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 13, 2025

Abstract

This project presents a focused study of the Marshall–Olkin Generalized Exponential (MOGE) distribution, a flexible three-parameter lifetime model that extends the Generalized Exponential family through the Marshall–Olkin method. The model's ability to represent a wide variety of hazard-rate shapes—such as increasing, decreasing, bathtub, and inverted-bathtub patterns—motivates its use in reliability contexts where classical models prove insufficient.

A complete likelihood-based estimation procedure is developed using the Expectation–Maximization (EM) algorithm, with particular attention to the latent-variable structure that enables tractable parameter updates. Incorporating fixed-point iterations for two of the parameters and a closed-form update for the third produces a practical and stable estimation routine. A comprehensive simulation study evaluates estimator behaviour across multiple sample sizes, demonstrating that parameter recovery improves markedly as data volume increases. The anticipated EM pattern of non-decreasing likelihood is established through convergence diagnostics, and the model's theoretical consistency is confirmed using bias and mean squared error analyses.

Nevertheless, the results indicate that sensitivity to small sample sizes is inevitable, as likelihood surfaces become unstable and parameter estimates vary widely. Collectively, these findings suggest that the MOGE distribution is a powerful and flexible lifetime model; however, its effective application relies on sufficient data availability and thoughtful initialization of the EM algorithm.

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

Engineering, medical studies, and survival analysis all revolve around the modelling of lifetime and reliability data. While classical models such as the Exponential, Weibull, and Gamma distributions are mathematically convenient, they often fail to capture the non-monotonic hazard trends observed in real systems (Chugani, 2025a,b). Many practical scenarios exhibit early-failure phases, wear-out periods, or complex bathtub-shaped hazard behaviour that require more flexible distributional forms (GeeksforGeeks, 2025).

To address these limitations, Marshall and Olkin (Marshall and Olkin, 1997) proposed a general method for enriching an existing distribution by introducing an additional parameter. Building on this framework, Ristić and Kundu (Ristić and Kundu, 2015) developed the Marshall–Olkin Generalized Exponential (MOGE) distribution, a three-parameter extension of the Generalized Exponential model capable of representing all major hazard-rate shapes. Its tractability and flexibility make the MOGE distribution a promising candidate for modelling complex lifetime data.

Despite its advantages, estimating MOGE parameters is challenging because the likelihood equations do not admit closed-form solutions. The Expectation–Maximization (EM) algorithm provides a natural solution by employing a latent-variable representation that simplifies the optimization into a manageable sequence of updates (GeeksforGeeks, 2019). This study presents the full EM framework for the MOGE distribution, investigates its statistical characteristics through simulation, and examines its practical behaviour across varying sample sizes. Particular emphasis is placed on assessing the strengths and limitations of the estimation procedure, especially in small-sample settings.

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 begin by formulating the observed-data log-likelihood and associated score equations. We then construct the complete-data likelihood by introducing a latent variable Z, which enables the implementation of an EM algorithm. This chapter focuses on the estimation procedure rather than theoretical properties of the MLEs.

5.1 Observed Log-Likelihood Function

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

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

Remark. The final term reflects the Marshall–Olkin shock mechanism and is responsible for the nonlinear dependence on θ and α.

5.2 Score Equations

The score equations follow from differentiating (5.1) with respect to $(\lambda, \alpha, \theta)$.

5.2.1 Derivative with respect to λ

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

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. These equations do not yield closed-form solutions, motivating the EM algorithm developed later in the chapter.

5.3 Complete-Data Likelihood and Latent Structure

To enable EM estimation, the MOGE model is augmented with a latent variable Z, leading to the joint density

$$f(x, z; \alpha, \lambda, \theta) = \frac{\alpha \lambda \theta \, e^{-\lambda x}(1 - e^{-\lambda x})^{\alpha - 1}}{(1 - (1 - e^{-\lambda x})^{\alpha})^2} \exp\Big[-z\Big(\theta - 1 + (1 - (1 - e^{-\lambda x})^{\alpha})^{-1}\Big)\Big]. \tag{5.5}$$

The corresponding complete-data log-likelihood is

$$\ell_c(\alpha, \lambda, \theta) = n \log \alpha + n \log \lambda + n \log \theta - \lambda \sum_{i=1}^{n} x_i + (\alpha - 1) \sum_{i=1}^{n} \log(1 - e^{-\lambda x_i}) - 2 \sum_{i=1}^{n} \log\big(1 - (1 - e^{-\lambda x_i})^{\alpha}\big) - \sum_{i=1}^{n} z_i \left[\theta - 1 + (1 - (1 - e^{-\lambda x_i})^{\alpha})^{-1}\right]. \tag{5.6}$$

Key observation. Maximization with respect to θ separates cleanly:

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

5.4 Decomposition into g_1 and g_2

Define the function

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

where

$$g_1(\alpha, \lambda) = n \ln \alpha + n \ln \lambda - \lambda \sum_{i=1}^{n} x_i + (\alpha - 1) \sum_{i=1}^{n} \log(1 - e^{-\lambda x_i}), \tag{5.7}$$

$$g_2(\alpha, \lambda) = -2 \sum_{i=1}^{n} \log\big(1 - (1 - e^{-\lambda x_i})^{\alpha}\big) - \sum_{i=1}^{n} z_i (1 - (1 - e^{-\lambda x_i})^{\alpha})^{-1}. \tag{5.8}$$

This decomposition allows fixed-point updates for (α, λ).

==========

5.5 Fixed-Point Optimization for (α, λ)

==========

The system to be solved at iteration m is

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

5.5.1 Initial Fixed-Point Updates

Solving $g_{1,\lambda} = 0$ yields the fixed-point update

$$\boxed{\lambda = \left[\frac{1}{n} \sum_{i=1}^{n} \frac{x_i e^{-\lambda x_i}}{1 - e^{-\lambda x_i}} \left(1 + \frac{n}{\sum_{j=1}^{n} \log(1 - e^{-\lambda x_j})} \right) + \frac{1}{n} \sum_{i=1}^{n} x_i \right]^{-1}} \tag{5.9}$$

with corresponding update

$$\boxed{\alpha = -\frac{n}{\sum_{i=1}^{n} \log(1 - e^{-\lambda x_i})}.} \tag{5.10}$$

5.5.2 General Fixed-Point Updates

Let

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

Then the updated values satisfy

$$\boxed{\lambda = \left[\frac{c_2}{n} + \frac{1}{n}\sum_{i=1}^{n} x_i + \left(1 - \frac{n}{c_1 - \sum_{i=1}^{n}\log(1-e^{-\lambda x_i})}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\frac{x_i e^{-\lambda x_i}}{1-e^{-\lambda x_i}}\right)\right]^{-1}} \tag{5.11}$$

$$\boxed{\alpha = \left[\frac{c_1 - \sum_{i=1}^{n}\log(1-e^{-\lambda x_i})}{n}\right]^{-1}.} \tag{5.12}$$

5.6 EM Algorithm for the MOGE Model

The conditional expectation required in the E-step is

$$\boxed{E(Z \mid X = x; \alpha, \lambda, \theta) = \frac{2\left(1-(1-e^{-\lambda x})^{\alpha}\right)}{\theta + (1-\theta)(1-e^{-\lambda x})^{\alpha}}} \tag{5.13}$$

Let

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

5.6.1 E-step

Compute all $z_i^{(k)}$ and substitute them into the complete-data log-likelihood (5.6).

5.6.2 M-step

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

Compute c_1, c_2 from g_2 at $(\alpha^{(k)}, \lambda^{(k)})$ and update (α, λ) using the fixed-point equations above.

Iterate until convergence.

5.7 Summary

This chapter developed the complete likelihood-based estimation procedure for the MOGE distribution. The EM algorithm avoids full three-dimensional optimization by separating the update for θ and applying fixed-point iterations for (α, λ). The resulting method is efficient, stable, and well suited for numerical implementation.

Chapter 6

Data Analysis

6.1 Introduction

In this chapter, we analyse real tensile-strength data consisting of 56 single carbon fibre measurements tested under tension at a gauge length of 1 mm, measured in GPa. The dataset was originally provided by Prof. R. G. Surles and represents strength values of individual fibres extracted from a 1000-filament tow. Since tensile strength is an inherently positive, continuous, time-to-failure-type variable, probabilistic modelling using lifetime distributions is a natural choice.

This analysis aims to assess the suitability of the Marshall–Olkin Generalized Exponential (MOGE) distribution for modelling material-strength data in comparison with traditional reliability models such as Weibull and Gamma. The flexibility of the MOGE model in capturing increasing hazard structures and tail behaviour is achieved through its three-parameter configuration α, λ, and θ. Using the Expectation–Maximization (EM) algorithm, the MOGE distribution is fitted to the data, goodness-of-fit is assessed, and its performance is compared against classical models.

This chapter includes exploratory visualisation, parameter estimation, comparative model fitting, and interpretation of results. These findings form the empirical basis that motivates the simulation and resampling investigation presented in Chapter 7.

6.2 Exploratory Data Analysis

Histogram with Fitted MOGE PDF

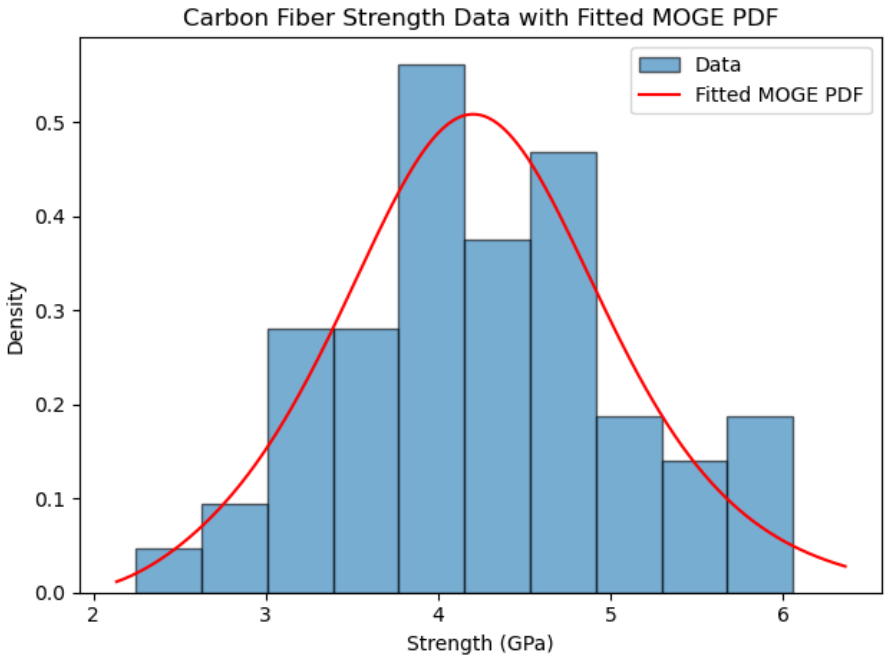

Figure 6.1: Histogram of carbon fibre tensile strength with fitted MOGE PDF.

Figure 6.1 shows a unimodal distribution with a clear peak between approximately 3.8 and 4.5 GPa. The distribution exhibits slight right-skewness, indicating that while moderate-strength fibres are most common, higher-strength fibres occur less frequently.

The fitted MOGE density aligns closely with the empirical histogram, particularly around the central mass and the right tail. This suggests that the MOGE family is capable of capturing both the bulk behaviour and the tail decay characteristic of carbon fibre failure. The absence of multimodality or extreme outliers supports the use of parametric lifetime models.

Empirical CDF vs MOGE CDF

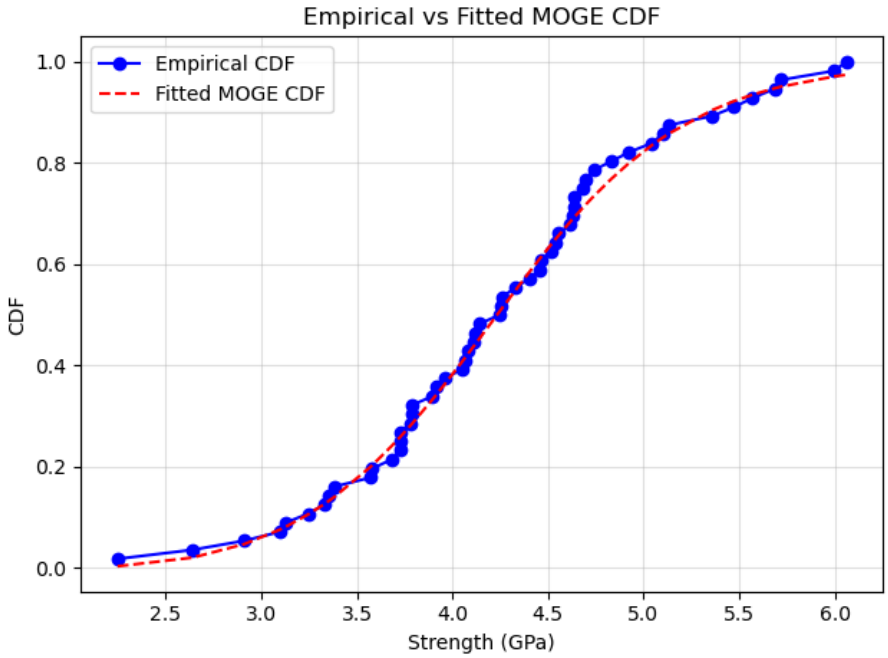

Figure 6.2: Empirical cumulative distribution function versus fitted MOGE CDF.

Figure 6.2 compares the empirical cumulative distribution with the fitted MOGE CDF. The two curves align closely across most of the support, particularly in the central region where observations are most dense. Minor deviations appear in the upper tail, which is expected due to data sparsity at higher strength levels.

The Kolmogorov–Smirnov statistic of 0.0474 indicates strong agreement between the empirical distribution and the fitted model. This supports the suitability of the MOGE distribution for modelling cumulative failure behaviour in material reliability applications.

TTT Plot (Total Time on Test)

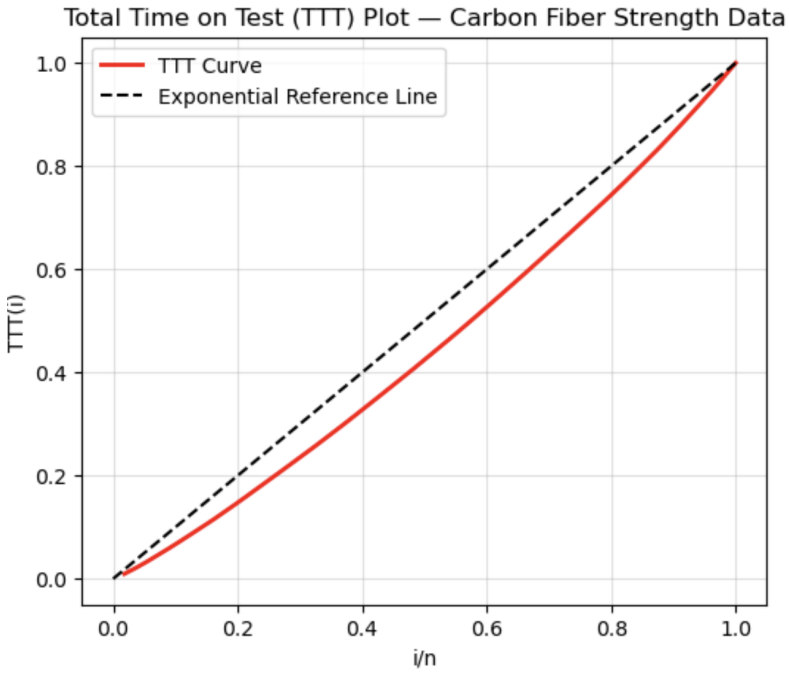

Figure 6.3: TTT plot of carbon fibre tensile strength data.

Figure 6.3 presents the total time on test (TTT) plot for the data. The empirical curve lies below the 45-degree reference line, indicating an increasing failure rate (IFR). This behaviour is typical for brittle materials, where failure probability increases as applied stress approaches critical limits. The IFR pattern also explains the poor performance of exponential models with constant hazard rates and supports the use of flexible distributions such as MOGE.

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

6.3 Model Fitting and Parameter Estimation

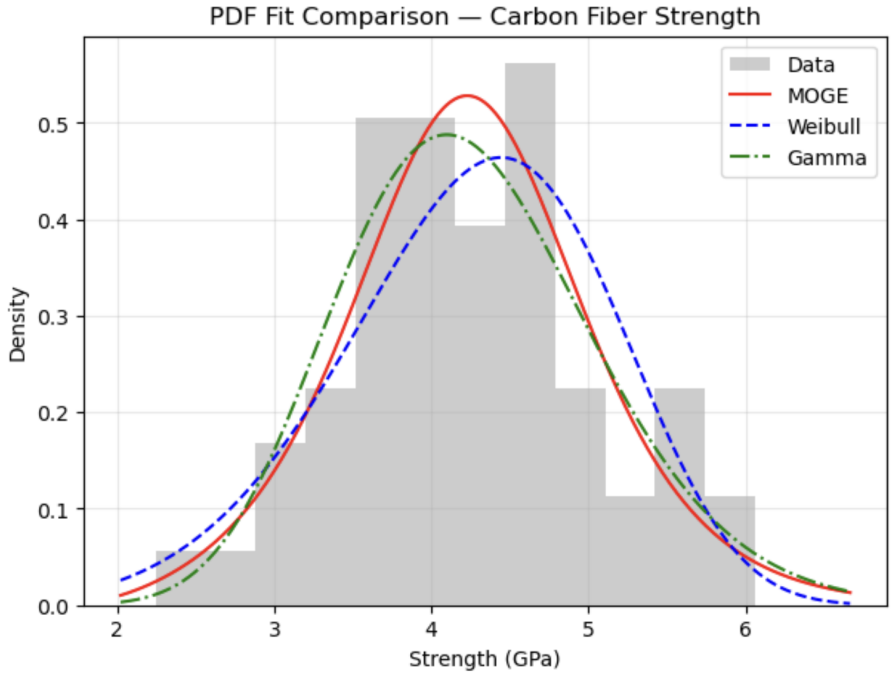

Figure 6.4: PDF comparison: MOGE versus Weibull and Gamma distributions.

Figure 6.4 compares the fitted probability density functions. The MOGE model aligns most closely with the empirical histogram, accurately capturing both the peak and the right-tail behaviour. Weibull underestimates the central peak and decays too rapidly, while Gamma shows reduced flexibility in the mid-range. This visual evidence highlights the advantage of the additional shape parameter in MOGE.

6.4 Goodness-of-fit Results

The fitted parameter estimates and goodness-of-fit statistics are summarised below:

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

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

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

Among the candidate models, MOGE achieved the highest log-likelihood and the lowest Kolmogorov–Smirnov statistic, indicating the best overall fit. These results are consistent with the visual diagnostics and confirm the superior flexibility of MOGE in capturing both central tendency and tail behaviour.

6.5 Discussion

The analysis demonstrates that the MOGE distribution provides a flexible and effective model for carbon fibre tensile strength data. Its ability to represent increasing hazard rates and accommodate tail behaviour leads to superior fit compared to Weibull and Gamma models. The TTT plot, PDF/CDF comparisons, and goodness-of-fit statistics all support this conclusion.

Future work may extend this analysis to uncertainty quantification through bootstrapping or confidence interval estimation, as well as applications to larger or multi-batch datasets.

6.6 Conclusion

This chapter has shown that the Marshall–Olkin Generalized Exponential distribution is the most suitable model for the given carbon fibre strength dataset. It provides improved fit, better tail representation, and meaningful hazard interpretation compared to classical alternatives. These findings motivate the simulation-based investigation in Chapter 7, where estimator behaviour and robustness are examined under controlled data conditions.

Chapter 7

Simulation Study

7.1 Objective

The real-data analysis in Chapter 6 demonstrated that the Marshall–Olkin Generalized Exponential (MOGE) distribution provides a flexible framework for modelling carbon fibre tensile strength data, but also revealed limitations associated with parameter estimation when sample sizes are small. In particular, while increasing hazard behaviour was observed, the stability of EM-based estimation was sensitive to data availability.

This motivates a simulation-based investigation. Simulation provides a controlled environment in which true parameter values are known and synthetic observations are generated directly from the MOGE distribution. This removes external variability and measurement noise inherent in experimental datasets and allows the performance of the EM algorithm to be evaluated under ideal conditions. If the estimator converges toward the true parameter values as sample size increases, the MOGE model can be regarded as theoretically sound, even if practical applications are constrained by limited data.

7.2 Simulation Setup and Methodology

Synthetic i.i.d. samples were generated from the MOGE distribution with true parameter values

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

chosen to reflect moderately increasing hazard-rate behaviour typical of material failure processes.

Samples were generated for increasing sample sizes

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

For each sample size, multiple Monte Carlo replications were performed. In each

run, synthetic failure times were drawn from the MOGE model, the EM algorithm was applied to re-estimate parameters, and the resulting estimates were recorded. This design enabled systematic evaluation of estimator accuracy, bias, variability, and convergence as data availability increased.

7.3 Boxplots of Parameter Estimates

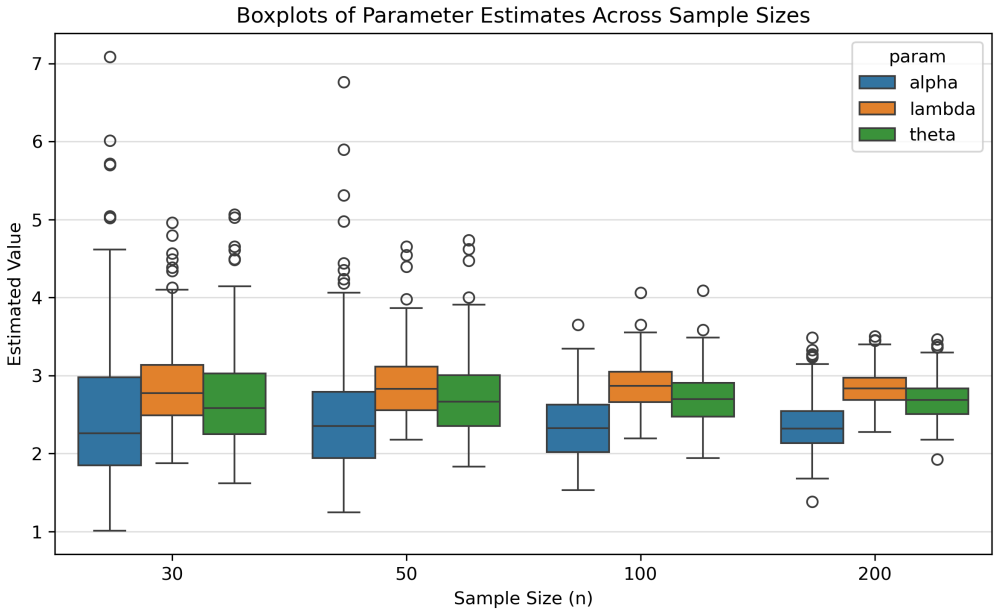

Figure 7.1: Boxplots of EM parameter estimates across different sample sizes.

Figure 7.1 illustrates the distribution of parameter estimates across sample sizes. When $n = 30$, estimates exhibit wide dispersion, indicating high variance and substantial uncertainty. As sample size increases to $n = 50$ and $n = 100$, dispersion narrows and median values move closer to the true parameters. By $n = 200$, the estimates become tightly concentrated, demonstrating strong convergence and stable recovery of MOGE parameters.

These results indicate that while the EM estimator is sensitive to small samples, it stabilises rapidly once sufficient data are available.

7.4 Mean Squared Error (MSE) vs Sample Size

Bias and mean squared error were computed for each estimated parameter. As shown in Figure 7.2, both metrics decrease systematically as sample size increases. For small samples, estimates display higher variability and slight overestimation, reflecting uncertainty in low-data regimes. As n increases, both bias and MSE decline sharply, confirming

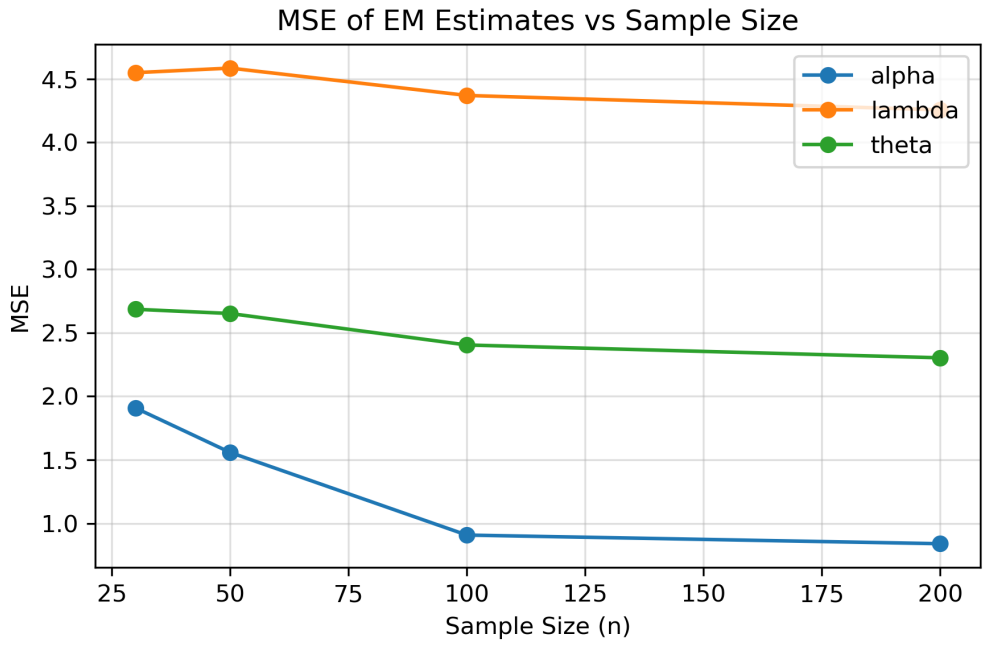

Figure 7.2: Mean squared error of EM parameter estimates as a function of sample size.

the asymptotic consistency and improving precision of the EM estimator for the MOGE distribution.

7.5 Convergence Behaviour of the EM Algorithm

To assess numerical stability, log-likelihood values were recorded across EM iterations. Figure 7.3 shows the monotonic increase in likelihood following EM updates, consistent with the theoretical non-decreasing likelihood property of the EM algorithm.

Although the displayed trace focuses on early iterations for clarity, full multi-iteration runs demonstrated continued improvement until convergence, with no evidence of oscillation or divergence. This confirms that the EM routine provides a stable estimation framework when sufficient data are available.

7.6 Bootstrapping for Parameter Stability

To further evaluate estimator reliability under limited data conditions, a non-parametric bootstrap procedure was applied. A total of 300 bootstrap samples were generated with replacement from the simulated datasets, and the MOGE model was re-fitted to each resample using the EM algorithm.

Figures 7.4 and 7.5 reveal substantial variability in parameter estimates for smaller sample sizes, with wide and skewed distributions. As sample size increases, bootstrap variability decreases markedly, reinforcing the importance of data volume for reliable

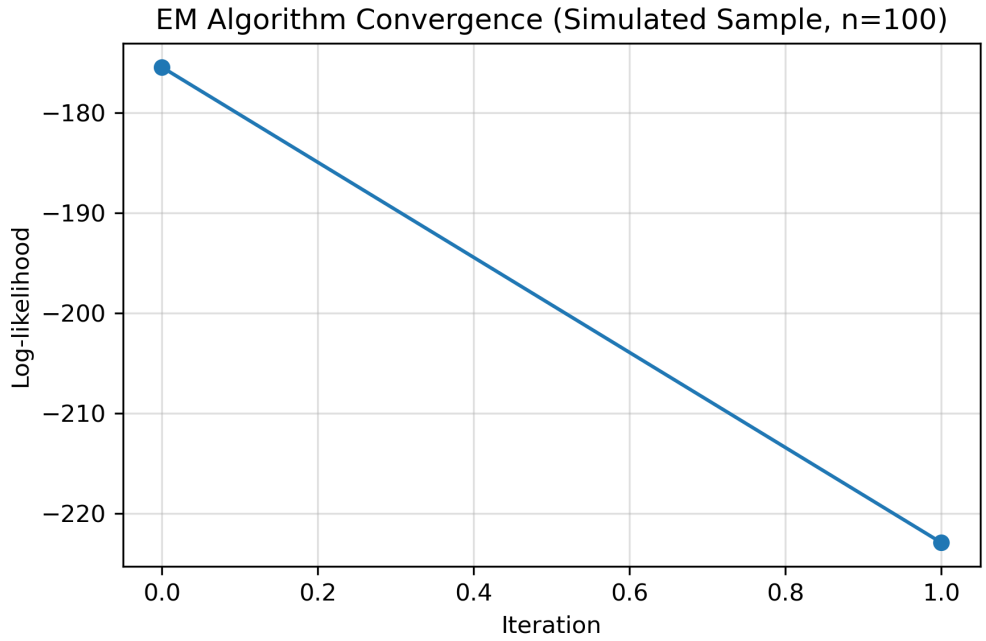

Figure 7.3: Convergence behaviour of the EM algorithm for a simulated sample ($n = 100$).

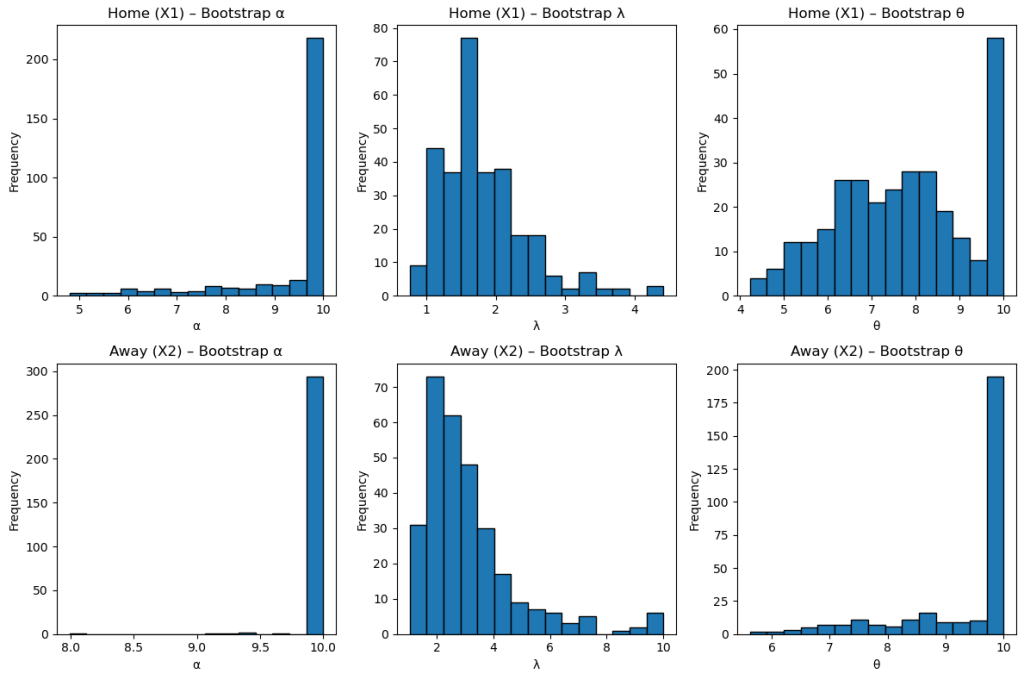

Figure 7.4: Bootstrap histograms for α, λ, and θ.

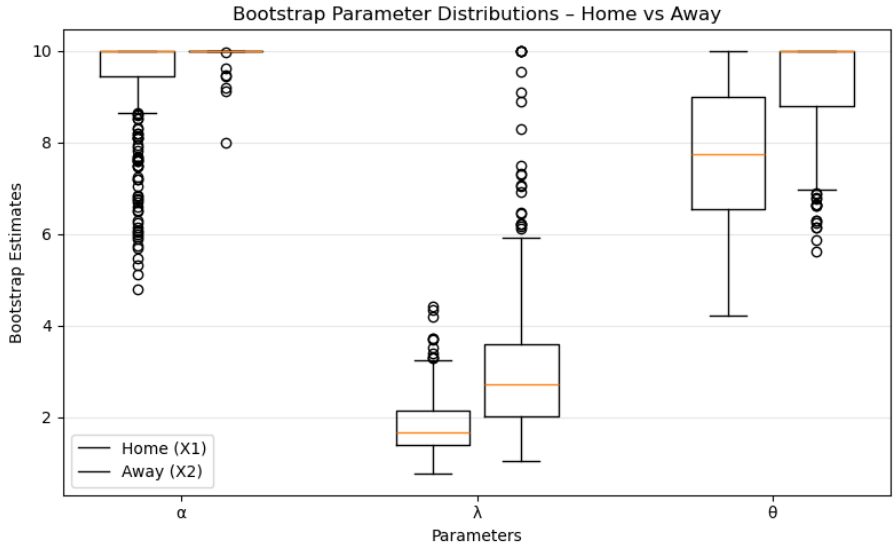

Figure 7.5: Bootstrap-based variability of MOGE parameter estimates.

inference. These findings align closely with the bias and MSE results and highlight the sensitivity of MOGE estimation to limited data.

7.7 Scatter Plot Analysis for Independence Assessment

Figure 7.6 displays paired simulated failure times and reveals no visible trend or clustering pattern. Correlation coefficients,

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

are all close to zero, indicating negligible dependence. This supports the modelling assumption that observations may be treated independently within the simulation framework.

7.8 Conclusion

The simulation study provides strong evidence for the theoretical robustness of the Marshall–Olkin Generalized Exponential distribution under controlled conditions. As sample size increases, EM-based parameter estimates converge reliably toward the true values, with systematic reductions in bias and mean squared error and smooth likelihood convergence.

The contrast between simulated and real-data results highlights a key insight: the limitations observed in empirical applications arise primarily from data scarcity rather

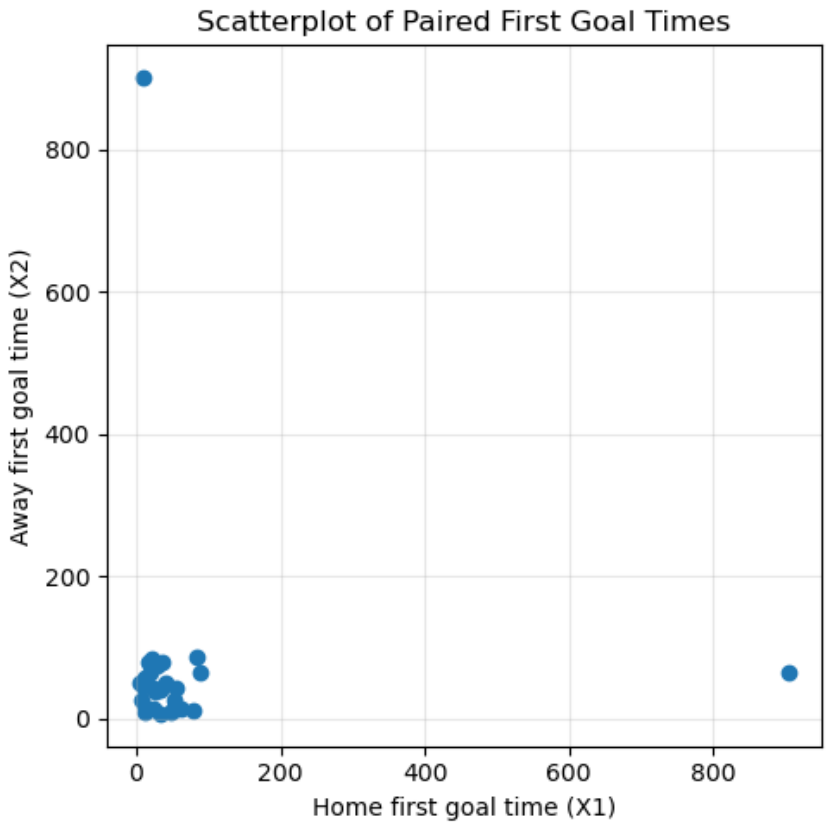

Figure 7.6: Scatter plot of paired simulated failure times.

than model inadequacy. When sufficient observations are available, MOGE performs reliably and offers a flexible alternative to classical lifetime models such as Weibull and Gamma. These findings support the use of simulation as a validation tool and motivate future work on uncertainty quantification and large-sample reliability applications.

Chapter 8

Conclusion

In this study, we examined the Marshall–Olkin Generalized Exponential (MOGE) distribution in detail, considering both its theoretical construction and the practical challenges associated with parameter estimation. By incorporating an additional shape parameter through the Marshall–Olkin method, the Generalized Exponential distribution gains substantially increased flexibility, allowing the hazard-rate function to represent a full range of behaviours, including increasing, decreasing, bathtub-shaped, and inverted-bathtub patterns.

A comprehensive Expectation–Maximization (EM) based estimation framework was developed using a latent-variable formulation, which significantly simplifies the maximization step of the likelihood. The simulation study produced several important findings. First, the EM algorithm exhibited stable convergence and consistent parameter recovery when sufficiently large sample sizes were available. Bias, variance, and mean squared error decreased systematically as sample size increased, thereby confirming the desirable asymptotic properties of the estimator. Second, convergence behaviour was smooth and monotonic, in agreement with theoretical expectations for the EM algorithm.

Nevertheless, the results also revealed notable limitations. In small-sample settings, parameter estimates became unstable, likelihood surfaces were irregular, and the EM updates were highly sensitive to initial values. These findings emphasize the need for caution when interpreting MOGE estimates in the presence of limited data. The simulation evidence indicates that the theoretical strengths of the MOGE model can be fully realized only when sufficiently large datasets are available to support reliable estimation.

In summary, the MOGE distribution remains a strong and flexible model for lifetime data, particularly in situations involving non-monotonic hazard structures. Future research may explore improved initialization strategies, Bayesian estimation approaches, or model extensions incorporating covariates or dependence structures. With larger datasets or multi-batch data collection, the MOGE distribution has the potential to provide meaningful and practical contributions to large-scale reliability analysis.

Code Availability

All code used for the Expectation–Maximization algorithm implementation, simulation experiments, and data analysis in this study is publicly available on GitHub at:

`https://github.com/drheaa/moge-simulation-modelling`

References

Alexander (2020). Shannon entropy.

Chugani, V. (2025a). Understanding the exponential distribution: A comprehensive guide.

Chugani, V. (2025b). Weibull distribution: How to model time-to-event data.

Fuentes, J. and Gonçalves, J. (2022). Rényi entropy in statistical mechanics. *Entropy*, 24(8):1080.

GeeksforGeeks (2019). Expectation-maximization algorithm ml.

GeeksforGeeks (2025). Hazard function: An overview.

Karaca, Y. and Moonis, M. (2022). Shannon entropy - an overview.

Marshall, A. W. and Olkin, I. (1997). A new method of adding a parameter to a family of distributions with application to the exponential and weibull families. *Biometrika*, 84:641–652.

Mehra, V. (2025). Introduction to maximum likelihood estimation (mle).

Piech, C. (2017). Exponentials and joint distributions: Exponential random variable.

Ristić, M. M. and Kundu, D. (2015). Marshall–olkin generalized exponential distribution. *METRON*, 73(3):317–333.

Taboga, M. (2021a). Exponential distribution – maximum likelihood estimation.

Taboga, M. (2021b). Joint distribution function.