University of Nevada, Reno

Statistical Analysis and Modeling of Claim Duration for Workers' Compensation Insurance

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Mathematics

by

Ryan M. Liebert

Dr. Anna Panorska / Thesis Advisor

December 2015

THE GRADUATE SCHOOL

We recommend that the thesis
prepared under our supervision by

RYAN M. LIEBERT

Entitled

Statistical Analysis And Modeling Of Claim Duration For Workers' Compensation Insurance

be accepted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

Anna Panorska, Ph.D, Advisor

Tomasz Kozubowski, Ph.D, Committee Member

Frederick Harris, Ph.D, Graduate School Representative

David W. Zeh, Ph.D., Dean, Graduate School

December, 2015

Abstract

Workers' compensation is a form of insurance that protects employees and business owners from the cost of injuries occurring in the workplace. The duration of time a workers' compensation claim remains open largely depends on the type and severity of the injury. This work focuses on statistical analysis and modeling of claim duration for a workers' compensation insurer. A data set for claim duration that included over two million claims spanning from 1915 to 1994 was analyzed. Exploratory data analysis revealed that the distribution of the data was multi-modal with a gap at approximately 55 years and a positive skew. Log linear analysis and modeling was used to understand the association between categorical variables. The EM algorithm was used to fit gamma and normal mixture models to the claim duration data. Log likelihood and AIC values were used to show that a normal mixture provided the best fit for the data. The likelihood ratio test was used to select the number of components in the mixture model. This test indicated the four component normal mixture model was the best model. The Komogorov-Smirnov goodness-of-fit test indicated that the selected model was identical to the population distribution that generated the data. Finally, standard errors of the parameter estimates were reported indicating that little uncertainty existed in the estimates.

Acknowledgements

I would like to thank the following:

- Dr. Anna Panorska for giving me the opportunity to work on this data set as well as her limitless patience and support throughout the project;

- Dr. Tomasz Kozubowski for his support throughout my course work and thesis. It was through taking his classes that I first came to appreciate the beauty of probability and statistics;

- Dr. Frederick Harris for being willing to be on my committee;

- Dr. Ilya Zaliapin for his suggestions regarding this work.

- The insurance company who supplied the data, and in particular a certain statistician who helped reinforce my appreciation and enjoyment of statistics;

Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Setting

Workers' compensation is a form of insurance that protects employees and business owners from the cost of injuries occurring in the workplace. For an employee this may include medical bills required to treat an injury, lost wages, and rehabilitation cost. For an employer, also called the insured, workers' compensation insurance protects from being liable for the injuries of the employee. A request for a workers' compensation insurer to indemnify the injuries of a worker is called a claim.

In the United States employers are required to purchase workers' compensation insurance in every state except Texas. Rules and requirements for employer coverage vary among states. The cost to the employer for workers' compensation coverage is called the premium. The premium charged to an employer is typically a function of the risk associated with the line of business, the amount of payroll being covered, and the safety record of a respective business.

Examples of workers' compensation claims range from being simple to complex. A simple example may be a secretary opening a package delivered to her office by the postal service. Suppose she is using a box cutter to open the package, her hand slips and she cuts herself. The cut requires medical attention so the secretary visits the doctor and has her cut treated. After treatment, she returns to work the same day. The employer files a claim and the insurer pays the medical bills only. This claim is referred to as a "medical only claim".

In another example consider an analyst. The analyst's work requires a large volume of typing each day. After years of work he begins feeling numbness and weakness

in his hand. He visits his doctor and is diagnosed with carpal tunnel syndrome. The doctor explains that his injury is related to his work as a analyst. The analyst is also instructed by his doctor that he must permanently reduce his work load by 10 hours a week. The workers' compensation insurer is liable for this injury, which is referred to as a "cumulative trauma injury". The workers compensation insurer is liable for the medical bills of the analyst as well the lost wages associated with the analyst's injury. The analyst's condition is referred to as "permanent partial disability."

The first example shows that a workers' compensation claim can be relatively minor and open for a short period of time. The second example shows that a workers' compensation claim can be more serious, requiring treatment and payment for long periods of time. Of course, even more serious claims are possible such as those that are classified as "permanent total disability" and "fatalities." The diversity of potential claims makes the risk associated with workers' compensation insurance relatively complicated to estimate. Since the insurer continues making payments to the claimant for the duration of time the claim is open, serious claims that are open for a long time are of particular interest to an insurer. This is because they may total large dollar amounts. If it is known what claims are most likely to be open for long periods of time then more experienced claims adjusters can be assigned those claims and implement best practices to reduce long-term cost. Knowing what claims are most likely to be open for long periods of time also helps planning the reserves necessary in the future for the company to remain solvent. Thus, explaining variability of the time a claim is open is an important and practical problem for any workers' compensation insurance company.

1.2 The Data and Research Questions

The data used in this work was derived from a data set provided by a workers' compensation insurance company. The data includes a total of 2,238,444 rows of information on different claims and 55 columns of variables that characterize those claims. The oldest claim in the data set had a date of injury of 1/1/1915. The youngest claim in the data set had a date of injury of 9/9/1994. Most of the variables were not relevant to the direct computation of the duration of time claims remained open. However, some variables provided insight into characteristics of claims open for long periods of time. Those variables were used to attempt to explain the variability in claim duration.

The duration of time claims remain open was computed by taking the difference between the date a claim was closed and the date of injury. The date of injury is defined as the date the insured reported that the injury took place. The date the claim was closed reflects the date when the claim was determined to require no further payment. In some states, such as Nevada, claims may be reopened if it is discovered that an injury requires additional treatment. If claims were reopened, the close date was updated accordingly. The date of injury was used as a proxy for the date the claim was open. It is a proxy because there can be a lag between the date of injury and the date the claim is reported to the insurer. The duration of time claims remain open was computed in R using the `strptime()` function to format the dates and the `difftime()` function to compute the time differences between dates. The key variables from the data set that were used in this analysis are

- Date of Injury,
- Close Date,
- Extent of Injury, and
- Claim Type.

Table 1.1 provides a glossary for the values that extent of injury and claim type take. Other variables were explored, but did not provide substantial insight into the variability present in claim duration. This may be primarily because entries for several variables were not consistently reported.

Table 1.1: A glossary for the values that extent of injury and claim type variables take.

Variable	Possible Values	Definition
Extent of Injury	1	Fatality
	2	Permanent total disability
	5	Permanent partial disability
	6	Temporary injury
	9	Medical only
Claim Type	NR	Non-reserve (Outstanding reserves unavailable)
	R	Reserved (Outstanding reserves available)
	RO	No financial data
	V	Void / null (Duplicate claims)

Minimal clean-up of the data was necessary. Of the 2,238,444 computed claim duration values 4,269 entries had NA (not available) values and 103 had values less than zero. The NAs were claims that had not yet closed. The 103 entries that were less than zero were taken to be the result of input error. NA and less than zero entries represented 0.2% of the total observations and were removed from the data set. After clean-up, the final data set included 2,234,072 observations of the duration of time the claims remained open. The original units of time was days, but this was converted to years by dividing values by 365.25 days. The additional quarter day is used to account for leap years.

The general questions addressed in this work are as follows (here we refer to the duration of time claims remain open as "data"):

1. What are the general statistical properties of the data such as the mean,

median, range, and variance?

2. Does the distribution of data exhibit any skewness?

3. What is the general distribution of the data?

4. Are there any interesting features of the distribution such as gaps or multiple modes?

5. Does extent of injury or claim type help explain any of the interesting properties of the distribution of the data such as gaps?

6. How can we model the data, and which is "the best" model?

To answer these questions the following statistical techniques were employed. Exploratory data analysis was used to understand general statistical properties of the data as well as its distribution. Log linear analysis of contingency tables was used to investigate how the duration of time claims remained open depended on extent of injury and claim type. Mixture models were determined to be the most appropriate models for the duration of time claims remain open. Mixture models were chosen because of the general properties of the distribution of the data. Those properties included multiple modes and a gap in the data. The Expectation Maximization (EM) algorithm was the modeling technique used to fit mixture models to the data. Model selection was done using log-likelihood values, Akaike information criterion (AIC), the likelihood ratio test, the Kolmogorov-Smirnov goodness-of-fit test, and standard errors of the parameter estimates.

1.3 Conclusions and Findings

The general conclusions and findings for the work are as follows. Exploratory data analysis showed that the mean duration of time claims remained open was 1.38 years

(about 504 days). The range of the duration of time claims remain open was 0.0027 years to 69.11 years (about 1 day to 25,202 days). The distribution of the data was shown to have a positive skew as the median was larger than the mean. The general distribution of data was multi-modal and there was a notable gap in the data separating claims open less than 55 years (about 20,089 days) and claims open longer than 55 years. Log linear analysis and modeling showed that the main effects and two-way interactions for both extent of injury and claim type were significant predictors of counts in the three-way contingency table. Only the three-way interaction between extent of injury, claim type, and an indicator for claims being open longer than 55 years was insignificant for predicting counts in the three-way contingency table. The EM algorithm was used to fit gamma and normal mixture models to the duration of time claims remained open. A normal mixture model with 4 components had the best log-likelihood and AIC values. The respective model was selected and validated by the likelihood ratio test and Kolmogorov-Smirnov goodness-of-fit test, respectively. Standard errors for the parameter estimates were relatively small.

This work is organized as follows. Chapter 1 provides an introduction to the setting, the data, the research questions, and general results. Chapter 2 provides the methods used to answer the questions presented in the introduction and outlines the exploratory data analysis used, analysis used on the contingency tables, the background for the EM algorithm used to fit mixture models to the data, and the statistical test used to select and validate the model. Chapter 3 presents the results of the analyses. Chapter 4 presents the final conclusions of the work and recommendations for future work.

Chapter 2 Methods

Exploratory data analysis, log linear analysis of contingency tables, model fitting with the EM algorithm, and statistical tests such as χ^2 tests of independence, likelihood ratio tests, and the Kolmogorov-Smirnov goodness-of-fit test were used to answer the questions listed in the introduction. Exploratory data analysis was used to answer questions regarding standard statistical properties of the data. Exploratory data analysis was also used to understand the general properties of the distribution of the data. Log linear analysis and modeling was used to explain the association between extent of injury, claim type, and claim being open longer or shorter than 55 years. The EM algorithm was used to fit mixture models to the data. Two types of mixture models were fit to the data, normal mixture models and gamma mixture models. Statistical tests such as the likelihood ratio test and Kolmogorov-Smirnov goodness-of-fit test were used for model selection and validation.

2.1 Exploratory Data Analysis

Exploratory data analysis required computation of descriptive statistics for the duration of time claims remain open. The `pastecs` library was used in `R` to compute the statistics. The particular function used was `stat.desc()`. Histograms and box-plots were also used to understand general properties of the distribution of the data. Histograms were used to visualize properties like skewness, multi-modality, and the existence of gaps. Box-plots were used to verify the presence of gaps and visualize the variability of the data.

2.2 Log linear analysis and modeling

Particular attention was given to characterizing claims that were open longer and shorter than 55 years. This was done because a gap in the distribution was found at that duration of time. Contingency tables were used to present the counts of claims based on extent of injury (E), claim type (C), and an indicator for claims being open longer than 55 years (O). Three-way and two-way contingency tables were constructed based on these variables. χ^2 tests of independence were performed on the contingency tables. The hypotheses were (Degroot and Schervish, 2001):

H_0 : The explanatory variables are independent

H_1 : The explanatory variables are dependent.

For example, the explanatory variables for the two-way contingency tables could be extent of injury and claim type, extent of injury and the indicator for claims being open longer than 55 years, or claim type and the indicator for claims being open longer than 55 years. Rejection of the null hypothesis for the two-way contingency table of extent of injury and the indicator for claims being open longer than 55 years would indicate that extent of injury is associated with a claim being open longer or shorter than 55 years. The test statistic for the χ^2 test of independence is computed as

$$T = \sum \frac{(Obs - Exp)^2}{Exp} \tag{2.1}$$

where Obs is the observed count and Exp is the expected count of a cell in the contingency table under H_0 (DeGroot and Schervish, 2001). Under H_0, T has a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom (DeGroot and Schervish, 2001). Here r is the number of rows in the contingency table and c is the number of columns.

Log linear models were then used to further analyze and model the main (three-way) contingency table. A log linear model is an example of a generalized linear model where the count of a cell is the response variable (Christensen, 1997). The response is assumed to be a Poisson random variable. Log linear models are often used to understand the association between categorical variables (Christensen, 1997). We refer to a log linear model that includes all main effects and interactions as the "saturated model". The saturated log linear model used to analyze the three-way contingency table for extent of injury, claim type, and the indicator for claims being open longer than 55 years has the following form:

$$log(\eta_{e,c,o}) = \lambda + \lambda_e^E + \lambda_c^C + \lambda_o^O + \lambda_{e,c}^{E,C} + \lambda_{e,o}^{E,O} + \lambda_{c,o}^{C,O} + \lambda_{e,c,o}^{E,C,O}, \tag{2.2}$$

where $\eta_{e,c,o}$ is the expected count for cell e, c, o, e refers to the possible values of extent of injury, c refers to the possible values of claim type, and o is an indicator for a claim being open longer than 55 years. In the model in (2.2) λ is the mean of the log of the expected counts, λ^E is a main effect term for extent of injury, λ^C is the main effect term for claim type, λ^O is the main effect term for the indicator for claims being open longer than 55 years, $\lambda^{E,C}$ is the interaction term for extent of injury and claim type, $\lambda^{E,O}$ is the interaction term for extent of injury and the indicator for claims being open longer than 55 years, $\lambda^{C,O}$ is the interaction term for claim type and the indicator for claim open longer than 55 years, and $\lambda^{E,C,O}$ is the interaction term for all three categorical variables. The saturated log linear model was used as a baseline in a backward model selection process. The backward selection process was used to determine significant associations between categorical variables extent of injury, claim type, and the indicator for claims open longer than 55 years. Backward model selection was based on the Aikaike Information Criterion (AIC) discussed later.

Model fit was tested using χ^2 deviance statistics. Deviance (D) is defined as

$-2(\mathcal{L}_{fitted} - \mathcal{L}_{saturated})$ and has a χ^2 distribution where the degrees of freedom are the difference between the number of parameters in the saturated model and the fitted model (Christensen, 1997). The hypothesis being tested is (Christensen, 1997)

H_0 : The log linear model fits well

H_1 : The log linear model does not fit well.

Fit was further analyzed by investigating a dissimilarity index. The dissimilarity index is another measure of how close the fitted counts are to the observed counts (Kuha and Firth, 2009). The dissimilarity index (R) presented in Agresti (2002) is computed as

$$R = \frac{\sum_{i=1}^{n} |\eta_i - \hat{\eta}|}{2n}. \tag{2.3}$$

The dissimilarity index estimates the smallest fraction of the sample being studied that would need to be adjusted to get a perfect fit (Kuha and Firth, 2009).

2.3 The Distribution of the Length of Time Claims Remain Open

Mixture distributions arise in a variety of disciplines. The EM algorithm can be used as an iterative method for solving the maximum likelihood problem of parameter estimation for mixture models (Gupta and Chen, 2010). The EM algorithm was first introduced by Sundberg in his dissertation (1971), and later developed in his published papers (1974 and 1976). In these works Sundberg provided an iterative method for solving the maximum-likelihood equation when data included missing values and arose from a distribution from the exponential family. The exponential family refers to a set of probability distributions with a probability density function

(pdf) of the form

$$p(\boldsymbol{x}|) = h(\boldsymbol{x})e^{\theta T(\boldsymbol{x})-A(\theta)}, \tag{2.4}$$

where θ is the vector of model parameters, $T(\boldsymbol{x})$ is the vector of sufficient statistics, A(θ) is the normalizing constant, which is independent of $\boldsymbol{x}$, and $h(\boldsymbol{x})$ is a known function of the data (DeGroot and Schervish, 2001). Sundberg (1974) presents several settings where the EM algorithm is applicable:

– Grouped and censored data, which is when information is observed about a class it belongs to, rather then x directly.

– Fitting mixture models, which refers models with a pdf of the form

$$p(x|\Theta) = \sum_{j=1}^{M} \alpha_j p_j(x|\theta_j), \tag{2.5}$$

where M is the number of components participating in the mixture distribution, p_j's belong to the exponential family for all j, α_j's are the mixing coefficients subject to the constraint $\sum \alpha_j = 1$, θ_j's are vectors of parameters for the p_j's, and $\Theta = \{\alpha_1, \ldots, \alpha_m, \theta_1, \ldots, \theta_m\}$. In this case the data that is missing indicates what distribution p_j generated a particular observation. Thus, a mixture distribution is a probability distribution where any given observations may be generated by one of two or more distinct probability distribution (p_j's) comprising the mixture distribution. The mixing coefficients then may be interpreted as the probabilities that an observation is generated by a particular p_j in the mixture distribution.

– Convolutions of data, which is when the sum of several observations is

observed rather than each particular observation.

– Folded distributions, which arise when only $|x|$ can be observed. Leone, Nelson, and Nottingham (1961) provide an application of folded distributions arising in US Air Force data where measurements were recorded without their algebraic sign.

– Incomplete data arising in multivariate statistical analysis where missing data occurred in a purely random fashion.

In this work we focus on the application of the EM algorithm to fitting mixture distributions.

To present the EM algorithm idea, we start with maximum likelihood estimation. The maximum likelihood method of parameter estimation is foundational to the EM algorithm. The maximum likelihood method proceeds in the following manner. Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be a sample of n independent and identically distributed (iid) observations taken from a population with probability density function $p(x|\theta)$. The likelihood function gives the likelihood of observing $\boldsymbol{x}$ under a given parameter θ.

Definition 2.3.1 (Likelihood function) *Suppose Ω is the parameter space of p and that $\theta \in \Omega$. Let $\mathcal{L} : \Omega \to \mathbb{R}^+$ such that*

$$\mathcal{L}(\theta|\boldsymbol{x}) = p(x_1|\theta) \cdots p(x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta), \tag{2.6}$$

then $\mathcal{L}$ is called the likelihood function of θ given $\boldsymbol{x}$.

The maximum likelihood method finds θ that maximizes $\mathcal{L}$ in (2.6). The value of θ that maximizes $\mathcal{L}$ is called the maximum likelihood estimator (MLE) of θ.

Definition 2.3.2 (Maximum likelihood estimator) *A parameter θ_{MLE} is called the maximum likelihood estimator of θ if*

$$\mathcal{L}(\theta_{MLE}|\boldsymbol{x}) = \max_{\theta\in\Omega} \big\{\mathcal{L}(\theta|\boldsymbol{x})\big\}. \tag{2.7}$$

If $\mathcal{L}$ is a differentiable function, then θ_{MLE} can be obtained by taking the first derivative of $\mathcal{L}$ with respect to θ, setting it to zero and solving for θ. Solutions are critical points and may correspond to maxima, minima, or saddle points of $\mathcal{L}$. Verification that a candidate for θ_{MLE} qualifies as a maximum involves checking if the function is concave down at θ_{MLE}. In most cases the maximum likelihood estimation problem is simplified by analyzing the log of the likelihood function since this transforms the product in (2.6) into a sum. Application of the log function preserves the value of θ_{MLE} since the log function is strictly increasing.

For mixture distributions (2.5) the likelihood function given $\boldsymbol{x}$ is

$$\mathcal{L}(\Theta|\boldsymbol{x}) = \prod_{i=1}^{n} p(x_i|\Theta) = \prod_{i=1}^{n} \Big(\sum_{j=1}^{M} \alpha_j p_j(x_i|\theta_j) \Big). \tag{2.8}$$

Accordingly, the log-likelihood function is

$$ln\big(\mathcal{L}(\Theta|\boldsymbol{x})\big) = ln\Big(\prod_{i=1}^{n} p(x_i|\Theta) \Big) = \sum_{i=1}^{n} ln\Big(\sum_{j=1}^{M} \alpha_j p_j(x_i|\theta_j) \Big). \tag{2.9}$$

Finding θ_{MLE} for (2.9) is simplified with the assumption that there exists unobserved data $\boldsymbol{y} = (y_1, \ldots, y_n)$, where y_i indicates which p_j generated the corresponding x_i ($y_i \in \{1, \ldots, M\}$). Under this assumption (2.7) is referred to as the incomplete-data likelihood function and

$$ln\big(\mathcal{L}(\Theta|\boldsymbol{x},\boldsymbol{y})\big) = ln\Big(\prod_{i=1}^{n} p(x_i,y_i|\theta_i)\Big) = ln\Big(\prod_{i=1}^{n}\frac{p(x_i,y_i|\theta_i)}{p(y_i)}p(y_i)\Big)$$
$$ln\Big(\prod_{i=1}^{n} p(x_i|y_i,\theta_i)p(y_i)\Big) = \sum_{i=1}^{n} ln\big(\alpha_{yi}p_{yi}(x_i|\theta_{yi})\big) \quad (2.10)$$

is called the complete-data likelihood function (Blimes, 1998). In the complete-data likelihood function the actual values of the y_i's are not known. Accordingly, (2.10) cannot be computed directly. Assuming that $\boldsymbol{y}$ is a realization of the random vector $\mathbf{Y}$ makes (2.10) a function of a random variable, and thus a random variable itself. By deriving a probability distribution for the unobserved data, and making an initial guess at the set of parameters $\Theta^{(m)}$, where m refers to the parameters at some iteration m, the expectation of (2.10) conditioned on $\boldsymbol{x}$ and $\Theta^{(m)}$ can be computed. The expectation of the conditional complete-data likelihood function is called the Q function, and is written

$$Q(\Theta|\Theta^{(m)}) = E\Big[ln\big(\mathcal{L}(\Theta|\boldsymbol{x},\boldsymbol{y})\big)\Big|\boldsymbol{x},\Theta^{(m)}\Big] = \int_{\mathcal{Y}} ln\mathcal{L}(\Theta|\boldsymbol{x},\boldsymbol{y})p(\boldsymbol{y}|\boldsymbol{x},\Theta^{(m)})d\boldsymbol{y}, \quad (2.11)$$

where $\mathcal{Y}$ is the support of $\mathbf{Y}$ (Blimes, 1998). It is shown in Blimes (1998) that the Q-function can also be written,

$$Q(\Theta|\Theta^{(m)}) = \sum_{l=1}^{M}\sum_{i=1}^{n} ln\big(\alpha_l p_l(x_i|\theta_l)\big)p(l|x_i,\Theta^{(m)})$$
$$\sum_{l=1}^{M}\sum_{i=1}^{n} ln(\alpha_l)p(l|x_i,\Theta^{(m)}) + \sum_{l=1}^{M}\sum_{i=1}^{n} ln\big(p_l(x_i|\theta_l)\big)p(l|x_i,\Theta^{(m)}), \quad (2.12)$$

where $l \in \{1,\ldots,M\}$ replaced the $y_i s$. In (2.12) $p(l|x_i,\Theta^{(m)})$ is the probability that component l in the mixture distribution generated observation x_i under the parameter estimates at iteration m. With this formulation of the Q-function it is possible to maximize with respect to the α_l parameters separately from the θ_l param-

eters. This is because the second term on the right sight of equation (2.12) does not depend on α_l, and the first term on the right side of equation (2.12) does not depend on θ_l. Application of the EM algorithm has the following general steps (Gupta and Chen, 2010):

E-step: Given $\boldsymbol{x}$ and current parameter values $\Theta^{(m)}$ derive the conditional probability distribution $p(\boldsymbol{y}|\boldsymbol{x}, \Theta^{(m)})$, and use it to evaluate the expected value of the conditional complete-data log likelihood function.

M-step: Find the set of parameters $\Theta^{(m+1)}$ that maximize the expected value of the conditional complete-data log likelihood function evaluated in the E-step. Use $\Theta^{(m+1)}$ in the next iteration of the E-step and repeat until a convergence threshold is exceeded. A convergence threshold (ε) is a user defined value that causes the algorithm to stop when the difference in log likelihood for sequential EM steps is less than ε.

It has been shown that the EM algorithm has the following properties. Gupta and Chen (2010) presented proofs of the monotonicity of the EM algorithm. That is, for each iteration the likelihood of the data under the new parameter estimates is greater than or equal to the likelihood under the previous parameter estimates. Given monotonicity, Gupta and Chen (2010) show that for each iteration the likelihood of the data under the new parameter estimates is strictly greater than the prior likelihood as long as the prior parameter estimates are not critical points. Although a proof for general convergence of the EM algorithm does not exist, it has been shown by Wu (1983) that under certain conditions (satisfied by the exponential family of distributions) the EM algorithm will converge to either a local minimum, maximum, or saddle point. By initializing the algorithm at different values of Θ one can compare

results and determine the most desirable set of parameters.

2.3.1 EM Estimators for Normal Mixture Models

Given the notation introduced in (2.12) we derive EM algorithm estimators for fitting normal mixture models to the data. We start with the assumption that p_l is the pdf of a normal distribution with mean μ_l and standard deviation σ_l, where $l \in \{1, \ldots, M\}$. Thus, $\Theta = \{\alpha_1, \ldots, \alpha_M, (\mu_1, \sigma_1), \ldots, (\mu_M, \sigma_M)\}$, and we have the following Q-function:

$$Q(\Theta|\Theta^{(m)}) = \sum_{l=1}^{M}\sum_{i=1}^{n} ln(\alpha_l)p(l|x_i, \Theta^{(m)}) + \sum_{l=1}^{M}\sum_{i=1}^{n} ln\big(p_l(x_i)\big)p(l|x_i, \Theta^{(m)}). \quad (2.13)$$

For the mixing coefficients α_l we use Lagrange multipliers to maximize Q with respect to the constraint $\sum_{l=1}^{M} \alpha_l = 1$. Accordingly, we need to maximize the following expression

$$S(\alpha_l, \lambda) \coloneqq \sum_{l=1}^{M}\sum_{i=1}^{n} ln(\alpha_l)p(l|x_i, \Theta^{(m)}) + \lambda\left(\sum_{l=1}^{M} \alpha_l - 1\right). \quad (2.14)$$

In this case we left out the second term on the right side of equation (2.13) because it did not depend on α_l. Next, we take the partial derivative of (2.14) with respect to α_l and set to zero to maximize:

$$\frac{\partial S}{\partial \alpha_l} = \sum_{i=1}^{n} \frac{p(l|x_i, \Theta^{(m)})}{\alpha_l} + \lambda = 0,$$

where summing both sides over l gives $\lambda = -n$ (Blimes, 1998).

Substituting $-n$ for λ and solving for $\alpha_l^{(m+1)}$ we obtain the following estimate for the mixing coefficients at iteration $m+1$:

$$\sum_{i=1}^{n} \frac{p(l|x_i, \Theta^{(m)})}{\alpha_l} + \lambda = 0$$
$$\sum_{i=1}^{n} \frac{p(l|x_i, \Theta^{(m)})}{\alpha_l} - n = 0$$
$$\sum_{i=1}^{n} p(l|x_i, \Theta^{(m)}) = \alpha_l \cdot n$$

$$\implies \alpha_l^{(m+1)} = \frac{1}{n} \sum_{i=1}^{n} p(l|x_i, \Theta^{(m)}) \tag{2.15}$$

Next, we derive estimators $\mu_l^{(m+1)}$ and $\sigma_l^{2,(m+1)}$. For the estimator of $\mu_l^{(m+1)}$ we take the partial derivative of (2.13) with respect to μ_l and solve after setting to zero:

$$\frac{\partial}{\partial \mu_l} \left[\sum_{l=1}^{M} \sum_{i=1}^{n} ln\big(p_l(x_i)\big) \underbrace{p(l|x_i, \Theta^{(m)})}_{\beta} \right]$$
$$= \frac{\partial}{\partial \mu_l} \left[\sum_{l=1}^{M} \sum_{i=1}^{n} \Big(- ln(\sqrt{2\pi}\sigma_l) - \frac{1}{2\sigma_l^2}(x_i - \mu_l)^2 \Big) \beta \right]$$
$$= \sum_{i=1}^{n} \frac{-1}{\sigma_l^2}(x_i - \mu_l)\beta = 0$$
$$\sum_{i=1}^{n} x_i\beta - \mu_l\beta = 0$$

$$\implies \mu_l^{(m+1)} = \frac{\sum_{i=1}^{n} x_i\beta}{\sum_{i=1}^{n} \beta} = \frac{\sum_{i=1}^{n} x_i p\big(l|x_i, \Theta^{(m)}\big)}{\sum_{i=1}^{n} p\big(l|x_i, \Theta^{(m)}\big)} \tag{2.16}$$

For the estimate of $\sigma_l^{2(m+1)}$ we take the partial derivative of (2.13) with respect

to $w = \sigma_l^{2,m+1}$ and solve for w after setting to zero:

$$\begin{aligned}
&\frac{\partial}{\partial w}\left[\sum_{l=1}^{M}\sum_{i=1}^{n} ln\big(p_l(x_i)\big)\underbrace{p(l|x_i,\Theta^{(m)})}_{\beta}\right] \\
=&\frac{\partial}{\partial \mu_l}\left[\sum_{l=1}^{M}\sum_{i=1}^{n}\Big(-\frac{1}{2}ln(2\pi)-\frac{1}{2}ln(w)-\frac{1}{2w}(x_i-\mu_l)^2\Big)\beta\right] \\
=&\sum_{i=1}^{n} -\frac{\beta}{2w}+\frac{\beta}{2w^2}(x_i-\mu_l)^2=0 \\
&\sum_{i=1}^{n}\beta w-\sum_{i=1}^{n}\beta(x_i-\mu_l)^2=0
\end{aligned}$$

$$\text{Thus, } w=\sigma_l^{2,(m+1)}=\frac{\sum_{i=1}^{N}\beta(x_i-\mu_l)^2}{\sum_{i=1}^{N}\beta}=\frac{\sum_{i=1}^{N}(x_i-\mu_l)^2 p(l|x_i,\Theta^{(m)})}{\sum_{i=1}^{N}p(l|x_i,\Theta^{(m)})} \tag{2.17}$$

Finally, the execution of the EM algorithm for normal mixture models consists of the following specific steps (Gupta and Chen, 2010),

Step 1: Initialize the algorithm by setting each parameter with a guess. Also, set a stopping threshold, ε, and compute the value for $ln\,\mathcal{L}^{(m)}$ based on initial parameter estimates where

$$ln\,\mathcal{L}^{(m)}=\frac{1}{n}\sum_{i=1}^{n} ln\Big(\sum_{l=1}^{M}\alpha_l^{(m)}p_l(x_i|\Theta^{(m)}\Big).$$

Step 2: (**E-Step**) For $l \in \{1,\ldots,M\}$, estimate the posterior probability that x_i was generated by the l^{th} distribution given our observed x_i and current param-

eter estimates,

$$p(l|x_i,\Theta^{(m)}) = \frac{\alpha_l^{(m)} p_l\big(x_i|\Theta^{(m)}\big)}{\sum_{l=1}^{M} \alpha_l^{(m)} p\big(x_i|\Theta^{(m)}\big)}.$$

Step 3: (**M-Step**) Use the posterior probabilities to derive the new parameter estimates that maximize the Q function,

$$\alpha_l^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n} p(l|x_i,\Theta^{(m)})$$

$$\mu_l^{(m+1)} = \frac{\sum_{i=1}^{n} x_i p\big(l|x_i,\Theta^{(m)}\big)}{\sum_{i=1}^{n} p\big(l|x_i,\Theta^{(m)}\big)}$$

$$\sigma_l^{2,(m+1)} = \frac{\sum_{i=1}^{n} \big(x_i-\mu_l\big)^2 p\big(l|x_i,\Theta^{(m)}\big)}{\sum_{i=1}^{n} p\big(l|x_i,\Theta^{(m)}\big)}$$

Step 4: Compute $ln\mathcal{L}^{(m+1)}$ and $|ln\mathcal{L}^{(m)} - ln\mathcal{L}^{(m+1)}|$. If $|ln\mathcal{L}^{(m)} - ln\mathcal{L}^{(m+1)}| < \varepsilon$, then terminate the iterations. If not then ε then store $ln\mathcal{L}^{(m+1)}$ and go to Step 2.

2.3.2 EM Estimators for Gamma Mixture Models

Given the notation introduced in (2.12) we derived estimators for fitting gamma mixture models to the data following Almhana et al. (2006) and Schwander and Nielsen (2013). We start with the assumption that p_l is the pdf of a gamma distribution with parameters k_l and β_l, where $k_l > 0$ is a shape parameter and $\beta_l > 0$ is a scale parameter. In this case $l \in \{1,\ldots,M\}$ and $\Theta = \{\alpha_1,\ldots,\alpha_m,(k_1,\beta_1),\ldots,(k_M,\beta_M)\}$. Further, p_l has the form:

$$p_l(x_i) = \frac{\beta_l^{k_l}}{\Gamma(k_l)} x_i^{k_l-1} e^{-x_i/\beta_l} \qquad l \in \{1,\ldots,M\}, \tag{2.18}$$

and

$$\Gamma(k_l) = \int_0^\infty z^{k_l-1}e^{-z}dz. \tag{2.19}$$

Derivation of $\alpha_l^{(m+1)}$ follows that of the normal mixture estimator with no change except that p_l is the pdf of a gamma random variable. Thus, $\alpha_l^{(m+1)}$ is updated by using (2.15) and the new definition of p_l. Derivation of the estimator for $\beta_l^{(m+1)}$ proceeds as follows (Almhana et al., 2006):

$$\begin{aligned}
\frac{\partial S}{\partial \beta_l} =& \frac{\partial}{\partial \beta_l}\left[\sum_{l=1}^{M}\sum_{i=1}^{n} ln(p_l(x_i))\underbrace{p(l|x_i,\Theta^{(m)})}_{\xi}\right] \\
=& \frac{\partial}{\partial \beta_l}\left[\sum_{l=1}^{M}\sum_{i=1}^{n}\Big(- ln(p_l(x_i)) + (k-1)ln(x_i) - k_l ln(\beta_l) - x_i/\beta_l\Big)\xi\right] \\
=& -\frac{k_l}{\beta_l}\sum_{i=1}^{n}\xi + \frac{1}{\beta_l^2}\sum_{i=1}^{n} x_i\xi = 0
\end{aligned}$$

$$\implies \beta_l^{(m+1)} = \frac{k_l^{(m)}\sum_{i=1}^{n}\xi}{\sum_{i=1}^{n} x_i\xi} = \frac{k_l^{(m)}\sum_{i=1}^{n} p(l|x_i,\Theta^{(m)})}{\sum_{i=1}^{n} x_i p(l|x_i,\Theta^{(m)})}. \tag{2.20}$$

Derivation of estimator for $k_l^{(m+1)}$ proceeds in a similar way (Almhana et al., 2006),

$$\begin{aligned}
\frac{\partial S}{\partial k_l} =& \frac{\partial}{\partial k_l}\left[\sum_{l=1}^{M}\sum_{i=1}^{n} ln(p_l(x_i))\underbrace{p(l|x_i,\Theta^{(m)})}_{\xi}\right] \\
=& \frac{\partial}{\partial k_l}\left[\sum_{l=1}^{M}\sum_{i=1}^{n}\Big(- ln(\Gamma(k_l)) + (k-1)ln(x_i) - k_l ln(\beta_l) - x_i/\beta_l\Big)\xi\right] \\
=& \sum_{i=1}^{n}\left[ln(x_i) - ln(\beta_l) - \frac{\partial}{\partial k_l} ln(\Gamma(k_l))\right]\xi
\end{aligned}$$

$$= \sum_{i=1}^{n} \left[ln(x_i) - ln(\beta_l) - \psi(k_l)) \right] \xi. \tag{2.21}$$

Setting (2.21) to zero and solving for k_l has no closed form expression, meaning there is no direct way to solve for $k_l^{(m+1)}$ (Almhana et al., 2006). Almhana et al. (2006) and Schwander and Nielsen (2013) use the fact that the EM algorithm is a gradient based algorithm, and therefore propose to update k_l in the direction of its gradient by a prescribed step size. In this case,

$$k_l^{(m+1)} = k_l^{(m)} + \frac{1}{k} G, \tag{2.22}$$

where

$$G = \frac{1}{n} \sum_{i=1}^{n} \left[ln(x_i) - ln(\beta_l^{(m)}) - \psi(k_l^{(m)}) \right] p(l|x_i, \Theta^{(m)}) \tag{2.23}$$

where $\psi(k)$ also has no explicit expression (Almhana et al., 2006). Almhana et al. (2006) suggest using

$$\psi(k) \approx ln(k - 0.5) + \frac{1}{24(x - 0.5)^2}, \tag{2.24}$$

which they show is a good approximation, particularly when $k \geq 2$. Implementation of the EM algorithm for a mixture of gamma distributions follows the steps outlined in section 2.3.1 with the exception that the gamma estimators are used. Algorithms for fitting normal mixture models and gamma mixture models were implemented using the `mixtools` library in R. The functions called were `normalmixEM()` and `gammamixEM()`.

2.4 Model Selection and Validation

After models were fit using the EM algorithm a variety of methods were used to select and validate them. Model selection was done through investigation of log-likelihood values, Akaike information criterion (AIC) values, likelihood ratio tests, and Kolmogorov-Smirnov goodness of fit tests. The tests were performed using parametric bootstrapping. The log-likelihood values provide a measure of the overall likelihood of generating a given data set under the proposed model, AIC values are a measure of the information carried in the model including a penalty for additional parameters. The likelihood ratio test is used to compare nested models (M versus $M+1$ where M is the number of component distributions in the mixture model), and provides a method of selecting the best model. The Kolmogorov-Smirnov goodness of fit test was used to verify the model was statistically identical to the population distribution that generated the data.

2.4.1 Akaike information criterion (AIC)

The AIC is typically computed using the equation

$$AIC = 2m - 2ln(\mathcal{L}), \tag{2.25}$$

where m is the number of parameters in the model and $\mathcal{L}$ is the likelihood function (Sakamoto et al., 1986). The second term in the equation is used to correct for bias (Sakamoto et al., 1986). The most desirable model will have a minimum AIC value and maximum $\mathcal{L}$ value. When fitting a model with the EM algorithm a definition different from (2.25) is required because the EM algorithm analyzes the expected log likelihood surface (Glosup and Axelrod, 1994). Sakamoto et al. (1986) define the AIC such that the expected log likelihood is $-1/2AIC$. Thus for computing the

AIC for models with parameters obtained by the EM algorithm

$$AIC = -2Q(\Theta, \hat{\Theta}), \tag{2.26}$$

is appropriate (Glosup and Axelrod, 1994). We used the formula in (2.26) for analysis.

2.4.2 Likelihood Ratio Test

The likelihood ratio test (DeGroot and Schervish, 2001) was used for model selection. The likelihood ratio test tests the following hypotheses (Degroot and Schervish, 2001):

$$H_0 : \Theta = \Theta_M$$

$$H_1 : \Theta = \Theta_{M+1}.$$

Rejection of H_0 means that the model with $M+1$ components is significantly better than a model with M components. In the testing of mixture models M starts at 2 and is increased until H_0 cannot be rejected. The likelihood ratio statistic

$$\lambda = \frac{\mathcal{L}(\Theta_M|\boldsymbol{x})}{\mathcal{L}(\Theta_{M+1}|\boldsymbol{x})} \tag{2.27}$$

can be used to test the above hypotheses (DeGroot and Schervish, 2001).

By applying $-2ln$ to both sides of (2.27) we get,

$$-2ln(\lambda) = 2\big(ln\mathcal{L}(\Theta_{M+1}|\boldsymbol{x}) - ln\mathcal{L}(\Theta_M|\boldsymbol{x})\big) \sim \chi^2_{dim(M+1)-dim(M)}. \tag{2.28}$$

which can be shown true when $\boldsymbol{x}$ is large. Computing the test statistic was approached with caution since (2.28) requires actual log likelihood values while the EM algorithm produces expected values of the log likelihood. To deal with this issue the distribution

of under the null hypothesis was discovered through a parametric bootstrap. The parametric bootstrap proceeds as follows:

1. We fit the claim duration data for M and $M+1$ components using the EM algorithm and set $Q^{obs} = 2(\mathcal{L}_{M+1} - \mathcal{L}_M)$

2. We simulate a new data set using the fitted M component model

3. We fit M and $M+1$ models to the simulated data using the EM algorithm and call these log likelihood values $\mathcal{L}^*_{M+1}$ and $\mathcal{L}^*_M$

4. We store $Q^* = 2(\mathcal{L}^*_{M+1} - \mathcal{L}^*_M)$

5. We repeat steps (2) to (4) 1,000 times and then take the p-value for the above hypothesis test to be the mean of $I(Q^* \geq Q^{obs})$ where I is the indicator function.

The result is a more robust test of the above hypotheses.

Model selection with the likelihood ratio test includes multiple comparisons. Multiple comparisons occur when a test involves multiple hypotheses that are dependent. For example, in the likelihood ratio test outlined here testing the null hypothesis that $\Theta = \Theta_3$ is dependent on the outcome of the tests for $H_0 : \Theta = \Theta_1$ versus $H_1 : \Theta = \Theta_2$ and $H_0 : \Theta = \Theta_2$ versus $H_1 : \Theta = \Theta_3$. The dependence arises because we need to first reject the null hypotheses that $\Theta = \Theta_1$ and $\Theta = \Theta_2$ before testing the null hypothesis that $\Theta = \Theta_3$. This distinction alters the error structure of the test.

When tests involve only a single test we focus on controlling type I error, which occurs when the null hypothesis is rejected even though it is true. The probability of type I error is given by the probability $\eta = P(\text{reject } H_0 \mid H_0 \text{ is true})$. When a test includes multiple comparisons special attention should be given to the propagation error, including type II error. Type II error occurs when the null hypothesis is not rejected although the alternative hypothesis is true. The probability of type II error

is denoted $\beta = P(\text{fail to reject } H_0 \mid H_1 \text{ is true})$. Figure 2.1 shows a logical tree of the possible outcomes if the null hypothesis that $\Theta = \Theta_4$ is the first null hypothesis that fails to be rejected. The β_is ($i \in \{1, 2, 3\}$) are the probability of committing a type II error for the $H_0 : \Theta = \Theta_M$ versus $H_1 : \Theta = \Theta_{M+1}$ tests denoted by the heptagons, and η is the probability of committing a type I error, or the significance level of the test.

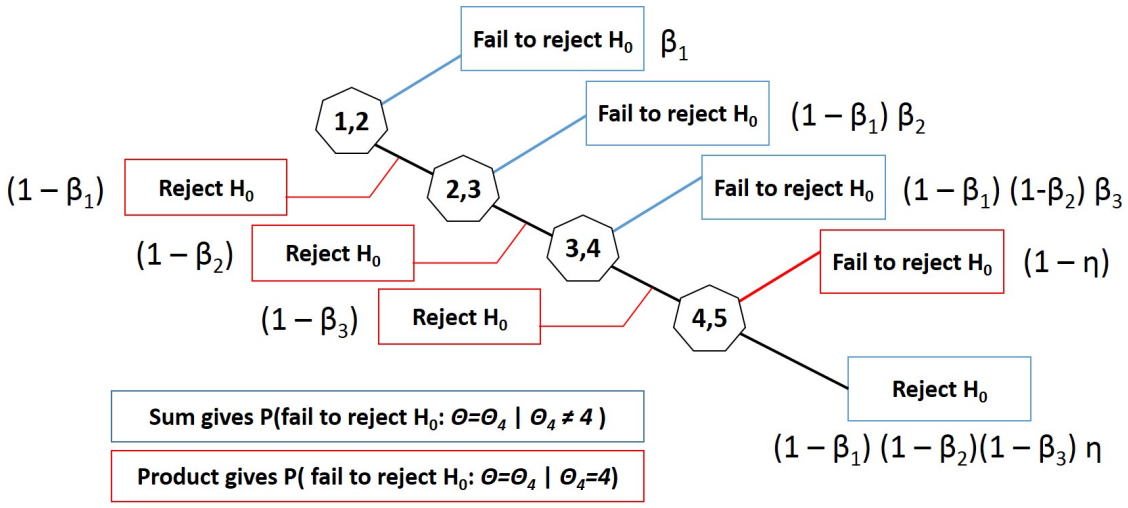

Figure 2.1: A logical tree for the likelihood ratio test when failing to reject the null hypothesis that $\Theta = \Theta_4$.

The overall error of the test when the first null hypothesis failing to be rejected is $\Theta = \Theta_4$ is given by the sum of the probabilities of the blue boxes. In this case, the probability that the selected model is correct is given by the product of the red boxes. We determined the β_is by using a parametric bootstrap and set the significance level of the test, η, so that

$$P(\text{ fail to reject } H_0 : \Theta = \Theta_M \mid \Theta \neq \Theta_M) = 0.05. \tag{2.29}$$

In particular, for the case when we fail to reject the null hypothesis that $\Theta = \Theta_4$ we

set η such that

$$P(\text{ fail to reject } H_0 : \Theta = \Theta_4 \mid \Theta \neq \Theta_4) =$$

$$\beta_1 + (1-\beta_1)\beta_2 + (1-\beta_1)(1-\beta_2)\beta_3 + (1-\beta_1)(1-\beta_2)(1-\beta_3)\eta = 0.05. \quad (2.30)$$

2.4.3 Kolmogorov-Smirnov Goodness-of-Fit Test

The Kolmogorov-Smirnov goodness-of-fit test (D'Agostino and Stephens, 1986) was used to determine if the selected mixture model was statistically identical to the distribution that generated the data. The Kolmogorov-Smirnov goodness-of-fit test utilizes the empirical cumulative distribution function which is defined as

$$\hat{F}(x) = n(x_i)/N, \quad (2.31)$$

where x_i is the ith observation, $n(x_i)$ is the number of observations less than or equal to x_i, and N is the total number of observations (Degroot and Schervish, 2001). The hypotheses for this test are (Degroot and Schervish, 2001)

$$H_0 : \hat{F}(x) = F(x)$$

$$H_1 : \hat{F}(x) \neq F(x).$$

The test statistic for the Kolmogorov-Smirnov goodness-of-fit test is computed as

$$D = \sup_{i \in \{1,\ldots,N\}} \left| F(x_i) - \hat{F}(x_i) \right|, \quad (2.32)$$

where F is the theoretical cumulative distribution function in H_0 (D'Agostino and Stephens, 1986). Critical values for the test are readily available in tables or in software packages like R.

To perform the test, the mean of 1,000 realizations of the Kolmogorov-Smirnov goodness-of-fit test statistic were obtained with a parametric bootstrap. The bootstrap is used because a bias may arise if the test is performed on the same data we used to fit the model.

2.4.4 Standard Error of the Parameter Estimates

After the model had been selected using log-likelihood values, AIC values, the likelihood ratio test, and the Kolmogorov-Smirnov goodness-of-fit test, analysis was performed for the parameter estimates. The distribution and standard errors of the parameters were obtained by another parametric bootstrap. The procedure reveals the degree of uncertainty associated with the final set of parameters. The bootstrap for the standard errors of the parameter estimates was implemented using the `boot.se` function in the `mixtools` package in R.

Chapter 3 Results

3.1 Exploratory Data Analysis of Claims Duration

Exploratory data analysis was used to answer the following questions:

- What are the general statistical properties of the data such as the mean, median, range, and variance?
- Is the distribution of the data symmetric or skewed?
- What is the general distribution of the population the data comes from?
- Are there any interesting features of the distribution such as gaps or multiple modes?

To answer the first question descriptive statistics were produced for the data. The descriptive statistics for all the claim duration data are presented in Table 3.1. The descriptive statistics for claims open less than 55 years and for claims open longer than 55 years are presented in Table 3.2 and Table 3.3, respectively.

As presented in Table 3.1, the mean of the entire data set was 1.38 years (about 504 days). The minimum (zeros omitted) was 0.0027 (about 1 day) and the max was 69.11 (about 25,242 days). Given the presented values for the min and max, the range for the data was 69.107 years (about 25,242 days). The median for the entire data set was 0.35 years (about 128 days). The data set has a positive skew. The coefficient of variation ($C_v = S/\bar{X}$) was 5.024. The coefficient of variation is a mean normalized measure of the dispersion of the data. 103 claims were reported to be open less than 0 years, which was assumed to be due to input error. These claims were removed from the data before computation of any descriptive statistics presented in Tables 3.1

through 3.3.

Table 3.1: Descriptive statistics for all claim lengths. Values are in years. The minimum was taken to be the smallest value other than zero.

Non-zero entries	Zero entries	"NA" entries	Min	Max
2,233,784	288	4,269	0.0027	69.11
Median	$\bar{X}$	S^2	S	C_v
0.353	1.377	47.833	6.916	5.024

From Table 3.2, there were 2,206,553 claims open less than 55 years. These claims had a median of 0.35 years (about 128 days), a mean of 0.62 years (about 226 days), and a standard deviation of 1.387 ($\sim$507 days). From Table 3.3, there were 27,444 claims open longer than 55 years. These claims had a median of 62.42 years (about 22,799 days), a mean of 62.19 years (about 22,715 days), and standard deviation of 3.15 years (about 115 days).

Table 3.2: Descriptive statistics for claims open less than 55 years. Values are in years.

Non-zero entries	Zero entries	"NA" entries	Min	Max
2,206,553	288	4,269	0.0027	53.500
Median	$\bar{X}$	S^2	S	C_v
0.350	0.623	1.925	1.387	2.23

A histogram for the duration of time claims remain open is presented in Figure 3.1. The histogram in top-left panel of Figure 3.1 is multi-modal, having two peaks in density occurring at values less than 0.5 years (about 183 days). In the histogram it can be seen that there is a clustering of claims open longer than 55 years, which

Table 3.3: Descriptive statistics for claims open longer than 55 years. Values are in years.

Non-zero entries	Zero entries	"NA" entries	Min	Max
27,444	0	0	55.18	69.11
Median	$\bar{X}$	S^2	S	C_v
62.417	62.186	9.908	3.148	0.0506

is shown in the top-right panel of Figure 3.1. The box plot in the bottom panel of Figure 3.1 conveys the fact that there is a gap between the claims open less than 55 years and the claims open longer than 55 years.

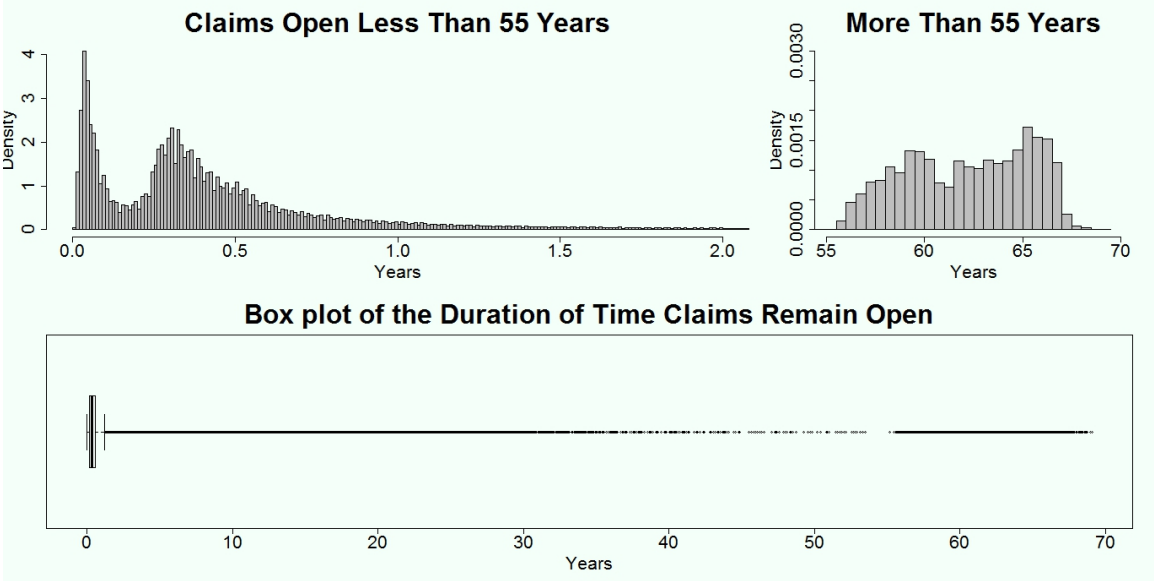

Figure 3.1: The histogram for the data representing the length of time claims remain open. The figure in the top-left panel shows claims open less than 55 years. The figure in the top-right panel shows claims open longer than 55 years.

3.1.1 Summary

The results of the exploratory data analysis are as follows:

– The mean of the entire data set was 1.38 years (504 days) and the range was 69.107 (about 25,241 days) (0 values omitted).

– The mean for the data open less than 55 years was 0.623 years (228 days) and the range was 53.507 years (19,543 days).

– The mean for the data open longer than 55 years was 62.186 years (about 22,713 days) and the range was 13.93 years (about 5,088 days).

– The histograms revealed that the general distribution of the data is positively skewed.

– The histogram and box-plot revealed that there is a gap in the distribution around 55 years. This suggests the claim duration data may be generated by a mixture distribution.

– The general distribution of the data is multi-modal. This also suggests the claim duration data is being generated by a mixture distribution.

3.2 Log Linear Analysis and Modeling

The three-way contingency table for extent of injury, claim type, and the indicator for claims open longer than 55 years is presented in Table 3.4. A χ^2 test of independence for Table 3.4 produced a χ^2 statistic of 32,225,119 with 39 degrees of freedom. The p-value for the test was less than 2.2e-16. The test reveals that there exists some dependence between extent of injury, claim type, and the indicator for claims being open longer than 55 years. The precise nature of this dependence is not revealed by the test, but it can be due to a three-way interaction, or one of the three two-way interactions.

Table 3.4: A 3-way contingency table for claim counts based on extent of injury, claim type, and the indicator for claims open longer than 55 years.

		Duration	
Extent of Injury	Claim Type	$\geq$ 55 yrs	$<$ 55 yrs
1	NR	0	0
	R	13	1,427
	RO	0	0
	V	0	0
2	NR	0	0
	R	1	874
	RO	0	0
	V	0	0
5	NR	0	22
	R	96	273,726
	RO	0	0
	V	2	0
6	NR	131	1,230,573
	R	39	1288
	RO	25,710	562,927
	V	1,336	3,150
9	NR	0	10
	R	6	132,453
	RO	0	0
	V	0	0

Two-way contingency tables were formed for extent of injury and claim type, extent of injury and the indicator for claims open longer than 55 years, and claim

type and the indicator for claims open longer than 55 years. The contingency tables for these three pairs of variables are presented in Table 3.5, 3.6, and 3.7, respectively.

Table 3.5: A contingency table for extent of injury codes with respect to claim type.

	Claim Type			
Extent of Injury	NR	R	RO	V
1	0	1,440	0	0
2	0	875	0	0
5	22	273,882	0	2
6	1,230,704	1,327	588,637	4,486
9	10	132,459	0	0

Table 3.6: A contingency table for extent of injury codes with respect to the duration of time claims remain open ($\leq$55 and $<$55)

	Duration	
Extent of Injury	$\geq$ 55 yrs	$<$ 55 yrs
1	13	1427
2	1	874
5	98	273,748
6	27,216	1,797,938
9	6	132,463

Table 3.7: A contingency table for claim type codes with respect to the duration of time claims remain open ($\leq$55 and $<$55)

	Duration	
Claim Type	$\geq$ 55 yrs	$<$ 55 yrs
N	131	1,230,605
R	155	409,768
RO	25,710	562,927
V	1,338	3,150

χ^2 tests of independence were performed on Tables 3.5 to 3.7. The results of the tests are presented in Table 3.8. The results indicate that the variables are dependent in all cases.

Table 3.8: Results for χ^2 tests of independence performed on Tables 3.4 to 3.6.

Description	χ^2	df	p-value
Extent of Injury / Claim Type	2,224,707	12	$<$ 2.2e-16
Claim Type / 55 yr indicator	98,241	3	$<$ 2.2e-16
Extent of Injury / 55 yr indicator	5,917	4	$<$ 2.2e-16

Next we built a log linear model for the three-way contingency table in Table 3.4. We began by fitting the saturated model and proceeded by progressively removing interactions until the AIC was improved. The saturated model is the model with all main effect terms, interaction terms, and a parameter for the intercept. The results for the saturated model and backwards selection process is presented in Table 3.9. The "*" between variables in the formulas in Table 3.9 indicates main effects and all lower order interactions. The ":" between variables in the formulas in Table 3.9 indicates interactions. The "-" is used to indicate that the respective interaction term was removed from the model with the lowest AIC.

Table 3.9: Results for backward selection of the log linear model for the three-way contingency table in Table 3.4.

Formula	χ^2	df	p-value	AIC
∼Extent.of.Inj*Claim.Type*Indicator	0	0	1	80
−Exent.of.Inj:Claim.type:Indicator	21.13217	12	0.04848	77.13
−Extent.of.Inj:indicator	390.3802	16	0	428
−Claim.Type:Indicator	63,228.54	15	0	63,279
−Extent.of.Inj:Claim.Type	2,097,405	24	0	2,097,437

Table 3.9 shows that a minimum AIC value was obtained after removing the three-way interaction from the saturated model. The AIC value obtained was 77.13. When two-way interactions were removed from the model without the three-way interaction the AIC value increased. The results show that the three-way interaction was insignificant. Since the AIC values serves as a measure of information carried in the model, where a lower AIC indicates more information, we can interpret the results in Table 3.9 in a way that orders the two-way interaction in terms of predictor im-

portance. For example, the models with two-way interactions Extent.of.Inj:Indicator and Claim.Type:Indicator alternatively removed had AIC scores of 428 and 63,279, respectively. Based on the AIC scores the former model is better. The AIC being better when Extent.of.Inj:Indicator is removed as compared to Claim.Type:Indicator implies that the Claim.Type:Indicator interaction is a more important predictor of counts in the contingency table. Following this logic the decreasing order of importance for the two-way interactions was extent of injury / indicator, claim type / indicator, and extent of injury / claim type.

The selected model based on the backward selection process was the saturated model less the three-way interaction for extent of injury, claim type, and the indicator for claims open longer than 55 years. The selected model failed to produce estimations of the coefficients. However, this is not of huge importance since the log linear model fit here is not intended to be predictive but rather it is used to explain the association between categorical variables in the contingency table. The p-value for the χ^2 statistic rejects the null hypothesis that the log linear model fits well at the 0.05 significance level. The results indicate that the model fit is unsatisfactory. The fitted values for the selected log linear model are presented in Table 3.10. The residuals are presented in Table 3.11.

The dissimilarity index (section 2.2, formula 2.3) was computed from the residuals in Table 3.11. The value was equal to 0.000003%. The interpretation is that 0.000003%, or 6.8251 observations would need to be adjusted to achieve a perfect fit.

Table 3.10: A 3-way contingency table of fitted claim counts based on the selected log linear model.

		Duration	
Extent of Injury	Claim Type	$\geq$ 55 yrs	$<$ 55 yrs
1	NR	0	0
	R	13.0076	1,426.9997
	RO	0	0
	V	0	0
2	NR	0	0
	R	1.0006	873.9498
	RO	0	0
	V	0	0
5	NR	0.00003	21.9998
	R	98.0470	273,723.9498
	RO	0	0
	V	1.9894	0.0158
6	NR	131	1,230,573
	R	36.9413	1290.0802
	RO	25,710	562,927
	V	1,337.9890	3,148.0106
9	NR	0.000002	10
	R	6.0035	132,452.9706
	RO	0	0
	V	0	0

Table 3.11: A 3-way contingency table for claim counts based on extent of injury, claim type, and the indicator for claims open longer than 55 years.

		Duration	
Extent of Injury	Claim Type	$\geq$ 55 yrs	$<$ 55 yrs
1	NR	0	0
	R	-2.1178e-3	8.3954e-6
	RO	0	0
	V	0	0
2	NR	0	0
	R	-5.8738e-4	6.5637e-6
	RO	0	0
	V	0	0
5	NR	-7.6546e-3	4.7276e-6
	R	-2.0745e-1	3.9187e-3
	RO	0	0
	V	0.0063	-0.0158
6	NR	2.7032e-6	-8.1525e-6
	R	3.3564e-1	-5.7931
	RO	0	0
	V	-5.4401e-2	3.5454e-2
9	NR	-1.8357e-3	2.6147e-7
	R	-1.4381e-3	8.0928e-5
	RO	0	0
	V	0	0

3.2.1 Summary

The results of the log linear analysis are as follows:

- χ^2 tests of independence on the three-way contingency table revealed that some dependence exists between extent of injury, claim type, and the indicator for claims being open longer than 55 years.

- χ^2 tests of independence on the two-way contingency tables revealed that some dependence exists between extent of injury and claim type, extent of injury and the indicator for claims open longer than 55 years, and claim type and the indicator for claims open longer than 55 years.

- Backward selection of a log linear model for the three-way contingency table revealed that the three-way interaction between extent of injury, claim type, and the indicator for claims open longer than 55 years was not significant. It also showed that all two-way interactions were significant.

- Based on the AIC scores, backward selection revealed that extent of injury and claim type was the most important two-way interaction, followed by claim type and the indicator for claims open longer than 55 years, and then followed by the extent of injury and the indicator for claims open longer than 55 years interaction.

- The selected log linear model was the saturated model less the three-way interaction for extent of injury, claim type, and the indicator for claims open longer than 55 years.

- The χ^2 statistic for the selected model produced a p-value of 0.04848. The p-value rejects the null hypothesis that the model fits well at the 0.05 significance level.

– Finally, a dissimilarity index was computed based on the residuals in Table 3.11. The dissimilarity statistic was computed to be 0.000003%, which has the interpretation that 0.000003% of the observations ($\sim$ 7 observations) would need to be adjusted to achieve a perfect fit.

3.3 Distribution of the Length of Time Claims Remain Open

3.3.1 Mixture of Normals

The EM algorithm was implemented in R using the package `mixtools` in order to model the distribution of the duration of time claims remain open. Fitting normal mixtures models to the data presented in Figure 3.1 was done in two ways. First, a series of arbitrary initial estimates for $\Theta^{(0)}$ were used to produce a set of results. Second, constraints on parameter values were enforced to see if a better fit could be attained through human intervention. The constraint was determined by partitioning the data and fitting a normal distribution to a subset. The constraint consisted of one mixture component being fixed as $\mathcal{N}(0.0484, 0.0218)$. Fitting the model with the constraint means that the expected value of the conditional log likelihood was maximized with respect to all parameters except the constrained parameters. Thus, they remained fixed throughout the EM algorithm process. Additional constraints were attempted but did not improve the results. For both cases, results include estimates of parameters for normal mixture models of up to 10 components. The key results are presented in Table 3.12. Estimated parameters for each model are presented in Appendix A. The convergence threshold was set to 10^{-8}. The convergence threshold corresponds to the difference between log likelihood values at iteration m and $m+1$ in the EM algorithm. When the difference is smaller than the convergence threshold, the algorithm stops. Table 3.12 shows the log-likelihood of the model increases when adding additional distributions. In both cases the log-likelihood greatly improved when the number of distributions increased from $M = 2$ to 4. Table 3.12 also includes AIC values for each model. The AIC values in Table 3.12 suggests that a 4 component mixture of normal distributions is preferred. Figure 3.2 presents a graph of the log-likelihood values presented in Table 3.12.

Table 3.12: Results for fitting normal mixture models to the data with the EM algorithm. The results include normal mixture models from M=2 to 10. Included in the table are the resulting log-likelihood values, and AIC values for both the models fitted with arbitrary $\Theta^{(0)}$ values and a constrained set of parameters in Θ.

EM Results	Arbitrary $\Theta^{(0)}$			Constrained Parameters in Θ		
	Iterations	$ln(\mathcal{L})$	AIC	Iterations	$ln(\mathcal{L})$	AIC
$M = 2$	24	-1,372,205	3,179,872	6	-3,287,218	4,413,099
3	29	-1,020,804	1,775,874	31	-1,032,436	1,872,931
4	106	-845,038	950,610	107	-753,700	944,691
5	148	-722,724	2,790,358	270	-669,941	2,122,928
6	155	-650,499	2,839,688	459	-650,751	2,780,242
7	994	-606,686	3,185,950	739	-626,074	2,883,754
8	541	-592,645	3,280,084	489	-623,680	2,892,131
9	1035	-585,723	3,799,290	922	-616,466	3,250,340
10	1444	-575,359	4,509,210	1,534	-610,309	4,313,817

Figure 3.3 presents histograms of the data along with plots of the fitted normal components for $M = 2$ to 10. The results in Figure 3.3 arose form arbitrary initial values of $\Theta^{(0)}$. Figure 3.4 presents plots of the empirical cumulative distribution function of the data versus the theoretical cumulative distribution function of the models plotted in Figure 3.3. A line for $y = x$ was also plotted to aid interpretation. Improvement is most noticeable when increasing the number of components from 2 to 4. Figure 3.5 shows plots that are similar to those in Figure 3.3 the difference being that the models were fit with a single set of constrained parameters in Θ. Figure 3.6 presents plots the empirical cumulative distribution function (ECDF) versus the theoretical cumulative distribution (TCDF) functions for the fitted models presented in Figure 3.5. Greatest improvement for the models fit with a single constrained set of parameters was when components were increased from M=2 to 4.

3.3.2 Mixture of Gammas

Fitting gamma mixture models to the data produced results that were sub-optimal when compared to those produced by normal mixture models. Figure 3.7 shows histograms for the data with curves plotted for fitted gamma mixture model components.

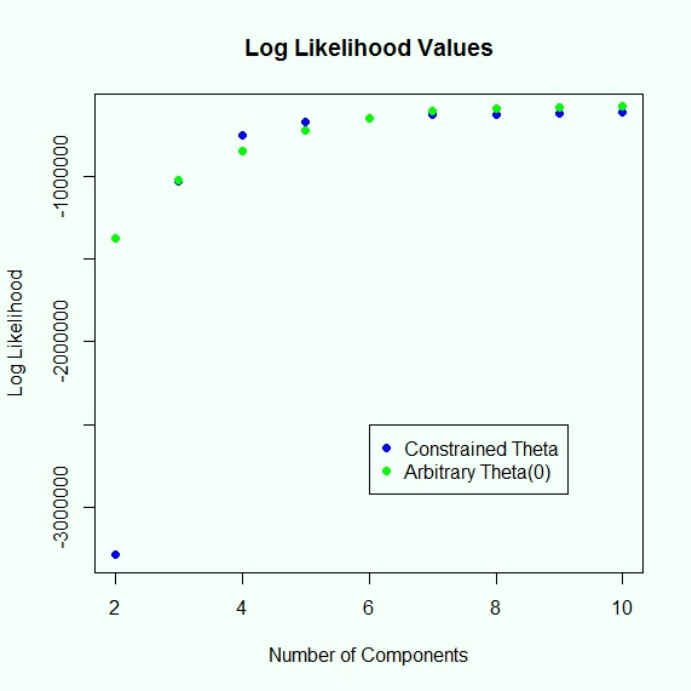

Figure 3.2: A plot of the log likelihood values for different models produced by the EM algorithm. The blue points refer to log likelihood values obtained wihthout any constraints on Θ. The green points refer to log likelihood values obtained with a constraint on Θ.

Table 3.13 presents the corresponding log likelihood values. When comparing the Log likelihood values in Table 3.13 to those in Table 3.12 we see that the fitted normal mixture models are more likely than the fitted gamma mixture models.

Table 3.13: Log likelihood values for gamma mixture models fit to the data.

	Parameters					
	M=2	3	4	5	6	7
Log Likelihood	-1,045,910	-881,089	-780,392	-694,172,	-709,307	-737,339

3.4 Model Selection and Validation

Initial results indicate that a mixture of normal distributions provide a better fit for the data than a mixture of gamma distributions. Log-likelihood values and computed

AIC values suggest the selection of the 4 component mixture of normals produced with a set of constrained parameters in Θ. Additional steps were taken to verify and validate this selection. Results of the likelihood ratio test using a bootstrapped likelihood ratio statistic are presented in Table 3.14. The results in Table 3.14 also indicate selection of a mixture of 4 normal distributions. The significance level (η) was set such that (2.32) was satisfied for $\beta_1 = 0$, $\beta_2 = 0.02$, and $\beta_3 = 0.01$. β_1 was zero because $H_0 : \Theta = \Theta_1$ was not tested for the constrained models. β_2 and β_3 where determined through parametric bootstrap. The significance level after accounting for multiple comparisons was $\eta = 0.021$. The test failed to rejected the null hypothesis that $\Theta = \Theta_4$. The $P($ fail to reject $H_0 : \Theta = \Theta_4 \mid \Theta = \Theta_4) = 0.95$ for this test.

Table 3.14: P-values computed using 1,000 realizations of the bootstrapped log of the likelihood ratio statistic.

p-values	Components (M, M+1)			
	1, 2	2, 3	3, 4	4, 5
Arbitrary $\Theta^{(0)}$	–	0	0.020	0.32
Contrained Θ	–	0	0.015	0.28

A bootstrapped Kolmogorov-Smirnov goodness-of-fit test was applied to mixture models with M =2 to 6 components. The mean of these realizations produced the p-values presented in Table 3.15.

Table 3.15: P-values computed using the mean of 1,000 realizations of the bootstrapped Kolmogorov-Smirnov goodness-of-fit test statistic.

p-values	Components (M)				
	2	3	4	5	6
Arbitrary $\Theta^{(0)}$	0.034	0.086	0.055	0.818	0.970
Contrained Θ	0.029	0.091	0.646	0.735	0.783

When the likelihood ratio statistic was computed using 1,000 realizations from a parametric bootstrap procedure a mixture of normal distributions with 4 components was selected. A Kolmogorov-Smirnov goodness-of-fit test was also performed to verify

that the models fitted as mixtures of normal distributions were statistically identical to the distribution generating the data. The Kolmogorov-Smirnov goodness-of-fit test showed that fitted normal mixture models with 3 or more components were statistically identical to the distribution generating the data. These results suggest the 4 component normal mixture model fit with a constrained parameter in Θ should be selected. The selected model is of the form

$$X \sim \boldsymbol{\alpha} \cdot (X_1, X_2, X_3, X_4)' \tag{3.1}$$

where $X_1 \sim \mathcal{N}(0.0484, 0.0218)$, $X_2 \sim \mathcal{N}(0.367, 0.155)$, $X_3 \sim \mathcal{N}(1.083, 0.570)$, $X_4 \sim \mathcal{N}(5.386, 5.341)$, and $\alpha \sim Mult(1, \pi)$ with $\pi = (0.185, 0.595, 0.183, 0.037)$. Where $multi(1, \pi)$ refers to a multinomial distribution that has a probability mass function

$$f(x_1, \ldots, x_k) = P(X_1 = x_1 \ldots X_k) = \begin{cases} \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \cdots p_k^{x_k} & \text{when } \sum_{i=1}^{k} x_i = n \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

as defined in DeGroot & Schervish (2001). Standard errors for the parameters of the model in (3.1) were estimated using 1,000 bootstrapped realizations. Histograms for α_i, μ_i, and σ_i are presented in Figures 3.8, 3.9, and 3.10, respectively. Table 3.16 presents the standard errors for each parameter. The small standard errors suggest the parameter estimates are reliable.

Table 3.16: Standard errors for the parameter estimates for the fitted normal mixture model with M=4 components. The model has a constrained parameter in Θ.

Standard Errors	Component			
	X_1	X_2	X_3	X_4
α	0.00031	0.00049	0.00042	0.00014
μ	3.9771e-5	0.00016	0.00147	0.0192
σ	3.574e-5	1.5401e-4	8.5974e-4	1.308e-2

3.4.1 Summary

Model selection and validation procedures produced a 4 component normal mixture model. The model was fit by using a single fixed parameter in Θ where $\mu_1 = 0.0484$ and $\sigma_1 = 0.0218$. This model was selected using a likelihood ratio test where the likelihood ratio statistic was produced with 1,000 parametric bootstrap realizations. After the model was selected it was shown to be identical to the distribution generating the data. This was done through a Kolmogorov-Smirnov goodness-of-fit test. The test statistic was also computed using 1,000 parametric bootstrap realizations. Finally, the standard errors of the final models parameters were computed using a bootstrap. The small standard errors presented in Table 3.17 show that uncertainty in the parameter estimates is small.

The selected model for claim duration X is

$$X \sim \boldsymbol{\alpha} \cdot (X_1, X_2, X_3, X_4)' \tag{3.3}$$

where $X_1 \sim \mathcal{N}(0.0484, 0.0218)$, $X_2 \sim \mathcal{N}(0.367, 0.155)$, $X_3 \sim \mathcal{N}(1.083, 0.570)$, $X_4 \sim \mathcal{N}(5.386, 5.341)$, and $\boldsymbol{\alpha} \sim Mult(1, \pi)$ with $\pi = (0.185, 0.595, 0.183, 0.037)$.

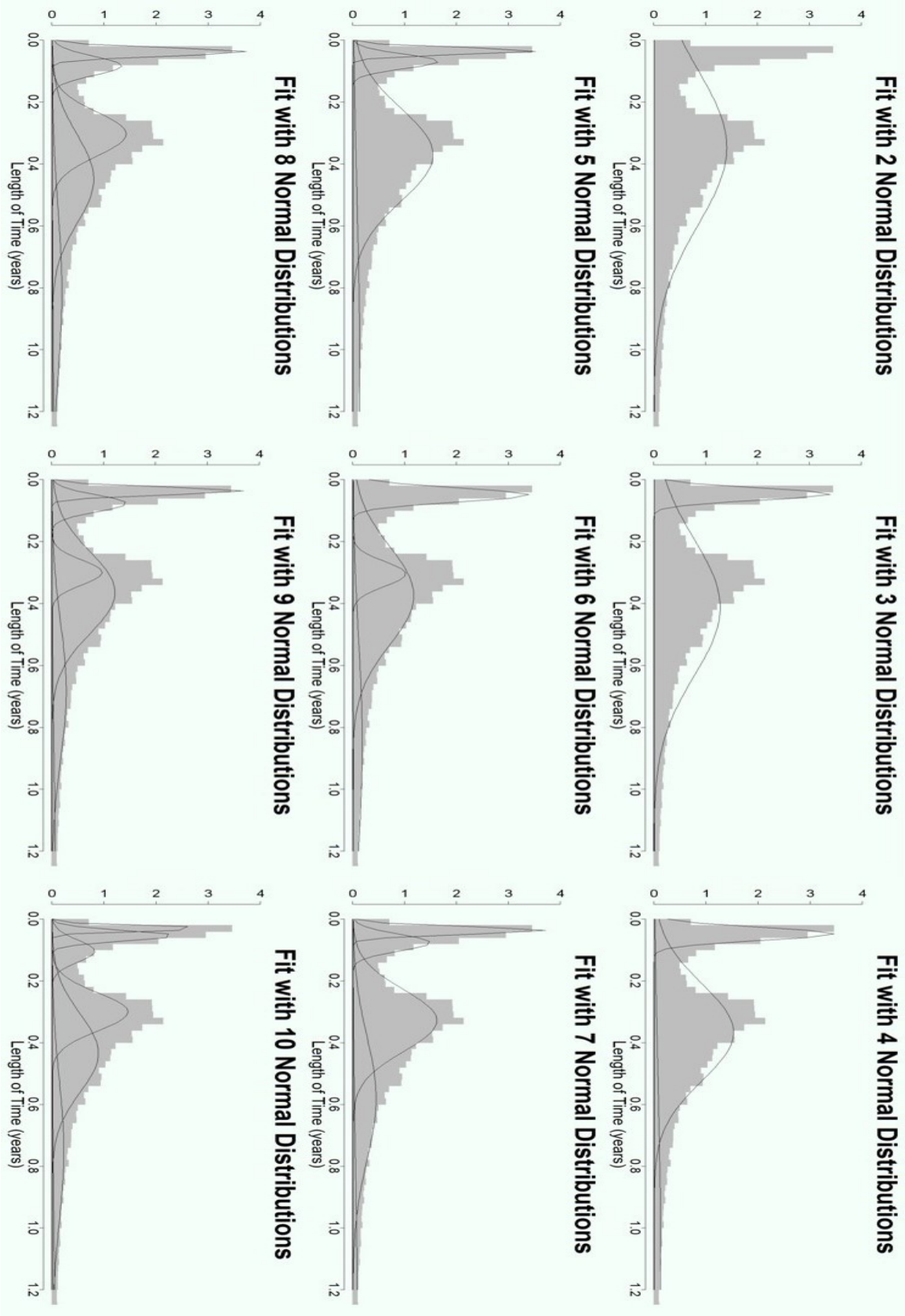

Figure 3.3: Density plots of the components of the mixutre models produced by the EM algorithm with arbitary intial values of Θ. The number of components range from 2 to 10. The data from Figure 2 is presented in grey.

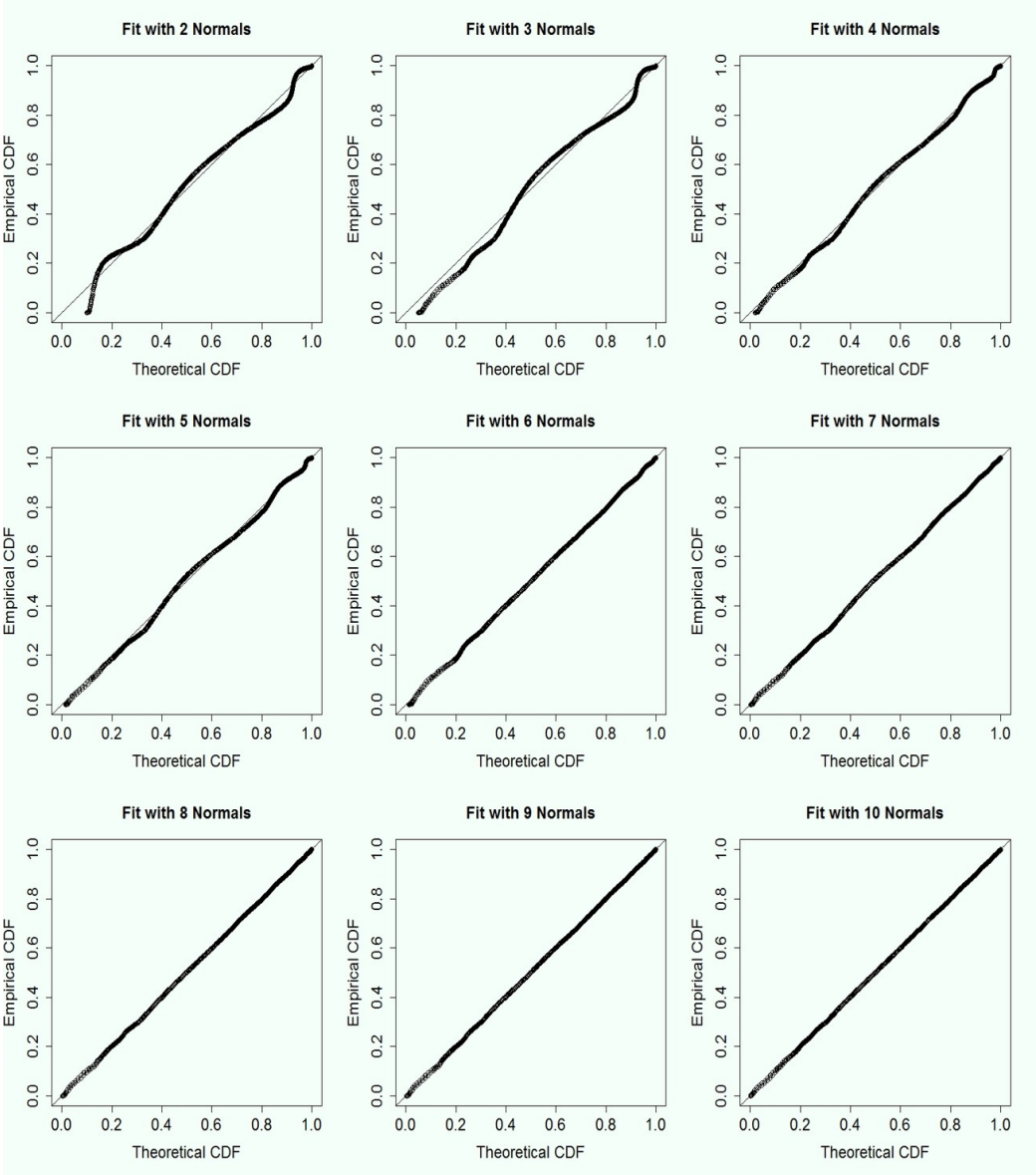

Figure 3.4: Plots of the empirical cumulative distribution function of the data versus the theoretical cumulative distribution function of the fitted models. A line for $y = x$ is also plotted to aid interpretation.

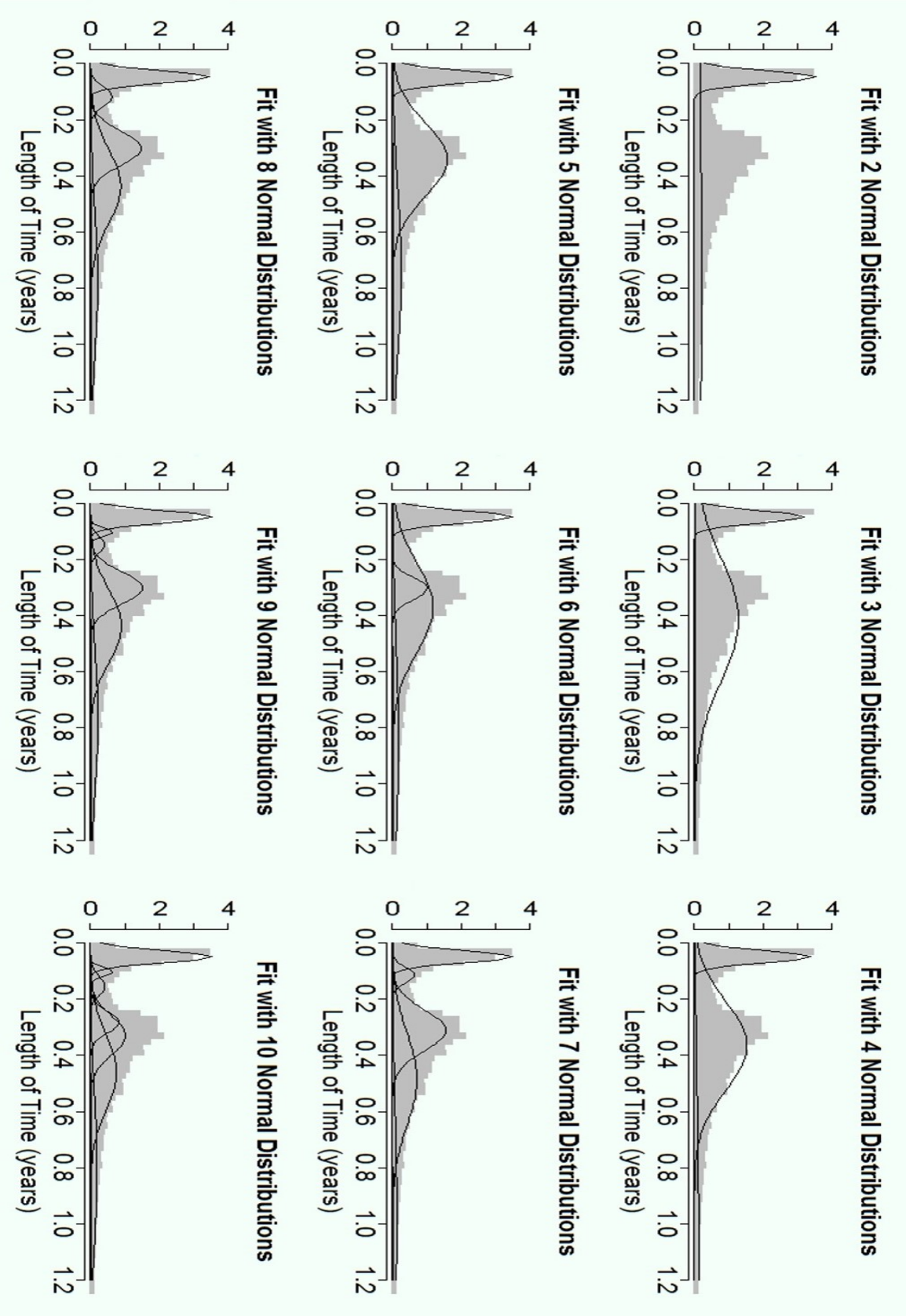

Figure 3.5: Density plots of the components of the mixutre models produced by the EM algorithm with a constrained set of parameters in Θ ($\mu_1 = 0.0484$ and $\sigma = 0.0218$). The number of components range from 2 to 10. The data from Figure 2 is presented in grey.

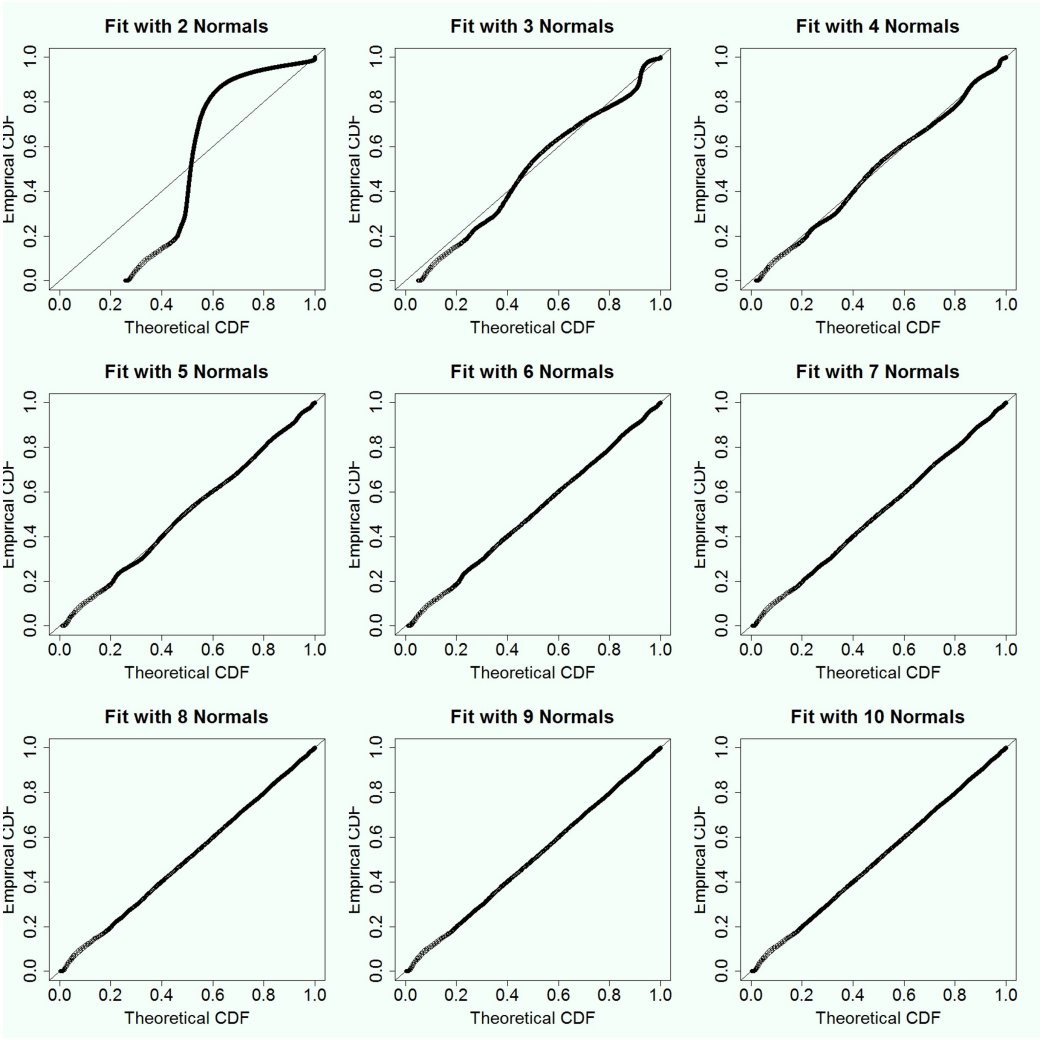

Figure 3.6: Plots of the empirical cumulative distribution function of the data versus the theoretical cumulative distribution function of the fitted models. A line for $y = x$ is also plotted to aid interpretation.

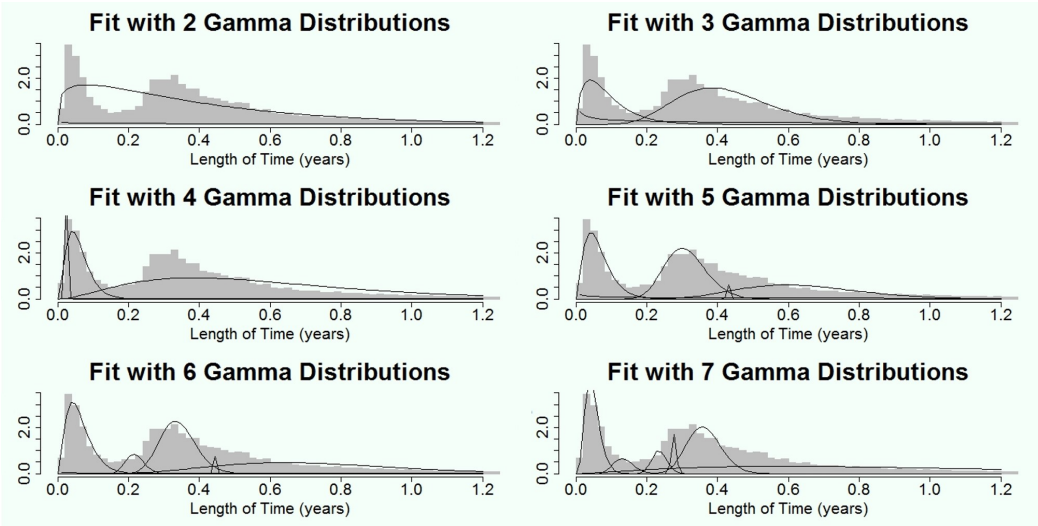

Figure 3.7: Plots of the histogram of the data with curves for the fitted gamma mixture models. The plots are for gamma mixtures with 2 to 7 components.

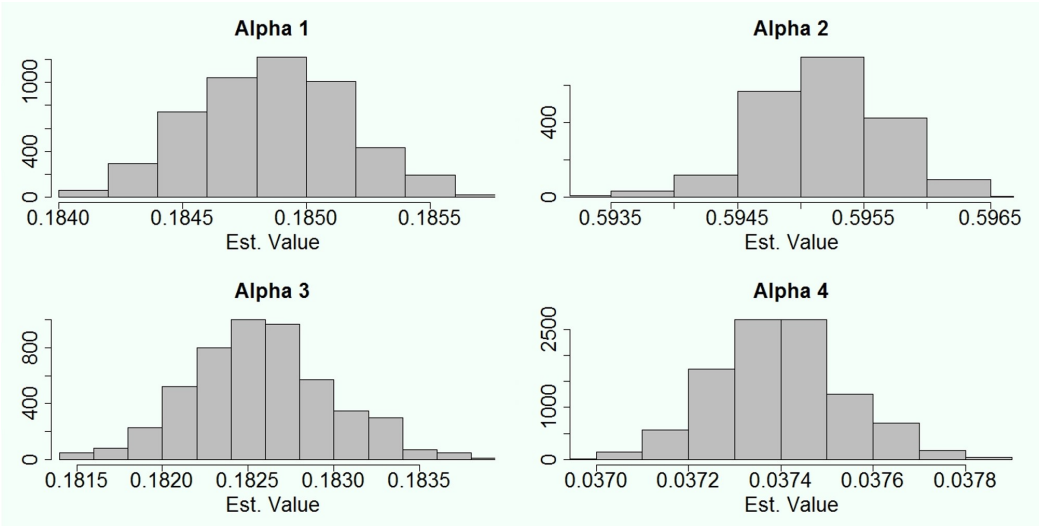

Figure 3.8: Histograms of 1,000 realizations from a parametric bootstrap on the weighting parameters α.

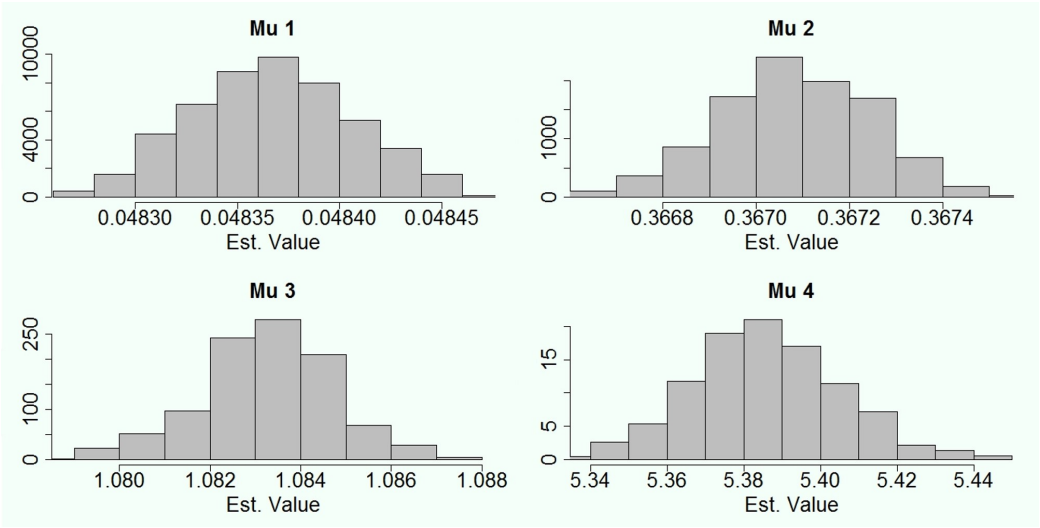

Figure 3.9: Histograms of 1,000 realizations from a parametric bootstrap on μ.

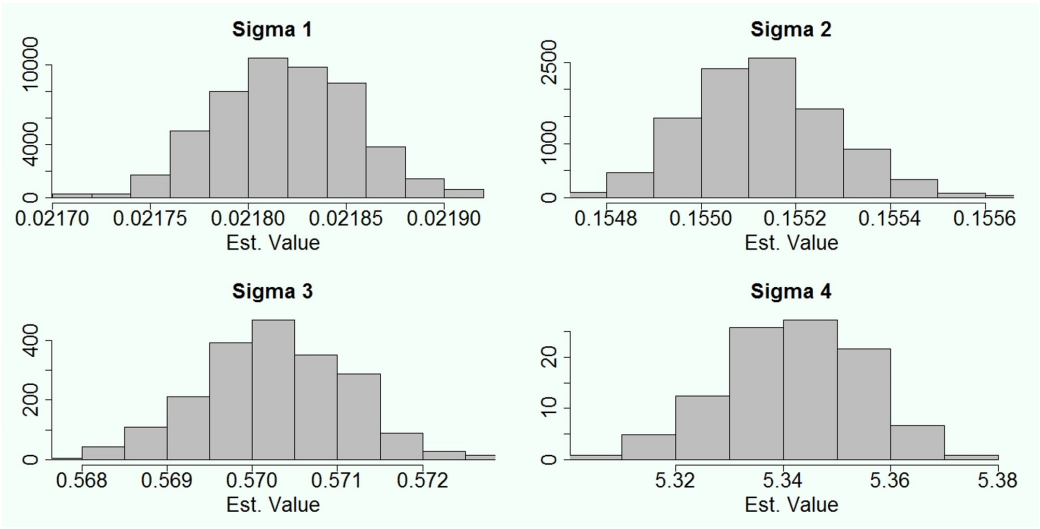

Figure 3.10: Histograms of 1,000 realizations from a parametric bootstrap on σ.

Chapter 4 Conclusions and Future Work

4.1 Conclusions

The questions addressed in this work were:

1. What are the general statistical properties of the data such as the mean, median, range, and variance?

2. Does the distribution of data exhibit any skewness?

3. What is the general distribution of the data?

4. Are there any interesting features of the distribution such as gaps or multiple modes?

5. Does extent of injury or claim type help explain any of the interesting properties of the distribution of the data such as gaps?

6. How can we model the data, and which is "the best" model?

The first four questions were answered through exploratory data analysis. The mean duration of time claims remained open for the entire data set was 1.38 years and the variance was 47.833. The non-zero minimum and maximum were 0.0027 and 69.11 years, respectively. The median was 0.35 years. The mean being greater than the median revealed that the data exhibited positive skew. Histograms examined in the exploratory data analysis section indicated that the general distribution of the data was likely a mixture distribution. This was believed due to the multiple modes and a gap at approximately 55 years.

The fifth question was answered through log linear analysis and modeling. χ^2 tests of independence were performed on the three-way contingency table and the two-way contingency tables for extent of injury, claim type, and the indicator for claims being open longer than 55 years. The tests of independence revealed that dependence between the explanatory variables existed in all cases. A backward selection process using the AIC for the log linear model of the three-way contingency table was used to determine which interactions were significant. The process revealed that all main effects and two-way interaction for extent of injury, claim type, and the indicator for claims open longer than 55 years were significant. The interpretation is that claims being open more or less than 55 years are dependent on both extent of injury and claim type. A significant association also exists between claim type and extent of injury. Only the three way interaction was determined to be insignificant in the log linear model. Thus, the selected log linear model was the saturated model less the three-way interaction term. The p-value for the deviance statistic of the model was 0.04848, which rejects the null hypothesis that the model fits well at the 0.05 significance level. Residual analysis was performed using a dissimilarity index. The analysis revealed 0.000003% of the observations (about 7) would need to be altered to obtain a perfect fit.

Finally, we answered the last question in the list. It was determined that the general distribution of the data could be fit well with normal and gamma mixture models. Normal and gamma mixture models were fit using EM algorithm functions in the `mixtools` library in R. Log-likelihood values and AIC values were used to determine whether the normal or gamma mixture models were to be preferred. The results indicated that a normal mixture model with a single set of constrained parameters was preferred. Model selection was done using 1,000 realization of the bootstrapped likelihood ratio statistic. The result of the test was that a normal mixture model with 4 components was optimal. After the model was selected a bootstrapped Kolmogorov-Smirnov goodness-of-fit test was performed to verify that the 4 component normal

mixture model fit the population distribution that generated the data. The result of the test indicated that the 4 component normal mixture model fit the population distribution that generated the data. Finally, a bootstrap was performed to obtain the standard errors of the parameter estimates. The standard errors of the parameter estimates for the selected 4 component normal mixture model indicated that there was little uncertainty in the parameter estimates. The selected model for claim duration X was:

$$X \sim \boldsymbol{\alpha} \cdot (X_1, X_2, X_3, X_4)' \tag{4.1}$$

where $X_1 \sim \mathcal{N}(0.0484, 0.0218)$, $X_2 \sim \mathcal{N}(0.367, 0.155)$, $X_3 \sim \mathcal{N}(1.083, 0.570)$, $X_4 \sim \mathcal{N}(5.386, 5.341)$, and $\alpha \sim Mult(1, \pi)$ with $\pi = (0.185, 0.595, 0.183, 0.037)$.

The advantage of knowing the analytical form of the distribution of claim duration is to be able to approximate the probability of a claim being open longer than a given duration of time. This is done using Monte Carlo simulation. As an example we simulated 100,000 realizations of the distribution presented in (3.3), we call it $\boldsymbol{x}$, and looked at the mean of $I(\boldsymbol{x} > 5yrs)$. In doing so we determined that the probability of a claim being open longer than 5 years is approximately 0.02.

4.2 Future Work

Analysis for future work includes looking at the relationship between claim duration and loss amount. This could be done in several ways. One way would be by looking at the conditional distributions of claim duration and loss amount with respect to extent of injury and claim type. It may be that the conditional distributions are statistically different, and if so this would explain variability in both claim duration and loss data. In addition to this a hierarchical cluster analysis could be used to determine which combinations of claims have duration and losses that are most similar. In both

cases the results could be used by the claims department to make initial estimates of losses based on the known characteristics of a claim. This information would assist in reserving funds for future claim payments.

Bibliography

[1] Agresti, A. (2002). Categorical Data Analysis, 2nd Ed. Wiley. New York, New York.

[2] Almhana, J., Liu, Z., Choulakian, V., McGorman, R. (2006). A recursive algorithm for gamma mixture models. *IEEE International Conference on Communications*, **1**: 197-202.

[3] Blimes, J. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. International Computer Science Institute. Available: `http://melodi.ee.washington.edu/people/bilmes/mypapers/em.pdf`. Accessed: Aug. 15, 2015.

[4] Christensen, R. (1997). Log-Linear Models and Logistic Regression, 2nd Ed. Springer Texts in Statistics. Springer-Verlag. New York, New York. ISBN: 978-1-4757-7113-8.

[5] D'Agostino, R.B. & Stephens, M.A (1986). Goodness-of-Fit Techniques. Marcel Dekker, INC: New York, New York.

[6] DeGroot, M.H. & Schervish, M.J. (2001). Probability and Statistics, third edition. Addison-Wesley. ISBN 978-0-201-52488-8.

[7] Glosup, J.G. & Axelrod, M.C. (1994). Use of the AIC with the EM algorithm: A demonstration of a probability model technique. Lawrence Livermore

National Laboratory. Joint Statistical Meeting, Toronto, Canada, August 15-18.

[8] Gupta, M.R. & Chen, Y. (2010). Theory and use of the EM algorithm. Foundation and Trends in Signal processing, **4**(3): 223-296.

[9] Kuha, J. and Firth, D. (2011). On the index of dissilimarity for lack of fit in loglinear and log-multiplicative mdoels. *Computational Statistics & Data Analysis*. **55**(1): pg 375-388.

[10] Leone, F.C., Nelson, L.S., and Nottingham, R.B. (1961). The folded normal distribtution. *Technometrics*. **3**(4): pg 543-550.

[11] Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986). Akaike information criterion statistics. D. Reidel Publishing Company, Dordrecht, The Netherlands.

[12] Schwander, O. & Nielsen, F. (2013). Fast learning of gamma mixture models with k-MLE. *Springer Berlin Heidelberg*, **7**(953): pg 235-249.

[13] Sundberg, R. (1971). Maximum likelihood theory and applications for distributions generated when observing a function of an exponential family variable. Doctoral dissertation. Institute of Mathematical Statistics, Stockholm University.

[14] Sundberg, R. (1974). Maximum likelihood theory and applications for incomplete data from an exponential family. *Scandinavian Journal of Statistics*. **1**(2): pg 49-58.

[15] Sundberg, R. (1976). An interative method for solution of the likelihood equations for incomplete data from exponential families. *Communications in Statistics - Simulation and Computation*. **5**(1): pg 55-64.

Appendix A EM Algorithm Parameter Estimates

Table A.1: Parameter estimates for normal mixture models

$\mathcal{N}$	Arbitrary $\Theta^{(0)}$			Constraint on Θ		
	$\hat{\alpha}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\alpha}$	$\hat{\mu}$	$\hat{\sigma}$
$M = 2$	0.8833	0.3466	2.7457	0.1920	0.0484	0.0218
	0.1167	0.2506	3.5603	0.8080	0.7637	1.5923
3	0.1686	0.0464	0.0197	0.1741	0.0484	0.0218
	0.7119	0.4132	0.2217	0.7062	0.4154	0.2203
	0.1195	2.7166	3.5214	0.1197	2.7140	3.5188
4	0.1822	0.0456	0.0210	0.1848	0.0484	0.0218
	0.1819	1.0865	0.5727	0.5952	0.3671	0.1551
	0.5986	0.3663	0.1562	0.1826	1.0834	0.5703
	0.0372	5.4008	5.3506	0.0374	5.3864	5.3415
5	0.1012	0.0345	0.0114	0.1911	0.0484	0.0218
	0.0966	0.0685	0.0233	0.5218	0.3429	0.1309
	0.5771	0.3704	0.1487	0.0781	2.0714	1.0219
	0.1864	1.0635	0.5549	0.1939	0.7468	0.2990
	0.0386	5.2914	5.2818	0.0151	8.7383	7.0831
6	0.1930	0.0493	0.0226	0.1903	0.0484	0.0218
	0.0666	2.3019	1.1143	0.1578	0.8537	0.3453
	0.4682	0.3749	0.1588	0.0991	0.3047	0.0391
	0.1015	0.3045	0.0395	0.4732	0.3729	0.1596
	0.1576	0.8538	0.3446	0.0665	2.3036	1.1155
	0.0131	9.4045	7.3701	0.0130	9.4113	7.3731
7	0.1181	0.0363	0.0127	0.1907	0.0484	0.0218
	0.1014	0.0761	0.0270	0.0442	0.1144	0.0273
	0.4055	0.3283	0.0998	0.2817	0.3114	0.0721
	0.2269	0.5807	0.2012	0.2894	0.4858	0.1617
	0.0432	2.9755	1.3443	0.1256	0.9894	0.3837
	0.0955	1.2198	0.4737	0.0568	2.5393	1.1945
	0.0095	11.0242	7.9902	0.0115	10.0129	7.6108

Table A.2: Parameter estimates for normal mixture models

$\mathcal{N}$	Arbitrary $\Theta^{(0)}$			Constraint on Θ		
	$\hat{\alpha}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\alpha}$	$\hat{\mu}$	$\hat{\sigma}$
$M = 8$	0.1033	0.1195	0.0308	0.1929	0.0484	0.0218
	0.1281	0.0378	0.0136	0.0226	0.1022	0.0138
	0.0276	3.7690	1.6130	0.0253	0.1417	0.0264
	0.2542	0.3031	0.0710	0.2557	0.3093	0.0653
	0.2858	0.4484	0.1422	0.3000	0.4727	0.1542
	0.1296	0.8158	0.2728	0.1320	0.9548	0.3692
	0.0641	1.6735	0.6290	0.0595	2.4700	1.1686
	0.0072	12.5666	8.5321	0.0120	9.3819	7.5390
9	0.1186	0.0366	0.0128	0.1537	0.0484	0.0218
	0.0988	0.0766	0.0274	0.0353	0.0292	0.0073
	0.4034	0.3648	0.1320	0.4591	0.3553	0.1490
	0.0703	1.2796	0.4354	0.0231	0.2692	0.0157
	0.0959	0.2996	0.0392	0.0627	0.3191	0.0323
	0.1563	0.6823	0.2235	0.1578	0.7347	0.2630
	0.0141	5.4880	2.3026	0.0702	3.5827	1.5536
	0.0383	2.5253	0.9247	0.0077	1.5611	0.5988
	0.0044	15.6288	9.5715	0.0304	12.2298	8.4200
10	0.0699	0.0296	0.0087	0.1941	0.0484	0.0218
	0.0988	0.0537	0.0166	0.0249	0.1032	0.0147
	0.0735	0.1034	0.0355	0.0318	0.1486	0.0297
	0.0120	6.1085	2.5796	0.2157	0.3002	0.0561
	0.2244	0.3001	0.0612	0.2766	0.4333	0.1200
	0.2848	0.4339	0.1276	0.1383	0.7334	0.2279
	0.1337	0.7507	0.2346	0.0662	1.3600	0.4598
	0.0647	1.3987	0.4746	0.0357	2.6804	0.9829
	0.0346	2.7622	1.0164	0.0126	5.8814	2.4762
	0.0037	16.7391	9.9495	0.0039	16.3284	9.8086

Table A.3: Parameter estimates for gamma mixture models

$\mathcal{G}$	Arbitrary $\Theta^{(0)}$		
	$\hat{\alpha}$	$\hat{\mu}$	$\hat{\sigma}$
$M = 2$	0.8992	1.2562	0.3167
	0.1008	0.7790	3.4309
3	0.2345	1.8052	0.0491
	0.5176	9.7836	0.0437
	0.2479	0.5322	4.2299
4	0.2142	3.2192	0.0186
	0.0387	42.405	0.0005
	0.6799	2.8597	0.2071
	0.0673	1.0087	4.7710
5	0.2376	2.7308	0.0236
	0.3217	27.401	0.0113
	0.0077	1.00e4	4.28e-5
	0.2622	13.369	0.0482
	0.1639	0.6756	4.4480
6	0.2449	2.8744	0.0219
	0.0548	70.234	0.0031
	0.2840	45.881	0.0074
	0.0063	1.84e4	2.41e-5
	0.3074	7.4362	0.0993
	0.1026	0.8451	4.6286
7	0.1996	5.3418	0.0090
	0.0462	22.585	0.0060
	0.0530	1.23e2	1.91e-3
	0.0294	1.60e3	1.74e-4
	0.2437	56.636	0.0064
	0.4007	2.4896	0.3853
	0.0274	1.4953	5.0030