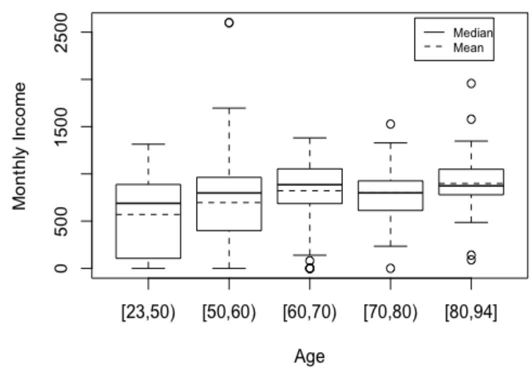

Figure 2.2: Boxplot illustrating the monthly income in each age group.

2.3 Benefit Grants with Monthly Income across Different Groups

The benefit grants change differently along the monthly income for applicants across different groups. In Figure 2.3, it is observed that the average benefit grants are generally the same for applicants who are veterans or who are spouses of living veterans. The same situation occurs between male and female applicants. However, a few female applicants, widowed applicants or those who are spouses of living veterans sought relatively high benefit grants. In addition, it is found that widowed applicants have low monthly income as well as low benefit grants sought.

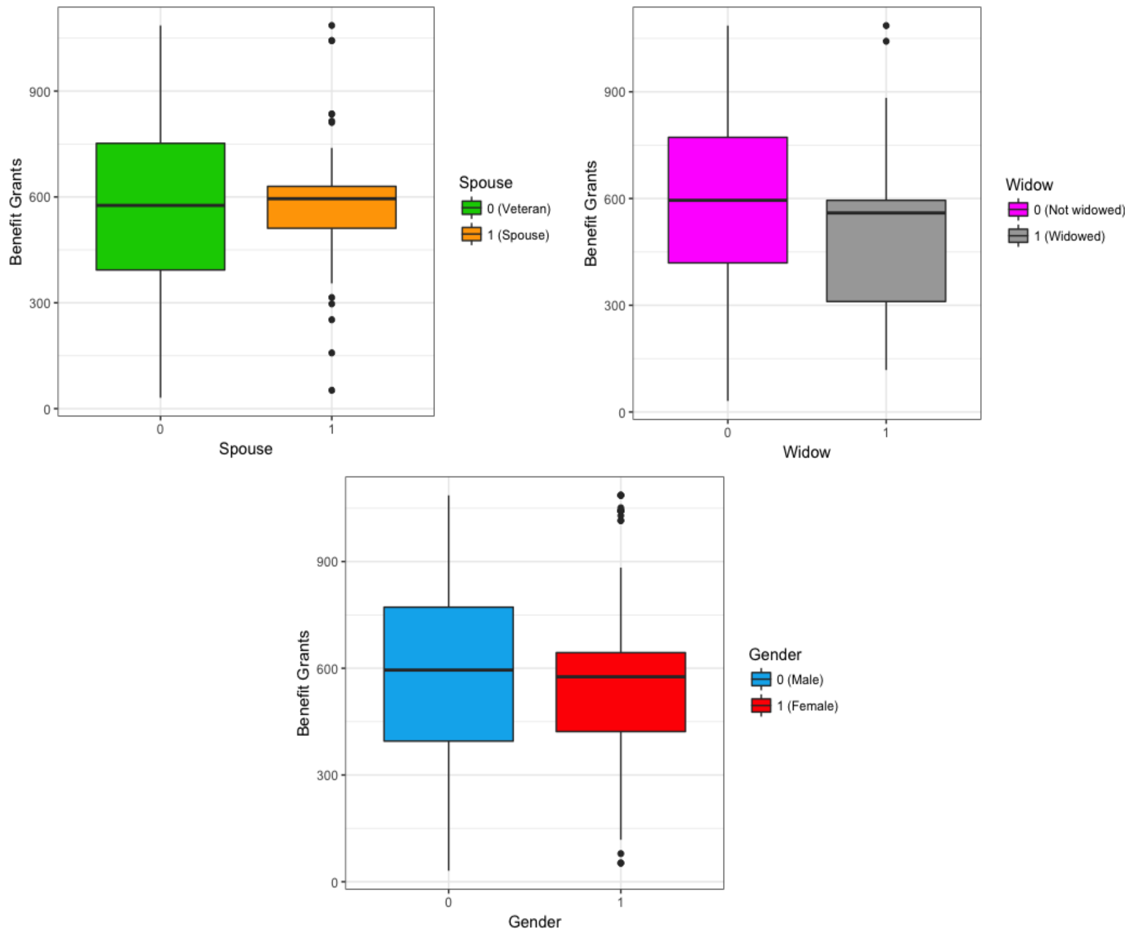

Figure 2.3: Boxplots illustrating the benefit grants for each value of the variables Spouse, Widow and Gender, respectively.

The plot of Figure 2.4 shows the second highest monthly income applicants in the age group between 60 to 70, which is observed from Figure 2.2, sought the highest benefit grants, while applicants over 80 with the highest monthly income sought relatively low benefit grants.

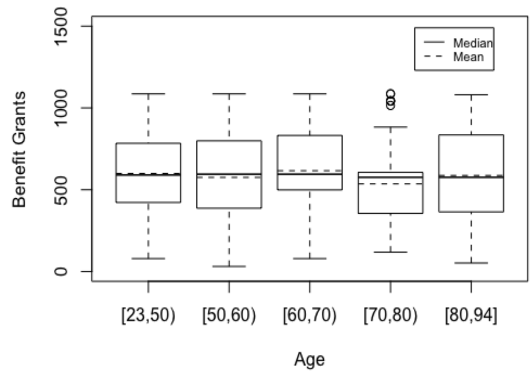

Figure 2.4: Boxplot illustrating the benefit grants in each age group.

Chapter 3

Methodology

3.1 Finite Mixture Models

Finite mixture models are considered as a powerful tool for clustering and classification in recent decades (McLachlan and Peel, 2000; McNicholas, 2016a). In direct applications of finite mixture models (see Titterington *et al.*, 1985), the objective is to model the data as a mixture probability distribution, with each mixture component corresponding to a cluster. However, some have argued that it is not always the case that a component corresponds to a cluster (see McNicholas, 2016a,b).

Let $\mathbf{X}$ be a random vector, which comes from a population with G subgroups. Then the density function of $\mathbf{X}$ can be written

$$f(\mathbf{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}|\boldsymbol{\theta}_g), \tag{3.1}$$

where $\pi_g > 0$ are called the mixing proportions, such that $\sum_{g=1}^{G} \pi_g = 1$, $f_g(\mathbf{x}|\boldsymbol{\theta}_g)$ is the gth component density, and $\boldsymbol{\vartheta} = (\pi_1, \pi_2, \ldots, \pi_G, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_G)$ denotes the vector of all parameters.

3.2 Cluster-Weighted Model (CWM)

Consider data of the form $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ so that each observation is a realization of the pair $(\mathbf{X}, Y)$ defined on some sample space Ω, where $\mathbf{X}$ is a vector of covariates and Y is a response variable. Assume that Ω can be partitioned into G groups, $\Omega_1, \Omega_2, \ldots, \Omega_G$. Extended from traditional finite mixture models, the CWM is a flexible family of mixture models for fitting the joint distribution $p(\mathbf{x}, y)$ of a random vector $(\mathbf{X}, Y)$ given by

$$p(\mathbf{x}, y) = \sum_{g=1}^{G} p(y|\mathbf{x}, \Omega_g) p(\mathbf{x}|\Omega_g) \pi_g, \tag{3.2}$$

where $\pi_g > 0$, $\sum_{g=1}^{G} \pi_g = 1$, $p(y|\mathbf{x}, \cdot)$ is the conditional distribution of Y given $\mathbf{X}$ and $p(\mathbf{x}|\cdot)$ is the marginal distribution of $\mathbf{X}$.

The original formulation of the CWM dates back to the linear Gaussian CWM (Gershenfeld, 1997), where $(\mathbf{X}, Y)$ is assumed to be real-valued, both $Y|\mathbf{X}$ and $\mathbf{X}$ follow a Gaussian distribution, and the relationship between Y and $\mathbf{X}$ is linear. Wedel (2002) called CWMs saturated mixture regression models. Ingrassia *et al.* (2012b) proposed that the model is considered as nested in the linear Gaussian CWM when both $p(y|\mathbf{x}, \Omega_g)$ and $p(\mathbf{x}|\Omega_g)$ are Gaussian. Moreover, Ingrassia *et al.* (2012a) proposed the Student-t CWM under the assumptions that both $Y|\mathbf{X}$ and $\mathbf{X}$ are t-distributed. Soon after, Ingrassia *et al.* (2014) defined a family of twelve CMWs, nested in the linear t-CWM, for model-based clustering. Subedi *et al.* (2013) developed a CWM analogue of the mixture of factor analyzers model and this, in turn, permits high-dimensional data. The polynomial Gaussian CWM was then proposed by Punzo (2014) to model a nonlinear distribution on $Y|\mathbf{X}$. Building on the work on Punzo and McNicholas (2016), Punzo and McNicholas (2017) consider a contaminated version of the CWM.

3.3 Generalized Linear Mixed CWMs

An extension is proposed by Ingrassia *et al.* (2015), where the conditional distribution $Y|\mathbf{X}$ is assumed to be the exponential family and the covariates $\mathbf{X}$ are assumed to be of mixed-type (continuous and finite discrete). In this case, a Gaussian distribution is used for continuous covariates in the model and the product of multinomial distributions is used for the finite discrete covariates.

Let $\mathbf{U}$ be a p-variate vector of continuous covariates and $\mathbf{V}$ be a q-variate vector of finite discrete covariates with levels $c_1, \ldots, c_q$. The vector of covariates $\mathbf{X} = (\mathbf{U}, \mathbf{V})$ is defined in d dimensions with $d = p + q$. The joint probability can be written as

$$p(\mathbf{x}, y; \boldsymbol{\vartheta}) = \sum_{g=1}^{G} q(y|\mathbf{x}; \boldsymbol{\xi}_g) p(\mathbf{u}; \boldsymbol{\psi}_g^*) p(\mathbf{v}; \boldsymbol{\psi}_g^{**}) \pi_g, \tag{3.3}$$

where $q(y|\mathbf{x}; \boldsymbol{\xi}_g)$ is the conditional distribution of Y given $\mathbf{X}$ with parameter $\boldsymbol{\xi}_g$, $p(\mathbf{u}; \boldsymbol{\psi}_g^*)$ is the marginal distribution of $\mathbf{U}$ with parameter $\boldsymbol{\psi}_g^*$, $p(\mathbf{v}; \boldsymbol{\psi}_g^{**})$ is the marginal distribution of $\mathbf{V}$ with parameter $\boldsymbol{\psi}_g^{**}$, π_g are as previously defined and $\boldsymbol{\vartheta}$ denotes the set of all parameters in the model.

The response variable Y is assumed to be generated from the exponential family in a generalized linear model, more related details can be found in McCullagh and Nelder

(1989); Wedel and DeSarbo (1995); Dobson and Barnett (2008). For each group Ω_g, the conditional distribution $q(y|\mathbf{x};\boldsymbol{\xi}_g)$ is modeled with the parameter vector $\boldsymbol{\beta}_g$ and an additional parameter λ_g, which is given by

$$q(y|\mathbf{x};\boldsymbol{\xi}_g) = q(y|\mathbf{x};\boldsymbol{\beta}_g,\lambda_g) = \exp\left\{\frac{y\eta(\mathbf{x};\boldsymbol{\beta}_g) - b[\eta(\mathbf{x};\boldsymbol{\beta}_g)]}{a(\lambda_g)} + c(y,\lambda_g)\right\}, \tag{3.4}$$

where $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are specified functions satisfying $\mathrm{E}(Y) = \mu = b'[\eta(\mathbf{x};\boldsymbol{\beta}_g)]$ and $\mathrm{Var}(Y) = b''[\eta(\mathbf{x};\boldsymbol{\beta}_g)]a(\lambda_g)$. In addition, $a(\lambda_g) > 0$, and $\eta(\mathbf{x};\boldsymbol{\beta}_g) = \eta_g = \boldsymbol{\beta}_g\mathbf{x}$ is the canonical function. There is a monotone and differentiable link function $g(\cdot)$ providing the relationship between a linear combination of unknown parameters $\boldsymbol{\beta}_g$ and the expected value μ_g, i.e. $\eta_g = g(\mu_g)$. For more details about modeling the conditional distributions of discrete responses, see Ingrassia *et al.* (2015). In this thesis, the continuous response variable Y is assumed to be generated from the Gaussian distribution.

Let $\mathbf{y}_1$ and $\mathbf{y}_2$ denote the vectors of uncensored and censored observations, respectively. Assuming $\tilde{\mathbf{y}}_2$ is the vector of unknown values which are censored to vector $\mathbf{y}_2$, $\mathbf{X}$ is considered to be partitioned into $\mathbf{X} = \binom{\mathbf{X}_1}{\mathbf{X}_2}$ corresponding to uncensored and censored observations. Under the assumption of the Gaussian distribution, and that $Y|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}_g, \sigma_g^2)$, then the conditional density $q(y|\mathbf{x};\boldsymbol{\beta}_g,\sigma_g)$ can be written as

$$(2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\left\{-\frac{[(\mathbf{y}_1 - \mathbf{X_1}\boldsymbol{\beta}_g)'(\mathbf{y}_1 - \mathbf{X_1}\boldsymbol{\beta}_g) + (\tilde{\mathbf{y}}_2 - \mathbf{X_2}\boldsymbol{\beta}_g)'(\tilde{\mathbf{y}}_2 - \mathbf{X_2}\boldsymbol{\beta}_g)]}{2\sigma^2}\right\}. \tag{3.5}$$

With respect to a vector of continuous covariates $\mathbf{U}$ following the Gaussian distribution, the density function $p(\mathbf{u};\boldsymbol{\psi}_g^*)$ is modeled with the mean $\boldsymbol{\mu}_g$ and covariance matrix $\mathbf{\Sigma}_g$, which is thus given by

$$p(\mathbf{u};\boldsymbol{\psi}_g^*) = \phi(\mathbf{u};\boldsymbol{\mu}_g,\mathbf{\Sigma}_g) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}_g|^{1/2}}\left\{\exp\Big(-\frac{1}{2}(\mathbf{u}-\boldsymbol{\mu}_g)^\top\mathbf{\Sigma}_g^{-1}(\mathbf{u}-\boldsymbol{\mu}_g)\Big)\right\}. \quad (3.6)$$

The density function $p(\mathbf{v};\boldsymbol{\psi}_g^{**})$ for a vector of q finite discrete covariates $\mathbf{V}$ is given by the product of q conditionally independent multinomial distributions of parameters $\boldsymbol{\alpha}_{gr}$, where $\boldsymbol{\alpha}_{gr} = (\alpha_{gr1},\ldots,\alpha_{grc_r})'$, $\alpha_{grs} > 0$ and $\sum_{s=1}^{c_r}\alpha_{grs} = 1$, $r = 1,\ldots,q$. For each observation, only one of the probabilities α_{grs} in $\boldsymbol{\alpha}_{gr}$ is valid. The density function $p(\mathbf{v};\boldsymbol{\psi}_g^{**})$ can then be expressed as $p(\mathbf{v};\boldsymbol{\alpha}_g)$, where $\boldsymbol{\alpha}_g = (\boldsymbol{\alpha}_{g1}',\ldots,\boldsymbol{\alpha}_{gq}')'$, see Ingrassia *et al.* (2015). The generalized linear mixed CWM can be written in the form

$$p(\mathbf{x},y;\boldsymbol{\vartheta}) = \sum_{g=1}^{G} q(y|\mathbf{x};\boldsymbol{\beta}_g,\lambda_g)\phi(\mathbf{u};\boldsymbol{\mu}_g,\mathbf{\Sigma}_g)p(\mathbf{v};\boldsymbol{\alpha}_g)\pi_g, \quad (3.7)$$

where $q(y|\mathbf{x};\boldsymbol{\beta}_g,\lambda_g)$, $\phi(\mathbf{u};\boldsymbol{\mu}_g,\mathbf{\Sigma}_g)$, $p(\mathbf{v};\boldsymbol{\alpha}_g)$ are density functions as defined above.

3.4 Likelihood and Parameter Estimation

The EM algorithm is used for parameter and incomplete data estimation. In this thesis, both the labels of observations and the censored data are considered as incomplete data to estimate. The EM algorithm iterates between two steps: an expectation step (E-step) and an maximization step (M-step). It is achieved as follows: consider n independent observations $(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)$. Then, the likelihood function can be written as

$$L(\boldsymbol{\vartheta}) = \prod_{i=1}^{n} p(\mathbf{x}_i,y_i,\boldsymbol{\vartheta}) = \prod_{i=1}^{n}\sum_{g=1}^{G} q(y_i|\mathbf{x}_i;\boldsymbol{\beta}_g,\lambda_g)\phi(\mathbf{u}_i;\boldsymbol{\mu}_g,\mathbf{\Sigma}_g)p(\mathbf{v}_i;\boldsymbol{\alpha}_g)\pi_g. \quad (3.8)$$

Define an indicator vector $\mathbf{z}_i = (z_{i1}, \ldots, z_{iG})'$ as the component membership of observation i, where

$$z_{ig} = \begin{cases} 1, & \text{if } (\mathbf{x}_i, y_i) \text{ comes from } \Omega_g, \\ 0, & \text{otherwise.} \end{cases} \tag{3.9}$$

In this case, the likelihood function L_c for complete-data $\{\mathbf{x}_i, y_i, \mathbf{z}_i; i = 1, \ldots, n\}$ can be written as

$$L_c(\boldsymbol{\vartheta}) = \prod_{i=1}^{n}\prod_{g=1}^{G}[q(y_i|\mathbf{x}_i;\boldsymbol{\beta}_g,\lambda_g)\phi(\mathbf{u}_i;\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)p(\mathbf{v}_i;\boldsymbol{\alpha}_g)\pi_g]^{z_{ig}}, \tag{3.10}$$

and the complete-data log-likelihood l_c is given by

$$\begin{aligned} l_c(\boldsymbol{\vartheta}) &= \prod_{i=1}^{n}\prod_{g=1}^{G}\ln[q(y_i|\mathbf{x}_i;\boldsymbol{\beta}_g,\lambda_g)\phi(\mathbf{u}_i;\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)p(\mathbf{v}_i;\boldsymbol{\alpha}_g)\pi_g]^{z_{ig}} \\ &= \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}[\ln q(y_i|\mathbf{x}_i;\boldsymbol{\beta}_g,\lambda_g) + \ln\phi(\mathbf{u}_i;\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g) + \ln p(\mathbf{v}_i;\boldsymbol{\alpha}_g) + \ln\pi_g]. \end{aligned} \tag{3.11}$$

The EM algorithm implemented is to estimate parameters and unknown observations z_{ig}. At the start of the EM algorithm, all parameters, the values of component membership and the censored data need to be initialized. The initialization of all parameters is denoted by $\boldsymbol{\vartheta}^{(0)}$. The EM algorithm iteratively alternates between E and M steps until parameter estimates are converged. In each $(k+1)$th iteration, the E-step calculates the expected value of $l_c(\boldsymbol{\vartheta})$ given the observed data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ and parameter estimates $\boldsymbol{\vartheta}^{(k)}$ from the previous kth iteration. The conditional expectation of missing data z_{ig} is defined by a new term τ_{ig}, thus, given by

$$\begin{aligned}\mathrm{E}_{\boldsymbol{\vartheta}^{(k)}}[z_{ig}|(\mathbf{x}_i, y_i)] &= \tau_{ig}^{(k)} \\ &:= \frac{q(y_i|\mathbf{x}_i;\boldsymbol{\beta}_g^{(k)},\lambda_g^{(k)})\phi(\mathbf{u}_i;\boldsymbol{\mu}_g^{(k)},\boldsymbol{\Sigma}_g^{(k)})p(\mathbf{v}_i;\boldsymbol{\alpha}_g^{(k)})\pi_g^{(k)}}{p(\mathbf{x}_i, y_i;\boldsymbol{\vartheta}^{(k)})}. \qquad (3.12)\end{aligned}$$

In the $(k+1)$th iteration of the M-step, it maximizes the conditional expectation of $l_c(\boldsymbol{\vartheta})$, i.e., $Q(\boldsymbol{\vartheta};\boldsymbol{\vartheta}^{(k)})$, with respect to $\boldsymbol{\vartheta}$, and obtains the new parameter estimates $\boldsymbol{\vartheta}^{(k+1)}$. Now, $Q(\boldsymbol{\vartheta};\boldsymbol{\vartheta}^{(k)})$ is written as

$$\begin{aligned}Q(\boldsymbol{\vartheta};\boldsymbol{\vartheta}^{(k)}) &= \sum_{i=1}^{n}\sum_{g=1}^{G}\tau_{ig}^{(k)}\ln\pi_g + \sum_{i=1}^{n}\sum_{g=1}^{G}\tau_{ig}^{(k)}\ln q(y_i|\mathbf{x}_i;\boldsymbol{\beta}_g,\lambda_g) \\ &+ \sum_{i=1}^{n}\sum_{g=1}^{G}\tau_{ig}^{(k)}\ln\phi(\mathbf{u}_i;\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g) + \sum_{i=1}^{n}\sum_{g=1}^{G}\tau_{ig}^{(k)}\ln p(\mathbf{v}_i;\boldsymbol{\alpha}_g). \qquad (3.13)\end{aligned}$$

The updates for π_g, $\boldsymbol{\mu}_g$, $\boldsymbol{\Sigma}_g$ and $\boldsymbol{\alpha}_{gr}$ can be computed by

$$\pi_g^{(k+1)} = \frac{\sum_{i=1}^{n}\tau_{ig}^{(k)}}{n}, \qquad (3.14)$$

$$\boldsymbol{\mu}_g^{(k+1)} = \frac{\sum_{i=1}^{n}\tau_{ig}^{(k)}\mathbf{u}_i}{\sum_{i=1}^{n}\tau_{ig}^{(k)}}, \qquad (3.15)$$

$$\boldsymbol{\Sigma}_g^{(k+1)} = \frac{\sum_{i=1}^{n}\tau_{ig}^{(k)}(\mathbf{u}_i-\boldsymbol{\mu}_g)^T(\mathbf{u}_i-\boldsymbol{\mu}_g)}{\sum_{i=1}^{n}\tau_{ig}^{(k)}}, \qquad (3.16)$$

$$\boldsymbol{\alpha}_{gr}^{(k+1)} = \frac{\sum_{i=1}^{n}\tau_{ig}^{(k)}v_i^{rs}}{\sum_{i=1}^{n}\tau_{ig}^{(k)}}. \qquad (3.17)$$

The updates for $\boldsymbol{\beta}_g$ and λ_g in (3.13) are computed in the following. Under the conditional Gaussian distribution in (3.5), the conditional density function for each observation $q(y_i|\mathbf{x}_i;\boldsymbol{\beta}_g,\lambda_g)$ is given by

$$q(y_i|\mathbf{x}_i;\boldsymbol{\beta}_g,\sigma_g)=(2\pi\sigma_g^2)^{-1/2}\exp\left\{-\frac{[(y_i-\mathbf{x}_i\boldsymbol{\beta}_g)^2]}{2\sigma_g^2}\right\}, \tag{3.18}$$

and the complete-data log-likelihood of the conditional density function is

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{g=1}^{G}\tau_{ig}^{(k)}\ln q(y_i|\mathbf{x}_i;\boldsymbol{\beta}_g,\lambda_g)\\&=\sum_{i=1}^{n}\sum_{g=1}^{G}\tau_{ig}^{(k)}\left[-\frac{1}{2}\ln(2\pi\sigma_g^2)-\frac{(y_i-\mathbf{x}_i\boldsymbol{\beta}_g)^2}{2\sigma_g^2}\right]\\&=\sum_{i=1}^{n}\tau_{ig}^{(k)}\left[-\frac{1}{2}\ln(2\pi\sigma_g^2)\right]-\frac{\sum_{i=1}^{n}\tau_{ig}^{(k)}(y_i-\mathbf{x}_i\boldsymbol{\beta}_g)^2}{2\sigma_g^2}\\&=\sum_{i=1}^{n}\tau_{ig}^{(k)}\left[-\frac{1}{2}\ln(2\pi\sigma_g^2)\right]-\frac{\sum_{i=1}^{n}\tau_{ig}^{(k)}(y_i^2-2\boldsymbol{\beta}'_g\mathbf{x}'_iy_i+\boldsymbol{\beta}'_g\mathbf{x}'_i\mathbf{x}_i\boldsymbol{\beta}_g)}{2\sigma_g^2}.\end{aligned} \tag{3.19}$$

Therefore, the conditional expectation of the complete-data log-likelihood function defined as the part of Q-function in (3.13), is given by

$$\sum_{i=1}^{n}\sum_{g=1}^{G}\tau_{ig}^{(k)}\left[-\frac{1}{2}\ln(2\pi\sigma_g^2)\right]-\frac{\sum_{i=1}^{n}\tau_{ig}^{(k)}(\tilde{y}_i^2-2\boldsymbol{\beta}'_g\mathbf{x}'_i\tilde{y}_i+\boldsymbol{\beta}'_g\mathbf{x}'_i\mathbf{x}_i\boldsymbol{\beta}_g)}{2\sigma_g^2}, \tag{3.20}$$

where $\tilde{y}_i$ and $\tilde{y}_i^2$ be the vectors of the expected values for the observations y_i and y_i^2, which are defined as

$$\tilde{y}_i=\begin{cases}y_i, & \text{if } y_i \text{ are uncensored } (y_i\in\mathbf{y}_1)\\ A, & \text{if } y_i \text{ are censored } (y_i\in\mathbf{y}_2)\end{cases}, \tag{3.21}$$

and

$$\tilde{y}_i^2 = \begin{cases} y_i^2, & \text{if } y_i \text{ are uncensored } (y_i \in \mathbf{y}_1) \\ B, & \text{if } y_i \text{ are censored } (y_i \in \mathbf{y}_2) \end{cases}. \tag{3.22}$$

Here, A and B respectively indicate the first and the second moments of the censored observations $\mathbf{y}_2$, conditional on the non-censored observations and current parameter estimates. Note that A and B are expressed as the weighted sums of A_g and B_g, respectively. Given the observed data and the current estimates of the parameters, the E-step is to calculate A and B written as

$$A = \mathrm{E}(\mathbf{Y}_2|\mathbf{y}_1, \boldsymbol{\tau}_g, \boldsymbol{\beta}_g, \sigma_g) = \sum_{g=1}^{G} \boldsymbol{\tau}_g A_g, \tag{3.23}$$

$$B = \mathrm{E}(\mathbf{Y}_2'\mathbf{Y}_2|\mathbf{y}_1, \boldsymbol{\tau}_g, \boldsymbol{\beta}_g, \sigma_g) = \sum_{g=1}^{G} \boldsymbol{\tau}_g B_g, \tag{3.24}$$

where $\boldsymbol{\tau}_g = (\tau_{1g}, \tau_{2g}, \ldots, \tau_{ng})$ are defined as the weighted parameters, such that $\sum_{g=1}^{G} \boldsymbol{\tau}_g = 1$, each gth component A_g and B_g are given by

$$A_g = \mathbf{X}_2\boldsymbol{\beta}_g + \sigma_g f\Big(\frac{\mathbf{y}_2 - \mathbf{X}_2\boldsymbol{\beta}_g}{\sigma_g}\Big), \tag{3.25}$$

$$B_g = \|\mathbf{X_2}\boldsymbol{\beta}\|^2 + \sigma_g(\mathbf{X_2}\boldsymbol{\beta}_g + \mathbf{z})' f\Big(\frac{\mathbf{y}_2 - \mathbf{X_2}\boldsymbol{\beta}_g}{\sigma_g}\Big) + n_2\sigma_g^2, \tag{3.26}$$

where $f(x) = \varphi(x)/\Phi(x)$,

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad \Phi(x) = \int_x^\infty \varphi(t)dt,$$

and $\|\cdot\|$ is the Euclidean norm.

To maximize (3.20) in the M-step of (3.13) given the current values A and B with respect to $\boldsymbol{\beta}_g$ and σ_g, and it is equivalent to independently maximize each of the G expressions. The updates for $\boldsymbol{\beta}_g$ and σ_g^2 are given by

$$\boldsymbol{\beta}_g^{(k+1)} = \frac{\sum_{i=1}^n \tau_{ig}^{(k)}(\mathbf{x}_i'\tilde{y}_i)}{\sum_{i=1}^n \tau_{ig}^{(k)}(\mathbf{x}_i'\mathbf{x}_i)}, \tag{3.27}$$

$$(\sigma_g^2)^{(k+1)} = \frac{\sum_{i=1}^n \tau_{ig}^{(k)}(\tilde{y}_i^2 - 2\boldsymbol{\beta}_g'\mathbf{x}_i'\tilde{y}_i + \boldsymbol{\beta}_g'\mathbf{x}_i'\mathbf{x}_i\boldsymbol{\beta}_g)}{\sum_{i=1}^n \tau_{ig}^{(k)}}. \tag{3.28}$$

3.4.1 EM Initialization

Choosing the starting values for the EM algorithm is an important procedure because it could heavily affect the convergence of the EM algorithm. The earlier work done for setting initial values can be found in Biernacki *et al.* (2003); Karlis and Xekalaki (2003). In this thesis, a standard approach used is to generate a random initialization of $\tau_i^{(0)} = (\tau_{i1}^{(0)}, \ldots, \tau_{iG}^{(0)})', i = 1, \ldots, n$, and values $\tau_{i1}^{(0)}, \ldots, \tau_{iG}^{(0)}$ generated are summing to one. This approach is called random soft initialization. Finally, the vector $\hat{\tau}_i^{(0)}$ which gives the highest observed-data log-likelihood among 10 repeated trials will be selected.

3.4.2 EM Convergence Criterion

The Aitken acceleration (Aitken, 1926) at the kth iteration is given by

$$a^{(k)} = \frac{l^{(k+1)} - l^{(k)}}{l^{(k)} - l^{(k-1)}}, \tag{3.29}$$

where $l^{(k)}$ is the log-likelihood value from the kth iteration. The Aitken acceleration used is to get the asymptotic estimate of the log-likelihood in the $(k+1)$th iteration given by Böhning *et al.* (1994) as

$$l_{\infty}^{(k+1)} = l^{(k)} + \frac{1}{1-a^{(k)}}(l^{(k+1)} - l^{(k)}). \tag{3.30}$$

Based on this, Lindsay (1995) proposed that the convergence of the EM algorithm is considered to be reached when the difference between the log-likelihood and its estimated asymptotic value is sufficiently small, which is written by

$$l_{\infty}^{(k+1)} - l^{(k+1)} < \epsilon, \tag{3.31}$$

where ϵ is small positive real number. McNicholas *et al.* (2010) suggested that the algorithm can be stopped when

$$l_{\infty}^{(k+1)} - l^{(k)} < \epsilon, \tag{3.32}$$

provided the difference is positive.

3.5 Model Selection

The criterion used for model selection is the Bayesian information criterion (BIC) (Schwarz *et al.*, 1978), which does not underestimate the number of components in a mixture model (Leroux *et al.*, 1992), and also consistently estimates the number of mixture components under certain regularity conditions (Keribin, 2000).

The form of the BIC is given by

$$\mathrm{BIC} = 2l_c(\hat{\boldsymbol{\vartheta}}) - \rho \log n, \tag{3.33}$$

where $l_c(\hat{\boldsymbol{\vartheta}})$ is the maximized log-likelihood, $\hat{\boldsymbol{\vartheta}}$ is the maximum likelihood estimate of $\boldsymbol{\vartheta}$, ρ is the number of free parameters needed to be estimated, and n denotes the number of observations. The model with the highest BIC is chosen as the best model to fit the data.

Chapter 4

Analysis of Veterans' Benefits Data

4.1 Generalized Linear Mixed CWMs

The generalized linear mixed CWM is implemented in the R package `flexCWM` (Mazza *et al.*, 2017) for grouping unlabelled observations into clusters. To deal with the incomplete (right-censored) data in the veterans' benefits dataset, the package `flexCWM` has been modified and the expected values of benefit grants are computed by using the EM algorithm in Section 3.4 . With respect to the continuous response variable, its conditional distribution is assumed to be Gaussian. Mixture components from $G = 1, \ldots, 6$ have been tried. To avoid convergence to a local maximum, the same procedure for model selection was repeated 10 times. The BIC is used to select the best model.

The best model selected consists of three components. The log-likelihood of the model is computed to be -11988 and the value of BIC computed is -24630. One component is considered to be identified as one cluster. Table 4.1 gives the details of the difference between clusters.

Table 4.1: The basic information in the discrete variables CencID, Gender, Spouse and Widow for each cluster.

	cluster 1		cluster 2		cluster 3	
Uncensored (0)	131	(51.0%)	116	(57.4%)	64	(55.2%)
Censored (1)	126	(49.0%)	86	(42.6%)	52	(44.8%)
Male (0)	257	(100%)	187	(92.6%)	2	(1.7%)
Female (1)	0	(0%)	15	(7.4%)	114	(98.3%)
Primary beneficiary(veteran) (0)	257	(100%)	201	(99.5%)	73	(62.9%)
Spouse (1)	0	(0%)	1	(0.5%)	43	(37.1%)
Not widowed (0)	255	(99.2%)	202	(100%)	68	(58.6%)
Widowed (1)	2	(0.8%)	0	(0%)	48	(41.4%)
Total	257	(44.7%)	202	(35.1%)	116	(20.2%)

It is observed that there are three clusters generated and each cluster includes 257, 202, 116 observations, respectively. Over 40% of each cluster is made up of censored observations, with cluster 1 containing the most. The clusters 1 and 3 can be respectively considered as the male and female group as cluster 1 are all males and cluster 3 consists of 98.3% females. Cluster 2 is also predominantly male (92.6%). Nearly all applicants grouped in clusters 1 and 2 are not widowed but are veterans. In addition, it is also observed that applicants who identify as spouses of living veterans or widowed are almost females in cluster 3.

Figure 4.1 shows further study on discovering the difference between clusters with respect to age, monthly income and benefit grants. The solid line and the dashed line in each boxplot represent the median and the mean, respectively.

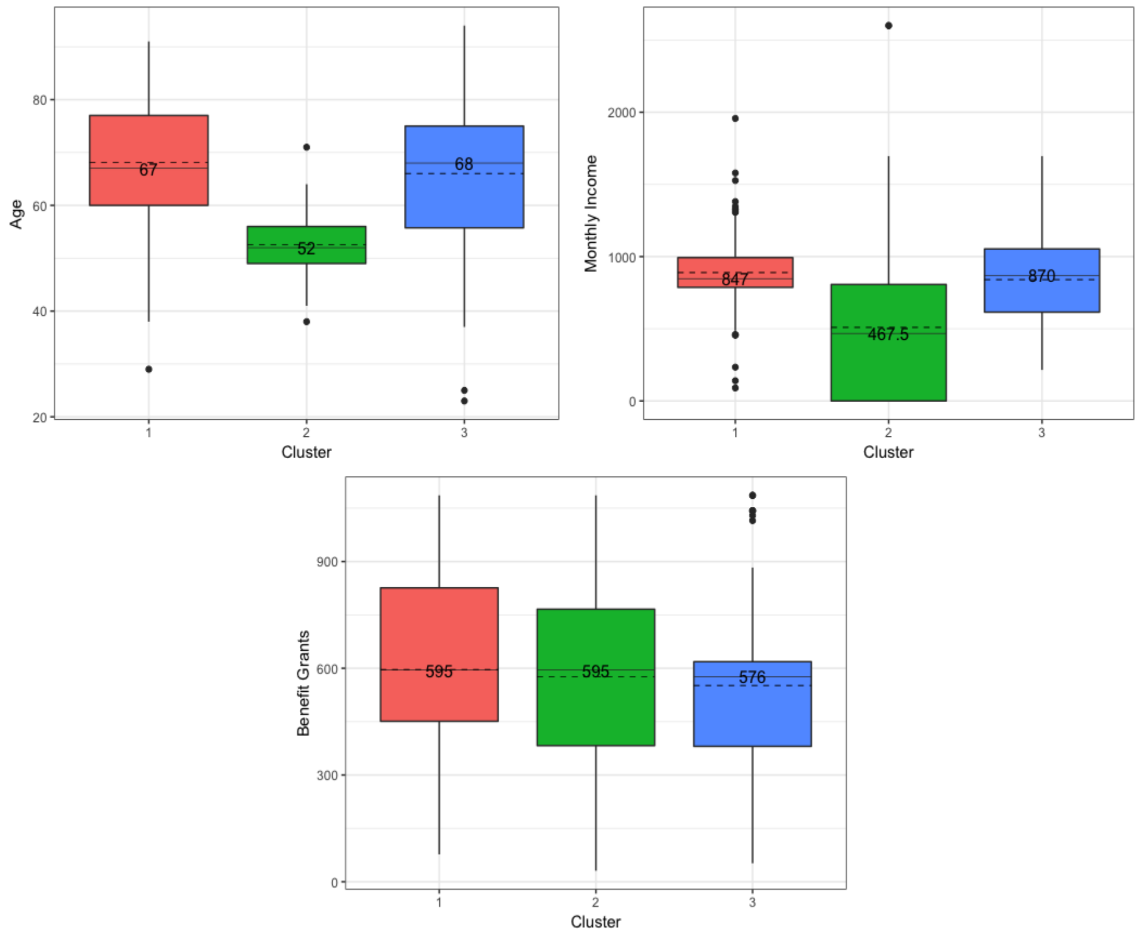

Figure 4.1: Boxplots illustrating the difference between clusters in age, monthly income and benefit grants, respectively.

It is observed that the female group (cluster 3) has the highest monthly income but the lowest benefit grants sought, which are observed to have a small age difference with the male group indicated as cluster 1. The average monthly income of the male group is slightly higher than that in the female group, and additionally, the male group is asking for the highest benefit grants among three clusters. In cluster 2, which is almost entirely male, applications came from the youngest applicants and 27.2% applications reported zero monthly income, which attributes the monthly income to be the lowest among three clusters. In addition, applicants in cluster 2 with the youngest age are likely to ask for high benefit grants.

The two plots of Figure 4.2 show that males grouped in cluster 1 have relatively high monthly income and benefit grants sought compared to those in cluster 2, while females grouped in cluster 3 also have relatively high monthly income but lower benefit grants than those in cluster 2.

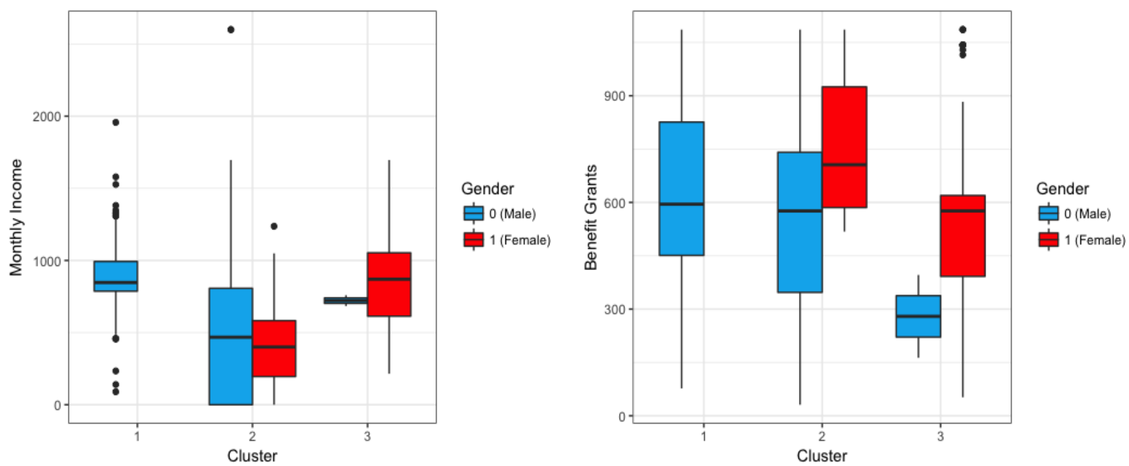

Figure 4.2: Boxplots illustrating the difference between clusters in monthly income (left) and benefit grants (right) with respect to the variable Gender.

In Figure 4.3, it is observed that applicants who are spouses of living veterans or who are not windowed have higher monthly income as well as benefit grants sought.

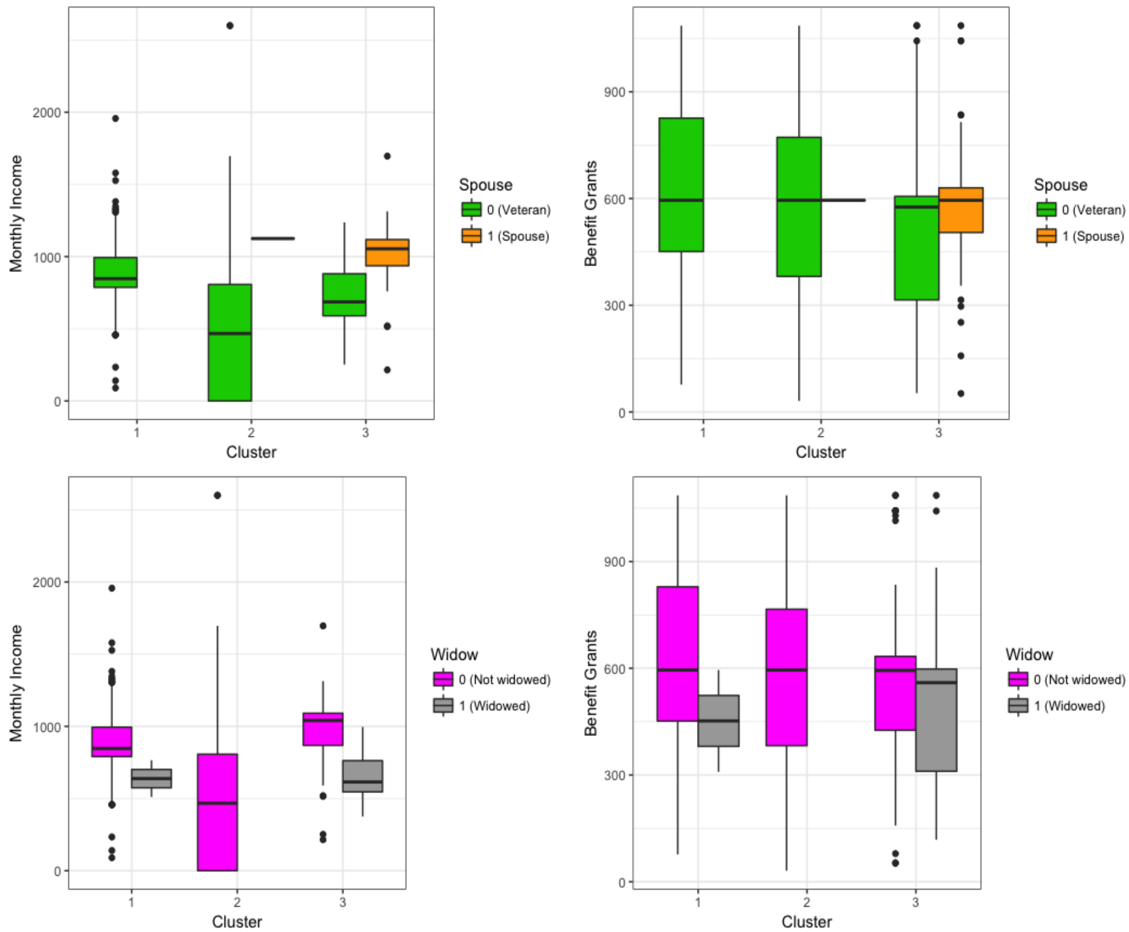

Figure 4.3: Boxplots illustrating the difference between clusters in monthly income (left) and benefit grants (right) with respect to the variables Spouse and Widow.

4.2 Model Evaluation

The expected benefit grants sought under the model are computed based on (3.25) and (3.23) in Section 3.4. Figure 4.4 shows there is a big difference between the true and the expected benefit grants sought and all censored observations are expected to have higher benefit grants than before. The solid line and the dashed line in each boxplot respectively represent the median and the mean of the benefit grants.

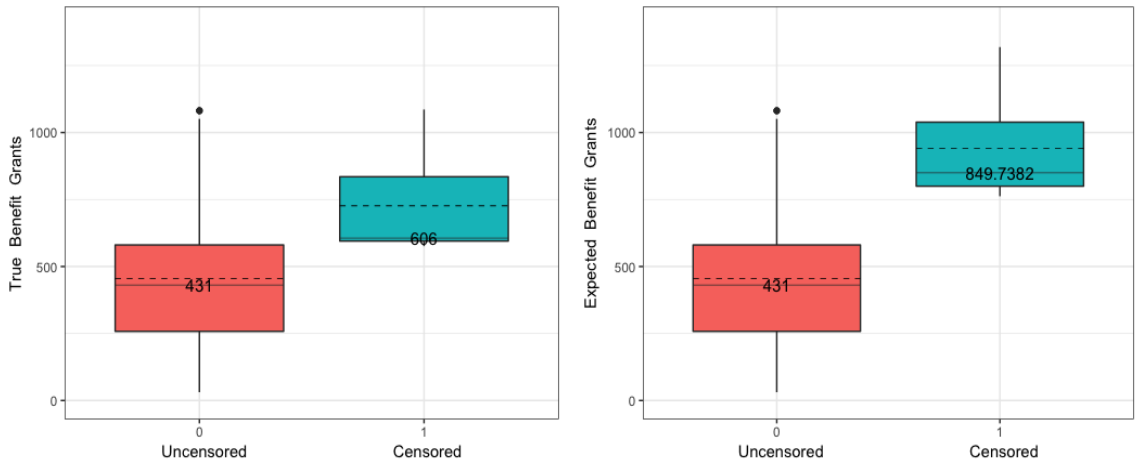

Figure 4.4: Boxplots illustrating the comparisons between true (left) and expected (right) benefit grants for the uncensored and censored groups.

Figure 4.5 shows that the difference between the true and the expected benefit grants with respect to each cluster. It is observed that the female group cluster 3 has the largest increase in benefit grants sought.

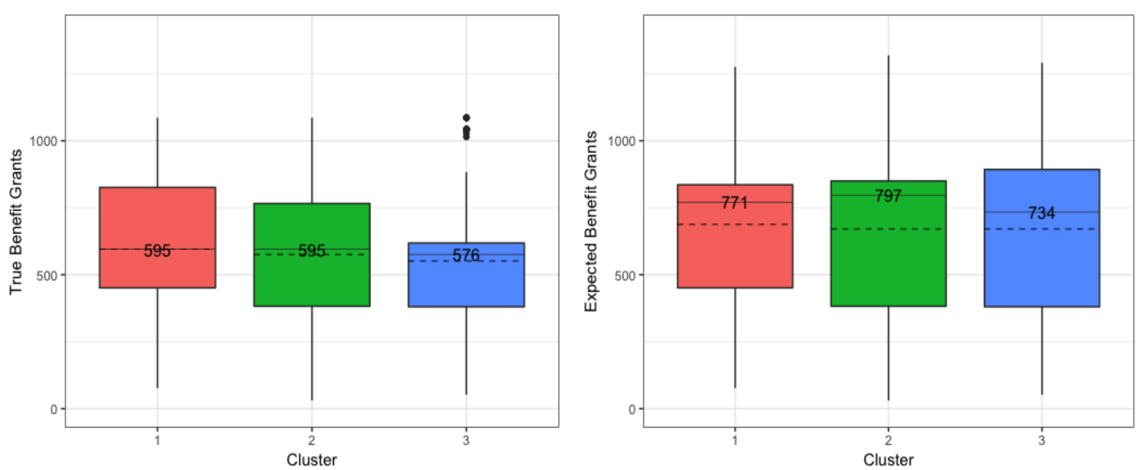

Figure 4.5: Boxplots illustrating the comparisons between true (left) and expected (right) benefit grants in each cluster.

Furthermore, the benefit grants sought for the female group cluster 3 are expected to have the highest average as well as the median by only considering the censored observations, which is followed by cluster 2 which includes a majority of male applicants and the male group cluster 1 (Figure 4.6).

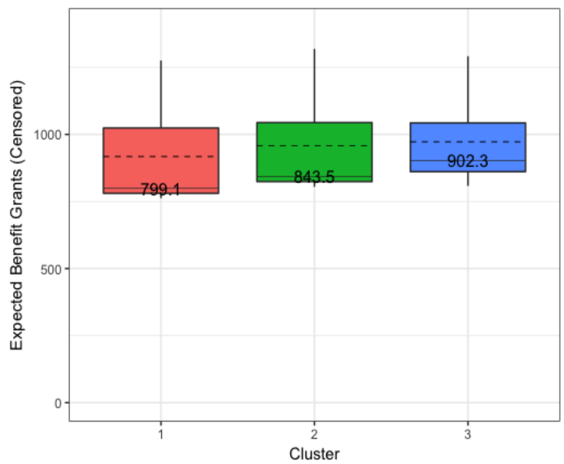

Figure 4.6: Boxplot illustrating the comparisons between expected benefit grants for censored observations in each cluster.

In addition, the two plots of Figure 4.7 show that the benefit grants sought of females are expected to increase more than males, which leads to a high benefit grants increase for the female group cluster 3.

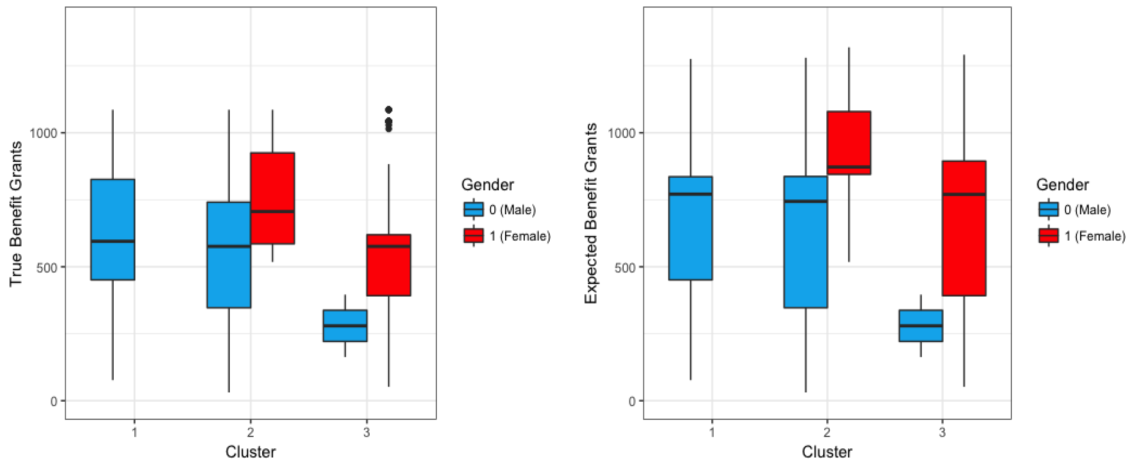

Figure 4.7: Boxplots illustrating the comparisons between true (left) and expected (right) benefit grants with respect to the variable Gender in each cluster.

Furthermore, in Figure 4.8, it is observed that the applicants who are spouses of living veterans are expected to be higher than veterans and widowed female applicants have relatively low benefit grants than those are not widowed.

Compared to the total amount of benefit grants observed, \$333,472, the total amount of benefit grants is expected to be \$390,172.2. Therefore, the government may need to provide more funds for benefit programs. The increase of benefit grants (\$56,799.2) are expected to be attributed \$23,596.4 to cluster 1, \$19,238.1 to cluster 2 and \$13,865.7 to cluster 3.

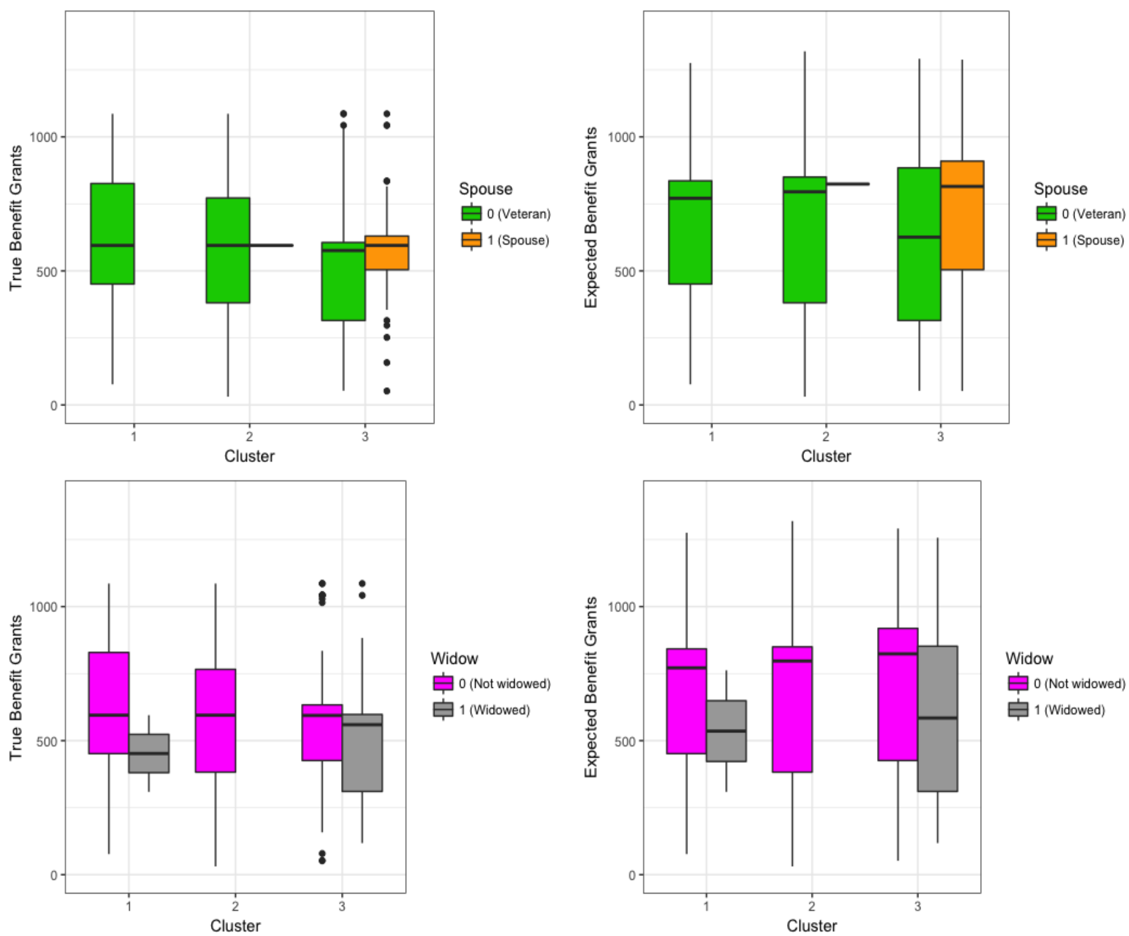

Figure 4.8: Boxplots illustrating the comparisons between true (left) and expected (right) benefit grants with respect to the variable Spouse and Widow in each cluster.

Chapter 5

Conclusions and Future Work

The veterans' benefits data are considered to be classified into three groups by using the generalized linear mixed CWM under the Gaussian distribution with censored data. One group includes male veterans who are almost all not widowed, one group consists of non-widowed veterans who are mostly male. Another group is made up of almost all female applicants and contains nearly all of the applicants who are spouses of living veterans or widows. It is observed that this last (i.e., female) group has the highest monthly income but the lowest benefit grants sought. The male group with relatively high monthly income has the highest benefit grants sought. The group including both female and male applicants (i.e., mostly male) has the lowest monthly income, as applicants recorded as zero monthly income are included therein.

In general, the benefit grants sought, as estimated through the EM algorithm, are expected to be larger than the observed values for the right-censored observations. The increase in benefit grants sought for female applicants is greater than males, as it is observed that the increase in benefit grants sought for cluster 3 (i.e., female group) is the largest. In addition, the benefit grants sought for applicants who are spouses of living veterans are still expected to be higher than veterans, and those for applicants who are not widowed are also expected to be higher than those widowed.

Besides, the application year also needs to be taken as an important consideration if the benefit grants for applicants are expected to be increased or not, as the money available may vary from year to year. In addition to the increasing benefit grants sought for applicants with incomplete (right-censored) data, those with uncensored benefit grants sought may also be considered to need more benefit grants, because it is possible that applicants asking for low benefit grants actually require more. Further analysis of these applicants — and all categories of applicants, in fact — is necessary to gain a full understanding of the veterans' benefits landscape.

In future work, besides the traditional linear CWM approach based on the Gaussian assumption, some linear CWMs based on Student's t or Gamma distributions will be taken into consideration. Further, a polynomial Gaussian CWM could also be tried. To deal with outliers, the use of contaminated CWM can be explored.

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.

Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics* & *Data Analysis*, **41**(3), 561–575.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B. G. (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.

Chauveau, D. (1995). A stochastic EM algorithm for mixtures with censored data. *Journal of Statistical Planning and Inference*, **46**(1), 1–25.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, pages 1–38.

Dobson, A. J. and Barnett, A. G. (2008). *An Introduction to Generalized Linear Models*. CRC press, Boca Raton.

Essenburg, T. J. and Hanson, L. K. (2014). *The New Faces of American Poverty: A Reference Guide to the Great Recession*, volume 1. ABC-CLIO, Santa Barbara.

Gershenfeld, N. (1997). Nonlinear inference and cluster-weighted modeling. *Annals of the New York Academy of Sciences*, **808**(1), 18–24.

Guiahi, F. (2001). Fitting loss distributions in the presence of rating variables. *Journal of Actuarial Practice*, **9**(1).

Ingrassia, S., Minotti, S. C., and Incarbone, G. (2012a). An EM algorithm for the Student-t cluster-weighted modeling. In *Challenges at the Interface of Data Analysis, Computer Science, and Optimization*, pages 13–21. Springer.

Ingrassia, S., Minotti, S. C., and Vittadini, G. (2012b). Local statistical modeling via a cluster-weighted approach with elliptical distributions. *Journal of Classification*, **29**(3), 363–401.

Ingrassia, S., Minotti, S. C., and Punzo, A. (2014). Model-based clustering via linear cluster-weighted models. *Computational Statistics & Data Analysis*, **71**, 159–182.

Ingrassia, S., Punzo, A., Vittadini, G., and Minotti, S. C. (2015). The generalized linear mixed cluster-weighted model. *Journal of Classification*, **32**(1), 85–113.

Karlis, D. and Xekalaki, E. (2003). Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics & Data Analysis*, **41**(3), 577–590.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 49–66.

Koul, H. L., Susarla, V., and Van Ryzin, J. (1981). Regression analysis with randomly right-censored data. *The Annals of Statistics*, pages 1276–1288.

Le, C. T. (1997). *Applied Survival Analysis*. Wiley, New York.

Leroux, B. G. *et al.* (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, **20**(3), 1350–1360.

Lindsay, B. G. (1995). Mixture models: Theory, Geometry and Applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, pages i–163. JSTOR.

Mazza, A., Punzo, A., and Ingrassia, S. (2017). Package 'flexCWM': Flexible Cluster-Weighted Modeling. *Version 1.7*.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons, New York, USA.

McNicholas, P. D. (2016a). *Mixture Model-Based Classification*. CRC Press, Boca Raton.

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

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

Miljkovic, T. and Barabanov, N. (2015). Modeling veterans' health benefit grants using the expectation maximization algorithm. *Journal of Applied Statistics*, **42**(6), 1166–1182.

Miller, R. G. (1976). Least squares regression with censored data. *Biometrika*, **63**(3), 449–464.

Miller Jr, R. G. (2011). *Survival Analysis*. John Wiley & Sons, New York.

Park, C. and Lee, S. B. (2012). Parameter estimation from censored samples using the expectation-maximization algorithm. *arXiv preprint arXiv:1203.3880*.

Punzo, A. (2014). Flexible mixture modelling with the polynomial Gaussian cluster-weighted model. *Statistical Modelling*, **14**(3), 257–291.

Punzo, A. and McNicholas, P. D. (2016). Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal*, **58**(6), 1506–1537.

Punzo, A. and McNicholas, P. D. (2017). Robust clustering in regression analysis via the contaminated Gaussian cluster-weighted model. *Journal of Classification*, **34**(2), 249–293.

Qin, J., Ning, J., Liu, H., and Shen, Y. (2011). Maximum likelihood estimations and EM algorithms with length-biased data. *Journal of the American Statistical Association*, **106**(496), 1434–1449.

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

Subedi, S., Punzo, A., Ingrassia, S., and McNicholas, P. D. (2013). Clustering and classification via cluster-weighted factor analyzers. *Advances in Data Analysis and Classification*, **7**(1), 5–40.

Titterington, D. M., Smith, A. F., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

U.S. Department of Veterans Affairs (2010). National Center for Veterans Analysis and Statistics. Available at https://www.va.gov/vetdata/veteran_population.asp.

U.S. Department of Veterans Affairs (2015). Veteran Poverty Trends. Available at https://www.va.gov/vetdata/docs/specialreports/veteran_poverty_trends.pdf.

Wedel, M. (2002). Concomitant variables in finite mixture models. *Statistica Neerlandica*, **56**(3), 362–375.

Wedel, M. and DeSarbo, W. S. (1995). A mixture likelihood approach for generalized linear models. *Journal of Classification*, **12**(1), 21–55.