MAXIMUM LIKELIHOOD ESTIMATION 0F PARAMETERS OF TRUNCATED AND CENSORED GAMMA AND EXPONENTIAL DISTRIBUTIONS USING THE EXPECTATION MAXIMISATION ALGORITHM

By

Namukolo Mukubesa

A dissertation prepared in partial fulfillment of the requirements for the degree of Master of Science in Statistics

THE UNIVERSITY OF ZAMBIA

LUSAKA

2024

COPYRIGHT

The copyright of this dissertation vests in the author. No quotation from it or information derived from it is to be published without full acknowledgment of the source. The dissertation is to be used for private study or non-commercial research purposes only.

Published by the University of Zambia (UNZA) in terms of the non-exclusive license granted to UNZA by the author.

DECLARATION

The work described in this Master of Science (MSc) dissertation was carried out under the supervision of Dr. Victor Nawa, Department of Mathematics and Statistics, University of Zambia, Lusaka.

The MSc dissertation represents original work by the author and has not otherwise been submitted in any form for any degree or diploma to any other University. Where use has been made of the work of others it is duly acknowledged in the text.

Signed:

....

Namukolo Mukubesa (Student)

....

Dr.Victor Nawa (Supervisor)

APPROVAL

This dissertation of Namukolo Mukubesa has been approved as fulfilling the requirements or partial fulfillment of the requirements for the award of Master of Science in Statistics by the University of Zambia.

Name of Examiner 1: ....

Signature:..

Date:...

Name of Examiner 2: ....

Signature:..

Date:...

Name of Examiner 3: ....

Signature:..

Date:...

Chairperson of Board of Examiners: ....

Signature:..

Date:...

ABSTRACT

This study aimed to determine the maximum likelihood (ML) estimates for the Gamma and Exponential distributions under conditions of left-truncated and right-censored data, employing the Expectation-Maximisation (EM) algorithm. Additionally, it sought to compare the performance of the EM algorithm with two alternative estimation techniques, namely the Newton-Raphson (NR) algorithm and the Stochastic Expectation-Maximisation (SEM) algorithm. Simulated left-truncated and right-censored data were generated from the Exponential and Gamma distributions. Performance comparisons among the algorithms were conducted based on simulations involving varying degrees of censoring, truncation, and sample sizes, utilizing metrics such as mean square error (MSE), bias, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). The validation of AIC and BIC was assessed under both candidate models (Gamma and Exponential distributions) using the EM algorithm. All computations were performed using R-4.3 version.

The EM algorithm consistently yielded estimates with low bias under moderate levels of censoring and truncation, indicating its accuracy in estimation. Comparison of the EM algorithm with the NR and SEM algorithms revealed similar estimates for larger sample sizes. In terms of convergence speed, both the EM and NR algorithms converged faster than the SEM algorithm. Moreover, as the sample size increased, bias and MSE decreased for all three algorithms.

Furthermore, the EM algorithm demonstrated superior performance in terms of MSE, bias, AIC, BIC, and convergence speed. However, in scenarios with high levels of censoring and truncation, all three algorithms encountered difficulties in accurately estimating parameters due to data loss.

Ultimately, the findings of this study are anticipated to advance statistical methodology and enhance the accuracy of parameter estimation techniques for censored and truncated data scenarios.

Key words : Truncation, censoring, left-truncated, right-censored, maximum likelihood estimation, expectation maximisation algorithm, stochastic expectation algorithm, Newton-Raphson algorithm, likelihood function, simulations, Gamma and Exponential distribution.

DEDICATION

I humbly dedicate this work to three exceptional groups of people: the females, orphans, and the determined.

To the females, I want to emphasize that there are no boundaries to how far you can reach in life. Your journey is a testament to your remarkable resilience, dedication, unwavering focus, and unyielding persistence in pursuing your dreams and aspirations.

To the orphans, our stories may have been marked by the absence of parents, but our strength lies in how we have triumphed over adversities. Despite facing numerous limitations, we have persevered and succeeded, proving that our experiences do not define us, but rather, it is our determination and accomplishments that truly shine.

To the determined souls, I offer my encouragement and unwavering support. This dedication is a testament to your relentless spirit and refusal to surrender to life's challenges. As you continue pushing forward, remember that success lies just beyond the horizon of perseverance. Set your sights on your dreams, for with your unwavering determination, you are destined to achieve them.

Finally, I dedicate this work to the cherished memory of my late mother, Alice Wakumelo Indopu. Though she is no longer with us, her spirit lives on in the pursuit of my dreams. She would have cherished seeing her second youngest daughter embrace her lifelong passion and embark on this fulfilling journey. This dedication is a tribute to her love, encouragement, and unwavering belief in my potential.

May this work serve as a beacon of hope and inspiration to all, reminding us that with determination, courage, and resilience, we can overcome any obstacle and achieve greatness in our lives.

ACKNOWLEDGMENTS

I am deeply humbled and sincerely grateful to the Almighty God for bestowing upon me good health and unwavering energy throughout my academic journey.

I extend my heartfelt appreciation and indebtedness to my esteemed supervisor, Dr. Nawa, whose invaluable mentorship, unyielding support, and exceptional expertise have been instrumental in shaping my academic growth and success. His guidance has been a beacon of light, leading me through the intricacies of research and academia.

My profound gratitude goes to Mr. Alick Banda for his generous sharing of knowledge and expertise in R-programming. His guidance has significantly enriched my understanding and proficiency in data analysis, a skill that will undoubtedly prove invaluable in my future endeavours.

I would like to express my heartfelt gratitude to Fr Kalaluka for his unwavering support and guidance throughout my journey. His presence and encouragement have been invaluable to me, and I am truly thankful for the mentorship and assistance he has provided.

To the lecturers in the mathematics and statistics department, I am immensely thankful for their unwavering dedication, passion, and the wealth of knowledge they imparted to me. Their commitment to nurturing students' intellectual growth has played a pivotal role in shaping me into the professional I am today.

I would also like to express my sincere thanks to EAUMP for the partial sponsorship, which eased the financial burden and allowed me to focus on my studies with determination and enthusiasm.

A special note of gratitude goes to Mr. Mwewa for his continuous encouragement and unwavering inspiration throughout this journey. His belief in my abilities has been a driving force in pushing me towards excellence.

To my dearest friends Memory, Mercy, Mannah, Caren, Twiza, Hannah, Watson and Nephat, your unwavering support during my studies has been a source of strength and motivation.

I am truly fortunate to have friends like you who have been there through thick and thin. Additionally, I extend my thanks to my colleagues Malakata, Nyirongo, Keith, Derrick, Agness, Bornwel, Phizzy, Carols, Emmanuel, Zielesa, Dominique, and Mr. Michael for their motivation and camaraderie.

Lastly, my heartfelt appreciation goes to my family for their unwavering prayers, constant support, and encouraging words. Their love and belief in me have been the pillars of strength that kept me going, even during the most challenging times.

To all those who have played a part in my academic journey, I am eternally grateful. Your contributions have left an indelible mark on my life, and I will carry the lessons and memories with me as I step into the future with confidence and gratitude. Thank you, from the bottom of my heart.

TABLE OF CONTENTS

Index of Notation

Below is a list of symbols that will be frequently used and a brief indication of their meaning.

MLE	Maximum likelihood estimate
ML	Maximum likelihood
SEM	Stochastic expectation maximisation
MM	Method of the moments
EM	Expectation maximisation
NR	Newton-Raphson
Trunc(%)	Truncation percentage
Cens(%)	Censoring percentage
MSE	Mean squared error
B	Bias
ALGO	Algorithms
AIC	Akaike information criterion
BIC	Bayesian information criterion
NI	number of iterations
Accu	Accuracy
LS	Least Squares
UMVUE	Unique minimum variance unbiased estimators
$\arg\max_{\boldsymbol{\theta}}$	Argument of maximum

List of Tables

CHAPTER 1

Introduction

This chapter gives the background of the study by introducing the exponential and gamma distributions, the statement of the problem, aim of the study, objectives of the study, research questions and significance of the study.

1.1 Background

The Exponential and Gamma are the commonly used distributions for modelling lifetime data. An exponential distribution is a continuous life time random variable and it is mainly used in reliability studies. The Gamma distribution is an extension of the Exponential distribution and it has two parameters.

The probability density function (pdf) for an Exponential random variable X taking values x in the interval $[0, \infty)$ with parameter $\lambda > 0$ is given by

$$f(x;\lambda) = \begin{cases} \dfrac{1}{\lambda}\, e^{-x/\lambda}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

The pdf for a gamma random variable X taking values of x in the interval $[0, \infty)$ with parameters $\alpha > 0$ and $\beta > 0$ is given by

$$f\,(x;\alpha,\beta) = \begin{cases} \dfrac{x^{\alpha-1}\, e^{-x/\beta}}{\beta^{\alpha}\,\Gamma(\alpha)}, & x > 0 \\ 0, & \text{otherwise} \end{cases}$$

where $\Gamma(z) = \int_0^\infty t^{z-1}\, e^{-t}\, dt$ is the Euler's Gamma function, $\alpha > 0$ is the shape parameter and $\beta > 0$ is the scale parameter (Johnson et al., 1994).

Estimation of parameters, which represent a population, based on the information contained in a sample is one of the fundamental problems in mathematical statistics. Estimating the parameters of a gamma and exponential distributions based on truncated and censored data is a common problem in life testing, quality control and other fields. Thus, in a case where the data is incomplete, truncated or censored distributions are used to estimate the parameters as opposed to the parent distributions. If we ignore truncation or censoring when analysing data and treat the incomplete data as complete, our estimates of population parameters will be biased and inconsistent (Emura and Michimae, 2022).

Maximum likelihood (ML) estimation is one of the methods which is used most extensively in statistical literature due to its flexibility, simplicity and optimality. However, computational burdens may arise when it comes to estimating parameters when the data is incomplete or hidden. In an event where data is incomplete, various methods such as the Newton-Raphson (NR), Fisher scoring, stochastic expectation maximisation (SEM) and Bayesian methods have been applied in obtaining ML-estimates of parameters (Mittal, 1984). The goal of this study is to estimate the parameters of the gamma and exponential distributions for left-truncated and right-censored data and also make comparisons of the performance of the expectation maximisation (EM) algorithm and other algorithms such as (NR and SEM).

1.2 Problem Statement

There are many instances where the analytical form of the ML-estimates cannot be obtained especially when the data is truncated or censored, hence numerical methods are required to find the estimates. ML-estimates can be difficult to find in the presence of hidden or incomplete data, and ignoring truncated and censored observations results in loss of potentially valuable information. It is therefore important to carry out an analysis which will incorporate the fact that the data is incomplete. A number of studies involving ML-estimation have been done on truncated or censored gamma and exponential distributions using various methods such as Newton-Raphson, Fisher scoring and Bayesian approaches. However, there is limited work on the ML-estimates for gamma and exponential distributions for left-truncated and right-censored data using the EM-algorithm. Hence this study estimated the ML-estimates for the gamma and exponential distributions for left-truncated and right-censored data using EM-algorithm, and compared the performances of the EM-algorithm with other commonly used algorithms (Newton-Raphson and SEM).

1.3 Aim of the Study

The aim of this study was to find the ML-estimates for the gamma and exponential distribution parameters for left-truncated and right-censored data using EM-algorithm, and compare the EM-algorithm with other commonly used algorithms.

1.4 Objectives of the Study

The objectives of the study were to:

(a) find the ML-estimates for exponential distribution for left-truncated and right-censored data using the EM-algorithm.

(b) find the ML-estimates for gamma distribution for left-truncated and right-censored data using the EM-algorithm.

(c) compare the performance of the EM-algorithm and other algorithms (Newton-Raphson and SEM) for the left-truncated and right-censored data for gamma and exponential distributions.

1.5 Research Questions

The following were the research questions

(a) How is the ML-estimate of the parameter of the left-truncted and right-censored Exponential distribution obtained using the EM-algorithm?

(b) How are the ML-estimates of the parameters of the left-truncted and right-censored Gamma distribution obtained using the EM-algorithm?

(c) How does the EM-algorithm compare with other algorithms such as NR and SEM in terms of mean square error (MSE), bias, Akaike information criterion (AIC) and Bayesian information criterion (BIC)?

1.6 Significance of the Study

The significance of this study lies in its potential to offer valuable guidelines for finding the ML-estimates of the parameters of the left-truncted and right-censored Exponential and Gamma distributions in situations where censored and truncated data are encountered. By employing the EM-algorithm and comparing it with other well-known algorithms, such as

NR and SEM, for finding ML-estimates ,this research will pave the way for more accurate and robust parameter estimates in the presence of incomplete data. The findings from this study will contribute to the advancement of statistical methodology and improve the accuracy of parameter estimation techniques for censored and truncated data scenarios. This is particularly important in real-world applications where data incompleteness is a common occurrence, such as in medical research, engineering, finance, and social sciences.

The study's outcomes can directly impact the decision-making process in various fields by providing more reliable and robust parameter estimates for gamma and exponential distributions. Researchers, practitioners, and analysts can apply the recommended guidelines to better understand and interpret data, leading to more informed conclusions and more accurate predictions.

Furthermore, this research may foster the adoption of the EM-algorithm as an effective tool for dealing with incomplete data in the context of maximum likelihood estimation. By highlighting the strengths and weaknesses of different algorithms, this study can inform researchers about the most appropriate approach for their specific data scenarios.

1.7 Limitations of the Study

- The study faced the limitation of not having the real life data which was left-truncated and right-censoring that followed the Gamma and Exponential distribution.
- The study faced the limitation in determining appropriate initial values for the statistical models used. This limitation can impact the accuracy and convergence of the estimation process, potentially leading to biased results. Further exploration and refinement of the initial value selection process would enhance the robustness of future analyses. The presence of substantial censoring and truncation in the data posed challenges for parameter estimation.

CHAPTER 2

Literature Review

Parameter estimation refers to the process of determining the values of unknown distribution parameters based on sample data. Various methods exist for estimating these parameters, including maximum likelihood (ML), least squares (LS), and method of moments (MM). In our study, we focused on ML-estimation due to its simplicity, flexibility, and optimality. While ML-estimation may have limitations when dealing with incomplete or hidden data, researchers have extensively explored censored and truncated distributions.

One notable development in addressing incomplete data was introduced by Hartley (1958). He devised a method applicable to any discrete distribution, effectively transforming the problem of ML-estimation for incomplete data into that of complete data. This transformation was achieved through the introduction of pseudo frequencies using a scheme of proportional allocation for missing or grouped classes. The iterative procedure provided by Hartley required no special aid tables for solving the estimating equations. Approximations of the asymptotic variances and covariances of distribution parameter estimates were obtained using certain first-order divided differences involving derivatives of the log-likelihood function (Hughes, 1962). While Hartley's method was originally designed for discrete distributions, Hughes (1962) extended it to both discrete and continuous cases, encompassing distributions such as Poisson, Gamma, Exponential, Binomial, Normal, Negative Binomial, Geometric, and Uniform. However, it should be noted that their approach only considered either truncated or censored data, not both as in our case.

Cohen (1950) examined the estimation of parameters for a truncated Gamma distribution using the method of moments. Due to the lack of closed-form expressions for the moments of the truncated Gamma distribution, obtaining moment estimates proved to be tedious, and no effort was made to evaluate their variances or efficiency. Des (1953) focused on estimating the mean and variance of Type III populations from singly and doubly truncated samples with known truncation points. He considered various types of incompleteness in his study, including truncation and censorship in single tails, truncation and censorship in both tails separately, and combined truncation and censorship in both tails. In our study, we employed the EM-algorithm on a dataset that was left-truncated and right-censored for both the Gamma and Exponential distributions.

Broeder (1955) investigated the case of a gamma distribution truncated in a single tail and employed the method of maximum likelihood to derive an estimating equation for the scale parameter when the other parameter was known. Broeder's work addressed each specific case of incompleteness considered, providing valuable insights into parameter estimation for truncated distributions.

Balakrishnan and Mitra (2012) made significant contributions to likelihood inference through the EM-algorithm for fitting several important lifetime distributions to left-truncated and right-censored data. Their study explored distributions such as generalized gamma, Weibull, log-normal, extreme value, and gamma distributions. While they primarily utilized the EM-algorithm, our study extends their work by comparing the EM-algorithm to other algorithms, namely Newton-Raphson (NR) and SEM algorithms. Gupta and Chen (2010) focused on the theory and application of the EM-algorithm, specifically in estimating Gaussian mixture models (GMMs) and hidden Markov models (HMMs). However, their study solely centered on the EM-algorithm, while our research seeks to expand the comparison to include other algorithms like SEM and NR. Johnson et al. (2014) explored maximum likelihood estimation from the left-truncated gamma distribution and compared ML-estimates to method of moment (MM) estimates. While their study exclusively dealt with left-truncated data, our study incorporates both left-truncated and right-censored data.

Shang (2021) conducted research on parameter estimation for the generalized gamma distribution based on left-truncated and right-censored data. In his study, he specifically focused on estimating the three-parameter generalized gamma distribution using the SEM algorithm. The study compared the SEM algorithm to existing estimation procedures based on the maximum likelihood (ML) method, such as the direct optimization method, profile likelihood method, and EM-algorithm, considering aspects of accuracy and stability. In contrast, our study compare the estimates obtained using EM-algorithm with SEM and NR algorithms in terms of mean squared error (MSE), bias, accuracy, AIC, and BIC.

Deemer and Votaw (1955) explored maximum likelihood estimation of parameters for truncated or censored exponential distributions encountered in bombing accuracy studies. Clark and Williams (1960) focused on ML-estimation of the scale parameter for an exponential distribution censored in the upper tail. Muslim (2008) considered the estimation of the mean of a truncated exponential distribution, deriving ML-estimates, unique minimum variance unbiased estimators (UMVUE), and providing modifications for ML-estimates, asymptotic variances, and confidence intervals. However, Muslim's study did not consider the EM-algorithm or SEM algorithm for estimating parameters for truncated exponential distributions. Furthermore, Muslim's analysis only considered truncated data, unlike our study, which incorporates both truncation and censoring.

Faris (2008) examined the estimation of the mean and unique minimum variance unbiased estimators of the truncated exponential distribution. He provided a modification for the maximum likelihood estimator, asymptotic variances, and asymptotic confidence intervals for the estimators.

Mittal (1984) addressed the problem of estimating parameters of truncated distributions (normal, exponential, gamma, and Weibull) using Bayesian methods. He proposed a new estimation procedure, the mixed estimator, combining the ML-estimator and the Bayes modal estimator. Emura and Shiu (2016) studied the estimation and model selection for left-truncated and right-censored lifetime data, comparing the EM-algorithm to the Newton-Raphson algorithm in terms of convergence performance. They also examined the performance of Akaike's information criterion (AIC) for model selection via simulations. However, our study extends this comparison to include the SEM-algorithm and applies it to the gamma and exponential distributions.

Fisher (1931) considered the normal distribution truncated in a single tail, showcasing that moment estimates for this problem are identical to the maximum likelihood estimates. Hald (1949) investigated the normal distribution censored or truncated in a single tail, deriving ML-estimating equations by transforming observations to have the point of truncation at the origin. Gjeddebaek (1949) focused on the completely censored normal distribution, employing ML to obtain estimating equations for the parameters. Chandrakant et al. (2019) examined the truncated normal distribution under progressive type II censoring, utilizing Bayesian approaches to obtain point and interval estimates of unknown parameters. However, our study uses the EM-algorithm on the exponential and gamma distributions to obtain ML-estimates for both distributions.

Balakrishnan and Mitra (2011) explored likelihood inference for log-normal data with left truncation and right censoring. They used the EM-algorithm to find ML-estimates of the lognormal model based on left-truncated and right-censored data. They evaluated the maximisation step of the EM-algorithm using two methods: Taylor series expansion approximation and the EM-gradient algorithm, comparing them through Monte Carlo simulation. Our study extends their work by comparing the EM-algorithm to the SEM algorithm in terms of MSE, bias, accuracy, AIC and BIC.

Mitra et al. (2021) focused on likelihood analysis and the SEM-algorithm for left-truncated and right-censored data from the Lehmann family of life distributions. They obtained ML-estimates, constructed asymptotic confidence intervals using the missing information principle, and assessed the effect of model misspecification within the Lehmann family of distributions through Monte Carlo simulations. Our study expands upon their research by applying it to the exponential and gamma distributions and comparing the SEM, EM, and NR algorithms. Additionally, our study employs AIC and BIC for model selection, differing from their likelihood-based and distance-based methods.

In our study, our primary objective was to obtain ML-estimates for the gamma and exponential distributions using the EM-algorithm and other algorithms such as Newton-Raphson and SEM.

CHAPTER 3

Preliminaries

3.1 The Incomplete data

Incomplete data refers to a dataset in which one or more variables contain missing values, leading to limitations in the observations or measurements available for analysis. These missing values can occur for various reasons, such as data collection errors, data entry issues, or deliberate omissions.

In the realm of statistical analysis, encountering incomplete data is a common and critical challenge. Various situations can give rise to incomplete data, including missing values, grouped data, censored data, and truncated data. For the purposes of this study, we focused specifically on the incompleteness resulting from right-censored and left-truncated data. Left-truncated data is the type of data where observation below a threshold are excluded from the sample. For each observation, there is a lower limit such that the data point would not have been included in the sample if it was less than the lower limit. On the other hand, right-censored data refers to data in which the observation for an individual ends before the event of interest occurs. This means that the exact time of the event is unknown for these individuals, but it is known that the event has not occurred up to a certain point in time when the observation ends. Censoring and truncation are special features of time-to-event data that restrict our ability to observe the data to the fullest and this causes difficulties in statistical analysis.

3.1.1 Censoring

Censoring occurs when we do not know the exact time-to-event for an included observation. Types of censoring include the following - left censoring, right censoring, and interval censoring. A lifetime X associated with a specific individual in a study is considered to be left censored if it is less than a censoring time C, that is, the event of interest has already occurred for the individual before that person is observed in the study at time C. For such individuals, we know that they have experienced the event sometime before time C, but their exact event time is unknown. The exact lifetime X would be known if and only if $X \geq C$. The data from a left-censored sampling scheme can be represented by pairs of random variables (T, ϵ), where T is equal to X if the lifetime is observed and ϵ indicates whether the exact lifetime X is observed or not ($\epsilon = 1$ if X is observed, $\epsilon = 0$ if X is not observed).

Right censoring occurs if an event of interest happens after the end of the study. For instance if we assume that there is a lifetime X and a fixed censoring time C, then the exact lifetime X of an individual will be known if and only if $X \leq$C. Right-censored data arise frequently in ecological studies of event times. For example, some animals may engage in a behaviour of interest only after the end of the study. Interval censoring occurs when the time-to-event for an observation is between two values but the exact time-to-event is unknown.

3.2 Censored Distribution

Censored distribution is a statistical concept that arises when some of the observations or events are not fully observed or reported. In such cases, the observed data is said to be censored and the distribution of the observed data is referred to as the censoring distribution. Censoring occurs when all possible values of the variable of interest are not fully observed or recorded because of limitations in data collection or measurement.

In the context of time-to-event data analysis, censoring commonly occurs when the event of interest such as (failure, occurrence of an event) has not been fully realized for some subjects by the end of the study or observation period. As a result, the event time for these subjects is unknown, and instead, we have partial information available, such as knowing that the event has not yet occurred (right-censoring) or knowing that it occurred within a certain time interval (interval-censoring).

The censoring distribution represents the proportion or frequency of data points that are censored at different time points or intervals. Understanding the censoring distribution is crucial for accurate and unbiased statistical analysis, particularly in survival analysis, where the goal is to estimate the probability of an event occurring over time.

The general formula for the censoring distribution depends on the type of censoring present in the dataset. There are three common types of censoring in time-to-event data analysis: right-censoring, left-censoring, and interval-censoring. Each type of censoring has a different formula for its respective censoring distribution. Let's explore the general formulas for each:

Right-Censoring Distribution

Right-censoring occurs when some subjects in the dataset have their event times unknown because the event has not yet occurred by the end of the study or observation period. The right-censoring distribution represents the proportion or frequency of data points that are right-censored at various time points. The general formula for the right-censoring distribution $(F_{c_R}(t))$ is given by $(F_{c_R}(t)) = \Pr(\mathrm{T} > \mathrm{t})$ where, $(F_{c_R}(t))$ is the right-censoring distribution function, representing the probability that the true event time (T) is greater than t (the censoring time).

For example, in a study tracking the time until a light bulb burns out, researchers follow 50 bulbs for 100 days. By the end of the study, 20 bulbs have burned out, providing complete data on their burnout times. However, the remaining 30 bulbs are still functioning, and their burnout times are unknown. These 30 bulbs represent right-censored data points because their event times (burnout) haven't occurred by the end of the study period.

Left-Censoring Distribution

Left-censoring occurs when the event of interest has occurred before the start of the study or data collection, and we only have information that the event time is less than a certain value. Left-censoring is less common than right-censoring. The general formula for the left-censoring distribution $(F_{c_L}(t))$ is given by $(F_{c_R}(t)) = \Pr(\mathrm{T} < \mathrm{t})$ where, $(F_{c_R}(t))$ is the left-censoring distribution function, representing the probability that the true event time (T) is less than t (the censoring time).

For example, assessing medication efficacy for symptom improvement, researchers enroll patients who began medication use before the study's start. Patient A, having started the medication 6 months prior, shows symptom improvement after 2 weeks of the study. However, the exact time of symptom improvement before the study is unknown, rendering Patient A's data left-censored.

Interval-Censoring Distribution

Interval-censoring occurs when the event time is only known to fall within a specific time interval rather than being precisely observed. This type of censoring is more complex and less common in practice. The general formula for the interval-censoring distribution $(F_{c_I}(t))$ is $(F_{c_I}(t)) = \Pr(a < T < b)$ where, $(F_{c_I}(t))$ is the interval-censoring distribution function, representing the probability that the true event time (T) lies between a and b (the censoring interval).

It should be noted that, the formula for right-censoring, left-censoring and interval-censoring are not specific mathematical expression like the CDF (cumulative distribution function). Instead, it is a concept that is used to represent the probability that the true event time (T) for censored subjects is either greater than, less than or between a given value t (the censoring time).

3.2.1 Truncation

Truncation of survival data occurs when only those individuals whose event time lies within a certain observational window (X_L, X_R) are observed. An individual whose event time is not in this interval is not observed and no information on this subject is available to the investigator. This is in contrast to censoring where there is partial information on each subject. They are mainly two types of truncation namely left truncation and right truncation. Left truncation occurs when X_R approaches ∞ and we only observe those individuals whose event time X exceeds the truncation time X_L. That is we observe X if and only if $X_L \leq X$. For lifetime data, right truncation occurs when X_L is 0 and we observe only those individuals whose event time X does not exceeds the truncation time X_R. That is we observe X if and only if $X < X_R$. Though not common in lifetime data, right truncation occurs when the values of a random variable can be observed only when they are smaller than an upperbound.

3.3 Truncated Distribution

A truncated distribution is a statistical concept that arises when certain observations or events are systematically excluded or restricted from a dataset based on specific criteria. Truncation differs from censoring, where data points are partially observed but not necessarily removed. In the context of data analysis, truncation occurs when data values do not meet certain criteria and are consequently excluded from the dataset. Understanding truncation is crucial for ensuring that data analysis accurately reflects the underlying population or phenomenon, particularly when specific inclusion or exclusion criteria are applied. There are two common types of truncation, left-truncation and right-truncation. Each type of truncation has a different formula for its respective truncation distribution.

Definition 3.3.1. Let X be a random variable that is distributed according to some probability density function, $f(x)$, with cumulative distribution function $F(x)$, then the truncated distribution g(x) restricted over the interval $y = (a, b]$ is given by

$$g\,(x) = \frac{f(x)}{F(b) - F(a)}, \qquad a < x \leq b$$

(Johnson et al., 1994).

Definition 3.3.2. Left-truncated distribution is a distribution where just the lower end has been removed and its formula is given by

$$g(x) = \frac{f(x)}{1 - F(y)}, \qquad x > y.$$

Definition 3.3.3. Right-truncated distribution is a distribution where just the upper end has been removed and its formula is given by

$$g(x) = \frac{f(x)}{F(y)}, \qquad x \leq y$$

.

3.4 The Exponential Distribution

An exponential distribution is a continuous probability model that is widely used in lifetime data. It is the only continuous probability distribution that has a constant failure rate. If X_i, where $i = 1, 2, 3, ..., n$ are independently and identically distributed exponential random variables with parameter λ, then the (pdf) of the sum of the random variables result into a gamma distribution with parameters n and λ .

Definition 3.4.1. A continuous random variable X taking values of x in the interval $[0, \infty)$ is said to have an exponential distribution with parameter λ if it has probability density function

$$f(x;\lambda) = \begin{cases} \frac{1}{\lambda} e^{-\frac{x}{\lambda}}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

where $\lambda > 0$. If X is an exponential random variable with parameter λ, then we say $X \sim EXP(\lambda)$.

Definition 3.4.2. If $X \sim EXP(\lambda)$ then the

(i) cumulative distribution function (cdf) denoted by $F\,(x)$ is given by

$$F(x;\lambda)=\begin{cases}1-e^{-\frac{x}{\lambda}}, & x\geq 0\\ 0, & x<0\end{cases}$$

(ii) realibility or survival function is given by

$$S(x)=1-F(x)=e^{-\frac{x}{\lambda}}.$$

(iii) moment generation function (mgf) of X is given by $M_X\,(t)=\frac{1}{1-\lambda t},\quad t<\frac{1}{\lambda}$.

(iv) moments of X, $k\in N$ are given by $E\big(X^k\big)=k!\,\lambda^k$.

Definition 3.4.3. If X is an exponential random variable with parameter λ, that is left truncated at a and right truncated at b then its pdf is given by

$$f(x;\lambda)=\begin{cases}\dfrac{\frac{1}{\lambda}e^{\frac{-x}{\lambda}}}{e^{\frac{-a}{\lambda}}-e^{\frac{-b}{\lambda}}}, & a<x\leq b\\ 0, & \text{otherwise}\end{cases}$$

Definition 3.4.4. If X is an exponential random variable with parameter λ, that is right truncated at b then its pdf is given by

$$f(x;\lambda)=\begin{cases}\dfrac{\frac{1}{\lambda}e^{\frac{-x}{\lambda}}}{1-e^{\frac{-b}{\lambda}}}, & 0<x\leq b\\ 0, & \text{otherwise}\end{cases}$$

Definition 3.4.5. If X is an exponential random variable with parameter λ, that is left truncated at a then its pdf is given by

$$f(x;\lambda)=\begin{cases}\dfrac{\frac{1}{\lambda}e^{\frac{-x}{\lambda}}}{e^{\frac{-a}{\lambda}}}, & x>a\\ 0, & \text{otherwise}\end{cases}$$

Definition 3.4.6. If X is an exponential random variable with parameter λ, that is left truncated at a and right truncated at b then its cdf is given by

$$F(x;\lambda)=\frac{e^{-a/\lambda}-e^{-x/\lambda}}{e^{-a/\lambda}-e^{-b/\lambda}},\quad a<x\leq b.$$

3.5 The Gamma Distribution

A gamma distribution is a continuous probability distribution that is widely used in different fields of science to model lifetime data. It arises naturally in processes where the waiting times between Poisson distributed events are relevant.

Definition 3.5.1. A continuous random variable X taking values of x in the interval $[0, \infty)$ is said to have a gamma distribution with parameters $\alpha > 0$ and $\beta > 0$ if it has probability density function

$$f\left(x; \alpha, \beta\right) = \begin{cases} \dfrac{x^{\alpha-1}\, e^{-x/\beta}}{\beta^{\alpha}\, \Gamma(\alpha)}, & x > 0 \qquad \text{(Johnson et al., 1994)} \\ 0, & \text{otherwise} \end{cases}$$

where $\Gamma(z) = \int_0^{\infty} t^{z-1}\, e^{-t}\, dt$ is the Euler's gamma function or the complete Euler function, $\alpha > 0$ is the shape parameter and $\beta > 0$ is the scale parameter. If X is a gamma random variable with parameters α and β, then we say $X \sim GAM(\alpha\,, \beta)$.

Definition 3.5.2. If $X \sim \text{GAM}\left(\alpha\,, \beta\right)$ then the

(i) cumulative distribution function (cdf) denoted by $F\left(x\right)$ is given by

$$F\left(x; \alpha, \beta\right) = \frac{\gamma\left(\alpha, x/\beta\right)}{\Gamma(\alpha)}$$

where

$$\gamma(a, z) = \int_0^z t^{a-1}\, e^{-t} dt,$$

is the lower incomplete gamma function.

(ii) moment generation function (mgf) of X is $M_X\left(t\right) = \left(1 - \beta\, t\right)^{-\alpha}, \quad t < \frac{1}{\beta}$.

(iii) moments of X, $k \in N$ are given by $E\big(X^k\big) = \frac{\beta^k \Gamma(k+\alpha)}{\Gamma(\alpha)}$.

Definition 3.5.3. If X is a gamma random variable with parameters α and β, that is left truncated at a and right truncated at b then its pdf is given by

$$f\left(x; \alpha, \beta\right) = \begin{cases} \dfrac{x^{\alpha-1}\, e^{-x/\beta}}{\beta^{\alpha}\, \left[\gamma\left(\alpha, \frac{b}{\beta}\right) - \gamma\left(\alpha, \frac{a}{\beta}\right)\right]}, & a < x < b \\ 0, & \text{otherwise} \end{cases}$$

Definition 3.5.4. If X is a gamma random variable with parameters α and β that is right truncated at b then its pdf is given by

$$f\left(x;\alpha,\beta\right)=\begin{cases}\dfrac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\left[\gamma(\alpha,\frac{b}{\beta})\right]}, & 0<x\leq b\\ 0, & \text{otherwise}\end{cases}$$

Where $\gamma\left(\alpha,\frac{b}{\beta}\right)=\left[\Gamma\left(\alpha\right)-\Gamma\left(\alpha,\frac{b}{\beta}\right)\right]$

Definition 3.5.5. If X is a gamma random variable with parameters α and β that is left truncated at a then its pdf is given by

$$f\left(x;\alpha,\beta\right)=\begin{cases}\dfrac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\left[\Gamma\left(\alpha\right)-\ \gamma(\alpha,\frac{b}{\beta})\right]}, & x\geq a\\ 0, & \text{otherwise.}\end{cases}$$

where $\Gamma(\alpha)-\gamma(\alpha,\frac{b}{\beta})=\Gamma(\alpha,\frac{b}{\beta})$ and $\Gamma(\alpha,\infty)=\int_{\infty}^{\infty}t^{\alpha-1}e^{-t}\,dt=0$.

Definition 3.5.6. If X is a gamma random variable with parameters α and β, that is left truncated at a and right truncated at b then its cdf is given by

$$F(x;\alpha,\beta)=K\beta\left[\Gamma(\alpha,\frac{a}{\beta})-\Gamma(\alpha,\frac{x}{\beta})\right]$$

where the constant K is

$$K=\frac{1}{\beta\left[\Gamma\left(\alpha,\frac{a}{\beta}\right)-\Gamma\left(\alpha,\frac{b}{\beta}\right)\right]}.$$

CHAPTER 4

Methodology

In this chapter, we present the theoretical background of the ML-estimation, EM-algorithm, SEM-algorithm, NR-algorithm and the measures of comparing the different algorithms.

4.1 Theoretical Background

This section will discuss the maximum likelihood (ML) functions for left-truncated and right-censored data, expectation maximisation, the application of expectation maximisation on gamma and exponential distribution, stochastic expectation maximisation and Newton-Raphson algorithm.

4.1.1 Maximum Likelihood Estimation

The ML-estimation and methods of moments (MM) are commonly used methods in the estimation of parameters. The first method is preferable because of its asymptotic properties and the second one is preferable for its simplicity. Other methods have been proposed, in particular graphical estimation methods and methods based on order statistics (Dubey, 1967). ML-estimation is the method of estimating the parameters of a probability distribution by maximising the likelihood function in order to make the observed data most probable for the statistical model. The ML-estimation has the limitation of making the assumption that the data is complete or fully observed which is not always the case (Dempster et al., 1977).

Let $\mathbf{X} = (X_1, ..., X_n)$ denote an observed random sample of size n on some random vector $\boldsymbol{X}$ from a population with pdf $f(x|\theta)$, where $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^T$ is the vector of unknown parameters of size p and the likelihood function is defined as

$$L\left(\boldsymbol{\theta}|\boldsymbol{x}\right)=\prod_{i=1}^{n} f\left(x_i|\boldsymbol{\theta}\right). \tag{4.1}$$

An estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ can be obtained as a solution of the likelihood equation $L\left(\boldsymbol{\theta}\right)$ or the log likelihood equation log $L\left(\boldsymbol{\theta}\right)$ given by

$$\frac{\partial \text{Log} L\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}=0.$$

The ML-estimates of the parameter $\boldsymbol{\theta}$ denoted as $\widehat{\boldsymbol{\theta}}=\widehat{\boldsymbol{\theta}}\left(x\right)$, is the value of θ at which $L\left(\boldsymbol{\theta}|\boldsymbol{x}\right)$ attains its maximum (Shang, 2021), i.e., $\widehat{\boldsymbol{\theta}}=\arg\max_{\boldsymbol{\theta}} L\left(\boldsymbol{\theta}|\boldsymbol{x}\right)$, where $\arg\max_{\boldsymbol{\theta}}$: is a mathematical notation that stands for "argument of the maximum". It refers to the input values that produce the maximum output of a function. Obtaining the ML-estimates can be a difficult task in practice. There are many instances where the analytical form of the ML-estimates cannot be obtained hence numerical methods are used.

4.1.2 Likelihood function for truncated and censored data

If $\mathbf{X}=\left(X_1,...,X_n\right)$ denotes an observed random sample of size n on some random vector $\boldsymbol{X}$ from a population with pdf $f\left(x|\boldsymbol{\theta}\right)$ and cdf $F\left(x|\boldsymbol{\theta}\right)$, where $\boldsymbol{\theta}=\left(\theta_1,...,\theta_p\right)^T$ is the vector of unknown parameters of size p. Then the likelihood function for observations that are

(a) left-truncated at $x=T$ is given by

$$L\left(\boldsymbol{x}|\boldsymbol{\theta}\right)=\frac{1}{1-F\left(T\right)}\prod_{i=1}^{n} f\left(x_i\right), \quad T\leq x_i, \tag{4.2}$$

(b) right-truncated at $x=T$ is given by

$$L\left(\boldsymbol{x}|\boldsymbol{\theta}\right)=\frac{1}{F\left(T\right)}\prod_{i=1}^{n} f\left(x_i\right), \quad T\geq x_i, \tag{4.3}$$

(c) doubly truncated at T_1 and T_2 is given by

$$L\left(\boldsymbol{x}|\boldsymbol{\theta}\right)=\frac{1}{F\left(T_2\right)-F\left(T_1\right)}\prod_{i=1}^{n} f\left(x_i\right), \quad T_1\leq x_i\leq T_2. \tag{4.4}$$

If $\mathbf{X}=\left(X_1,...,X_n\right)$ denotes an observed random sample of size n on some random vector $\boldsymbol{X}$

from a population with pdf $f(x|\boldsymbol{\theta})$ and cdf $F(x|\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^T$ is the vector of unknown parameters of size p. Then the likelihood function for observations that are

(a) left censored at $x = T$ is given by

$$L(\boldsymbol{x}|\boldsymbol{\theta}) = k\,[F(T)]^c \prod_{i=1}^{n} f(x_i), \quad T \leq x_i, \tag{4.5}$$

(b) right censored at $x = T$ is given by

$$L(\boldsymbol{x}|\boldsymbol{\theta}) = k\,[1 - F(T)]^c \prod_{i=1}^{n} f(x_i), \quad T \geq x_i, \tag{4.6}$$

(c) doubly censored at T_1 and T_2 is given by

$$L(\boldsymbol{x}|\boldsymbol{\theta}) = k\,[F(T_1)]^{c_1}\,[1 - F(T_2)]^{c_2} \prod_{i=1}^{n} f(x_i), \quad T_1 \leq x_i \leq T_2, \tag{4.7}$$

where k denotes constants that do not depend on the parameters, n denotes the number of complete observations and c_j denotes the number of observations censored at T_j (Cohen, 1991).

Let X be a lifetime variable, C be the censoring time variable and ν_i denote the truncation indicator such that

$$\nu_i = \begin{cases} 0, & \text{if the } i^{th} \text{ unit is truncated} \\ 1, & \text{if the } i^{th} \text{ unit is not truncated.} \end{cases}$$

Let δ_i denote the censoring indicator such that

$$\delta_i = \begin{cases} 0, & \text{if the } i^{th} \text{ unit is censored} \\ 1, & \text{if the } i^{th} \text{ unit is not censored.} \end{cases}$$

Balakrishna and Mitra (2012) expressed the likelihood function for the left-truncated and right-censored data as

$$L\left(\boldsymbol{x};\boldsymbol{\theta}\right)=\prod_{i\in S_1}\bigg(f_X\left(x_i\right)\bigg)^{\delta_i\nu_i}\bigg(1-F_X\left(x_i\right)\bigg)^{(1-\delta_i)\nu_i}$$
$$\times\prod_{i\in S_2}\left(\frac{f_X\left(x_i\right)}{1-F_X\left(\tau_i^L\right)}\right)^{\delta_i(1-\nu_i)}\left(\frac{1-F_X\left(x_i\right)}{1-F_X\left(\tau_i^L\right)}\right)^{(1-\delta_i)(1-\nu_i)}$$

where τ_i^L denotes the left-truncation time, S_1 denotes the index set corresponding to the units which are not left truncated, S_2 denotes the index set corresponding to the units which are left-truncated and $\boldsymbol{\theta}$ denotes the vector of unknown parameters.

The log-likelihood can be expressed as

$$\operatorname{Log}L\left(\boldsymbol{x};\boldsymbol{\theta}\right)=\sum_{i=1}^{n}\delta_i\nu_i\log f_X\left(x_i\right)+\sum_{i=1}^{n}\left(1-\delta_i\right)\nu_i\log\left(1-F_X\left(x_i\right)\right)$$
$$+\sum_{i=1}^{n}\delta_i\left(1-\nu_i\right)\ \log\left(\frac{f_X\left(x_i\right)}{1-F_X\left(\tau_i^L\right)}\right)+\sum_{i=1}^{n}\left(1-\delta_i\right)\left(1-\nu_i\right)\ \log\left(\frac{1-F_X\left(x_i\right)}{1-F_X\left(\tau_i^L\right)}\right).$$

If $X\sim\ \mathrm{GAM}\left(\alpha\,,\beta\right)$ then

$$L\left(\boldsymbol{x};\boldsymbol{\theta}\right)=\prod_{i\in S_1}\left(\frac{x_i^{\alpha-1}e^{-\frac{x_i}{\beta}}}{\beta^{\alpha}\Gamma\left(\alpha\right)}\right)^{\delta_i\nu_i}\left[\frac{\Gamma\left(\alpha,\frac{x_i}{\beta}\right)}{\Gamma\left(\alpha\right)}\right]^{(1-\delta_i)\nu_i}$$
$$\times\prod_{i\in S_2}\left[\frac{x_i^{\alpha-1}e^{\frac{-x_i}{\beta}}}{\beta^{\alpha}\Gamma\left(\alpha\right)}\frac{\Gamma\left(\alpha\right)}{\Gamma\left(\alpha,\frac{\tau_i^L}{\beta}\right)}\right]^{\delta_i(1-\nu_i)}\left[\frac{\Gamma\left(\alpha,\frac{x_i}{\beta}\right)}{\Gamma\left(\alpha\right)}\frac{\Gamma\left(\alpha\right)}{\Gamma\left(\alpha,\frac{\tau_i^L}{\beta}\right)}\right]^{(1-\delta_i)(1-\nu_i)}$$

and the log-likelihood for left-truncated and right-censored data can be expressed as

$$
\begin{aligned}
\operatorname{Log} L(\boldsymbol{x};\boldsymbol{\theta}) = &\sum_{i=1}^{n}\left[\delta_i\nu_i \log\left(\frac{x_i^{\alpha-1}e^{-x_i/\beta}}{\beta^{\alpha}\Gamma\left(\alpha\right)}\right)\right] + \sum_{i=1}^{n}\left[(1-\delta_i)\nu_i \log\left(\frac{\Gamma\left(\alpha, x_i/\beta\right)}{\Gamma\left(\alpha\right)}\right)\right] \\
&+ \sum_{i=1}^{n}\delta_i(1-\nu_i)\ \log\left(\frac{x_i^{\alpha-1}e^{-x_i/\beta}}{\beta^{\alpha}\Gamma\left(\alpha\right)}\ \frac{\Gamma\left(\alpha\right)}{\Gamma\left(\alpha, \tau_i^L/\beta\right)}\right) \\
&+ \sum_{i=1}^{n}(1-\delta_i)(1-\nu_i)\ \log\left(\frac{\Gamma\left(\alpha, x_i/\beta\right)}{\Gamma\left(\alpha\right)}\ \frac{\Gamma\left(\alpha\right)}{\Gamma\left(\alpha, \tau_i^L/\beta\right)}\right) \\
= &\sum_{i=1}^{n}\Big[\delta_i\nu_i(\alpha-1)\log x_i + \delta_i(\alpha-1)\log x_i - \delta_i\nu_i(\alpha-1)\log x_i - \delta_i\nu_i\frac{y_i}{\beta} - \delta_i\frac{x_i}{\beta} \\
&+ \delta_i\nu_i\frac{x_i}{\beta} - \nu_i\log\Gamma\left(\alpha\right) - \delta_i\nu_i\alpha\log\beta - \delta_i k\log\beta + \delta_i\nu_i\alpha\log\beta - \delta_i\nu_i\log\Gamma\left(\alpha\right) \\
&+ \delta_i\nu_i\log\Gamma\left(\alpha\right) + \nu_i\log\Gamma\left(\alpha, x_i/\beta\right) - \delta_i\nu_i\log\Gamma\left(\alpha, x_i/\beta\right) + \nu_i(1-\delta_i)\log\Gamma\left(\alpha, x_i/\beta\right) \\
&+ (1-\delta_i)\log\Gamma\left(\alpha, x_i/\beta\right) - \nu_i(1-\delta_i)\log\Gamma\left(k, x_i/\beta\right) - \delta_i(1-\nu_i)\log\Gamma\left(\alpha, \tau_i^L/\beta\right) \\
&- (1-\nu_i)\log\Gamma\left(\alpha, \tau_i^L/\beta\right) + \delta_i(1-\nu_i)\log\Gamma\left(k, \tau_i^L/\beta\right)\Big]
\end{aligned}
$$

$$
\begin{aligned}
= &\sum_{i=1}^{n}\Big[\delta_i(\alpha-1)\log x_i - \delta_i\frac{x_i}{\beta} - \delta_i\alpha\log\beta - \nu_i\log\Gamma\left(\alpha\right) + (1-\delta_i)\log\Gamma\left(\alpha, x_i/\beta\right) \\
&- (1-\nu_i)\log\Gamma\left(\alpha, \tau_i^L/\beta\right)\Big] \\
= &\sum_{i=1}^{n}\left[\delta_i\left\{(\alpha-1)\log x_i - \frac{x_i}{\beta} - \alpha\log\beta\right\} + (1-\delta_i)\log\Gamma\left(\alpha, x_i/\beta\right)\right] \\
&- \sum_{i=1}^{n}\left[\nu_i\log\Gamma\left(\alpha\right) + (1-\nu_i)\log\Gamma\left(\alpha, \tau_i^L/\beta\right)\right].
\end{aligned}
$$

Similarly if $X \sim \text{Exp}(\lambda)$ then the log-likelihood can be expressed

$$
\begin{aligned}
\text{Log}\, L(\boldsymbol{x}, \lambda) =& \sum_{i=1}^{n} \Big[\delta_i \nu_i \log\left(\frac{1}{\lambda} e^{-x_i/\lambda}\right) - \delta_i \nu_i \log\left(e^{-x_i/\lambda}\right) + \nu_i \log\left(e^{-x_i/\lambda}\right) \\
& + \delta_i (1-\nu_i) \log\left(\frac{\frac{1}{\lambda} e^{-x_i/\lambda}}{e^{-\tau_i^L/\lambda}}\right) + (1-\nu_i) \log\left(\frac{e^{-x_i/\lambda}}{e^{-\tau_i^L/\lambda}}\right) \\
& - \delta_i (1-\nu_i) \log\left(\frac{e^{-x_i/\lambda}}{e^{-\tau_i^L/\lambda}}\right)\Big] \\
=& \sum_{i=1}^{n} \Big[- \delta_i \nu_i \log\lambda - \delta_i \nu_i \frac{x_i}{\lambda} + \delta_i \nu_i \frac{x_i}{\lambda} - \nu_i \frac{x_i}{\lambda} - \delta_i (1-\nu_i) \log\lambda - \delta_i (1-\nu_i) \frac{x_i}{\delta} \\
& + \delta_i (1-\nu_i) \frac{\tau_i^L}{\lambda} - (1-\nu_i) \frac{x_i}{\lambda} + (1-\nu_i) \frac{\tau_i^L}{\lambda} + \delta_i (1-\nu_i) \frac{x_i}{\lambda} - \delta_i (1-\nu_i) \frac{\tau_i^L}{\lambda}\Big] \\
=& \sum_{i=1}^{n} \Big[- \delta_i \nu_i \log\lambda - \nu_i \frac{x_i}{\lambda} - \delta_i \log\lambda + \delta_i \nu_i \log\lambda - \frac{x_i}{\lambda} + \nu_i \frac{x_i}{\lambda} + (1-\nu_i) \frac{\tau_i^L}{\lambda}\Big] \\
=& \sum_{i=1}^{n} \Big[- \delta_i \log\lambda - \frac{x_i}{\lambda} + (1-\nu_i) \frac{\tau_i^L}{\lambda}\Big] \\
=& \sum_{i=1}^{n} \Big[\delta_i \log\left(\frac{1}{\lambda}\right) - \frac{x_i}{\lambda}\Big] + \sum_{i=1}^{n} (1-\nu_i) \frac{\tau_i^L}{\lambda}.
\end{aligned}
$$

Differentiating the log-likelihood with respect to λ yields

$$
\frac{\partial L(\boldsymbol{x}, \lambda)}{\partial \lambda} = \sum_{i=1}^{n} \left[-\frac{\delta_i}{\lambda} + \frac{x_i}{\lambda^2} - (1-\nu_i) \frac{\tau_i^L}{\lambda^2} \right]
$$

setting $\frac{\partial L(\lambda)}{\partial \lambda} = 0$ yields

$$
\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} (1-\nu_i) \tau_i^L}{\sum_{i=1}^{n} \delta_i}.
$$

4.1.3 Expectation-Maximisation Algorithm (EM-Algorithm)

The EM-algorithm is a broadly applicable approach to the iterative computation of ML-estimates and it is useful in a variety of incomplete-data problems (Mclachlan, 2004). Dempster et al., (1977) synthesized earlier formulations of the EM-algorithm and presented a general formulation of finding ML-estimates for a variety of problems and provided an initial catalogue of problems where this method could be profitably applied. Since then the EM-algorithm has been applied in a variety of general statistical problems such as resolution of mixtures, multiway contingency tables, variance components estimation, factor analysis, as well as in specialized applications in such areas as genetics, medical imaging, and neural networks (Mclachlan, 2004).

The EM-algorithm is an iterative method that attempts to find the ML-estimate of a parameter θ. It iterates through two steps to obtain estimates. The first step is an Expectation (E) step, in which missing values are filled-in with a guess, that is, an estimate of the missing value, given the observed values in the data. The second step is a Maximisation (M) step, in which the completed data from the E-step are processed using ML estimation as though they were complete data.

Let $\boldsymbol{Z} = (Z_1, ..., Z_n)$ denote the incomplete observed data z, having probability density function (pdf) $h\left(.|\boldsymbol{\theta}\right)$, where $\boldsymbol{\theta} = (\theta_1, ..., \theta_r)'$ is a vector of unknown parameters with parameter space Ω and $\boldsymbol{W} = (W_1, ..., W_m)$ be a vector containing the unobservable or missing data. Let $\boldsymbol{x}$ denote a vector containing the complete data and $f_c\left(\mathbf{x}; \boldsymbol{\theta}\right)$ denote the pdf for the random vector $\boldsymbol{X}$ corresponding to the complete data vector $\boldsymbol{x}$, then the ML-estimate via EM- algorithm is obtained from the complete log-likelihood

$$\text{Log}\, L_c\left(\boldsymbol{\theta}; \mathbf{x}\right) = \log f_c\left(\mathbf{x}; \boldsymbol{\theta}\right).$$

In a nutshell, the EM algorithm solves the problem of the incomplete-data likelihood indirectly by proceeding iteratively using the complete-data log-likelihood function, $Log\, L_c\left(\boldsymbol{\theta}; \mathbf{x}\right)$. The EM-algorithm operates in two distinct steps, the expectation step which is referred to as the E-step and maximasation step or “M-step” (Dempster et al., 1977) . These two steps (E and M) are repeated until the change in the estimates from one iteration to the next does not exceed a set limit (suitable stoping rule). The E-step calculates the expected log-likelihood, and the M-step maximizes the expected log-likelihood with respect to the parameters. In the E-step, we obtain the objective function

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = E\big[\log\, L_c\left(\boldsymbol{\theta}|\boldsymbol{z}, \boldsymbol{W}\right)|\boldsymbol{\theta}^{(k)}, \boldsymbol{z}\big]$$

which is the conditional expectation of the complete log-likelihood given the observed data $\boldsymbol{z}$ and the parameter vector $\boldsymbol{\theta}^{(k)}$ is the estimate obtained on the k^{th} iteration. In the M-step, we maximize the objective function $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ with respect to θ to obtain the updated estimate as

$$\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta}}\; Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}).$$

The E- and M-steps are repeated until convergence. Convergence may be determined, by using a suitable stopping rule like $||\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}|| < \epsilon$ for some $\epsilon > 0$

Gamma distribution

Balakrishnan and Mitra (2012) studied the EM-algorithm for GAM $(\alpha\,,\beta)$ distribution based on the left-truncated and right-censored data. In their study the complete data likelihood for gamma distribution in the absence of censoring was expressed as

$$L\left(\mathbf{t};\boldsymbol{\theta}\right)=\prod_{i\in S_1}\left(\frac{t_i^{\alpha-1}e^{-\frac{t_i}{\beta}}}{\beta^{\alpha}\Gamma\left(\alpha\right)}\right)\times\prod_{i\in S_2}\left(\frac{t_i^{\alpha-1}e^{\frac{-t_i}{\beta}}}{\beta^{\alpha}\Gamma\left(\alpha,\frac{\tau_i^L}{\beta}\right)}\right)$$

where τ_i^L denotes the left truncation time, t is a lifetime variable, S_1 denotes the index set corresponding to the units which are not left truncated, S_2 denotes the index set corresponding to the units which are left truncated. Thus the log-likelihood function becomes

$$\log L_c\left(\mathbf{t},\boldsymbol{\theta}\right)=\sum_{i=1}^{n}\left[(\alpha-1)\log\,t_i-\frac{t_i}{\beta}-\alpha\,\log\,\beta\right]-\sum_{i=1}^{n}\left[\nu_i\log\,\Gamma(\alpha)+(1-\nu_i)\,\log\,\Gamma\left(\alpha,\frac{\tau_i^L}{\beta}\right)\right].$$

In the E- step the objective function $Q(\boldsymbol{\theta},\,\boldsymbol{\theta}^{(k)})$ is expressed as

$$\begin{aligned}Q(\boldsymbol{\theta},\,\boldsymbol{\theta}^{(k)})=&E_{\theta^{(k)}}\left[Log\,L_c\left(t,\boldsymbol{\theta}\right)|z,\boldsymbol{\delta}\right]\\=&\sum_{i:\delta_i=1}(\alpha-1)\;\text{log}t_i+\sum_{i:\delta_i=0}(\alpha-1)\,E_{1i}^{(k)}-\sum_{i:\delta_i=1}\frac{t_i}{\beta}-\sum_{i:\delta_i=0}\frac{E_{2i}^{(k)}}{\beta}\\&-n\alpha\,\log\,\beta\,-\sum_{i=1}^{n}\left[\nu_i\,\log\,\Gamma\left(\alpha\right)+\,(1-\nu_i)\,\log\,\Gamma\left(\alpha,\,\frac{\tau_i^L}{\beta}\right)\right],\end{aligned}$$

where $\boldsymbol{Z}=(z_1,z_2,...,z_n)'$ is the incomplete observed data vector, $z_i=\min\left(t_i,c_i\right)$, $\boldsymbol{\delta}=(\delta_1,...\delta_n)'$ denotes the vector of censoring indicators, c_i is the right censoring point, $E_{1i}^{(k)}=E_{\theta^{(k)}}[\log T_i|T_i>z_i]$ and $E_{2i}^{(k)}=E_{\theta^{(k)}}[T_i|T_i>z_i]$ (Balakrishnan and Mitra, 2012). Where $E_{1i}^{(k)}=E_{\theta^{(k)}}[\log T_i|T_i>z_i]$ represents the expected value of the logarithm of the survival time for each observation that is not censored, given that the true survival time is greater than the observed minimum of the true survival time and the censoring time. It captures the expected contribution of each uncensored observation to the complete data log-likelihood function under the current parameter estimates at kth iteration.

$E_{2i}^{(k)}=E_{\theta^{(k)}}[T_i|T_i>z_i]$ computes the expected value of the survival time itself for each uncensored observation, given the observed data and current parameter estimates. It provides an estimate of the expected survival times, which can be useful for various analyses, such as estimating median survival time or making predictions about future survival times. It also provides additional information about the distribution of survival times beyond just the likelihood of failure. While $E_{1i}^{(k)}=E_{\theta^{(k)}}[\log T_i|T_i>z_i]$ focuses on the likelihood aspect,

$E_{2i}^{(k)} = E_{\theta^{(k)}}[T_i | T_i > z_i]$ provides insights into the actual survival times of the observations.

In the M-step the $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ function is maximised in order to determine

$$\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}).$$

According to Balakrishnan and Mitra (2012) the first order derivatives with respect to α and β gives the following

$$\frac{\partial Q}{\partial \beta} = \frac{1}{n\alpha}\left[\sum_{i:\delta_i=1} t_i + \sum_{i:\delta_i=0} E_{2i}^{(k)} - \sum_{i=1}^{n} \frac{(1-\nu_i)\left(\tau_i^L\right)^{\alpha} e^{-\frac{\tau_i^L}{\beta}}}{\beta^{\alpha-1}\Gamma\left(\alpha,\, \frac{\tau_i^L}{\beta}\right)}\right],$$

and

$$\begin{aligned}\frac{\partial Q}{\partial \alpha} = & \sum_{i:\delta_i=1} \text{log} t_i + \sum_{i:\delta_i=0} E_{1i}^{(k)} - \text{nlog}\beta - n\psi\left(\alpha\right) \\ & - \sum_{i=1}^{n} (1-\nu_i)\left[\log\left(\tau_i^L/\beta\right) - \log\left(\tau_i^L/\beta\right)/\left(1 - e^{-\tau_i^L/\beta}\sum_{p=0}^{\infty}\frac{\left(\tau_i^L/\beta\right)^{\alpha+p}}{\Gamma\left(\alpha+p+1\right)}\right)\right. \\ & \left. + e^{-\tau_i^L/\beta}\sum_{p=0}^{\infty}\frac{\left(\tau_i^L/\beta\right)^{\alpha+p}\psi\left(\alpha+p+1\right)}{\Gamma\left(\alpha+p+1\right)}\Big/\left(1 - e^{-\tau_i^L/\beta}\sum_{p=0}^{\infty}\frac{\left(\tau_i^L/\beta\right)^{\alpha+p}}{\Gamma\left(\alpha+p+1\right)}\right)\right].\end{aligned}$$

The equations $\frac{\partial Q}{\partial \beta} = 0$ and $\frac{\partial Q}{\partial \alpha} = 0$ can be solved numerically for β and α to obtain the estimates $\beta^{(k+1)}$ and $\alpha^{(k+1)}$. The E-step and M-step are then repeated till convergence is achieved to the desired level of accuracy (Balakrishnan and Mitra, 2012). It should be noted that, the parameter estimates obtained using the EM-algorithm from left truncated and right censored data will be the ML-estimates of the gamma distribution parameters. These estimates can be used to make predictions and calculate other quantities of interest, such as survival probabilities and hazard rates.

Exponential

Since exponential is a special case of a gamma distribution with $\alpha = 1$ its complete data log-likelihood is given by

$$\text{log}L_c\left(\boldsymbol{t},\, \boldsymbol{\lambda}\right) = \sum_{i=1}^{n}(1-\nu_i)\left(\frac{\tau_i^L}{\lambda}\right) + \sum_{i=1}^{n}\left[\log\left(\frac{1}{\lambda}\right) - \frac{t_i}{\lambda}\right].$$

In the E-step the objective function $Q\left(\boldsymbol{\lambda},\boldsymbol{\lambda}^{(\boldsymbol{k})}\right)$ is expresssed as

$$\begin{aligned} Q\left(\boldsymbol{\lambda},\boldsymbol{\lambda}^{(\boldsymbol{k})}\right) =& E_{\lambda^{(r)}}\left[\log L_c\left(t,\lambda\right)|z,\boldsymbol{\delta}\right] \\ =& -\left(\sum_{i:\delta_i=1}^{n}\frac{t_i}{\lambda}+\sum_{i:\delta_i=0}^{n}\frac{E_{2i}^{(k)}}{\lambda}\right)-\text{n}\log\lambda+\sum_{i=1}^{n}\left(1-\nu_i\right)\frac{\tau_i^L}{\lambda}. \end{aligned}$$

In the M-step the $Q(\boldsymbol{\lambda},\boldsymbol{\lambda}^{(k)})$ function is maximised with respect to λ

$$\boldsymbol{\lambda}^{(k+1)} = \arg\max_{\boldsymbol{\lambda}}\ Q(\boldsymbol{\lambda},\boldsymbol{\lambda}^{(k)}).$$

Differentiating $Q(\boldsymbol{\lambda},\boldsymbol{\lambda}^{(k)})$ with respect to λ yields

$$\frac{\partial Q}{\partial\lambda}=\sum_{i:\delta_i=1}\frac{t_i}{\lambda^2}+\sum_{i:\delta_i=0}\frac{E_{2i}^{(k)}}{\lambda^2}-\frac{n}{\lambda}-\sum_{i=1}^{n}\left(1-\nu_i\right)\frac{\tau_i^L}{\lambda^2}$$

and setting $\frac{\partial Q}{\partial\lambda}=0$ the estimate parameter for λ becomes

$$\hat{\lambda}=\frac{\sum_{i:\delta_i=1}t_i+\sum_{i:\delta_i=0}E_{2i}^{(k)}-\sum_{i=1}^{n}\left(1-\nu_i\right)\tau_i^L}{\text{n}}$$

where $E_{2i}^{(k)}=E_{\theta^{(k)}}[T_i|T_i\ >z_i]$.

The parameter estimate obtained using the EM-algorithm from left truncated and right censored data will be the ML-estimate of the exponential distribution parameter. This estimate can be used to make predictions and calculate other quantities of interest, such as survival probabilities and hazard rates. Note that in the case of the exponential distribution, the shape parameter is fixed to 1, so only the scale parameter λ needs to be estimated.

4.1.4 Stochastic EM algorithm

The stochastic expectation maximisation (SEM) algorithm is the stochastic version of the EM-algorithm which addresses the limitations of the EM-algorithm (Shang, 2021). In the SEM-algorithm the analytical computation of the conditional expectation involved in $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(\boldsymbol{k})})$ is replaced by simulating an unobserved sample. If $\boldsymbol{W}=(w_1,...,w_n)'$ is the vector containing the unobservable or missing data, $\boldsymbol{\theta}^{(k)}=\left(\theta_1^{(k)},\theta_2^{(k)},...,\theta_r^{(k)}\right)$ is the value of the parameter at the k^{th} iteration and $\boldsymbol{Z}=(Z_1,...,Z_n)$ is the incomplete observed data , then the conditional distribution of W_i given the observed data is given by

$$f_{W_i|W_i>z_i}\left(w_i|w_i > z_i, \boldsymbol{\theta}^{(\boldsymbol{k})}\right). \tag{4.8}$$

The SEM-algorithm consists of two steps namely the stochastic expectation (SE) step and the maximisation (M) step. In the SE-step we randomly generate observations $\boldsymbol{w}^{(k)} = (w_1^{(k)}, ..., w_n^{(k)})$ from the conditional distribution in equation (3.8) in order to replace the lifetime of the i^{th} censored unit. After the random observations generated from the conditional distribution have been imputed we obtain a pseudo-complete data. And it is this pseudo-complete data that we use to obtain the pseudo-complete log-likelihood function

$$Q\left(\boldsymbol{\theta}, \boldsymbol{w}^{(k)}\right) = \log L\left(\boldsymbol{\theta}|z, w^{(k)}\right). \tag{4.9}$$

In the M-step we maximise the pseudo-complete log-likelihood in equation (3.9) in order to obtain the updated estimate $\boldsymbol{\theta}^{(k+1)}$ (Mitra, Kundu and Balakrishnan, 2017). Lastly, the SE and the M steps are alternatively repeated for the large cycles or iterations and the number of iterations to consider are user defined. If there are N iterations, one obtains a sequence of estimates $\boldsymbol{\theta}^{(k)}$ where $k = 1, 2, ..., N$. It should be noted that the first few iterations are discarded for burn-in and an average over the remaining iterations is taken so as to get the final estimate $\hat{\boldsymbol{\theta}}$.

Gamma distribution

The SEM-algorithm for gamma distribution with parameters (α, β) is based on replacing the missing data with the random observations drawn from the conditional distribution, this is to say in the SE-step we randomly draw values of $W^{(k)}$ from the conditional distribution in order to generate the pseudo-complete data. According to Shang (2021), the conditional distribution is given by

$$f_{W_i|W_i>z_i}(w) = \begin{cases} \dfrac{w_i^{\alpha-1} e^{-w_i/\beta}}{\beta^{\alpha}\Gamma\left(\alpha, \frac{\tau_i^L}{\beta}\right)}, & w > z_i \\ 0, & \text{otherwise} \end{cases}$$

where i^{th} unit is right censored. The imputation is implemented by generating the random variates using the inverse transformation method (Shang, 2021). Once the random observations have been generated we then have the pseudo-complete data. The pseudo-complete data is then used to obtain the pseudo-complete log-likelihood function. Thus the function can then be expressed as

$$
\begin{aligned}
Q\left(\boldsymbol{\theta}|\boldsymbol{w}^{(k)}\right) =& \log L\left(\theta|\boldsymbol{t^{(k)}}\right) \\
=& \sum_{i=1}^{n} \nu_i \log f_W\left(t_i^{(k)}\right) + \sum_{i=1}^{n}(1-\nu_i) \log \frac{f_W\left(t_i^{(k)}\right)}{1-F_W(\tau_i^L)} \\
=& \sum_{i=1}^{n} \nu_i \log\left(\frac{\left(t_i^k\right)^{\alpha-1} e^{-t_i^{(k)}/\beta}}{\beta^\alpha \Gamma\left(\alpha\right)}\right) + \sum(1-\nu_i)\log\left(\frac{\left(t_i^k\right)^{\alpha-1} e^{-t_i^{(k)}/\beta}}{\beta^\alpha\Gamma\left(\alpha\right)} \frac{\Gamma\left(\alpha\right)}{\Gamma(\alpha, \frac{\tau_i^L}{\beta})}\right) \\
=& \sum \nu_i \left((\alpha-1)\log t_i^{(k)} - \frac{t_i^{(k)}}{\beta} - \alpha\log\beta - \log\Gamma\left(\alpha\right)\right) \\
&+ \sum(1-\nu_i)\left((\alpha-1)\log t_i^{(k)} - \frac{t_i^{(k)}}{\beta} - \alpha\log\beta - \log\Gamma\left(\alpha\right) - \log\left(\alpha, \frac{\tau_i^L}{\beta}\right)\right) \\
=& \sum\Bigg[\nu_i(\alpha-1)\log t_i^{(k)} - \nu_i\frac{t_i^{(k)}}{\beta} - \nu_i\alpha\log\beta - \nu_i\log\Gamma\left(\alpha\right) - \nu_i(\alpha-1)\log t_i^{(k)} + \nu_i\frac{t_i^{(k)}}{\beta} \\
&+ \nu_i\alpha\log\beta + \nu_i\log\Gamma\left(\alpha\right) - (1-\nu)\log\left(\alpha, \frac{\tau_i^L}{\beta}\right) + (\alpha-1)\log t_i^{(k)} - \frac{t_i^{(k)}}{\beta} \\
&- \alpha\log\beta - \log\Gamma\left(\alpha\right)\Bigg] \\
=& \sum\Bigg[-\nu_i\log\Gamma\left(\alpha\right) + \nu_i\log\Gamma\left(\alpha\right) - (1-\nu_i)\log\Gamma\left(\alpha, \frac{\tau_i^L}{\beta}\right) + (\alpha-1)\log t_i^{(k)} - \frac{t_i^{(k)}}{\beta} \\
&- \alpha\log\beta - \log\Gamma\left(\alpha\right)\Bigg] \\
=& \sum\left[(\alpha-1)\log t_i^{(k)} - \frac{t_i^{(k)}}{\beta} - \alpha\log\beta\right] - \sum\left[\nu_i\log\Gamma\left(\alpha\right) + (1-\nu_i)\log\Gamma\left(\alpha, \frac{\tau_i^L}{\beta}\right)\right]
\end{aligned}
$$

where $\boldsymbol{t^{(k)}}$ is the pseudo complete data and $t_i^{(k)} = w_i^{(k)}$ if censored. In the M-step we choose the initial values for $\left(\alpha_0^{(k)}, \beta_0^{(k)}\right)$ and use an approximate optimisation method to obtain an updated estimate $\boldsymbol{\theta}^{(k+1)}$ which maximises the pseudo-complete log-likelihood function. Lastly, the SE and the M steps are alternatively repeated for large cycles or iterations and the number of iterations to consider are user defined. If there are N iterations, one obtains a sequence of estimates $\boldsymbol{\theta}^{(k)}$ where $k = 1, 2, ..., N$. It should be noted that the first few estimates are discarded for burn-in and an average over the remaining estimates is taken so as to get the final estimate $\hat{\boldsymbol{\theta}}$. Suppose the first A estimates are discarded for burn-in and an average over the remaining estimates is taken so as to get the final estimate $\hat{\boldsymbol{\theta}}$. Thus

$$\hat{\boldsymbol{\theta}} = \frac{1}{N-A}\sum_{k=A+1}^{N} \boldsymbol{\theta}^{(k)}.$$

Exponential distribution

The SEM-algorithm for exponential distribution with parameter $\boldsymbol{\lambda}$ is based on replacing the missing data with the random observations drawn from the conditional distribution, this is to say in the SE-step we randomly draw values of $W^{(k)}$ from the conditional distribution in order to generate the pseudo-complete data. The conditional distribution of W_i given the observed data is given by

$$f_{W_i|W_i>z_i}(w) = \begin{cases} \dfrac{e^{-\left(w-\tau_i^L\right)/\lambda}}{\lambda}, & w > z_i \\ 0, & \text{otherwise} \end{cases}$$

where i^{th} unit is right censored. The imputation is implemented by generating the random variates using the inverse transformation method (Shang, 2021). Once the random observations have been generated we then have the pseudo-complete data. The pseudo-complete data is then used to obtain the pseudo-complete log-likelihood function. Thus the function can then be expressed as

$$\begin{aligned}
Q\left(\lambda|\boldsymbol{w^{(k)}}\right) &= \log L\left(\lambda|\boldsymbol{t^{(k)}}\right) \\
&= \sum_{i=1}^{n} \nu_i \log f_W\left(t_i^{(k)}\right) + \sum_{i=1}^{n}(1-\nu_i)\frac{\log f_W\left(t_i^{(k)}\right)}{1-F_W(\tau_i^L)} \\
&= \sum_{i=1}^{n} \nu_i \log\left(\frac{e^{-t_i^{(k)}/\lambda}}{\lambda}\right) + \sum_{i=1}^{n}(1-\nu_i)\log\left(\frac{e^{-t_i^{(k)}/\lambda}}{\lambda\, e^{-\tau_i^L/\lambda}}\right) \\
&= \sum\left[\nu_i \log\left(\frac{e^{-t_i^{(k)}/\lambda}}{\lambda}\right) - \nu_i \log\left(\frac{e^{-t_i^{(k)}/\lambda}}{\lambda}\right) + \log\left(\frac{e^{-t_i^{(k)}/\lambda}}{\lambda}\right) - (1-\nu_i)\log e^{-\tau_i^L/\lambda}\right] \\
&= \sum\left[\log\left(\frac{e^{-t_i^{(k)}/\lambda}}{\lambda}\right) - (1-\nu_i)\log e^{-\tau_i^L/\lambda}\right] \\
&= \sum\left(\log\left(\frac{1}{\lambda}\right) - \frac{t_i^{(k)}}{\lambda} + (1-\nu_i)\frac{\tau_i^L}{\lambda}\right) \\
&= \sum\left[\log\left(\frac{1}{\lambda}\right) - \frac{t_i^{(k)}}{\lambda}\right] + \sum(1-\nu_i)\frac{\tau_i^L}{\lambda}
\end{aligned}$$

where $\boldsymbol{t}^{(\boldsymbol{k})}$ is the pseudo complete data and $t_i^{(k)} = w_i^{(k)}$ if censored. In the M-step we choose the initial values for $\lambda_0^{(k)}$ and use an approximate optimisation method to obtain an updated estimate $\boldsymbol{\lambda}^{(k+1)}$ which maximises the pseudo-complete log-likelihood function. Lastly, the SE and the M steps are alternatively repeated for the large cycles or iterations and the number of iterations to consider are user defined. If there are N iterations, one obtains a sequence of estimates $\boldsymbol{\lambda}^{(k)}$ where k=1,2,...,N. It should be noted that the first few estimates are discarded for burn-in then an average over the remaining estimates is taken so as to get the final estimate $\hat{\boldsymbol{\lambda}}$. Suppose the first A estimates are discarded for burn-in then an average over the remaining estimates is taken so as to get the final estimate $\hat{\boldsymbol{\lambda}}$. Thus

$$\hat{\boldsymbol{\lambda}} = \frac{1}{N-A} \sum_{k=A+1}^{N} \boldsymbol{\lambda}^{(k)}.$$

4.1.5 Newton-Raphson algorithm

The Newton-Raphson algorithm is an iterative procedure that can be used to calculate ML-estimates. The basic idea behind the algorithm is to construct the likelihood function around some initial parameter value $\boldsymbol{\theta}^{(0)}$ and adjust the parameter value to that which maximises the likelihood function. According to Emura and Michimae (2022), this algorithm starts with some initial parameter value $\boldsymbol{\theta}^{(0)}$ and then the sequence follows

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \left(\frac{\partial^2 \log L\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} \right)^{-1} \frac{\partial \log L\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} | \boldsymbol{\theta}^{(k)}, k = 0, 1, 2, ...$$

Iterate until the parameter value convergences to $||\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}|| \approx 0$ (Emura and Michimae, 2022).

The Newton-Raphson algorithm is a powerful method for finding roots of nonlinear equations and has many practical applications in science, engineering, and finance. However, it is not always guaranteed to converge to a root, and can sometimes converge to a local minimum or maximum of the function. Careful selection of the initial guess is therefore important to ensure convergence to the desired root.

4.2 Practical Application

In this section we provide an exposition of how we estimated the ML-estimates for gamma and exponential distributions for left-truncated and right-censored data using EM-algorithm and also how we compared the performance of the EM-algorithm to other algorithms, namely Newton-Raphson (NR) and Stochastic expectation maximisation (SEM).

Comparisons of the performance of the EM-algorithm and other algorithms (SEM and NR) were based on simulation for various degrees of censoring and sample sizes using mean square error (MSE), bias, accuracy, Akaike Information Criterion (AIC) and the Bayesian Information Criterion(BIC). We implemented the EM-algorithm, NR-algorithm, and SEM-algorithm for both gamma and exponential distributions using left-truncated and right-censored data. We chose a set of true parameter values for the distributions, and simulated multiple datasets of left-truncated and right-censored data using these parameter values. For each simulated dataset, we estimated the parameters using the EM-algorithm, NR-algorithm, and SEM-algorithm, and recorded the estimated parameter values. The mean square error (MSE) provided a measure of the average squared difference between the estimated and true parameter values. It is calculated using the formula:

$$\text{MSE} \ = \sum_{i=1}^{m} \left(\frac{\left(\hat{\lambda}_i - \lambda\right)^2}{m} \right)$$

where $\hat{\lambda}$ is the estimated parameter value, λ is the true parameter value, and m is the number of simulated datasets.

Bias measures the systematic deviation between the estimated and true parameter values. It is computed as the difference between the expected value of the estimated parameter and the true parameter value:

$$\text{Bias} \ = \ \text{E}\left(\hat{\lambda}\right) - \lambda$$

Accuracy assesses the accuracy of the estimated parameters by quantifying the proportion of estimates within a specified tolerance range around the true parameter values. It can be expressed as a percentage:

$$\text{Accuracy} \ = \left(1 - \frac{|\hat{\lambda} - \lambda|}{\lambda}\right) * 100$$

The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are statistical measures used for model selection and comparison. They consider both the goodness of fit and the complexity of the model. The formulas for AIC and BIC are as follows:

$$\begin{aligned} \text{AIC} \ &= -2 * \text{log-likelihood } + 2 * \text{ number of parameters} \\ \text{BIC} \ &= -2 * \text{log-likelihood } + \log(n) * \text{ number of parameters} \end{aligned}$$

where log-likelihood is the logarithm of the likelihood function, and n is the sample size. A small AIC or BIC is indicative of a better model.

By conducting simulations with various degrees of censoring and sample sizes, we evaluated the performance of the EM-algorithm, SEM -algorithm, and NR-algorithm using these metrics. This analysis provided insights into their accuracy, precision, computational efficiency, model selection capabilities, and robustness in estimating the parameters of the gamma and exponential distributions for left-truncated and right-censored data. All the estimations in this study were based on the simulated data from exponential and gamma distributions. All computations were done using R 4.3 version.

CHAPTER 5

Results, Analysis and Discussion

This chapter presents the analysis and discussion of our results. The focus was on estimating the ML-estimates for gamma and exponential distributions when the data was left-truncated and right-censored and making comparisons of the performances of the EM-algorithm to other commonly used algorithms.

5.1 Parameter estimation for exponential and gamma distributions using the EM-algorithm

In this section we used the EM-algorithm to estimate the parameters of the exponential and gamma distributions for different sample sizes and truncation percentages based on 100 simulations.

Table 5.1: Exponential distribution - sample size(n), truncation(trunc), censoring(Cens), parameter (λ), estimate ($\hat{\lambda}$), bias(B), mean square error (MSE) and accuracy(Accu) in 100 simulations.

λ n	Trunc(%) Cens(%)	$\hat{\lambda}$	B($\hat{\lambda}$)	MSE ($\hat{\lambda}$)	Accu
λ=0.8	05 ; 50	0.89	0.091	0.008	0.88
	30 ; 50	1.16	0.36	0.13	0.54
$n = 50$	15 ; 10	0.6	-0.2	0.047	0.786
	60 ; 30	0.35	-0.53	0.29	0.4
λ=0.9	05 ; 50	0.78	-0.116	0.013	0.87
	30 ; 50	1.06	0.16	0.026	0.819
$n = 50$	15 ; 10	1.01	0.11	0.01	0.86
	60 ; 30	0.43	-0.47	0.237	0.51

Table 5.1: Exponential distribution - sample size(n), truncation(trunc), censoring(Cens), parameter (λ), estimate ($\hat{\lambda}$), bias(B), mean square error (MSE) and accuracy(Accu) in 100 simulations for n= 100 and 1000.

$\lambda = 0.8$	05 ; 50	0.86	0.06	0.00376	0.91
	30 ; 50	1.15	0.35	0.12	0.58
$n = 100$	15 ; 10	0.975	0.17	0.03	0.798
	60 ; 30	0.47	-0.42	0.18	0.45
$\lambda = 0.9$	05 ; 50	0.81	-0.08	0.007	0.9
	30 ; 50	1.056	0.15	0.02	0.82
$n = 100$	15 ; 10	0.115	0.115	0.013	0.87
	60 ;30	0.54	-0.43	0.21	0.56
$\lambda = 0.8$	05 ; 50	0.803	0.0039	0.000015	0.99
	30 ; 50	1.08	0.28	0.08	0.606
$n = 1000$	15 ; 10	0.90	0.107	0.01	0.86
	60 ; 30	0.57	-0.33	0.11	0.58
$\lambda = 0.9$	05 ; 50	0.83	0.07	0.0049	0.92
	30 ; 50	1.03	0.13	0.0169	0.86
$n = 1000$	15 ; 10	0.99	0.09	0.008	0.9
	60 ; 30	0.61	-0.39	0.15	0.61

Table 4.1 show ML-estimates for exponential distribution with parameter λ when the data was simulated from an exponential distribution using the EM-algorithm under 3 different sample sizes (50, 100 and 1000) and four different truncation (Trunc) and censoring (Cens) percentages. Where $n = 50$ is regarded as small sample size, $n = 100$ is slightly large and $n = 1000$ is large sample. The table shows that the EM-algorithm produced minimal MSE and bias when the left-truncation percentage was 5%, outperforming scenarios of $15\%, 30\%$, and 60% truncation. This observation suggests that lower degrees of truncation could result in more accurate parameter estimation. Additionally, increasing the sample size was associated with a decrease in mean squared error (MSE) and bias. Furthermore, the accuracy tended to be good in cases where the sample size was large and the truncation percentage low. This indicates that large sample size and low left-truncation percentage tend to yield more precise and less biased parameter estimates.

The EM-algorithm resulted in elevated bias and MSE in cases of high censoring and high truncation percentages. This indicates that the algorithm is more likely to converge reliably with larger samples, but high levels of censoring and truncation can still pose challenges.

This is so because high levels of truncation 60% resulted in a loss of most of the data, which, in turn, led to biased estimates of the parameters being studied. Overall, these findings suggest that the performance of the EM algorithm is influenced by the levels of censoring, truncation and sample size. Moderate levels of censoring and low truncation, along with larger sample sizes, tend to yield more accurate and less biased estimates.

Table 5.2: Gamma distribution - sample size(n), truncation(trunc), censoring(Cens), parameters (α , β), estimates ($\hat{\alpha}$, $\hat{\beta}$), bias(B), mean square error (MSE), and accuracy (Accu) based on 100 simulations.

n (α, β)	Trunc(%) Cens(%)	$\left(\hat{\alpha}, \hat{\beta}\right)$	B $(\hat{\alpha})$, B $(\hat{\beta})$	MSE$(\hat{\alpha})$, MSE$(\hat{\beta})$	Accu(α) , Accu(β)
	05 ; 50	1.11 , 0.61	-0.88 , -0.39	0.99 , 0.16	0.56 , 0.629
n=50	30 ; 50	0.97 , 0.54	-1.03 , -0.46	1.08 , 0.26	0.48 , 0.57
(2 , 1)	15 ; 10	1.17 , 0.69	-0.83 , -031	0.71 , 0.096	0.59 , 0.689
	60 ; 30	0.67 , 0.4	-1.53 , -0.7	2.35 , 0.49	0.34 , 0.39
n=50	05; 50	3.12 , 2.78	-1.89 , 0.789	3.42 , 0.64	0.63 , 0.61
	30 ; 50	3.01 , 2.87	-1.99 , 0.91	4.1 , 0.79	0.6 , 0.47
	15 ; 10	3.53 , 2.56	-1.46 , 0.58	2.21, 0.30	0.70 , 0.71
$(5, 2)$	60 ; 30	2.4 , 1.2	-2.8, -0.99	7.84 , 0.98	0.47 , 0.55
n=100	05; 50	1.61 , 0.91	-0.39 , -0.09	0.161 , 0.008	0.8 , 0.91
	30 ; 50	1.5 , 0.87	-0.49 , -0.126	0.62 , 0.14	0.75 , 0.87
$(2, 1)$	15 , 10	1.7 , 0.93	-0.31 , -0.069	0.094 , 0.0047	0.871 , 0.94
	60 ; 30	0.87 , 0.6	-1.15 , -0.39	1.42 , 0.17	0.45 , 0.54
n=100	05 ; 50	3.41 , 2.54	-1.62, , 0.57	2.5 , 0.31	0.68 , 0.71
$(5, 2)$	30 ; 50	3.11 , 2.81	-1.88 , 0.84	3.51 , 0.68	0.63 , 0.58
	15 ; 10	4.21 , 2.37	-0.83 , 0.41	0.64 , 0.21	0.84 , 0.81
	60 ; 30	2.8 , 1.31	-2.2 , -0.77	4.61 , 0.60	0.56 , 0.59
n=1000	05 ; 50	1.75 , 0.94	-0.25 , -0.06	0.072 , 0.0038	0.87 , 0.87
	30 ; 50	1.54 , 0.856	-0.45 , -0.14	0.2 , 0.02	0.77 , 0.87
$(2, 1)$	15; 10	1.81 , 0.96	-0.21 , -0.039	0.037 , 0.015	0.91 , 0.97
	60 ; 30	1.13 , 0.7	-0.7 , -0.28	0.5 , 0.099	0.59 , 0.65
n =1000	05 ; 50	4.13 , 2.34	-0.89 , 0.37	0.77 , 0.17	0.83 , 0,81
	30 ; 50	3.81 , 2,51	-1.2 , 0.53	1.51 , 0.28	0.77 , 0.74
	15 ; 10	4.8 , 2.19	-0.23 , 0.3	0.051 , 0.08	0.96 , 0.86
$(5, 2)$	60 ; 30	3.0 , 1.35	-1.99 , -0.68	4.01 , 0.476	0.61 , 0.67

Table 4.2 shows the ML-estimates for gamma distribution with parameters (α , β) when the data was simulated from a gamma distribution using the EM-algorithm under 3 different sample sizes (50, 100 and 1000) and four different truncation(Trunc) and censoring (Cens) percentages. The table shows that the EM-algorithm produced estimates with low bias under low and moderate levels of truncation and censoring (15% , 10%) and (05% , 50%). High levels of censoring and truncation led to a loss of most of the data, resulting in biased parameter estimates. Increasing the sample size resulted in decreased bias and mean squared error (MSE), suggesting that large samples yields more precise and accurate parameter estimates. Overall, estimating parameters for the gamma distribution using the EM algorithm was more challenging compared to the exponential distribution due to the complexity of determining appropriate initial values for multiple parameters.

5.2 Comparisons of the algorithms

Under this section we compared the EM-algorithm, SEM-algorithm and NR-algorithm using the MSE, bias, AIC, BIC, convergence speed and accuracy.

Table 5.3: Sample size(n), truncation(trunc), censoring(Cens), parameter (λ), algorithms (ALGO), estimate ($\hat{\lambda}$), bias(B), mean square error (MSE), Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), average number of iterations (NI) and accuracy (Accu) for the exponential distribution using three different algorithms (EM, SEM, and NR) based on 100 simulations.

n λ	Trunc(%) Cens (%)	ALGO	$\hat{\lambda}$	B($\hat{\lambda}$)	MSE ($\hat{\lambda}$)	AIC	BIC	NI	Accu
n=50	(05;50)	EM	0.89	0.091	0.008	71.5	72.1	5	0.88
		SEM	0.92	0.11	0.0126	74.5	75.1	8	0.876
		NR	0.92	0.112	0.014	72.9	73.8	5	0.87
λ=0.8	(15; 10)	EM	0.62	-0.2	0.047	72.35	72.65	8	0.786
		SEM	0.61	-0.29	0.089	74.7	75.4	11	0.75
		NR	0.59	-0.23	0.049	73.3	74.1	8	0.776
$n = 50$	(05; 50)	EM	0.78	-0.116	0.013	72.35	72.65	5	0.87
		SEM	0.752	-0.68	0.0201	74.7	78.9	11	0.75
		NR	0.776	-0.134	0.0179	75.3	76.1	6	0.776
$\lambda = 0.9$	(60; 30)	EM	0.43	-0.47	0.23	76.1	76.35	8	0.52
		SEM	0.41	-0.49	0.245	80.2	80.6	14	0.49
		NR	0.4	-0.51	0.25	80.5	80.8	7	0.48

Table 5.3: Sample size(n), truncation(trunc), censoring(Cens), parameter (λ), algorithms (ALGO), estimate ($\hat{\lambda}$), bias(B), mean square error (MSE), Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), average number of iterations (NI) and accuracy (Accu) for the exponential distribution using three different algorithms (EM, SEM, and NR) based on 100 simulations for $n = 100$ and $n = 1000$.

n λ	Trunc(%) Cens (%)	ALGO	$\hat{\lambda}$	B($\hat{\lambda}$)	MSE ($\hat{\lambda}$)	AIC	BIC	NI	Accu
$n = 100$	$(05; 50)$	EM	0.86	0.06	0.004	207.8	210.45	6	0.92
		SEM	0.864	0.071	0.0049	207	207.9	11	0.912
		NR	0.878	0.074	0.00547	210.8	211.5	9	0.9
λ=0.8	$(15; 10)$	EM	0.975	0.17	0.037	267.5	270.1	5	0.78
		SEM	0.964	0.194	0.039	277.9	279.3	27	0.789
		NR	0.976	0.19	0.042	278.1	280.5	8	0.74
$n = 100$	$(05; 50)$	EM	0.81	-0.08	0.007	206.1	208.7	6	0.92
		SEM	0.84	-0.09	0.037	211.2	213.1	16	0.89
		NR	0.809	-0.095	0.0082	210.83	211.2	8	0.9
	$(60; 30)$	EM	0.54	-0.43	0.2	213	213.4	9	0.57
λ=0.9		SEM	0.539	-0.45	0.24	216	216.8	16	0.55
		NR	0.52	-0.4	0.287	217.7	218.2	10	0.51
$n = 1000$	$(05; 50)$	EM	0.83	0.0039	0.00001	2108.5	2113.44	9	0.98
		SEM	0.802	0.009	0.00008	2129.3	2134.2	22	0.92
		NR	0.84	0.039	0.00156	2132.9	2134.8	13	0.899
λ=0.8	$(15; 10)$	EM	0.9	0.107	0.001	2656.1	2661.0	5	0.86
		SEM	0.91	0.115	0.0017	2656.2	2661.1	18	0.83
		NR	0.94	0.152	0.0196	2660.2	2662.4	8	0.823
$n = 1000$	$(05; 50)$	EM	0.83	-0.07	0.0049	2043.4	2043.9	5	0.92
		SEM	0.86	-0.08	0.0052	2045.4	2045.8	16	0.90
		NR	0.84	-0.074	0.0051	2044.1	2045.3	8	0.914
$\lambda = 0.9$	$(60; 30)$	EM	0.61	-0.38	0.14	2090	2091.1	11	0.61
		SEM	0.611	-0.389	0.145	2091.1	2092	18	0.62
		NR	0.6	-0.41	0.156	2096.1	2096	9	0.59

Table 4.3 shows the results of the comparisons of the 3 different algorithms, when the data was simulated from the exponential distribution. The table shows that the EM-algorithm, NR algorithm and the SEM-algorithm produced similar estimates, but the EM-algorithm and the NR-algorithm demonstrated faster convergence than the SEM-algorithm. The EM-algorithm outperformed other algorithms in terms of MSE, bias, AIC, BIC, accuracy and convergence speed in most simulations. Additionally, the algorithms performed better under low truncation and moderate censoring percentage $(05\%, 50\%)$. Based on these findings, it can be concluded that the EM-algorithm is a favourable choice for handling left-truncated and right-censored data from an exponential distribution, especially when dealing with moderate levels of censoring and truncation. Its faster convergence, accuracy, and better performance in various evaluation metrics make it a preferred algorithm for such scenarios.

Table 5.4: Sample size(n), truncation(trunc), censoring(Cens), parameters (α , β), algorithms (ALGO), estimates ($\hat{\alpha}$, $\hat{\beta}$), bias(B), mean square error (MSE), Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), average number of iterations (NI) and accuracy (Accu) for the gamma distribution using three different algorithms (EM, SEM, and NR) based on 100 simulations for n=50.

n	Trunc(%)	ALGO	$\hat{\alpha}$	B($\hat{\alpha}$)	MSE($\hat{\alpha}$),	AIC	BIC	NI	Accu($\hat{\alpha}$),
(α, β)	Cens(%)		$\hat{\beta}$	B($\hat{\beta}$)	MSE($\hat{\beta}$)				Accu($\hat{\beta}$)
$n = 50$	05 ; 50	EM	1.11	-0.88	0.99	279	280	15	0.56
			0.61	-0.39	0.16				0.61
		SEM	1.08	-0.9	1.02	282	284	28	0.52
			0.59	-0.42	0.2				0.57
		NR	0.99	-0.97	1.02	282.1	285	17	0.49
			0.6	-0.42	0.19				0.59
(2 , 1)	15; 10	EM	1.17	-0.83	0.71	259	260.1	15	0.59
			0.69	-0.31	0.96				0.69
		SEM	1.178	-0.85	0.71	262	263.4	26	0.58
			0.67	-0.32	0.96				0.67
		NR	1.179	-0.85	0.78	264	265.5	13	0.55
			0.63	0.38	0.99				0.66
$n = 50$	05; 50	EM	3.01	-1.89	3.402	316	318	19	0.63
			2.78	0.789	0.64				0.61
		SEM	3.12	-1.90	3.42	316.2	319	32	0.6
			2.80	0.79	0.67				0.59
$(5, 2)$		NR	3.08	-1.99	3.49	319.8	320	16	0.59
			2.82	0.83	0.68				0.58
	60; 30	EM	2.4	-2.8	7.86	338	338.4	15	0.47
			1.2	-0.99	0.98				0.57
		SEM	2.37	-2.61	8.1	341	342.1	22	0.45
			1.19	-1.1	1.3				0.56
		NR	2.37	-2.61	8.1	339.2	340	13	0.445
			1.16	-1.13	1.37				0.54

Table 5.4: Sample size(n), truncation(trunc), censoring(Cens), parameters (α , β), algorithms (ALGO), estimates ($\hat{\alpha}$, $\hat{\beta}$), bias(B), mean square error (MSE), Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), average number of iterations (NI) and accuracy (Accu) for the gamma distribution using three different algorithms (EM, SEM, and NR) based on 100 simulations for n=100.

n (α, β)	Trunc(%) Cens(%)	ALGO	$\hat{\alpha}$ $\hat{\beta}$	B($\hat{\alpha}$) B($\hat{\beta}$)	MSE($\hat{\alpha}$), MSE($\hat{\beta}$)	AIC	BIC	NI	Accu($\hat{\alpha}$), Accu($\hat{\beta}$)
$n = 100$	$05; 50$	EM	1.69	-0.312	0.1	543	544.2	18	0.8
			0.91	-0.07	0.0059				0.90
$(2, 1)$		SEM	1.67	-0.33	0.109	544	545.8	29	0.79
			0.92	-0.08	0.006				0.9
		NR	1.63	-0.5	0.291	544.2	546	17	0.77
			0.9	-0.14	0.012				0.89
	$15; 10$	EM	1.7	-0.31	0.094	527	529	17	0.871
			0.93	-0.069	0.0047				0.94
		SEM	1.71	-0.28	0.079	533	533.5	26	0.88
			0.96	-0.04	0.0016				0.995
		NR	1.69	-0.32	0.0961	531	532	18	0.863
			0.95	-0.067	0.003				0.95
$n = 100$	$(05; 50)$	EM	3.41	-1.62	2.5	628	630	18	0.70
			2.34	0.332	0.108				0.71
		SEM	3.40	-1.63	2.568	635	636	31	0.68
			2.32	0.33	0.104				0.712
$(5, 2)$		NR	3.397	-1.67	2.79	634	635.6	18	0.674
			2.53	0.59	0.348				0.70
	60 ; 30	EM	2.8	-2.2	4.61	651	652.1	12	0.56
			1.31	0.77	0.6				0.59
		SEM	2.81	-2.19	4.6	653	654.3	12	0.57
			1.3	-0.78	0.64				0.58
		NR	2.83	-2.3	4.7	654.7	656	10	0.54
			1.29	-0.8	0.76				0.56

Table 5.4: Sample size(n), truncation(trunc), censoring(Cens), parameters ($\alpha\,,\beta$), algorithms (ALGO), estimates ($\hat{\alpha}\,,\hat{\beta}$), bias(B), mean square error (MSE), Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), average number of iterations (NI) and accuracy (Accu) for the gamma distribution using three different algorithms (EM, SEM, and NR) based on 100 simulations for $n = 1000$.

n (α, β)	Trunc(%) Cens(%)	ALGO	$\hat{\alpha}$ $\hat{\beta}$	B($\hat{\alpha}$) B($\hat{\beta}$)	MSE($\hat{\alpha}$), MSE($\hat{\beta}$)	AIC	BIC	NI	Accu($\hat{\alpha}$), Accu($\hat{\beta}$)
n=1000	05; 50	EM	1.75	-0.25	0.072	5458	5459	18	0.87
			0.94	-0.06	0.0038				0.87
$(2, 1)$		SEM	1.74	-0.27	0.073	5461	5463	28	0.879
			0.94	-0.061	0.0038				0.864
		NR	1.72	-0.278	0.079	5461.2	5463	19	0.84
			0.93	-0.063	0.0049				0.85
	15; 10	EM	1.81	-0.21	0.037	5433	5437	17	0.91
			0.96	-0.039	0.015				0.97
		SEM	1.809	-0.23	0.039	5435	5439	28	0.9
			0.949	-0,041	0.018				0.95
		NR	1.72	-0.29	0.041	5437	5440	15	0.89
			0.947	-0.042	0.018				0.95
$n = 1000$	05; 50	EM	4.13	-0.89	0.77	6027	6032	17	0.83
			2.34	0.37	0.177				0.81
$(5, 2)$		SEM	4.137	-0.86	0.72	6029	6034.8	26	0.84
			2.48	-0.389	0.17				0.79
		NR	4.14	-0.78	0.69	6027.8	6033	15	0.867
			2.32	0.33	0.169				0.83
	60; 30	EM	3.0	-1.99	4.01	6050	6056	13	0.61
			1.35	-0.68	0.476				0.67
		SEM	3.01	-1.99	4.03	6053	6057	21	0.61
			1.349	-0.681	0.47				0.65
		NR	3.04	-2.0	4.1	6053.6	6058	13	0.6
			1.32	-0.71	0.59				0.63

Table 4.4 show the results of the comparisons of the 3 different algorithms (EM, SEM and NR), when the data was simulated from the gamma distribution under 3 different sample sizes (50, 100 and 1000) and three different truncation (Trunc) and censoring (Cens) percentages. The tables show that the EM-algorithm consistently outperformed the other algorithms

across various metrics, such as MSE, bias, accuracy, convergence speed, AIC, and BIC. As the level of truncation escalated to higher percentages 60% all the algorithms performed poorly. This finding underscores the importance of carefully considering the extent of truncation and censoring. For low truncation and censoring levels, increased sample size led to improved parameter estimates for all the algorithms. In terms of convergence the SEM-algorithm took longer to converge than the EM-algorithm and NR- algorithm.

5.3 Model selection validation

Under this section we validated the efficiency of the AIC and BIC under two candidate models - the gamma and exponential distributions.

Table 5.5: Model selection validation: True model versus fitted model using AIC and BIC.

n	Parameters	True model	Fitted model	AIC	BIC
50	1.2	Exponential	Exponential	248	250
			Gamma	290	294
	(0.5 ,2)	Gamma	Exponential	182.7	184
			Gamma	130.1	130.4
100	1.2	Exponential	Exponential	485	488.08
			Gamma	593	598
	(0.5 , 2)	Gamma	Exponential	365	367
			Gamma	280	282
1000	1.2	Exponential	Exponential	4824	4829
			Gamma	5915	5925
	(0.5 ,2)	Gamma	Exponential	3634	3639.7
			Gamma	2739	2742

Table 4.5 validates the model selection performance of AIC and BIC when the data was simulated from both gamma and exponential models under different sample sizes. When the data was simulated from the exponential distribution, we observed that the exponential

distribution yielded the smallest values for both AIC and BIC. This indicates that the exponential distribution provided the best fit to the simulated data according to the AIC and BIC criteria. When the data was simulated from the gamma distribution, we observed that the gamma distribution yielded the smallest values for both AIC and BIC. This finding shows that the gamma distribution provided the best fit to the simulated data according to the AIC and BIC criteria.The findings reinforce the notion that the AIC and BIC are valuable tools for model selection, as they can effectively capture the goodness-of-fit and determine the most suitable distribution for a given dataset.

5.4 Discussion

The findings from this study demonstrate that the EM-algorithm performed well in finding maximum likelihood (ML) estimates for the Exponential distribution, when the sample size was large and the level of left-truncation low (5%). It should be noted that the EM-algorithm performed poorly under small sample sizes and high truncation percentages, indicating that accurate estimation may be challenging under such conditions. This finding is consistent with previous studies, such as the work conducted by Balakrishnan and Mitra (2012) and Emura (2016), which also highlighted the limitations of small sample sizes and high levels of truncation on parameter estimation using the EM-algorithm. Estimating ML-estimates for the gamma distribution using the EM-algorithm posed greater challenges compared to the exponential distribution due to the gamma distribution's additional parameter. Determining appropriate initial values for the parameters was a challenge in achieving reliable estimates. The algorithm generally performed well with larger sample sizes, but performed poorly with small sample sizes. The impact of censoring and truncation on the accuracy of parameter estimation was similar to that observed in the exponential distribution case, with high levels of censoring and truncation leading to escalated bias in estimates due to the loss of data. This finding agrees with the results of the study conducted by Shang (2021), highlighting the challenges associated with estimating gamma distribution parameters using the EM-algorithm. Comparing the performance of the EM-algorithm with other algorithms (Newton-Raphson and SEM) for both exponential and gamma distributions showed that all the three algorithms produced similar estimates. Additionally all the algorithms are initial value sensitive.

CHAPTER 6

Conclusion and Recommendations

6.1 Conclusion

In conclusion, this study focused on finding ML-estimates for the Exponential and Gamma distributions using the EM-algorithm, as well as comparing its performance with the SEM and NR algorithms. The findings revealed valuable insights and provided important implications for parameter estimation under various scenarios of censoring and truncation.

For the exponential distribution, the EM-algorithm performed well and produced estimates with low bias when the levels of censoring and truncation were moderate. However, challenges were observed with small sample sizes. The impact of high levels of censoring and truncation was also evident, resulting in estimates with high bias due to loss of data.

The EM-algorithm generally performed well with larger sample sizes but faced difficulties with smaller sample sizes. The impact of censoring and truncation on parameter estimation in the gamma distribution was similar to the findings in the Exponential distribution case.

A number of observations were made when the EM-algorithm was compared with the SEM and NR algorithms. Across the board, these algorithms grapple with the constraints of small sample sizes, urging a cautious interpretation of results under such circumstances. However, the EM-algorithm emerges as the standout performer, surpassing its counterparts in multiple critical metrics including mean squared error (MSE), bias, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), convergence speed, and overall accuracy. While the SEM-algorithm often produces comparable estimates to the EM-algorithm and NR-algorithm, but it lags behind in terms of convergence speed.

Additionally, our utilization of the AIC and BIC criteria demonstrated their efficacy in guiding model selection. These criteria consistently pointed towards the Exponential and Gamma distributions as the most suitable models when they indeed represented the true underlying distributions for our simulated data. These results corroborate findings from prior research.

In conclusion, this study contributes to the field of parameter estimation for censored and truncated data, highlighting the strengths and limitations of the EM-algorithm, SEM, and NR algorithms. The findings are helpful for researchers and practitioners working with censored and truncated data, guiding them in selecting appropriate estimation methods and addressing challenges associated with sample size, censoring, and truncation.

6.2 Recommendations

In light of the study's findings, several recommendations can be made. Firstly, further exploration of initialization techniques for the EM-algorithm, SEM, and NR algorithms should be conducted to improve their performance and address challenges associated with determining suitable initial values. Additionally, future research should investigate the performance of different algorithms under various censoring and truncation scenarios such as doubly truncated or interval censoring to gain a comprehensive understanding of their strengths and limitations.

Finally, future research can consider estimating parameters for censored and truncated data using the Bayesian approach. Bayesian methods offer a different perspective and can provide abroad understanding of parameter estimation and model selection. Comparing the Bayesian approaches with the EM, SEM and NR algorithms will contribute to a comprehensive understanding of different estimation techniques and their respective strengths.

REFERENCES

[1] Balakrishnan, N. and D. Mitra. 2011. “Likelihood inference for lognormal data with left truncation and right censoring with an illustration.” Journal of Statistical Planning Inference, 141, 3536-3553.

[2] Balakrishnan, N. and D. Mitra. 2012. “Likelihood Inference Based on Left truncated and right Censored lifetime data.” Institute of Electrical and Electronics Engineers Transactions on Reliability, 62, 679-688.

[3] Balakrishnan, N. and M. H. Ling. 2013. “Expectation maximization algorithm for one shot device accelerated life testing with weibull lifetimes, and variable parameters over stress.” Institute of Electrical and Electronics Engineers Transactions on Reliability, 62(2), 537–551.

[4] Balakrishnan, N., H. Y. So and M. H. Ling. 2016. “EM Algorithm for one-shot device testing with competing risks under weibull distribution.” Institute of Electrical and Electronics Engineers Transactions on Reliability, 65(2), 973–991.

[5] Baikunth, G., 1975. “Unbiased estimates of reliability for the truncated gamma distribution.” Scandinavian Actuarial Journal, 3, 181–186.

[6] Blight, B. J. N., 1970. “Estimation from a censored sample for the exponential family.” Biometrika, 57, 389-395.

[7] Broeder, D. G. G., 1955. “On parameter estimation for truncated Pearson Type III distributions.” Annals of Mathematical Statistics, 26,659-663.

[8] Carter, W. H. and R.H MYERS. 1973. “Maximum likelihood estimation from linear combinations of discrete probability functions.” Journal of the American Statistical Assocciation, 68, 203-206.

[9] Chandrakant, L., Y. M. Tripathi and M.K. Rastogi. 2019. “Estimating the parameters of a truncated normal distribution under progressive type censoring.” Communications in Statistics Simulation and Computation, 10, 16.

[10] Clark, C. E. and G.T. Williams. 1960. “Estimates from censored samples.” Annals of the Institute of Mathematical Statistics, Tokyo, 12,209-220.

[11] Cohen, A. C., Jr. 1949. “On estimating the mean and standard deviation of truncated normal distributions.” Journal of the American Statistical Association, 44,518-525.

[12] Cohen, A. C., Jr. 1950a. “ Estimating parameters of type III population from truncated samples.” Journal of the American Statistical Association, 45,336-349.

[13] Cohen, A. C., Jr . 1950b. “Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples.” Annals of Mathematical Statistics, 21,557-569.

[14] Deemer, W. L. and D. F. Votaw. 1955. “Estimation of parameters of truncated or censored exponential distributions”. Annals of Mathematics and Statistics, 26, 498-504.

[15] Dempster, A. P., N. Laird and D.B. Rubin. 1977. “Maximum likelihood from incomplete data via the EM algorithm.” Journal of the Royal Statistical Society, 39(1), 1–38.

[16] Des, Raj., 1953. “Estimation of the parameters of Type III populations from truncated samples.” Journal of the American Statistical Association, 48, 336-349.

[17] Emura, T., S. Shiu and S. Matsui. 2016. “Estimation and model selection for left-truncated and right-censored lifetime data with application to electric power transformers analysis.” Communication in Statistics-Simulation and Computation, 45(9), 3171-3189.

[18] Faris, M. A., 2008. “Estimation of mean of truncated exponential distribution”, Journal of mathematics and statistics, 4(4), 284-288.

[19] Fisher, R. A., 1931. “Properties and applications of functions.” British Association for the Advancement of Science, 1, 26-35.

[20] Gjeddebaek, N. F., (1949). “Contribution to the study of grouped observations I. Application of the method of maximum likelihood in case of normally distributed observations.” Skandinavisk Aktuarietidskrift, 32, 135-159.

[21] Gupta, A. K., 1952. “Estimation of the mean and standard deviation of a normal population from a censored sample.” Biometrika, 39, 260-273.

[22] Harter, H. L., 1967. “Maximum-likelihood estimation of the parameters of a four-parameter generalized gamma population from complete and censored samples.” Technometrics, 9(1), 159–165.

[23] Hartley, H. 0., 1958. “Maximum likelihood estimation from incomplete data.” Biometrics, 14, 174-194.

[24] Hald, A., 1949. “Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point.” Skandinavisk Aktuarietidskrift, 32, 119-134.

[25] Hughes, E. J., 1962. ”Maximum likelihood estimation of distribution parameters from incomplete data.” Ph.D. Thesis, Lowa State University. 2089.https://lib.dr.iastate.edu/rtd/2089

[26] Hegde, L. M., and R. C. Dahiya. 1989. “Estimation of the parameters of a truncated gamma distribution,” Communications in Statistics - Theory and Methods, 18 (2), 561–577.

[27] Johnson, W. R. and V.D. Kliche and P.L. Smith. 2014. “Maximum Likelihood Estimation from a left-truncated Gamma distribution.” Quaterly Journal of the Royal Meterological Society, 140, 1245-1256.

[28] Krane, S. A., 1957. “Maximum likelihood estimation from incomplete data for continuous distributions.” Ph.D. Thesis, Iowa State University of Science and Technology.

[29] Meeker, W. Q. and L.A. Escobar. 1998. “Statistical Methods for Reliability Data.” John Wiley & Sons.

[30] Mitra, D., D. Kundu and N.Balakrishnan. 2021. “Likelihood analysis and stochastic EM algorithm for left truncated right censored data and associated model selection from the Lehmann family of life distributions.” Japanese Journal of Statistics and Data Science, 4(2), 1019-1048.

[31] Mittal, M. M., 1984. “Estimating the parameters of truncated distributions.” Ph.D, Dissertation, Mathematics and Statistics Old Dominion University.

[32] Pearson, K. and A. Lee. 1908. “On the generalized probable error in multiple normal correlation.” Biometrika, 6, 59-68.

[33] Philippe, A., 1997. “Simulation of right and left truncated gamma distributions by mixtures.” Statistics and Computing, 7 (3), 173–181.

[34] Shang, X. and H. K. T. Ng. 2021. “On parameter estimation for the generalised gamma distribution based on left-truncated and right-censored data”. Computer and Mathematical methods, 3.

[35] Zaninetti, L., 2014. “A right and left truncated gamma distribution with application to the stars.” Advanced Studies in Theoretical Physics, 23, 1139-1147.

Appendices

6.3 Appendix A

The code for finding the ML-estimate, MSE, AIC, BIC and bias for left-truncated and right-censored data using the EM-algorithm.

```
set.seed()

# Define the parameters
n<-
lambda <-

# Generate exponential data
data <- rexp(n, rate = lambda)
data
# Sort the data
sorted_data <- sort(data)
sorted_data
# Define the left truncation and right censoring positions
ten_percent <- 0.1 * n
a <- sorted_data[ten_percent]
a
sorted_data <- sorted_data[-c(1:ten_percent)]
sorted_data
sorted_n <- length(sorted_data)
fifty_percent <- 0.5 * sorted_n
b <- sorted_data[fifty_percent]
```

```
sorted_data[fifty_percent:sorted_n] <- sorted_data[fifty_percent]
sorted_data
# Define the log-likelihood function for the left-truncated and right-censored data
log_likelihood <- function(lambda, sorted_data, a, b) {
log_like <- sum(log(dexp(sorted_data, rate = lambda) / (1 - pexp(a, rate = lambda))))
sum(log(pexp(sorted_data, rate = lambda) / pexp(b, rate = lambda)))
return(log_like)
}

# E-step: Calculate expected values of missing data
e_step <- function(lambda, sorted_data, a, b) {
# Calculate conditional probabilities for left-truncated data
probabilities <- dexp(sorted_data, rate = lambda) / (1 - pexp(a, rate = lambda))

# Calculate conditional survival probabilities for right-censored data
probabilities <- probabilities * pexp(sorted_data, rate = lambda) / pexp(b, rate = la

return(probabilities)
}

# M-step: Update parameter estimates
m_step <- function(probabilities, sorted_data) {
lambda <- sum(sorted_data * probabilities) / sum(probabilities)
return(lambda)
}

# EM algorithm
em_algorithm <- function(sorted_data, a, b, max_iter = , tol = ) {
# Initialize parameters
lambda <- 1 / mean(sorted_data)  # Initial estimate for rate parameter
iter <- 0  # Number of iterations

# Start timer
start_time <- proc.time()

# Iterate until convergence or maximum number of iterations reached
```

```
for (iter in 1:max_iter) {
# E-step
probabilities <- e_step(lambda, sorted_data, a, b)

# M-step
lambda_new <- m_step(probabilities, sorted_data)

# Check convergence
if (abs(lambda - lambda_new) < tol) {
break
}

lambda <- lambda_new  # Update parameter estimate
}

# End timer
end_time <- proc.time()
elapsed_time <- end_time - start_time

return(list(lambda = lambda, iterations = iter, time = elapsed_time))
}

# Initialize variables to store results
num_simulations <- 100
lambda_estimates <- numeric(num_simulations)
aics <- numeric(num_simulations)
bics <- numeric(num_simulations)
biases <- numeric(num_simulations)
mses <- numeric(num_simulations)
accuracies <- numeric(num_simulations)
num_iterations <- numeric(num_simulations)
elapsed_times <- numeric(num_simulations)
space_complexities <- numeric(num_simulations)
# Loop over the number of simulations
for (i in 1:num_simulations) {
# Simulate left-truncated and right-censored data
```

```
t <- sorted_data

# Estimate MLE using the EM algorithm
em_result <- em_algorithm(t, a, b)
lambda_hat <- em_result$lambda

# Calculate log-likelihood
log_like <- log_likelihood(lambda_hat, t, a, b)

# Calculate number of parameters
k <- 1

# Calculate AIC and BIC
aic <- -2 * log_like + 2 * k
bic <- -2 * log_like + k * log(length(t))

# Estimate MLE using the EM algorithm
em_result <- em_algorithm(t, a, b)
lambda_hat <- em_result$lambda
iterations <- em_result$iterations
time <- em_result$time

# Calculate additional statistics
bias <- lambda_hat - lambda
accuracy <- 1 - abs(lambda_hat - lambda) / lambda
space_complexity <- sum(object.size(list(t = t, lambda = lambda, em_result = em_resul

# Store the results
lambda_estimates[i] <- lambda_hat
biases[i] <- bias
mses[i] <- mse
accuracies[i] <- accuracy
aics[i] <- aic
bics[i] <- bic
num_iterations[i] <- iterations
elapsed_times[i] <- time[["elapsed"]]
```

```
space_complexities[i] <- space_complexity
}

# Calculate the average estimate and statistics
average_estimate <- mean(lambda_estimates)
average_bias <- mean(biases)
average_mse <- mean(mses)
average_accuracy <- mean(accuracies)
average_iterations <- mean(num_iterations)
average_elapsed_time <- mean(elapsed_times)
average_space_complexity <- mean(space_complexities)
average_aic <- mean(aics)
average_bic <- mean(bics)

# Print the results
cat("Average MLE estimate for lambda:", average_estimate, "\n")
cat("Average bias:", average_bias, "\n")
cat("Average mean square error:", average_mse, "\n")
cat("Average accuracy:", average_accuracy, "\n")
cat("Average number of iterations:", average_iterations, "\n")
cat("Average elapsed time:", average_elapsed_time, "seconds\n")
cat("Average AIC:", average_aic, "\n")
cat("Average BIC:", average_bic, "\n")
```

6.4 Appendix B

The code for model selection for both Gamma and Exponental distributions for left truncated and right-censored data using the EM-algorithm.

```
###############################
library(truncdist)
library(MASS)

set.seed()  # Set seed for reproducibility
```

```
n <-   # Sample size
#rate <-   # Rate parameter of exponential distribution
a <-   # Left truncation point
b <-4# 1  # Right censoring point
lambda_true <-  # True rate parameter
u <- rexp(n, rate = lambda_true) + a  # Generate left-truncated data
d <- pexp(u, rate = lambda_true)  # Generate right-censored data
data <- ifelse(runif(n) <= d, u, b)

#shape_true <-  # True shape parameter
#rate_true <-  # True rate parameter
#u <- rgamma(n, shape = shape_true, rate = rate_true) + a  # Generate left-truncated
#d <- pgamma(u, shape = shape_true, rate = rate_true)  # Generate right-censored data
#data <- ifelse(runif(n) <= d, u, b)  # Combine left-truncated and right-censored dat


# Function to perform EM algorithm for gamma distribution
em_gamma <- function(data, a, b) {
n <- length(data)
iterations <- 100  # Number of iterations for EM algorithm

# Initialize parameters
shape <-
scale <-

# EM algorithm
for (i in 1:iterations) {
# E-step
weights <- pgamma(data, shape, scale) / (pgamma(b, shape, scale, lower.tail = FALSE)

# M-step
shape <- sum(weights) / sum(weights * (data - a))
scale <- sum(weights) / sum(weights * (b - data))
}

return(list(shape = shape, scale = scale))
```

```
}

# Function to perform EM algorithm for exponential distribution
em_exponential <- function(data, a, b) {
n <- length(data)
iterations <-   # Number of iterations for EM algorithm

# Initialize parameter
rate <- 1

# EM algorithm
for (i in 1:iterations) {
# E-step
weights <- dexp(data, rate) / (pexp(b, rate) - pexp(a, rate))

# M-step
rate <- sum(weights) / sum(weights * (b - a))
}

return(list(rate = rate))
}

# Perform EM algorithm and obtain parameter estimates for each distribution
gamma_est <- em_gamma(data, a, b)
exp_est <- em_exponential(data, a, b)

# Calculate log-likelihood for each model
loglik_gamma <- sum(dgamma(data, shape = gamma_est$shape, scale = gamma_est$scale, lo
loglik_exp <- sum(dexp(data, rate = exp_est$rate, log = TRUE))

# Calculate the number of parameters for each model
num_params_gamma <-
num_params_exp <-

# Calculate AIC for each model
aic_gamma <- -2 * loglik_gamma + 2 * num_params_gamma
```

```
aic_exp <- -2 * loglik_exp + 2 * num_params_exp

# Calculate BIC for each model
n <- length(data)
bic_gamma <- -2 * loglik_gamma + num_params_gamma * log(n)
bic_exp <- -2 * loglik_exp + num_params_exp * log(n)

# Model selection based on AIC
best_model_aic <- which.min(c(aic_gamma, aic_exp))

# Model selection based on BIC
best_model_bic <- which.min(c(bic_gamma, bic_exp))

# Print the results
cat("Model Selection Performance:\n")
cat("Gamma Model AIC:", aic_gamma, "\n")
cat("Exponential Model AIC:", aic_exp, "\n")
cat("Gamma Model BIC:", bic_gamma, "\n")
cat("Exponential Model BIC:", bic_exp, "\n")

cat("Best Model based on AIC:")
if (best_model_aic == 1) {
cat("Gamma\n")
} else if (best_model_aic == 2) {
cat("Exponential\n")
}

cat("Best Model based on BIC:")
if (best_model_bic == 1) {
cat("Gamma\n")
} else if (best_model_bic == 2) {
cat("Exponential\n")
```