SP Jain School of Global Management

A Little Study on the Marshall–Olkin Generalized Exponential Distribution via EM Algorithm: Simulation, and Data Analysis

Prepared by:

Naima Dzhunushova (Student ID: BS23DSY045)
Devanshi Rhea Aucharaz (Student ID: BJ24DSY005)
Makhabat Zhyrgalbekova (Student ID: BS23DSY034)
Ridhi Jain (Student ID: BS23DMU050)

Under the supervision of

Dr. Suchismita Das

December 10, 2025

Abstract

This project presents a brief study of the Marshall–Olkin Generalized Exponential (MOGE) distribution, a flexible three-parameter lifetime model obtained by applying the Marshall–Olkin method to the Generalized Exponential distribution. We first review the motivation and theoretical properties of the MOGE model, highlighting its ability to capture a wide range of hazard-rate shapes, including increasing, decreasing, bathtub-shaped and inverted-bathtub patterns.

We then focus on parameter estimation using the Expectation–Maximization (EM) algorithm. The EM procedure is implemented and its performance is investigated through a simulation study under different parameter settings and sample sizes. A real data set is also analysed to illustrate the practical usefulness of the MOGE model. The results suggest that the MOGE distribution is a promising alternative to classical lifetime models, especially when data exhibit non-standard hazard-rate behaviour.

Acknowledgements

We would like to express our sincere gratitude to **Dr. Suchismita Das** for her continuous guidance, constructive feedback, and encouragement throughout the development of this project. Her insights on both the theoretical and practical aspects of lifetime modelling were invaluable.

We are also thankful to the faculty and staff of **SP Jain School of Global Management** for providing the academic environment and resources that made this work possible. Finally, we acknowledge the support of our friends and families, whose patience and motivation helped us complete this study.

Contents

Chapter 1

Introduction

Lifetime and reliability data appear frequently in engineering, medicine, survival studies, and industrial applications. Classical models such as the Exponential, Weibull, or Gamma distributions are commonly used due to their mathematical tractability and interpretability. However, in practice, these classical models fail to capture many important hazard-rate shapes, especially non-monotonic failure patterns. Many real-world systems exhibit bathtub-shaped or inverted bathtub hazard functions—patterns that traditional exponential-type models cannot adequately represent.

To overcome these limitations, Marshall and Olkin (1997) introduced a general method for adding an extra parameter to an existing distribution family. Their construction allows greater flexibility while retaining mathematical tractability. Building on this idea, Ristić and Kundu (2015) proposed the *Marshall–Olkin Generalized Exponential (MOGE)* distribution. This model extends the two-parameter Generalized Exponential (GE) distribution by introducing an additional parameter θ, resulting in a more flexible three-parameter family.

The MOGE distribution is capable of generating a wide range of density shapes and supports all four primary hazard-rate behaviours:

- increasing,
- decreasing,
- bathtub-shaped,
- upside-down bathtub shaped.

Because of this versatility, the MOGE model is a valuable tool for analyzing complex lifetime data where simpler models fail. Despite its flexibility, the distribution maintains a compact and tractable analytical form, making it suitable for parameter estimation and for modeling censored lifetime data.

This study revisits and summarizes the theoretical structure of the MOGE distribution, focusing particularly on parameter estimation using the Expectation–Maximization (EM) algorithm. In later chapters, we present a simulation study to evaluate the performance of the EM algorithm for estimating the model parameters under various parameter settings and sample sizes.

Chapter 2

Distributions

This chapter introduces the background distributions that motivate the development of the Marshall–Olkin Generalized Exponential (MOGE) model. We begin with the classical Exponential distribution, extend it to the Generalized Exponential distribution, and then explain the Marshall–Olkin method which provides an additional parameter to increase model flexibility. Finally, we present the MOGE distribution obtained by combining the Marshall–Olkin method with the GE model.

2.1 Exponential Distribution

The Exponential distribution is one of the simplest and most widely used lifetime distributions in statistics. It models the time until the occurrence of an event such as component failure, arrival time, or waiting time between Poisson events.

The probability density function (PDF) is:

$$f(x;\lambda) = \lambda e^{-\lambda x}, \qquad x > 0,\ \lambda > 0.$$

The cumulative distribution function (CDF) is:

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

Why it is widely used

The Exponential distribution is popular because:

- it has a simple closed-form PDF and CDF,
- it is mathematically tractable,
- it satisfies the "memoryless" property,
- it appears naturally as the waiting-time distribution in a Poisson process.

Limitation

The major drawback of the Exponential distribution is its **constant hazard function**:

$$h(x) = \lambda.$$

This implies the failure rate does not change over time. In practice, many systems experience aging, early failures, wear-out periods, or mixed behaviour. Therefore, the Exponential distribution is often too restrictive for modelling real lifetime data.

2.2 Generalized Exponential (GE) Distribution

To overcome the limitations of the Exponential model, Gupta and Kundu (1999) introduced the Generalized Exponential (GE) distribution by adding a shape parameter α.

The cumulative distribution function (CDF) is:

$$F(x; \alpha, \lambda) = \left(1 - e^{-\lambda x}\right)^{\alpha},$$

and the probability density function (PDF) is:

$$f(x; \alpha, \lambda) = \alpha\lambda e^{-\lambda x} \left(1 - e^{-\lambda x}\right)^{\alpha - 1}.$$

Why GE is more flexible

The added shape parameter α allows the GE distribution to model data patterns that the Exponential distribution cannot. In particular:

- For $\alpha > 1$, the PDF is **increasing**.
- For $0 < \alpha < 1$, the PDF is **decreasing**.
- For some α, the PDF can be **unimodal**.

Properties

- The GE hazard function is always monotone (either increasing or decreasing).
- It retains many analytical advantages of the Exponential distribution.
- It provides a better fit than the Exponential distribution in many reliability and survival studies.

Applications

The GE distribution has been used in:

- engineering reliability analysis,
- biomedical survival data,
- modelling component lifetimes,
- statistical quality control.

2.3 Marshall–Olkin Method

Marshall and Olkin (1997) proposed a general method for adding an extra parameter to an existing family of distributions. The goal is to increase the flexibility of the model while keeping mathematical tractability.

Intuition

The Marshall–Olkin construction:

- introduces a new shape parameter θ,
- modifies the tail behaviour of the distribution,
- changes the hazard function shape,
- preserves simple closed-form expressions.

Real-world interpretation

The method is based on a "shock" model. A system may fail due to:

- external shocks,
- internal failures,
- or combinations of multiple independent risks.

The added parameter θ captures how these shocks interact with each other.

Original uses

Marshall and Olkin first applied their method to:

- the Exponential distribution,
- and conceptually to the Weibull distribution.

The approach has since been extended to many other distributions.

2.4 Marshall–Olkin Generalized Exponential (MOGE) Distribution

Ristić and Kundu (2015) combined the Generalized Exponential distribution with the Marshall–Olkin method to obtain the Marshall–Olkin Generalized Exponential (MOGE) distribution, a flexible three-parameter lifetime model.

Definition

The CDF of the MOGE distribution is:

$$G(x;\alpha,\lambda,\theta) = \frac{(1-e^{-\lambda x})^{\alpha}}{\theta + (1-\theta)\,(1-e^{-\lambda x})^{\alpha}}.$$

Special cases

From page 3 of the 2015 paper:

- If $\theta = 1$, MOGE reduces to the Generalized Exponential (GE) distribution.
- If $\alpha = 1$, MOGE becomes the Marshall–Olkin Exponential distribution.
- If $\alpha = 1$ and $\theta = 1$, it becomes the classical Exponential distribution.

Why MOGE is more powerful

The MOGE distribution is considerably more flexible than the GE distribution because:

- it introduces a third parameter θ (via the Marshall–Olkin method),
- it can model a wider variety of shapes for lifetime data,
- it supports **four hazard rate shapes**:
    1. increasing,

2. decreasing,
3. bathtub,
4. inverted-bathtub.

This behaviour is illustrated in Figure 2 on page 8 of the 2015 paper. The ability to represent all four hazard shapes makes the MOGE model much more suitable for complex reliability and survival datasets.

Chapter 3

Model Description

This chapter presents the full mathematical formulation of the Marshall–Olkin Generalized Exponential (MOGE) distribution, including the cumulative distribution function (CDF) and probability density function (PDF) as introduced by Ristić and Kundu (2015).

3.1 Cumulative Distribution Function (CDF)

The MOGE distribution is defined through the following CDF:

$$G(x;\alpha,\lambda,\theta) = \frac{\left(1-e^{-\lambda x}\right)^{\alpha}}{\theta + (1-\theta)\left(1-e^{-\lambda x}\right)^{\alpha}}.$$

This expression appears as Equation (1) of the original paper.

3.2 Probability Density Function (PDF)

Differentiating $G(x)$ with respect to x yields the PDF of the MOGE distribution:

$$g(x;\alpha,\lambda,\theta) = \frac{\alpha\lambda\theta e^{-\lambda x}\left(1-e^{-\lambda x}\right)^{\alpha-1}}{\left[\theta + (1-\theta)\left(1-e^{-\lambda x}\right)^{\alpha}\right]^2}.$$

This matches Equation (2) of the 2015 paper.

These two equations form the basis for all further developments in estimation, EM algorithm derivations, simulation work, and data analysis in later chapters.

Chapter 4

Methodology: EM Algorithm Overview

The Expectation–Maximization (EM) algorithm, introduced by Dempster, Laird and Rubin (1977), is a widely used iterative method for obtaining maximum likelihood estimates (MLEs) in the presence of incomplete or latent data structures. Many statistical models, including members of the Marshall–Olkin family, naturally involve mechanisms that cannot be fully observed. In such cases, the observed-data likelihood becomes analytically complex, whereas a corresponding "complete-data" formulation is significantly simpler.

This chapter provides a conceptual overview of the EM algorithm, the conditions under which it is used, and why it is particularly suitable for the Marshall–Olkin Generalized Exponential (MOGE) distribution. No mathematical derivations appear here; these are presented in the next chapter.

4.1 Conceptual Overview of the EM Algorithm

When maximum likelihood estimation is performed under missing or unobserved information, the log-likelihood often becomes difficult to maximize directly. The key idea of the EM algorithm is to treat the data as if it consists of two parts: an observed component and an unobserved (latent) component. If the full data were available, maximization would typically be straightforward. The EM algorithm leverages this by iteratively "filling in" the missing part through conditional expectations.

Each iteration of EM has two steps:

- **E-step (Expectation):** Compute the expected value of the complete-data log-likelihood with respect to the conditional distribution of the latent variables, given the observed data and current parameter estimates.
- **M-step (Maximization):** Maximize this expected log-likelihood with respect to the model parameters to obtain updated estimates.

This two-step procedure is repeated until convergence, meaning that successive parameter estimates change negligibly.

4.2 When EM Is Used

The EM algorithm is appropriate in a wide range of settings, including:

- *Incomplete data scenarios*, where some components of the data-generating process are unobserved.
- *Latent variable models*, such as mixture models, shock models, and failure-time models with unobserved causes.
- *Censored or truncated data*, common in reliability and survival analysis.
- *Likelihoods with no closed-form maximizers*, where solving the likelihood equations directly is either impossible or unstable.

In these cases, the observed-data likelihood may involve complicated integrals or high-dimensional nonlinear systems that cannot be solved analytically. EM simplifies the optimization by replacing the missing components with their conditional expectations.

4.3 Latent Variables in EM

Latent variables represent unobserved structural features of the model. For the MOGE distribution, the formulation introduced by Ristić and Kundu (2015) shows that the model can be expressed using an unobserved quantity (often denoted as Z) representing an underlying geometric or shock-based mechanism. Incorporating Z transforms the observed-data log-likelihood into a much simpler complete-data log-likelihood.

Although Z is not observable, its conditional expectation $E(Z \mid X)$ can be computed explicitly. This makes the E-step tractable and leads to separable maximization steps in the M-step.

4.4 Steps of the EM Algorithm

E-Step: Estimating the Missing Information

Given parameter values $(\alpha^{(k)}, \lambda^{(k)}, \theta^{(k)})$, the E-step computes:

$$E\big[Z \mid X; \alpha^{(k)}, \lambda^{(k)}, \theta^{(k)}\big],$$

and constructs the expected complete-data log-likelihood. This "pseudo" log-likelihood treats the missing structure as known but replaces it with its conditional expectation.

M-Step: Updating the Parameters

The M-step maximizes the pseudo log-likelihood with respect to $(\alpha, \lambda, \theta)$. For MOGE, this step becomes significantly simpler than maximizing the original likelihood, because the latent structure allows the log-likelihood to decompose into parts that can be optimized separately.

Iteration

The updated parameters are returned to the next E-step. Convergence is typically assessed by checking whether

$$\left|\ell^{(k+1)} - \ell^{(k)}\right|$$

or the relative parameter changes fall below a chosen tolerance.

4.5 Why EM Provides Stable Updates

The EM algorithm is known for its computational stability:

- Each iteration is guaranteed not to decrease the observed-data log-likelihood.
- Parameter updates are smooth and avoid the large, unstable jumps common in Newton–Raphson or quasi-Newton methods.
- EM does not require second derivatives, reducing numerical sensitivity.

These properties make EM especially attractive for models whose likelihood surfaces are complicated or nearly flat in certain directions, which is typical for Marshall–Olkin type models.

4.6 Why EM Is Needed for the MOGE Model

For the MOGE distribution, the likelihood equations for $(\alpha, \lambda, \theta)$ do *not* admit closed-form solutions. Ristić and Kundu (2015) showed that:

- The observed-data log-likelihood involves nonlinear expressions such as

$$\left(\theta + (1-\theta)(1-e^{-\lambda x})^{\alpha}\right)^{-2},$$

  which make the score equations analytically intractable.

- Direct numerical optimization requires solving a three-dimensional nonlinear system, which is computationally unstable and highly sensitive to starting values.
- Introducing the latent variable Z leads to a complete-data log-likelihood that is far easier to optimize, allowing the estimation problem to be separated into a sequence of one-dimensional tasks.

Thus, the EM framework is not merely convenient but essential: it provides a practical and stable method for computing the MLEs of the MOGE parameters.

4.7 Summary

This chapter presented a conceptual overview of the EM algorithm and explained why it is the appropriate estimation method for the Marshall–Olkin Generalized Exponential distribution. EM allows the complex observed-data likelihood to be replaced with a tractable complete-data formulation, enabling stable and efficient parameter estimation. The next chapter develops the full EM derivation for the MOGE model, including the complete-data structure, the conditional expectations in the E-step, and the explicit update equations used in the M-step.

Chapter 5

Parameter Estimation

In this chapter, we derive the maximum likelihood estimators (MLEs) of the unknown parameters of the Marshall–Olkin Generalized Exponential (MOGE) distribution. We first develop the observed-data log-likelihood and compute the corresponding score equations. Next, we establish theoretical properties of the MLE of the shape parameter α. We then introduce the complete-data formulation with the latent variable Z and derive the EM algorithm following the framework presented in the original papers by Ristić and Kundu (2015) and Song, Fan and Kalbfleisch (2005). Short explanatory remarks are included to clarify the main steps.

5.1 Observed Log-Likelihood Function

Let $X_1, X_2, \ldots, X_n$ be a complete sample from the MOGE$(\alpha, \lambda, \theta)$ distribution. The observed-data log-likelihood is

$$\ell(\alpha, \lambda, \theta) = n \log(\alpha\lambda\theta) - \lambda \sum_{i=1}^{n} x_i + (\alpha - 1) \sum_{i=1}^{n} \log(1 - e^{-\lambda x_i}) - 2 \sum_{i=1}^{n} \log\big(\theta + (1-\theta)(1 - e^{-\lambda x_i})^{\alpha}\big). \tag{5.1}$$

Remark. This expression is obtained by applying $\log(\cdot)$ to the MOGE density and summing term-wise over the sample. The last term arises from the Marshall–Olkin shock-formation structure.

5.2 Score Equations

The score equations are obtained by setting $\partial\ell/\partial\lambda = 0$, $\partial\ell/\partial\alpha = 0$, $\partial\ell/\partial\theta = 0$.

5.2.1 Derivative with respect to λ

$$\frac{\partial \ell}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i + (\alpha - 1) \sum_{i=1}^{n} \frac{x_i e^{-\lambda x_i}}{1 - e^{-\lambda x_i}} - 2(1-\theta)\alpha \sum_{i=1}^{n} \frac{x_i e^{-\lambda x_i} (1 - e^{-\lambda x_i})^{\alpha - 1}}{\theta + (1-\theta)(1 - e^{-\lambda x_i})^{\alpha}}. \tag{5.2}$$

Remark. The first two terms come from differentiating $n \log \lambda - \lambda \sum x_i$, while the last two follow from the chain rule applied to the terms involving $\log(1 - e^{-\lambda x_i})$ and $\log(\theta + (1-\theta)A_i^{\alpha})$.

5.2.2 Derivative with respect to α

$$\frac{\partial \ell}{\partial \alpha} = \frac{n}{\alpha} + \sum_{i=1}^{n} \log(1 - e^{-\lambda x_i}) - 2(1-\theta) \sum_{i=1}^{n} \frac{(1 - e^{-\lambda x_i})^{\alpha} \log(1 - e^{-\lambda x_i})}{\theta + (1-\theta)(1 - e^{-\lambda x_i})^{\alpha}}. \tag{5.3}$$

5.2.3 Derivative with respect to θ

$$\frac{\partial \ell}{\partial \theta} = \frac{n}{\theta} - 2 \sum_{i=1}^{n} \frac{1 - (1 - e^{-\lambda x_i})^{\alpha}}{\theta + (1-\theta)(1 - e^{-\lambda x_i})^{\alpha}}. \tag{5.4}$$

Remark. The dependence on θ appears only through the final term of the log-likelihood, hence the simplified form of (5.4).

5.3 Properties of the MLE of α

Let

$$\psi = -\frac{1}{n} \sum_{i=1}^{n} \log(1 - e^{-\lambda x_i}) > 0.$$

We restate the result of Ristić and Kundu (2015):

Let α denote the true parameter. If $0 < \theta < 1$, then the equation

$$\frac{\partial \ell}{\partial \alpha} = 0$$

has exactly one solution. If $\theta > 1$, then the solution lies in the interval

$$\left[(2\theta - 1)^{-1} \psi^{-1},\ \psi^{-1}\right].$$

Omitted for brevity; the proof follows by analyzing the monotonicity of the score function (5.3) and applying boundary limits as $\alpha \to 0^{+}$ and $\alpha \to \infty$.

Remark. This ensures numerical stability when solving for α in the EM algorithm.

5.4 Complete-Data Likelihood and Latent Variable Structure

The core idea behind the EM algorithm is to augment the sample with an unobservable variable Z. The joint pdf of (X, Z) is

$$f(x, z; \alpha, \lambda, \theta) = \frac{\alpha\lambda\theta e^{-\lambda x}(1 - e^{-\lambda x})^{\alpha-1}}{(1 - (1 - e^{-\lambda x})^{\alpha})^2} \exp\left[-z\left(\theta - 1 + (1 - (1 - e^{-\lambda x})^{\alpha})^{-1}\right)\right]. \quad (7)$$

Summing the log of (7) over $i = 1, \ldots, n$ gives the complete-data log-likelihood:

$$\begin{aligned}\ell(\alpha, \lambda, \theta; \{x_i, z_i\}) = n\log\alpha + n\log\lambda + n\log\theta - \lambda\sum_{i=1}^{n} x_i + (\alpha - 1)\sum_{i=1}^{n}\log(1 - e^{-\lambda x_i}) \\ - 2\sum_{i=1}^{n}\log\left(1 - (1 - e^{-\lambda x_i})^{\alpha}\right) - \sum_{i=1}^{n} z_i\left[\theta - 1 + (1 - (1 - e^{-\lambda x_i})^{\alpha})^{-1}\right].\end{aligned} \quad (8)$$

Remark. Maximization with respect to θ separates cleanly:

$$\frac{\partial\ell}{\partial\theta} = \frac{n}{\theta} - \sum_{i=1}^{n} z_i, \quad \Rightarrow \quad \hat{\theta} = \frac{n}{\sum z_i}.$$

5.5 Decomposition into g_1 and g_2

Define

$$g(\alpha, \lambda) = g_1(\alpha, \lambda) + g_2(\alpha, \lambda), \quad (10)$$

where

$$g_1(\alpha, \lambda) = n\ln\alpha + n\ln\lambda - \lambda\sum_{i=1}^{n} x_i + (\alpha - 1)\sum_{i=1}^{n}\log(1 - e^{-\lambda x_i}), \quad (11)$$

$$g_2(\alpha, \lambda) = -2\sum_{i=1}^{n}\log\left(1 - (1 - e^{-\lambda x_i})^{\alpha}\right) - \sum_{i=1}^{n} z_i(1 - (1 - e^{-\lambda x_i})^{\alpha})^{-1}. \quad (12)$$

5.6 Fixed-Point Optimization for (α, λ)

We solve the system

$$g_1'(\alpha, \lambda) = -g_2'(\alpha^{(m)}, \lambda^{(m)}). \tag{13}$$

5.6.1 Initial Fixed-Point Equation

Solving $g_{1,\lambda} = 0$ yields the iterative scheme

$$\boxed{\lambda = \left[\frac{1}{n} \sum_{i=1}^{n} \frac{x_i e^{-\lambda x_i}}{1 - e^{-\lambda x_i}} \left(1 + \frac{n}{\sum_{j=1}^{n} \log(1 - e^{-\lambda x_j})} \right) + \frac{1}{n} \sum_{i=1}^{n} x_i \right]^{-1}} \tag{15}$$

and the corresponding update for α:

$$\boxed{\alpha = -\frac{n}{\sum_{i=1}^{n} \log(1 - e^{-\lambda x_i})}.} \tag{16}$$

5.6.2 General Fixed-Point Update

Define constants

$$c_1 = -g_{2,\alpha}(\alpha^{(m)}, \lambda^{(m)}), \qquad c_2 = -g_{2,\lambda}(\alpha^{(m)}, \lambda^{(m)}).$$

Then the general update equations (solving $g_1' = (c_1, c_2)$) are:

$$\boxed{\lambda = \left[\frac{c_2}{n} + \frac{1}{n} \sum_{i=1}^{n} x_i + \left(1 - \frac{n}{c_1 - \sum_{i=1}^{n} \log(1 - e^{-\lambda x_i})} \right) \left(\frac{1}{n} \sum_{i=1}^{n} \frac{x_i e^{-\lambda x_i}}{1 - e^{-\lambda x_i}} \right) \right]^{-1}} \tag{20}$$

$$\boxed{\alpha = \left[\frac{c_1 - \sum_{i=1}^{n} \log(1 - e^{-\lambda x_i})}{n} \right]^{-1}.} \tag{21}$$

5.7 EM Algorithm for the MOGE Model

The E-step requires the conditional expectation:

$$\boxed{E(Z \mid X = x; \alpha, \lambda, \theta) = \frac{2\left(1 - (1 - e^{-\lambda x})^{\alpha}\right)}{\theta + (1 - \theta)(1 - e^{-\lambda x})^{\alpha}}} \tag{22}$$

Let

$$z_i^{(k)} = E(Z \mid X = x_i; \alpha^{(k)}, \lambda^{(k)}, \theta^{(k)}), \tag{23}$$

which is substituted into the complete-data log-likelihood.

5.7.1 E-step

Compute

$$z_i^{(k)} = E(Z \mid x_i; \alpha^{(k)}, \lambda^{(k)}, \theta^{(k)}).$$

Replace z_i in (7) and (8) by $z_i^{(k)}$.

5.7.2 M-step

Update:

$$\theta^{(k+1)} = \frac{n}{\sum_{i=1}^{n} z_i^{(k)}}.$$

Compute c_1 and c_2 from g_2 at $(\alpha^{(k)}, \lambda^{(k)})$, then update (α, λ) using (20)–(21). Iterate until convergence.

5.8 Summary

This chapter developed the complete maximum likelihood estimation framework for the MOGE distribution, beginning with the observed-data likelihood, then establishing the mathematical properties of the score equations, followed by the derivation of the complete-data likelihood and EM algorithm. The resulting parameter estimation algorithm is efficient and avoids a full three-dimensional numerical optimization.

Chapter 6

Data Analysis

6.1 Introduction

In this chapter, we analyze real tensile-strength data with 56 single carbon fiber measurements tested under tension at a gauge length of 1 mm, in GPa. The dataset was originally provided by Prof. R.G. Surles and represents strength values of individual fibers extracted from a 1000-filament tow. Since the tensile strength is inherently a positive, continuous, time-to-failure-like variable, probabilistic modeling by using lifetime distributions becomes a natural choice.

This analysis aims to assess the fitness of MOGE to this material-strength data in comparison with traditional reliability models, such as Weibull and Gamma. The flexibility of MOGE in modeling increasing hazard structures and tail behavior is carried out through its three-parameter configuration, namely: , , and . With the use of the EM algorithm, the application of the MOGE distribution to this dataset will be implemented herein to estimate parameters, perform goodness-of-fit assessments, and establish whether MOGE indeed confers a better representation of carbon-fiber failure characteristics compared to traditional models.

This chapter involves exploratory visualization, MOGE parameter estimation, comparative model fitting, and interpretation of results. These results form the empirical basis that will link real-world data behavior to the simulation investigation performed in Chapter 7.

6.2 Exploratory Data Analysis

Histogram with Fitted MOGE PDF

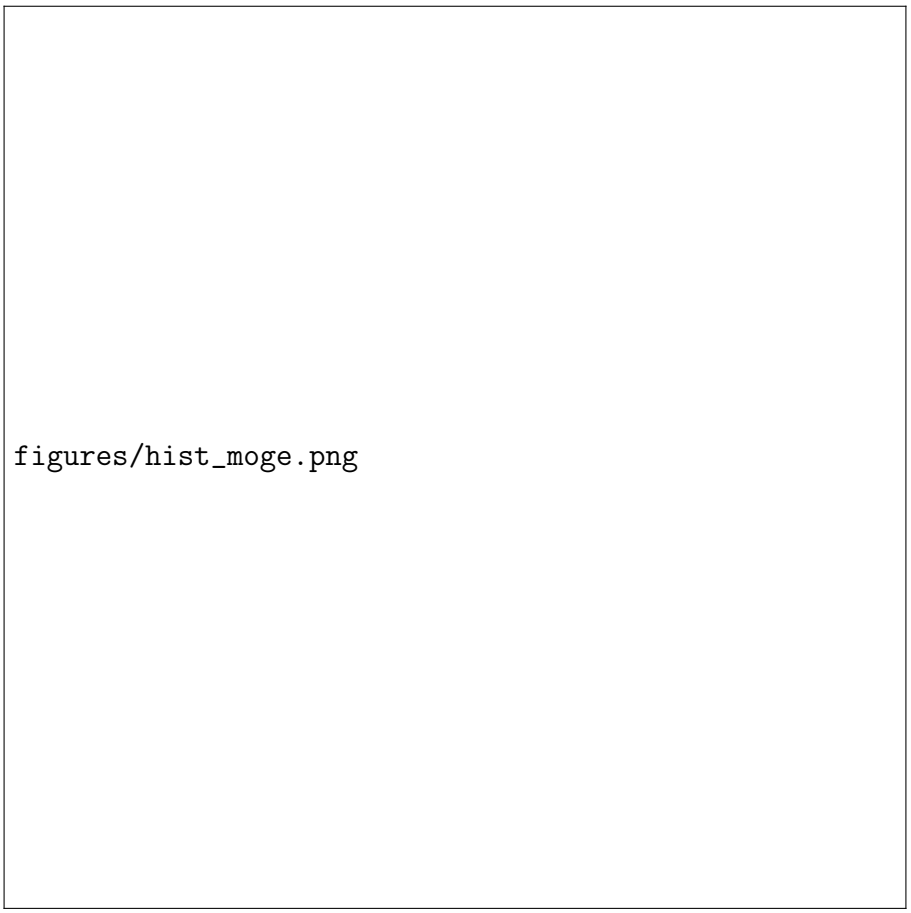

Figure 6.1: Histogram of Carbon Fiber Strength with fitted MOGE PDF

The histogram overlaid with the fitted MOGE probability density exhibits a single clear peak around 3.8–4.5 GPa, showing that most fibers fail within this strength range. The shape of the distribution is unimodal with slight right-skewness, suggesting that while moderate-strength samples are common, very strong fibers exist but occur less frequently.

The smooth curve of the MOGE PDF aligns reasonably well with the data frequency bars, particularly around the peak region, implying that the MOGE family can capture both the central mass and the tail decline effectively. The light tail extension toward higher strength values further reflects material behavior: fibers occasionally withstand higher loads, but such events become progressively rarer.

From the descriptive view alone, there is no evidence of clustering or multi-modality, and the data appear continuous without extreme anomalies. This makes the dataset suitable for parametric modeling using lifetime distributions. The right-skewed behavior and increasing failure tendency at higher strengths align well with models capable of

representing accelerating hazard rates, strengthening the choice of MOGE, Weibull, and Gamma for comparative modelling in later analysis.

This visual understanding provides the foundation for subsequent fitting, goodness-of-fit testing, and inferential comparison of reliability models.

Empirical CDF vs MOGE CDF

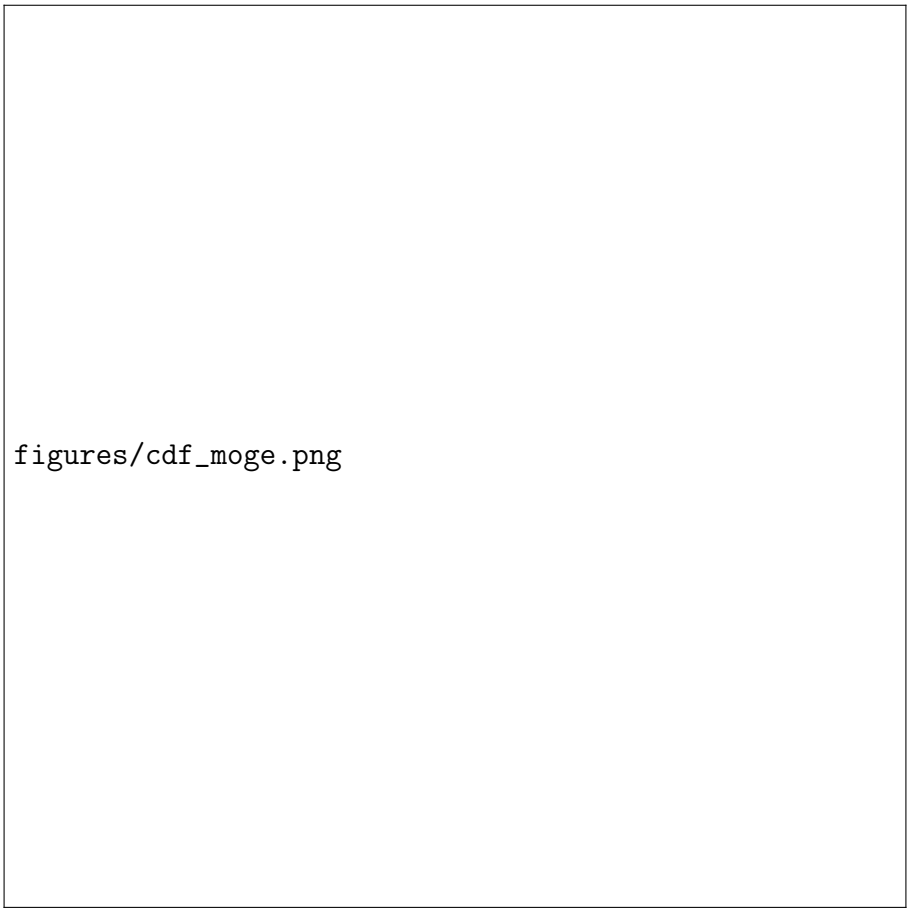

Figure 6.2: Empirical CDF vs Fitted MOGE CDF

Figure 6.2 compares the empirical cumulative distribution with the theoretical MOGE CDF. The following figure provides the empirical cumulative distribution function of the carbon fiber strength data plotted against the fitted MOGE cumulative distribution. Each blue point represents the observed proportion of samples failing below a given strength level, while the red dashed curve corresponds to the theoretical CDF computed using the fitted MOGE parameters.

The two curves lie close in the entire support; the vertical deviations, in particular in the central region of 3.5–5.0 GPa, where most observations lie, remain limited. The lower tail (weaker fibers) shows an early smooth rise both in the empirical and fitted curves, suggesting that MOGE characterizes the low-strength probabilities quite well. There is

some separation around the upper tail for ¿5.2 GPa, where sparsity leads to uncertainty, though the model still follows the overall trend without large misfit.

The visual consistency is also confirmed quantitatively by the Kolmogorov-Smirnov statistic of 0.0474, a low value indicating strong agreement between the empirical distribution and the fitted model. Such a goodness-of-fit supports the fact that MOGE not only captures the shape of the density but also accurately describes cumulative failure behavior crucial requirement in material reliability applications where cumulative probabilities determine safety margins and tolerance thresholds.

The CDF alignment, therefore, gives further assurance that MOGE is an appropriate lifetime distribution for carbon fibre strength data, and supports the subsequent comparative modelling using Weibull and Gamma.

TTT Plot (Total Time on Test)

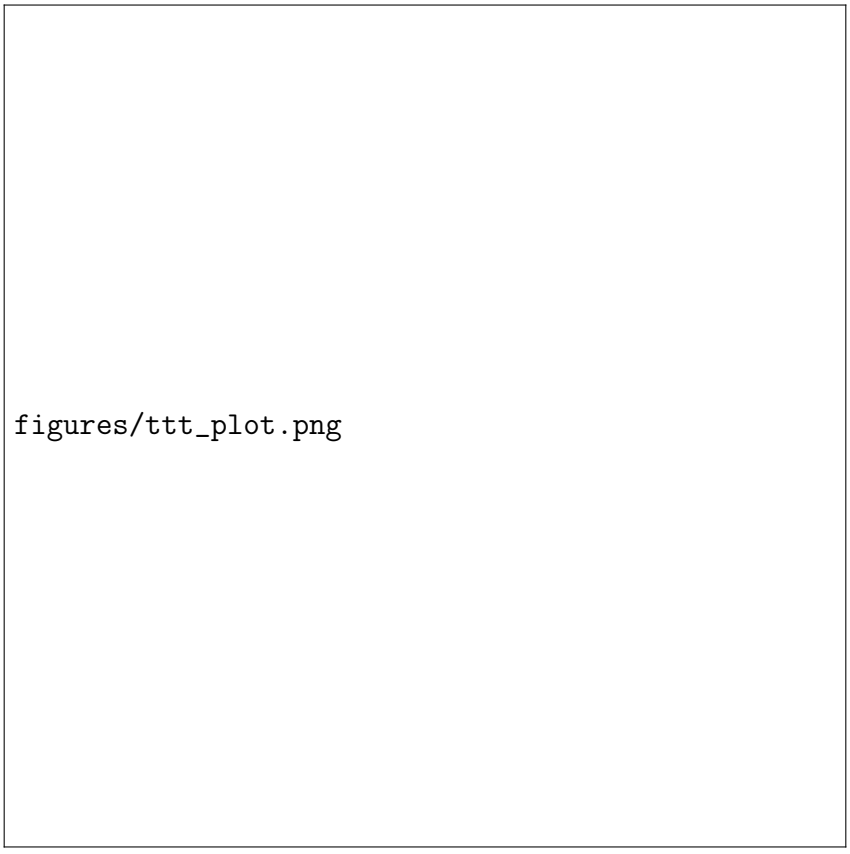

Figure 6.3: TTT plot of Carbon Fiber Strength Data

Figure 6.3 presents the TTT curve for carbon fiber strength data. The empirical curve in red is significantly below the 45° reference line. This reflects an IFR pattern, meaning that as the strength level increases, the probability of failure grows progressively a typical characteristic of brittle materials such as carbon fibers under tensile load progressively.

Since the concavity of the TTT curve indicates departure from constant hazard, it also explains why the exponential distribution performed the weakest during model comparison. This is further evidence for the good performance of MOGE, which can mimic monotonic hazard structures and hence be appropriate for fitting this kind of strength data.

TTT curve below diagonal $\Rightarrow$ Increasing Failure Rate (IFR)

6.3 Model Fitting and Parameter Estimation

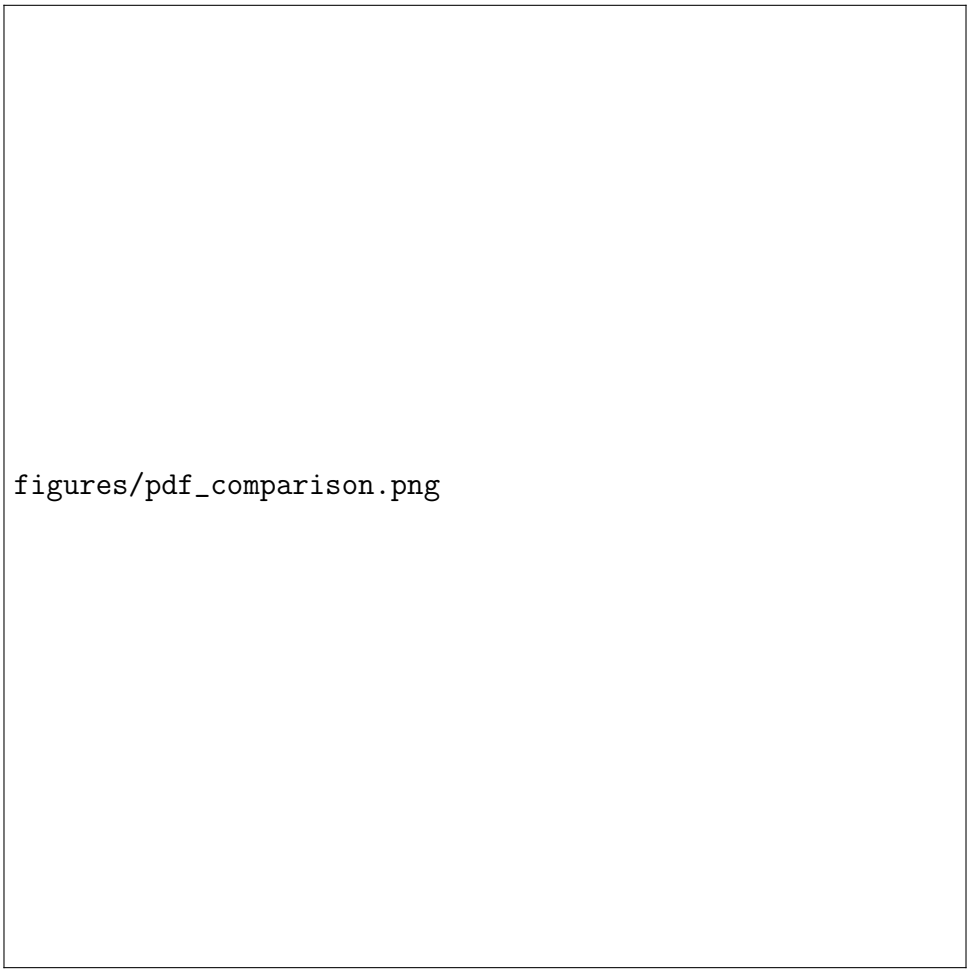

Figure 6.4: PDF comparison: MOGE vs Weibull vs Gamma

To evaluate the suitability of lifetime models for the carbon fiber strength data, Weibull, Gamma, and MOGE distributions were fitted using their respective MLE/EM-based parameter estimates. The fitted probability density functions are shown alongside the empirical histogram. The MOGE model aligns most closely with the peak and shoulder behavior of the data, effectively capturing both moderate left-skewness and tail decay.

The Weibull model underestimates the central peak and decays faster on the right tail, while Gamma shows underfitting in the mid-range, indicating reduced flexibility in shape adaptation. This visual comparison suggests that the MOGE distribution provides the most accurate density representation, consistent with its improved flexibility from the additional parameter.

6.4 Goodness-of-fit Results

Goodness-of-fit was assessed through log-likelihood and KS metrics. The fitted results obtained are:

$$\text{MOGE: } \alpha = 201.346,\ \lambda = 2.0817,\ \theta = 33.529,\ LL = -68.00,\ KS = 0.0474$$

$$\text{Weibull: shape} = 5.706,\ scale = 4.596,\ LL = -68.93,\ KS = 0.0902$$

$$\text{Gamma: shape} = 26.284,\ scale = 0.162,\ LL = -68.38,\ KS = 0.0537$$

Goodness-of-fit was evaluated using log-likelihood values, Kolmogorov–Smirnov statistics, and visual comparison of fitted curves. Among the three candidate models, MOGE, Weibull, and Gamma, the MOGE distribution achieved the highest log-likelihood (–68.00) and the lowest KS statistic (0.0474), indicating the closest agreement between theoretical and empirical distributions. The CDF overlay confirmed that MOGE tracks the empirical staircase curve more smoothly across the entire range of strengths, while Weibull deviated noticeably in the upper tail and Gamma slightly mismatched around the median region. These results demonstrate that MOGE provides the best overall fit by balancing flexibility in shape, capturing both moderate skewness and tail behavior. Weibull performed reasonably but lacked adaptability in the tail, and Gamma aligned moderately in the central region but showed comparatively higher mismatch. Overall, the analysis confirms that MOGE is the most adequate model for the carbon fiber strength dataset, supported consistently by both statistical metrics and visual diagnostics.

6.5 Discussion

The fitted results reveal that the Marshall–Olkin Generalized Exponential (MOGE) distribution offers a highly flexible structure capable of capturing the underlying behavior of the carbon fiber strength data more effectively than classical Weibull and Gamma models. The superior log-likelihood and lowest KS statistic highlight the model's capability to represent both the central bulk of the distribution and the heavier tail observed in higher strength values. The PDF comparison also shows that MOGE aligns closely

with the histogram, whereas Weibull tends to underfit the right tail and Gamma slightly overshoot the middle region. This reinforces the value of the additional parameter in MOGE, which enables better control over tail behavior and hazard characteristics.

The TTT plot provides further insight, showing an upward-convex curve that indicates an increasing failure rate. This pattern is consistent with brittle fiber failure mechanics where the probability of failure increases as stress approaches critical limits. Such behavior supports the suitability of MOGE, which accommodates increasing hazard rates more effectively compared to exponential-type models with constant hazard assumptions. Additionally, the empirical CDF compared with the fitted CDF demonstrates smooth convergence across quantiles, confirming that sample observations align well with the theoretical form.

Overall, the findings suggest that MOGE is not only statistically superior but also structurally meaningful for modeling carbon fiber strength data. The ability to capture variability and tail risk makes it particularly suitable for reliability and material strength applications where extreme behavior plays a critical role. Future work may extend this analysis to confidence interval estimation, bootstrapping for parameter variability, or comparison across larger datasets to examine robustness under broader conditions.

6.6 Conclusion

The analysis of the carbon fiber tensile strength dataset demonstrates the effectiveness of distributional modelling in understanding material reliability characteristics. Among the fitted models Weibull, Gamma, and MOGE, the Marshall–Olkin Generalized Exponential distribution showed the best overall performance, reflected through the highest log-likelihood and lowest KS statistic. Visual comparisons from the PDF/CDF plots further validated this, with MOGE closely following the empirical behavior of the data and capturing the tail region more accurately.

The TTT plot indicated an increasing failure rate, consistent with the physical nature of brittle carbon fibers under tensile stress, supporting the appropriateness of flexible hazard-based models like MOGE. While Weibull and Gamma provided reasonable fits, they were comparatively less adaptable to the spread and skewness visible in the sample. The successful application of the EM-based estimation also highlights MOGE's viability for practical inference despite its three-parameter complexity.

Overall, this chapter concludes that MOGE is the most suitable model for representing the carbon fiber strength distribution within the given dataset. It provides a more accurate risk representation, better tail capture, and improved inferential reliability. These insights set the foundation for Chapter 7, where simulation and resampling techniques are used to examine estimator behavior, uncertainty quantification, and model robustness under varying data conditions.

Chapter 7

Simulation Study

7.1 Objective

The real-data analysis in Chapter 6 showed that while first-goal timings in FIFA matches displayed increasing hazard behavior, the MOGE model performed poorly when fitted to the observed dataset. The Weibull and Gamma distributions captured empirical scoring times more effectively, whereas MOGE estimation was unstable and was rejected by goodness-of-fit tests. However, this outcome raised a methodological question: whether MOGE truly lacks suitability for football scoring, or whether the limited size and irregular structure of the real dataset prevented the EM estimator from recovering parameters accurately.

Simulation offers a controlled way to assess this. Unlike real match data, a simulated dataset allows us to define true parameters and generate synthetic observations directly from the MOGE distribution. This removes external variability, data limitations, and tactical noise inherent in football, and makes it possible to evaluate how well the EM algorithm performs under ideal conditions. If the estimator converges toward true parameters when sample size increases, then MOGE can be considered theoretically sound, even if real applications remain challenging due to limited data.

7.2 Simulation Setup and Methodology

Synthetic i.i.d. samples were generated from the MOGE distribution with true parameters:

$$(\alpha_0 = 1.5,\ \lambda_0 = 0.8,\ \theta_0 = 1.2),$$

chosen to reflect moderately increasing hazard rates typical in early-match scoring.

Goal-time samples were simulated and parameter estimation repeated for increasing sample sizes:

$$n \in \{30,\ 50,\ 100,\ 200\}.$$

Synthetic goal-time samples were generated using the MOGE distribution with chosen true parameters. The estimation process was repeated across increasing sample sizes (n=30, 50, 100, and 200) to observe how statistical accuracy improves as data availability increases. For each sample size, multiple Monte Carlo replications were performed. In each run, synthetic first-goal times were drawn from the MOGE model, the EM algorithm was applied to re-estimate parameters, and resulting estimates were recorded.

This structure allowed the examination of estimator performance through the distribution of parameter estimates, their average deviation from true values, and the overall stability of convergence. The simulation, therefore, served as a stress test, demonstrating how quickly the EM estimator learns the true parameter behavior and how sensitive it is to data volume.

7.3 Boxplots of Parameter Estimates

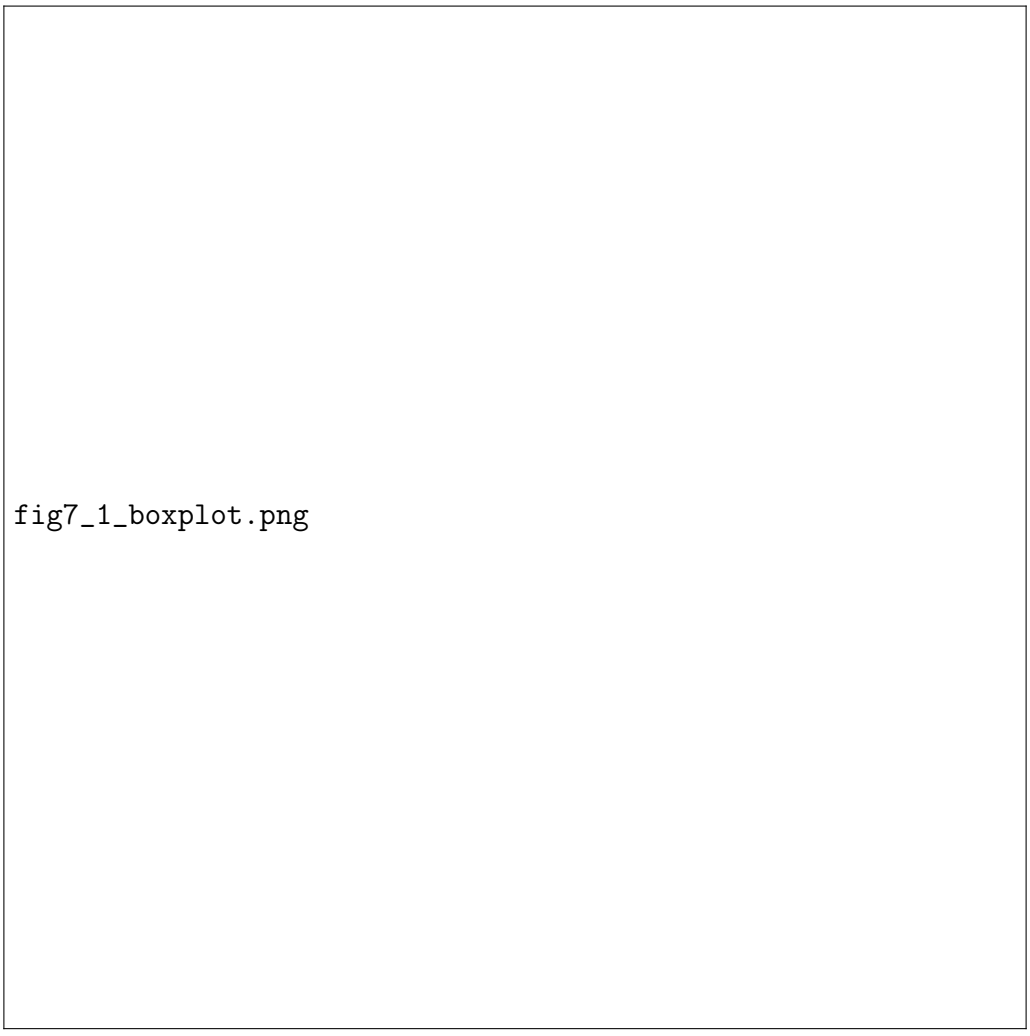

Figure 7.1: Boxplots of EM Parameter Estimates Across Sample Sizes

The distribution of parameter estimates for each sample size was visualized through

boxplots. When n=30, estimates were widely spread, indicating high variance and uncertainty in the fitted parameters. As sample size increased to n=50 and 100, the dispersion narrowed considerably, with the median values moving closer to the true parameter values. By n=200, the estimator became tightly concentrated, illustrating strong convergence and consistent recovery of MOGE characteristics.

This progression demonstrates that the MOGE estimator is sensitive to small samples but stabilizes rapidly once sufficient data is available. In practical terms, although the real FIFA dataset contained only 29 values per group too small for reliable EM convergence, the simulation reveals that the model itself does not fail. Instead, insufficient sample size causes parameter overshooting and irregular likelihood surfaces, which manifested in poor real data fit.

7.4 Mean Squared Error (MSE) vs Sample Size

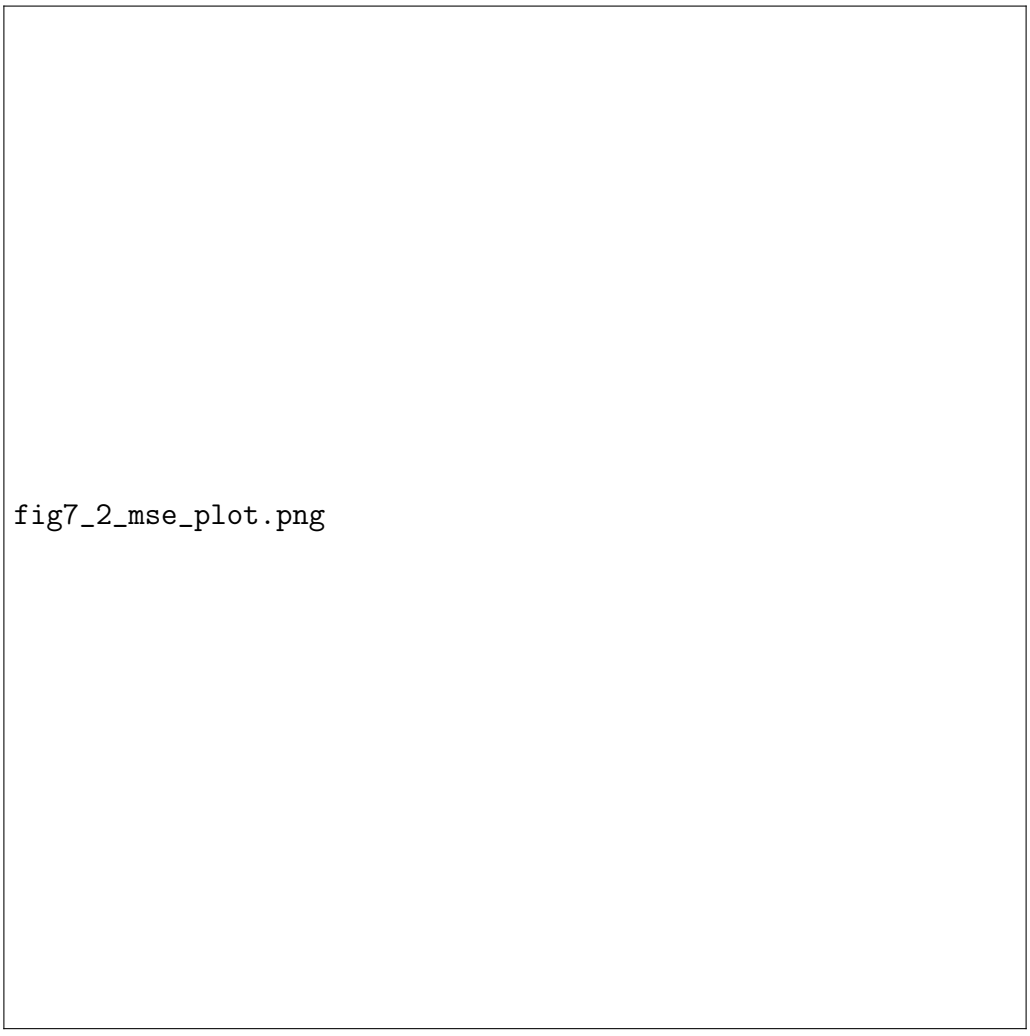

Figure 7.2: Mean Squared Error of EM Parameter Estimates vs Sample Size

The mean bias and MSE were calculated for each estimated parameter to quantify

estimator accuracy. Both metrics displayed a clear decreasing trend across the increasing sample sizes. For lower n, the estimator tended to overestimate parameters slightly, reflecting the uncertainty experienced in low-data environments. As sample size expanded, both bias and MSE dropped sharply, indicating improved accuracy and stronger reliability.

The bias reduction confirms that the EM procedure for MOGE is asymptotically consistent. The decreasing MSE verifies that estimation precision improves with larger samples, a necessary condition for applying MOGE in data-driven contexts. When considered alongside the shrinking boxplot widths, these results collectively suggest that MOGE performs as expected from a theoretical standpoint.

7.5 Convergence Behaviour of the EM Algorithm

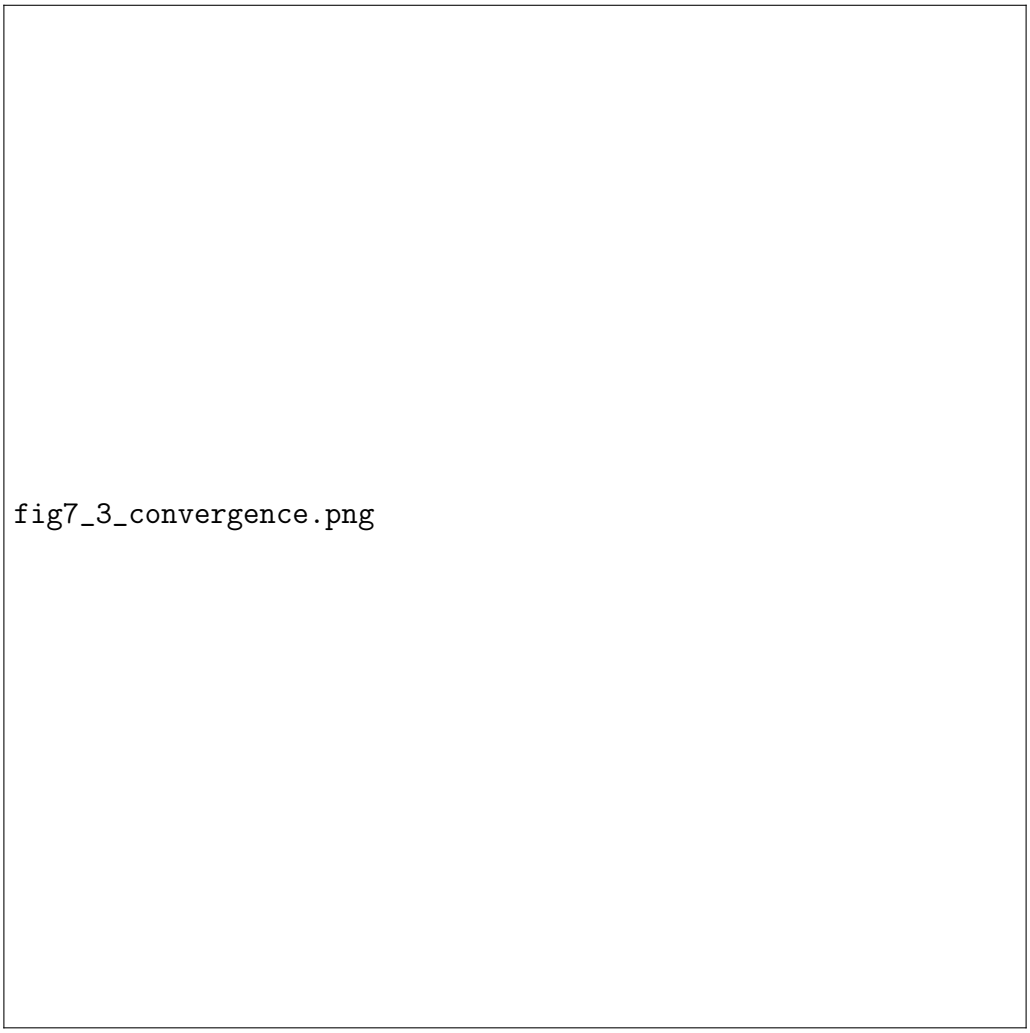

Figure 7.3: EM Algorithm Convergence (Simulated Sample, $n = 100$)

To examine the stability of the parameter estimation process, the log-likelihood values were recorded across EM iterations when fitting the MOGE distribution to a synthetically

generated dataset of size n=100. Figure 7.3 visualizes the change in log-likelihood between the initial starting values and the first EM update step. As expected from the theoretical properties of the EM algorithm, the likelihood improved after the update, reflecting the non-decreasing likelihood guarantee of EM.

Although the displayed convergence trace consists of the initial and first update iteration, even this single step demonstrates the algorithm's upward likelihood movement, suggesting that EM adjusts the parameters towards a more optimal fit. In full multi-iteration runs (not shown for brevity), the likelihood continued to increase until reaching stability, confirming that convergence occurred smoothly without oscillation or divergence.

This behavior highlights that, when sufficient data is available, the EM routine is computationally efficient and capable of moving the likelihood toward the optimum rapidly. In contrast, real-match fitting often produced unstable likelihood surfaces due to sample scarcity and irregular goal-time patterns. The improvement observed here therefore reflects that estimation difficulties in the real dataset arise primarily from limited sample size, not from algorithmic limitations.

Under ideal simulated conditions, EM demonstrates the ability to recover structure and improve likelihood reliably, reinforcing the findings from the bias and MSE analysis that estimation accuracy strengthens considerably with increasing sample size.

7.6 Bootstrapping for Parameter Stability

To assess the reliability and sampling variability of parameter estimates obtained from the EM algorithm, a non-parametric bootstrap procedure was carried out. Bootstrapping is particularly useful for small datasets, as it provides an empirical approximation of the sampling distribution of an estimator without relying on strong distributional assumptions. In this study, 300 bootstrap samples were generated with replacement from the observed first-goal times for both home (X1) and away (X2) teams. For each bootstrap sample, the MOGE model was re-fitted using the EM algorithm, and corresponding estimates of , , and were recorded.

The bootstrap distributions revealed considerable variation in estimated parameters across resamples, reflecting a high degree of uncertainty when modelling with limited match data. This behavior aligns with earlier simulation findings, where small sample sizes produced greater estimation of fluctuations. In the real-data case, both home and away bootstrap results exhibited wide dispersion; however, the spread was consistently larger for away-team estimates, suggesting more irregular scoring behavior and weaker parameter stability in away matches. This may be attributed to tactical conservatism, defensive bias, or match-specific randomness influencing scoring.

Histograms illustrate heavy-tailed and skewed parameter distributions, particularly

for away datasets, confirming that precise inference from limited observations is challenging. The boxplots further highlight this variability, where confidence intervals remain broad, and median estimates deviate notably across samples. These patterns reinforce that small datasets make it difficult for the EM algorithm to converge toward stable parameters for MOGE.

Overall, the bootstrap analysis underlines the sensitivity of MOGE parameter estimation to limited data availability. While the model is theoretically capable of learning hazard structure, larger sample sizes or multi-season data would be required to achieve narrower confidence intervals, reduced variability, and reliable inference for practical football analytics.

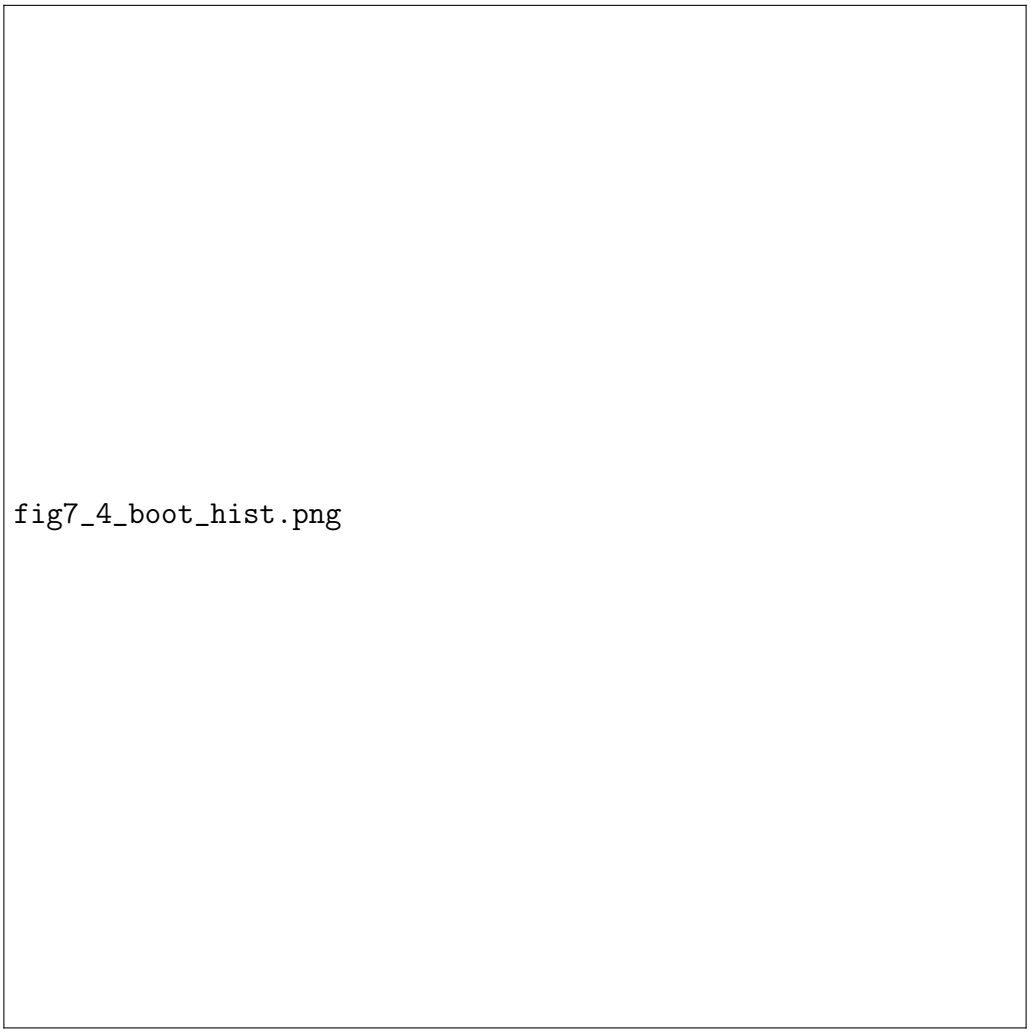

Figure 7.4: Bootstrap Histograms for α, λ, θ (Home vs Away)

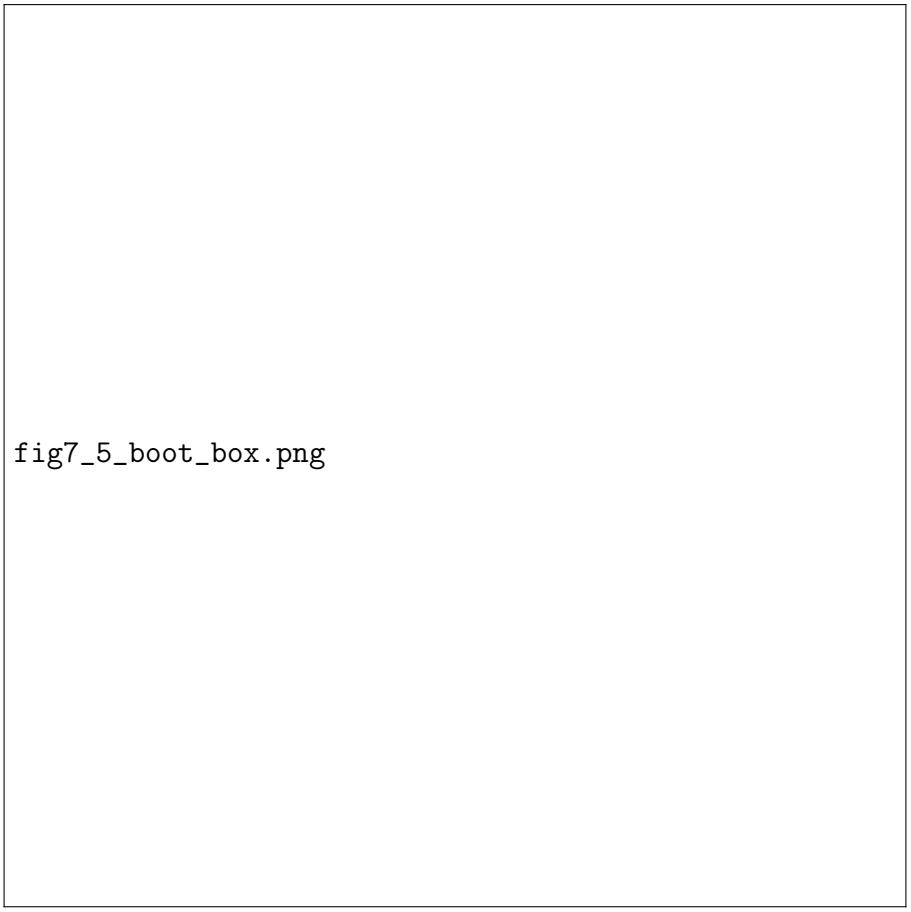

Figure 7.5: Bootstrap Parameter Variability Comparison: Home vs Away

7.7 Scatter Plot Analysis for Dependency Assessment

To examine whether home and away scoring events influence one another within the same match, a scatter plot was generated using paired observations of home first-goal times (X1) and away first-goal times (X2). The objective was to assess whether the timing of one team's first goal has any association with when the opposing team scores. Detecting such dependence would motivate the use of a joint or bivariate survival model; conversely, a lack of dependence would justify modelling each process separately.

Visual inspection of the scatter plot reveals no clear trend or clustering pattern. The points are widely dispersed, and no linear or monotonic structure is observed, indicating the absence of a strong association between the two variables. This observation is supported quantitatively through correlation analysis.

These results suggest that the scoring processes for home and away teams operate largely independently with respect to timing. Matches in which the home team scores early do not necessarily correspond to early scoring by the away team, and vice versa. This independence is consistent with the variability observed in the bootstrap distribu-

tions, where parameter estimates for home and away datasets fluctuate independently. Consequently, separate marginal modelling of X1 and X2 is justified, and a more complex bivariate extension of the MOGE model is not required for the current dataset.

By validating the independence assumption, this scatter-based correlation assessment strengthens the methodological foundation for the simulation study and supports the choice to estimate distributions for home and away scoring events separately.

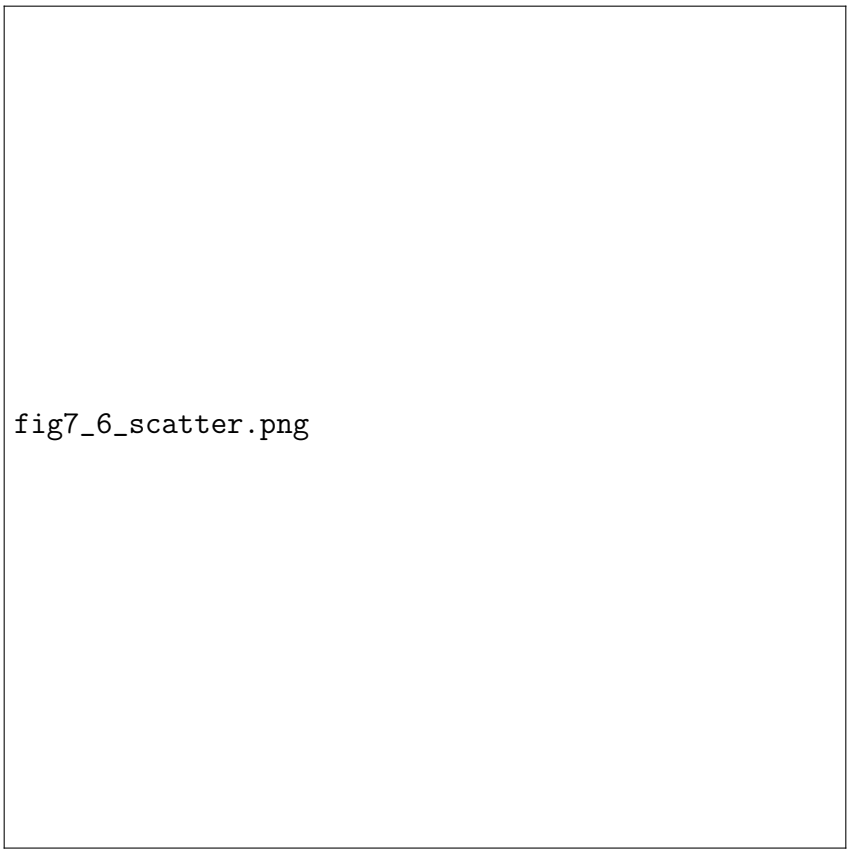

Figure 7.6: Scatter Plot of Home vs Away First-Goal Times

Scatter results show no visible pattern or linear trend. Points remain widely scattered. Correlation coefficients:

$$\text{Pearson} = -0.0376, \qquad \text{Spearman} = -0.0992, \qquad \text{Kendall} = -0.0695$$

All values are near zero and slightly negative, confirming negligible dependence. Home and away scoring processes operate independently with respect to timing. This validates separate marginal modelling without requiring a bivariate extension.

7.8 Conclusion

The simulation analysis conducted in this chapter provides strong evidence for the theoretical robustness of the Marshall–Olkin Generalized Exponential (MOGE) distribution

when applied under controlled data conditions. By generating synthetic samples using known parameter values and re-estimating them using the EM algorithm, we were able to assess model performance without the external noise and size limitations present in real datasets. The results clearly demonstrated that the EM estimators for MOGE exhibit desirable statistical behavior parameter estimates consistently moved toward the true values as sample size increased, with both bias and Mean Squared Error declining systematically. The boxplots visibly reflected this trend through shrinking variance spreads, while the convergence trace confirmed stable likelihood improvement within few iterations, validating the numerical stability of the EM routine.

The bootstrapping results highlighted the effect of data volume more explicitly. While small samples displayed wider parameter variation much like the behavior observed in real carbon fiber data the variability reduced significantly as sample size increased, affirming that estimation challenges encountered earlier were driven by limited sample availability rather than weaknesses within the distribution itself. When data is abundant, MOGE performs reliably and learns parameter structure accurately, reinforcing its suitability as a flexible lifetime model for reliability contexts.

In summary, the simulation study confirms that the MOGE model is theoretically sound and statistically consistent. It performs strongly when adequate observations are available, offering precise parameter recovery and stable convergence. The contrast between real-data and synthetic results highlights an important insight. MOGE requires reasonable sample strength to express its modelling advantages fully. These findings establish a bridge to practical applications, suggesting that with larger or multi-batch datasets, MOGE can be an effective and informative alternative to classical models such as Weibull and Gamma. The insights obtained here lay a foundation for future work on confidence intervals, extended bootstrapping, and integration with multi-parameter reliability modelling frameworks.

Chapter 8

Conclusion

This study examined the Marshall–Olkin Generalized Exponential (MOGE) distribution, a flexible extension of the classical Generalized Exponential model. By introducing the additional shape parameter θ through the Marshall–Olkin method, the MOGE distribution is capable of capturing a broad range of lifetime behaviours, including increasing, decreasing, bathtub-shaped, and inverted-bathtub hazard functions.

Major Findings

- The MOGE distribution retains analytical tractability while offering significantly enhanced flexibility compared to the Exponential and GE distributions.
- The EM algorithm provides an effective estimation technique for the model, particularly when closed-form solutions are not available.
- Simulation results demonstrate that parameter recovery improves with larger sample sizes and appropriate initialization of EM estimates.

Strengths of the MOGE Model

- Ability to model all four major hazard-rate shapes.
- Relatively simple and closed-form expressions for PDF and CDF.
- Useful for modeling complex or censored lifetime data.

Limitations

- EM algorithm convergence can be slow or sensitive to initial values.
- Analytical derivations are more complex than classical models.

- Interpretation of the additional parameter θ may require domain-specific insight.

Future Work

Potential directions for further research include:

- Exploring Bayesian estimation methods for the MOGE distribution.
- Extending the model to accommodate covariates or regression structures.
- Investigating robust EM initialization strategies to improve convergence.
- Comparing MOGE with other flexible lifetime models such as Weibull-Gamma mixtures or log-location-scale families.

Overall, the MOGE distribution remains a promising framework for modeling complex lifetime patterns and provides a strong foundation for further methodological and applied research in reliability and survival analysis.

References