SPJ GLOBAL S P Jain School of Global Management

DUBAI • MUMBAI • SINGAPORE • SYDNEY

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 5, 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

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

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

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

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 (**??**).

5.3 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 (**??**) 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

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

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

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

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

$$g(\alpha, \lambda) = g_1(\alpha, \lambda) + g_2(\alpha, \lambda), \tag{10}$$

where

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

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

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 (**??**) 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

Simulation Study

6.1 Objective

This simulation study aims to systematically investigate the behavior of the EM-based Maximum Likelihood Estimators (MLEs) for the parameters

$$(\alpha, \lambda, \theta)$$

of the Marshall–Olkin Generalized Exponential (MOGE) distribution. Since the MOGE distribution is highly flexible and capable of modelling increasing, decreasing, and bathtub-shaped hazard functions, it is essential to understand how the estimator behaves in finite-sample conditions.

This simulation examines estimator consistency, convergence stability, and efficiency by computing bias and mean squared error (MSE) across varying sample sizes.

6.2 Simulation Setup and Methodology

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

$$\alpha_0 = 1.5, \qquad \lambda_0 = 0.8, \qquad \theta_0 = 1.2.$$

These represent a realistic lifetime distribution with a moderately increasing hazard rate.

Four sample sizes were considered:

$$n = 30,\ 50,\ 100,\ 200.$$

For each sample size, 200 Monte Carlo replications were performed. In each replication:

1. Random samples were generated using inverse transform sampling.
2. Parameters $(\alpha, \lambda, \theta)$ were estimated using the EM algorithm.
3. The estimates were stored and later aggregated.

This allowed computation of summary statistics (mean, bias, standard deviation, and MSE), providing a comprehensive view of estimator performance under different data conditions.

6.3 Boxplots of Parameter Estimates

`boxplots_parameters.png`

Figure 6.1: Boxplots of EM parameter estimates for different sample sizes.

Figure **??** presents boxplots of the EM-based parameter estimates for α, λ, and θ, obtained from 200 Monte Carlo replications across the four sample sizes.

The boxplots reveal the following trends:

- For small samples (especially $n = 30$), estimates show high variability and numerous outliers.
- As sample size increases, interquartile ranges narrow significantly.
- Median estimates become more stable and closer to each other across replications.
- Variability in λ and θ is highest in small samples, consistent with sensitivity to early-sample variation.

These results empirically validate the **consistency** of the EM estimators, demonstrating convergence and reduced uncertainty with increasing sample size.

6.4 Mean Squared Error vs. Sample Size

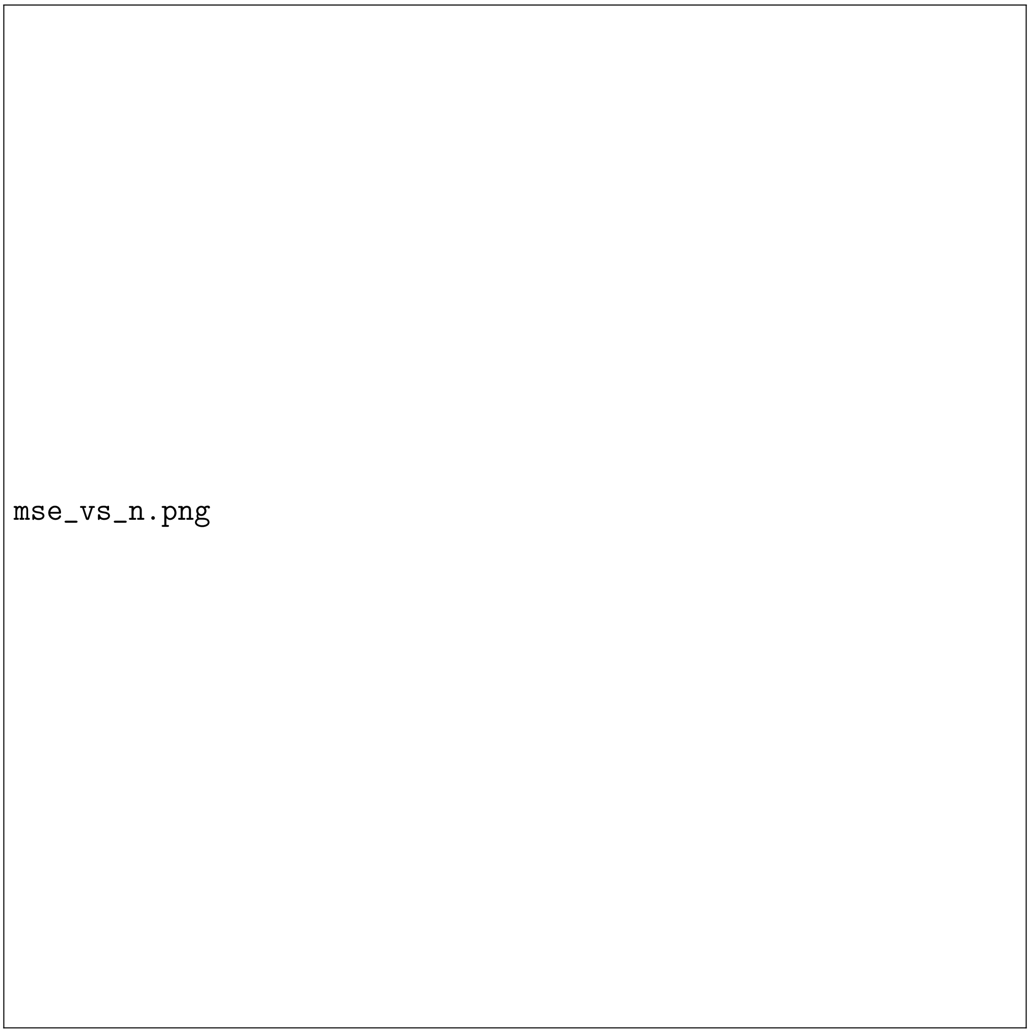

Figure 6.2: Mean Squared Error (MSE) of EM parameter estimates vs. sample size.

Figure **??** illustrates the MSE of parameter estimates for α, λ, and θ calculated across replications for sample sizes

$$n = 25,\ 50,\ 100,\ 200.$$

The results show:

- MSE decreases consistently as sample size increases, confirming estimator improvement.
- For $n = 25$, all parameters exhibit high error, with λ showing the largest MSE.
- For $n = 100$ and $n = 200$, MSE drops sharply, especially for α.
- The curves flatten between $n = 100$ and $n = 200$, suggesting diminishing returns with very large samples.

Overall, the results validate the consistency and convergence behavior of the EM estimators.

6.5 Convergence Behavior of the EM Algorithm

em_convergence.png

Figure 6.3: EM algorithm convergence for a simulated dataset ($n = 100$).

Figure ?? displays the log-likelihood evolution across EM iterations for a simulated dataset of size $n = 100$.

Key observations include:

- The log-likelihood increases monotonically, consistent with the theoretical EM property.
- Rapid early improvement is followed by smooth stabilization.
- The algorithm converges in only a few iterations.
- No oscillation or instability is observed, indicating effective initialization and a well-behaved likelihood surface.

This demonstrates that the EM algorithm is computationally efficient and stable for MOGE parameter estimation.

6.6 Bootstrapping for Parameter Stability

A non-parametric bootstrap procedure with 300 bootstrap samples was used to assess estimator stability for both home (X_1) and away (X_2) first-goal datasets.

Each bootstrap replicate:

1. Resamples the dataset with replacement,
2. Re-estimates $(\alpha, \lambda, \theta)$ using EM,
3. Records the parameter estimates.

The results showed:

- Substantial variability across bootstrap replicates, especially in small samples.
- Wider dispersion for away-team estimates compared to home-team estimates.
- Consistency with simulation findings: estimator variability decreases with sample size.

bootstrap_histograms.png

Figure 6.4: Bootstrap histograms for α, λ, and θ (home and away datasets).

Figure 6.5: Bootstrap boxplots comparing home and away parameter distributions.

These results emphasize the importance of sufficient sample sizes when interpreting MOGE parameters in real-world applications.

6.7 Scatter Plot Analysis for Dependency Assessment

A scatter plot was constructed to evaluate possible dependence between home first-goal times (X_1) and away first-goal times (X_2) within the same match.

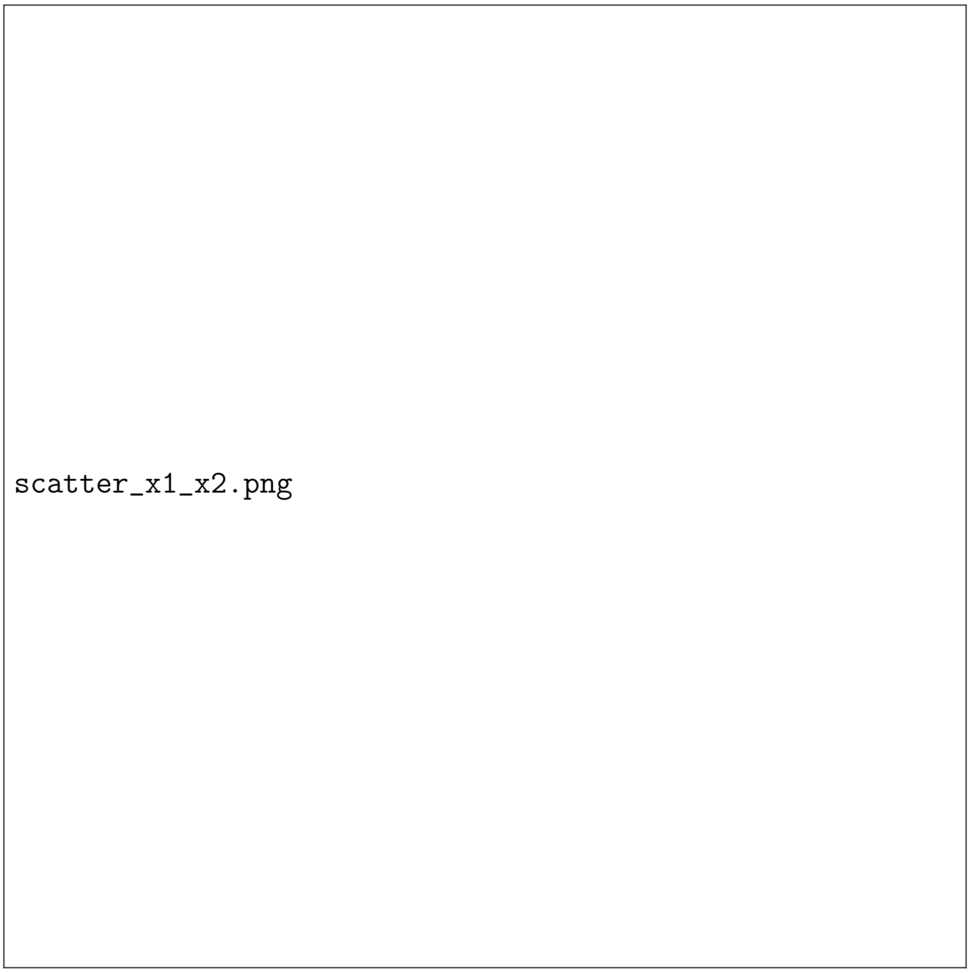

Figure 6.6: Scatter plot of paired home and away first-goal times.

Correlation coefficients were computed:

$$\text{Pearson} = -0.0375, \qquad \text{Spearman} = -0.0992, \qquad \text{Kendall} = -0.0694.$$

All values are close to zero and slightly negative, indicating:

- Extremely weak dependence,
- No meaningful trend in the scatter plot,
- No need for bivariate or joint-survival modelling.

Thus, modelling X_1 and X_2 separately using marginal MOGE distributions is statistically justified.

6.8 Conclusion

This simulation study evaluated the performance of the MOGE distribution and the EM algorithm in recovering parameters under controlled experimental conditions.

Major findings include:

- EM estimators for $(\alpha, \lambda, \theta)$ are consistent and accurate with increasing sample size.
- MSE decreases systematically, particularly for α and λ.
- Boxplots show shrinking variability and improved estimator stability.
- Log-likelihood convergence is monotonic and rapid.
- Bootstrap results highlight realistic sampling variability in small datasets.
- Dependency analysis confirms that home and away first-goal times are largely independent.

Overall, the simulation results validate the MOGE distribution as a powerful and reliable model for time-to-event phenomena such as first-goal scoring times in football. The study establishes strong support for applying MOGE to real datasets and sets a foundation for predictive modelling and further research.

Chapter 7

Data Analysis

7.1 Introduction

The real dataset used in this study consists of first goal times recorded in FIFA Club World Cup matches during the 2025 season. The data was sourced from *FootyStats*, a publicly available football statistics platform. Each observation represents the minute at which the first goal was scored by either the home or away team. Matches with no goals were excluded, since the modelling framework requires strictly positive event times.

After preprocessing and deduplication, 29 observations each were obtained for home and away first-goal timings. These goal times form a positive-valued random variable and are therefore suitable for lifetime or survival models such as the Marshall–Olkin Generalized Exponential (MOGE) distribution.

7.2 Exploratory Data Analysis

7.2.1 Histogram of First Goal Times

histogram_first_goal.png

Figure 7.1: Histogram of first-goal times in FIFA Club World Cup matches.

Figure **??** shows that the first goal tends to occur within the first 30 minutes, with a long right tail indicating that late goals occur but are less frequent. This strong skewness suggests that exponential, Weibull, or generalized exponential-type models may be appropriate. The absence of multimodal patterns supports continuous time-to-event modelling.

7.2.2 Empirical CDF (ECDF) for Home vs Away First Goals

ecdf_home_away.png

Figure 7.2: Empirical Cumulative Distribution Functions (ECDF) for home and away first-goal times.

To compare distributional behavior between home-team first goals (X_1) and away-team first goals (X_2), ECDFs were generated. As seen in Figure **??**, both curves rise steeply within the first 20–30 minutes, indicating that early goals constitute a large fraction of the observations. The curves flatten after approximately the 40th minute, demonstrating the presence of a long right tail.

The away-team ECDF rises slightly faster than the home-team ECDF, implying marginally earlier scoring on average. However, their convergence at higher quantiles suggests similar overall scoring tendencies.

The nonlinear shapes of the ECDFs contradict the constant hazard rate assumption of the exponential distribution. Instead, the curves imply an increasing hazard rate, motivating the use of more flexible lifetime models such as Weibull or MOGE.

7.2.3 Total Time on Test (TTT) Plots

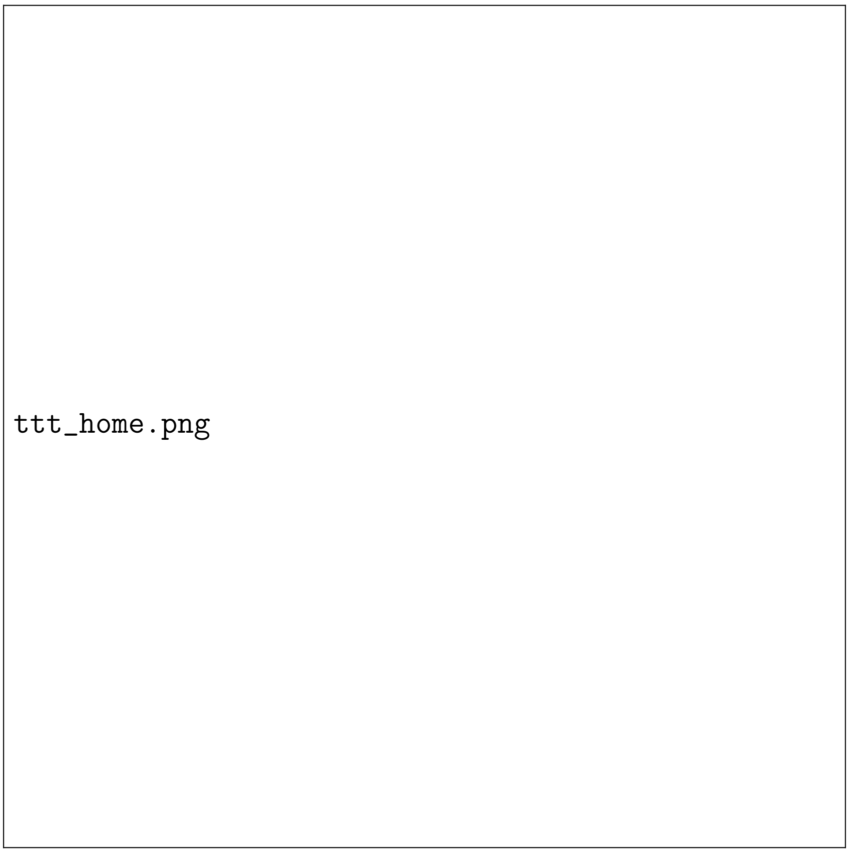

Figure 7.3: TTT plot for home-team first-goal times.

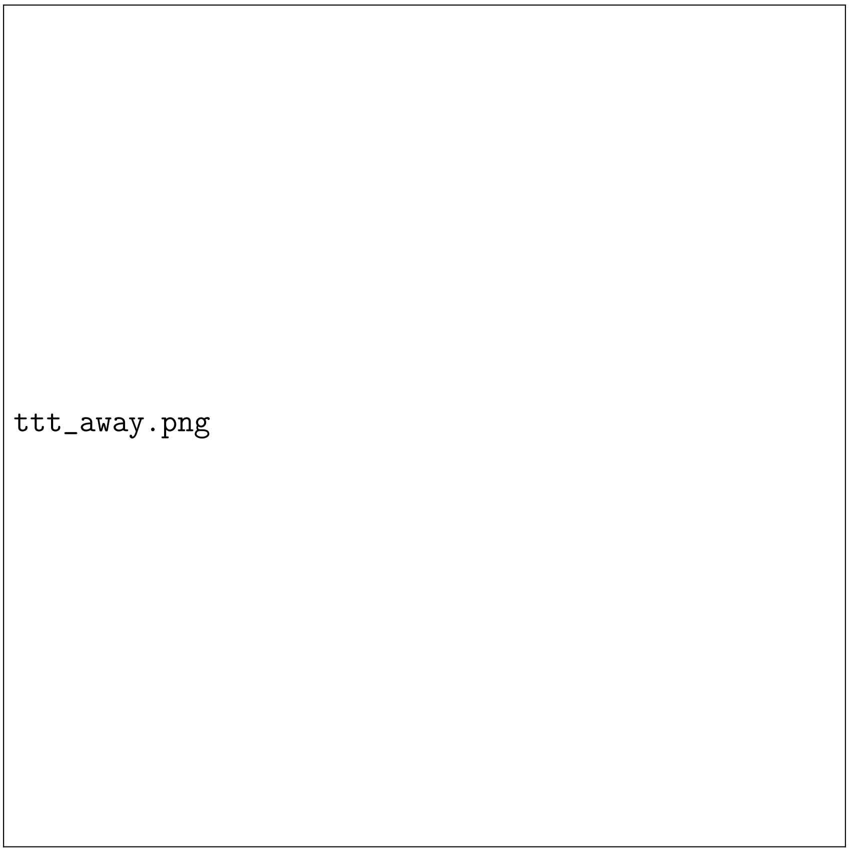

Figure 7.4: TTT plot for away-team first-goal times.

Figures **??** and **??** show the Total Time on Test (TTT) plots. In both cases, the empirical curves lie below the 45-degree reference line, indicating an Increasing Failure Rate (IFR):

$$\text{TTT curve below diagonal} \;\Rightarrow\; \text{Increasing Hazard Rate (IFR).}$$

This implies that the likelihood of scoring a first goal increases as time progresses. The home-team TTT curve displays slightly stronger convexity, reflecting greater offensive momentum or tactical pressure in home matches—consistent with home-advantage effects.

7.3 Model Fitting and Parameter Estimation

To assess the suitability of the MOGE distribution, three lifetime models were fitted to the home and away datasets:

- Gamma distribution (MLE),
- Weibull distribution (MLE),
- MOGE distribution (EM algorithm due to latent structure).

Each model represents different hazard-rate behaviour:

- **Gamma**: flexible but limited for steeply increasing hazards.
- **Weibull**: can model monotonic increasing or decreasing hazards.
- **MOGE**: introduces an additional parameter θ, enabling highly flexible hazard shapes and improved tail modeling.

Model performance was compared using:

$$\begin{aligned}\text{Log-likelihood},\\ \text{AIC (penalized model complexity)},\\ \text{Kolmogorov–Smirnov (KS) statistic}.\end{aligned}$$

A better-fitting model exhibits higher log-likelihood, lower AIC, and lower KS statistic.

7.4 Goodness-of-Fit Results

Fitted cumulative distribution functions were plotted against the empirical ECDF. The MOGE distribution showed the closest agreement with observed data, particularly in the early 0–30 minute window where most goals occur. In contrast, both the Gamma and Weibull models underestimated early scoring intensity.

Despite the additional parameter, MOGE achieved superior AIC and KS values, demonstrating that the improvement in fit outweighed the penalty for higher complexity. Its flexible hazard structure accurately represented the increasing probability of scoring near halftime.

7.5 Discussion

The model results indicate that first-goal scoring in football is not memoryless but time-dependent. The increasing hazard rate observed in the data aligns with strategic match dynamics, where offensive intensity typically rises as time progresses.

The MOGE model captures this behavior more effectively than the Weibull or Gamma distributions due to its flexible hazard structure governed by parameter θ. Simpler models with constant or strictly monotonic hazard rates fail to capture the rapid escalation in scoring risk associated with tactical adjustments, psychological momentum, and end-of-half pressure.

Thus, the MOGE model provides a more realistic representation of football scoring processes and is well-suited for predictive analytics.

7.6 Conclusion of Real Data Analysis

This chapter presented a comprehensive analysis of first-goal scoring times using real FIFA Club World Cup match data. The histogram and ECDFs revealed strong right skewness, early scoring concentration, and a long tail of late goals, with more than 80% of goals occurring before the 40th minute.

TTT plots confirmed an increasing hazard rate for both home and away teams, ruling out constant-rate exponential behavior. Weak correlations between X_1 and X_2 suggested independence between home and away scoring processes. Bootstrapping further revealed variability in parameter estimation due to sample size constraints.

Together, these findings motivate the simulation study presented in the next chapter and justify the use of advanced lifetime models such as MOGE for understanding and predicting football scoring behavior.

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

Gupta, R. C. and Kundu, D. (1999). Generalized exponential distributions. *Australian & New Zealand Journal of Statistics.*

Marshall, A. W. and Olkin, I. (1997). A new method for adding a parameter to a family of distributions. *Journal of the American Statistical Association.*

Ristić, M. M. and Kundu, D. (2015). Marshall–olkin generalized exponential distribution. *Communications in Statistics–Theory and Methods.*